The influence of meteorological variables and lockdowns on COVID-19 cases in urban agglomerations of Indian cities

Asha B. Chelani^{1,3} · Sneha Gautam^{2,3}

Accepted: 15 December 2021 / Published online: 25 January 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Coronavirus has been identified as one of the deadliest diseases and the WHO has declared it a pandemic and a global health crisis. It has become a massive challenge for humanity. India is also facing its fierceness as it is highly infectious and mutating at a rapid rate. To control its spread, many interventions have been applied in India since the first reported case on January 30, 2020. Several studies have been conducted to assess the impact of climatic and weather conditions on its spread in the last one and half years span. As it is a well-established fact that temperature and humidity could trigger the onset of diseases such as influenza and respiratory disorders, the relationship of meteorological variables with the number of COVID-19 confirmed cases has been anticipated. The association of several meteorological variables has therefore been studied in the past with the number of COVID-19 confirmed cases. The conclusions in those studies are based on the data obtained at an early stage, and the inferences drawn based on those short time series studies may not be valid over a longer period. This study attempted to assess the influence of temperature, humidity, wind speed, dew point, previous day's number of deaths, and government interventions on the number of COVID-19 confirmed cases in 18 districts of India. It is also attempted to identify the important predictors of the number of confirmed COVID-19 cases in those districts. The random forest model and the hybrid model obtained by modelling the random forest model's residuals are used to predict the response variable. It is observed that meteorological variables are useful only to some extent when used with the data on the number of the previous day's deaths and lockdown information in predicting the number of COVID-19 cases. Partial lockdown is more important than complete or no lockdown in predicting the number of confirmed COVID-19 cases. Since the time span of the data in the study is reasonably large, the information is useful to policymakers in balancing the restriction activities and economic losses to individuals and the government.

Keywords COVID—19 · Lockdown · Random forest model · Hybrid model · Meteorological parameters · India

1 Introduction

For the past year and a half, novel Coronavirus (COVID-19) has spread throughout almost every country (Ambade et al. 2021; Chelani and Gautam 2021; Gautam et al. 2022).

The nations are facing the wrath of the disease with outbreaks along with seasonal variations. In India, over 32,474,773 cases and 435,050 deaths have been reported (Worldometer 2021). The temporal variations in the number of COVID-19 cases observed during January 2020 till date across the regions indicate the presence of seasonality (Gautam et al. 2020a, b; Gautam 2020a,b; Chelani and Gautam 2021). It is a well-established fact that temperature and humidity could trigger the onset of diseases such as influenza and respiratory disorders (Shaman and Kohn 2009; Golakota et al. 2021; Gautam and Trivedi 2020; Humbal et al. 2019, 2018). Due to the low temperature in winter, a high number of flu or influenza cases are usually witnessed, whereas in summer, fewer cases are generally

✉ Sneha Gautam
gautamsneha@gmail.com; snehagautam@karunya.edu

¹ Air Pollution Control Division, Nagpur, India

² National Environmental Engineering Research Institute (CSIR-NEERI), Nehru Marg, Nagpur 440020, India

³ Department of Civil Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu 641114, India

observed (Damette et al. 2021; Chen et al. 2021). The association of several meteorological variables with the number of confirmed COVID-19 cases has therefore been analyzed in various studies (Cole et al. 2020; Conticini et al. 2020; Gautam et al. 2020, a, b; Zhu et al. 2020; Yao et al. 2020; Ambade et al. 2021; Gautam et al. 2021a, b, c). It has been observed that cities with an average temperature of $< 10^{\circ}\text{C}$ and lower humidity have more chances of spreading the virus than cities with higher temperatures (Sajadi et al. 2020; Araujo and Naimi 2020). In China, a temperature increase of 1°C in temperature is associated with a 4.861% increase in the daily confirmed COVID-19 cases (Xie and Zhu 2020). Few studies, however, have established the insignificant or negative effect of meteorological factors on COVID-19 confirmed cases (Yao et al. 2020; Liu et al. 2020; Shi et al. 2020; Méndez-Arriaga 2020; Wu et al. 2020; Yuan et al. 2021; Bisht et al. 2022). Studies on the association between the meteorological parameters and COVID-19 have provided mixed results and do not provide empirical evidence or confirm the statistical significance of the association (Kerr et al. 2021). In India, the warmer temperatures are usually observed in most of the regions for many days, which was the reason for the anticipation that the disease would not spread to the tune of other areas having cooler climates (Chen et al. 2021; Gautam et al. 2021b, c). However, the havoc created by the virus in India is not unknown (BBC 2021). Even in the warmer period, a large number of cases have been reported (<https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus>).

In the above studies, although the spatial horizon was quite large, the temporal horizon was quite limited due to the non-availability of the data on the number of COVID-19 confirmed cases. The outcome based on the small datasets available over a limited time period may lead to erroneous conclusions (Kerr et al. 2021). A period of almost one and a half years has passed since the onset of the pandemic, so a time series of the number of COVID-19 cases with reasonable data size can be obtained over an area. With the availability of data over a reasonably long time period, a rigorous study can be conducted on the relationship between meteorological factors and the number of COVID-19 confirmed cases to assess the role of the former in governing the disease spread. In addition, the effect of policy interventions such as complete or partial lockdowns implemented to prevent the spread of the virus can also be assessed. In India, the complete lockdown was initiated for March 24, 2020, which was later relaxed on April 14, 2020 in phase-wise manner. During the complete lockdown, all the activities except the essential services were completely halted. Partial lockdowns were imposed after April 14, 2020 with relaxation in phases (URL1 2020;

URL2 2020; URL3 2020; URL4 2020; URL5 2020; URL6 2020; URL7 2020; URL8 2020). It is interesting to know the influence of complete and partial lockdowns in the cities on the containment of the spread of the disease. As people gather for condolence meetings, although fewer in numbers due to government restrictions, the number of deaths that occurred due to COVID-19 on the previous day may influence the spread of the virus. With the data on meteorological factors such as daily mean temperature, wind speed, relative humidity, dew point temperature, number of previous day's deaths and interventions, i.e., lockdown variables in 18 districts across India, a random forest model is applied with the number of confirmed COVID-19 cases as a response variable and others as predictors from March 24, 2020 to June 15, 2021. Breiman (2001) defines random forest (*RF*) as a supervised learning model based on the concept of decision trees. The model is shown to outperform many machine-learning based models such as Naïve Bayes, Logistic Regression, Single Decision Tree, and Artificial Neural Network (Yu et al. 2016). In Kotschieder et al. (2011), they showed the applicability of the model in the presence of noise in the data. The number of confirmed COVID-19 cases may be dependent on the previous day's number of cases. It is desirable to include the time series of the previous day's cases in the *RF* model. As one of the objectives of the study is to identify the variables or predictors that have more importance on the response variable, it is probable that the inclusion of the previous day's cases may supersede the other predictors, and the importance of these predictors will not be clearly visible. To avoid this, the residuals of the *RF* model are modelled as an autoregressive process of order 1 (*AR1*). The results of both the models are then combined to obtain the hybrid model. This way, the variable importance using the *RF* model can be distinctly achieved. A hybrid model will provide more reliable predictions as these are shown to outperform single models in various studies (Díaz-Robles et al. 2008; Chelani and Devotta 2006). The study will be useful in decision making and estimating the number of COVID-19 cases.

2 Method

2.1 Data collection and model use

Figure 1 shows the districts that are included in the study. The monthly mean of the total number of cases reported for each district divided by the corresponding population since the onset of the disease till June 15, 2020 is plotted in Fig. 2. The COVID-19 data are obtained from COVID-19-India (2021). Although the lockdown has been imposed since March 2020, the number of confirmed cases across

Fig. 1 Location of districts in map of India

India is not available. Based on the availability of the data, the number of confirmed cases from 18 districts in different states of India from April 26, 2020 to June 15, 2021 are considered.

The time series of the number of confirmed cases for each district is normalized by dividing it by the population of the corresponding district. The newly formed time series

is considered the response variable, which is modelled as a function of meteorological variables such as temperature, wind speed, relative humidity, and dew point, lockdown variable, which is a categorical dummy variable, and the number of deaths reported on the previous day. The meteorological data are obtained from Weather Underground (2021). The policy interventions such as complete

Fig. 2 Monthly variations in total number of confirmed COVID-19 cases (divided by population of the district) since April 26, 2020 to June 15, 2021 in various districts of India

Fig. 2 continued

and partial lockdown are incorporated into the model as dummy variables with the following descriptors as:

- Complete lockdown—The complete lockdown was initiated in India starting with the ‘Janta Curfew’ on

March 22, 2020. The complete lockdown was announced on March 24, 2020 until May 31, 2020. During this phase, all non-essential services and factories were suspended, except for essential services such as grocery stores and vegetable sellers, which were

allowed to remain open for a particular duration of time. Traffic movement was restricted due to strict compliance by the police and local governments.

- Partial lockdown—Essential and non-essential activities were permitted for a few hours until 2PM or 4PM or 8PM.
- Unlock—On June 1, 2020, the central government announced unlock 1.0 with fewer restrictions, which was followed by a series of unlock phases such as unlock 2.0, unlock 3.0, unlock 4.0 and unlock 5.0, all of which were extended until November 30, 2020. All the phases of unlock are incorporated into the model as one dummy variable.
- Complete upliftment of the lockdown or no lockdown—From December 1, 2020, the restrictions were uplifted till the beginning of the second wave of COVID-19, which was initiated on April 5, 2021.

From April 5, 2021 to April 30, 2021, a partial lockdown was imposed, which became stricter from May 1, 2021 to May 31, 2021. The lockdown phases were followed in this study as imposed by the central government only. The model incorporates the lockdown variables as a categorical dummy variable, with complete lockdown as *Lk1*, unlock as *Lk2*, partial lockdown as *Lk3* and no lockdown as *Lk4*. The four dummy variables are therefore introduced in the model code with binary values as described above. The other variables such as mean daily temperature ($^{\circ}\text{C}$), humidity (%), wind speed (m s^{-1}), dew point temperature ($^{\circ}\text{C}$) for the respective district are used in the model as *Temp*, *RH*, *WS*, and *Dew*. The number of deaths in the previous few days may have had an influence on the spread of the virus. The number of deaths that occurred on the previous day is included in the model as it is assumed that the effect of the previous day's deaths is inclusive in the current day's deaths due to autocorrelation. It is denoted as *Death_1* in the model.

The reported cases may be dependent on the previous number of cases due to temporal correlations. The lagged variables of confirmed cases are, however, not included in the *RF* model to avoid the tautological effect on the model. The inclusion of the previous day's confirmed cases may overpower the model and therefore have more importance than other features. Therefore, the model is developed only with meteorological variables, number of previous day's deaths and intervention exogenous variables. The meteorological variables may also be auto-correlated. The lags of meteorological variables are not included in the model as it is assumed that the values observed on a particular day are inclusive of the effect of the previous day's values, so any effect on the response variable of those previous day's values shall be taken care of by the present day's observations.

2.2 Random forest model

Random forest (*RF*) is an ensemble method that involves the random forest of decision trees (Breiman 2001). It is a supervised learning algorithm that constructs decision trees based on the training data set. The combination of decision trees minimizes the out of bag error based on the training data sets. Studies have shown its applicability even in the presence of noise in the data (Kontschieder et al. 2011). Many nonlinear and high-dimensional complex classification and regression problems have been solved by applying *RF* (Yu et al. 2016). In the case of a regression problem, the predictions are obtained by the random selection of a number of predictor variables in the decision tree. The best solution is determined based on the number of nodes and the variables in the nodes of the fully grown tree. Every training set is fed to each decision tree (Breiman 2001, 2002; Kane et al. 2014; Moore et al. 2019; Izquierdo-Verdiguier and Zurita-Milla 2020). The number of variables or features to construct the model is selected with a specified value. For many regression problems, only 1/3rd of the total number of variables is often used (Liaw and Wiener 2002). The usual practise for selecting the number of trees is to randomly select an initial value from 10 to 1000 or higher and then choose the one based on model performance criteria. A higher number of trees, however, slows the learning process. A discussion of other selection criteria based on cross-validation and tuning is given in Stone (1974). The decision tree is constructed for the training cases with the specified nodes of the tree. The training set is selected and trained, whereas the remaining cases are used to estimate the error. So, for each case, out of bag error estimates are provided. The random forest model has an advantage over other supervised learning models for classification and regression problems as it avoids over-fitting samples and provides the solution based on averaging. Each variable has its importance in the overall model performance, which can be shown based on the Gene index for classification and mean square error (MSE) for regression problems. In this study, the *RF* model is applied using the library 'randomForest' in *R4.0.0* (R Development Core Team 2010). The data was divided into training and testing sets with a ratio of 85:15 for each district. The number of trees and the random selection of the number of variables were adjusted according to the mean square error as a cost function. The bootstrapping with a sample size of 500 was done to arrive at an optimum model. The 'Çaret' library in *R* was used to carry out bootstrapping of the training samples. The optimum number of trees and the number of variables with a minimum mean square error (*MSE*) are observed to be 80 and 7. The *RF* model helps in ranking the relative importance of the

predictors in modelling the response variable. For variable importance, $\%IncMSE$ is computed in the *R* package, which is the increase in *MSE* of predictions of out of bag samples as a result of permutations of the variables. $\%IncMSE$ is used to rank the important features of the response variable. The model is run each time a split on one predictor is conducted. An important predictor usually causes a large change in the *MSE*.

2.3 AR1 model

The *RF* model in this study uses the meteorological and dummy lockdown variables as predictors to predict the COVID-19 confirmed cases as a response variable. The intent of the study is to predict the response variable as well as identify the important predictor. If one includes the historical confirmed cases in the predictor set, it will overshadow the effect of other variables. To avoid this, the effective way is to model the residuals of a *RF* model, which can capture the temporal relationships and provide more reliable model performance than relying on a single model fitted to the datasets (Díaz-Robles et al. 2008; Chelani and Devotta 2006). The autoregressive model of order 1 (*ARI*) model is therefore fitted to the residuals of the *RF* model. The details of the model are given below.

Let r_t represents the residuals obtained by fitting the *RF* model to the number of confirmed cases for the training set.

$$r_t = y_t - \widehat{RF}_t \quad (1)$$

where $\widehat{RF}_t$ is the forecast value of the *RF* model for time t . *ARI* model can be fitted as in Eq. (2);

$$r_t = \varphi_0 + \varphi_1 r_{t-1} + \varepsilon_t \quad (2)$$

Like *RF* model, the coefficients are estimated for residuals of the training set and the predictions are obtained for residuals of the testing set.

2.4 Hybrid model

The predictions obtained by the *RF* and *ARI* models are combined to enhance the model performance. The hybrid methodology is based on the combination of the linear autocorrelation model and the random forest model (Díaz-Robles et al. 2008; Chelani and Devotta 2006), which can be given as;

$$\widehat{y}_t = \widehat{r}_t + RF_t \quad (3)$$

where r_t denotes the *ARI* model fitted to residuals of the *RF* model as given in Eq. (2), and RF_t denotes the *RF* model of the confirmed cases. First the *RF* model is fitted to the data and the residuals are obtained, which are then modelled as the *ARI* process. Let the forecast from the *ARI* model be

denoted as $\widehat{r}_t$. The new forecasts can therefore be obtained as

$$\widehat{y}_t = \widehat{RF}_t + \widehat{r}_t \quad (4)$$

To evaluate the performance of the models, the error statistics such as correlation between observed and predicted cases, root mean square error (*RMSE*), and normalized root mean square error (*NRMSE*) are utilized. These test statistics can be obtained as,

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \widehat{y}_t)^2} \quad (5)$$

$$NRMSE = \frac{RMSE}{1/n \sum_{t=1}^n y_t} \quad (6)$$

where y_t is the observed and $\widehat{y}_t$ is the predicted data and n is the total number of data points.

3 Results and discussion

The descriptive statistics of the number of confirmed cases and meteorological variables is given in Table 1. The mean daily temperature, wind speed, relative humidity, and dew point temperature range from 23.3 ± 6.7 °C to 35.8 ± 12.5 °C, 0.4 ± 0.2 ms⁻¹ to 3.1 ± 1 ms⁻¹, $44.7 \pm 24.3\%$ to $72.9 \pm 13.5\%$, and 7.08 ± 3.82 °C to 21.69 ± 5.02 °C, respectively. The number of confirmed cases ranged between 0 and 2036 per million, whereas the number of deaths ranged between 0 and 184 per million in all the districts during the study period.

The monthly variations in confirmed cases divided by the corresponding population of the district are given in Fig. 2, which shows that April and May have witnessed outbreaks due to the second wave of COVID-19 in all the districts. In Nagpur, Patiala, Thane, Mumbai, Ludhiana, and Pune, a high number of confirmed cases have also been observed in March. A few spikes are also observed in other months. Further, the correlation analysis of meteorological variables, the number of deaths and the number of confirmed cases is given in Table 2. The correlation between the number of deaths and confirmed cases is statistically significant in all the districts except in Mumbai and Patna. The correlation of the number of confirmed cases with temperature is significant in all the districts except in Chandrapur, Thane, and Ujjain. The correlation between relative humidity and the number of confirmed cases is mostly negative and significant except in Kolkata and Nagpur. In Mumbai and Thane, the correlation with relative humidity is positive. The correlation of the number of confirmed cases with wind speed is mostly negative and insignificant. Wind speed has been characterized as one of

Table 1 Descriptive statistics of number of confirmed cases and meteorological variables

S no.	District	Temp			RH			WS			Dew			Death_1			Cases*		
		Mean	±	SD	Mean	±	SD	Mean	±	SD	Mean	±	SD	Mean	±	SD	Mean	±	SD
1	Chandrapur	28.4	±	2.2	49.1	±	12.4	0.4	±	0.2	18.24	±	3.56	1.7	±	4.5	95.8	±	172.2
2	Chennai	32.0	±	3.5	72.8	±	8.3	2.1	±	1.4	26.0	±	3.31	2.7	±	4.3	178.5	±	211.7
3	Delhi	26.0	±	5.9	56.4	±	14.5	0.8	±	0.3	17.3	±	5.67	3.2	±	4.7	181.5	±	284.8
4	Dewas	29.3	±	1.3	51.0	±	23.1	3.1	±	1.0	19.45	±	4.9	0.1	±	0.3	11.9	±	19.2
5	Faridabad	26.6	±	5.8	53.0	±	18.7	0.7	±	0.3	17.16	±	5.95	0.9	±	1.3	132.6	±	202.4
6	Jodhpur	31.6	±	5.3	44.7	±	24.3	0.8	±	0.4	20.57	±	8.02	0.7	±	1.5	73.2	±	122.1
7	Kanpur	35.7	±	2.5	60.4	±	14.7	1.7	±	0.8	27.8	±	3.34	1.0	±	1.5	43.5	±	84.6
8	Kolkata	26.9	±	4.4	71.1	±	14.7	0.8	±	0.7	21.14	±	5.79	0.8	±	0.7	49.3	±	65.3
9	Ludhiana	24.6	±	7.5	65.6	±	20.5	0.5	±	0.2	17.74	±	6.43	1.4	±	1.7	59.4	±	88.6
10	Mumbai	25.9	±	2.0	72.9	±	13.5	0.5	±	0.1	20.49	±	3.47	3.0	±	4.1	138.6	±	161.7
11	Nagpur	24.2	±	6.4	49.1	±	11.9	0.5	±	0.1	14.14	±	7.11	4.1	±	7.0	255.1	±	388.6
12	Patiala	23.3	±	6.7	69.7	±	21.4	0.5	±	0.2	17.29	±	6.34	1.7	±	2.3	60.9	±	77.8
13	Patna	25.6	±	5.8	70.7	±	16.7	0.5	±	0.2	19.7	±	5.74	1.0	±	9.0	60.1	±	103.7
14	Pune	35.8	±	12.5	48.6	±	14.8	0.5	±	0.2	21.52	±	8.46	4.0	±	6.3	264.9	±	313.0
15	Rohtak	25.5	±	7.9	51.7	±	9.9	1.1	±	0.6	16.37	±	7.68	1.1	±	2.9	58.5	±	86.1
16	Thane	26.8	±	4.2	71.8	±	12.5	0.5	±	0.7	7.08	±	3.82	2.2	±	3.9	123.8	±	131.3
17	Ujjain	31.4	±	1.4	51.6	±	21.1	2.8	±	1.3	21.69	±	5.02	0.2	±	0.6	22.8	±	40.1
18	Varanasi	25.3	±	6.7	57.1	±	23.5	1.1	±	0.9	18.16	±	6.18	0.6	±	0.9	55.7	±	118.7

Cases* refers the number of confirmed cases per million

Death_1: number of deaths due to COVID-19 on previous day, *WS* wind speed, *Temp* temperature, *RH* relative humidity, *Dew* dew point temperature

Table 2 Correlation between meteorological variables and number of confirmed cases

S. no.	District	Death_1	Temp	RH	WS	Dew
1	Chandrapur	0.44*	−0.01	−0.18*	−0.03	−0.13*
2	Chennai	0.55*	0.13*	−0.22*	0.05	0.11*
3	Delhi	0.80*	0.12*	−0.34*	−0.03	−0.05
4	Dewas	0.17*	0.23*	−0.3*	−0.02	−0.22*
5	Faridabad	0.72*	0.14*	−0.29*	0	−0.04
6	Jodhpur	0.88*	0.21*	−0.23*	−0.04	0
7	Kanpur	0.60*	0.11*	−0.47*	0.15*	−0.34*
8	Kolkata	0.69*	0.27*	−0.06	−0.03	0.17*
9	Ludhiana	0.77*	0.25*	−0.4*	0.02	0.04
10	Mumbai	0.08	0.42*	0.23*	0.23*	0.42*
11	Nagpur	0.31*	0.24*	0.05	0.02	0.22*
12	Patiala	0.79*	0.26*	−0.41*	0.06	0
13	Patna	0.06	0.29*	−0.48*	0.21*	0.01
14	Pune	0.16*	0.15*	−0.43*	0.09	0.13*
15	Rohtak	0.71*	0.12*	−0.24*	0.13*	0.03
16	Thane	0.11*	0.10	0.13*	−0.13*	0.05
17	Ujjain	0.29*	−0.06	−0.44*	0.00	−0.39*
18	Varanasi	0.67*	0.40*	−0.17	0.16*	0.26*

Death_1: number of deaths due to COVID-19 on previous day, *WS* wind speed, *Temp* temperature, *RH* relative humidity, *Dew* dew point temperature, *: significant at $p = 0.05$

Table 3 Variable importance matrix

Variable	Description	%IncMSE
Death_1	Number of deaths due to COVID-19 on previous day	39.6
Temp	Temperature	22.7
WS	Wind speed	20.3
Dew	Dew point temperature	16.8
RH	Relative humidity	14.8
Lk3	Partial lockdown	11.3
Lk2	Unlock	9.3
Lk1	No lock down	6.9
Lk4	Complete lockdown	0

the factors in defining the ventilation coefficient of an area (Goyal and ChalapatiRao 2007). It has been observed in the past that high wind speeds and good ventilation are associated with a lower number of COVID-19 cases. The relationship in this study is, however, negative and not significant. The relationship between dew point and confirmed cases is sporadic with a negative correlation in Chandrapur, Dewas, Kanpur, and Ujjain and a positive correlation in Chennai, Kolkata, Mumbai, Nagpur, Pune, and Varanasi. When one looks at the scatter plot of confirmed cases and meteorological parameters, it can be seen that there is an inverse parabolic relation between confirmed cases and temperature, relative humidity, and dew point. The number of cases increases with these factors, and beyond a certain point, the number of confirmed cases starts decreasing. The number of confirmed cases start decreasing with the temperature at 31 ± 0.5 °C, relative humidity at $51.7 \pm 0.6\%$ and dew point at 21 ± 0.4 °C, respectively. The relationship between the number of confirmed cases and wind speed was exponentially decreasing with an initial increase.

Using the *RF* model, the percentage increase in MSE was calculated to assess the importance of the variables. The variable importance ranking shown in Table 3 suggests that the major governing factor of the number of confirmed COVID-19 cases is the number of deaths that occurred on the previous day. Plausibly, this is expected as the correlation of confirmed cases with the number of previous day's deaths is observed to be significant and highest as compared to the correlation with other variables (See Table 2). Temperature is the second most important factor influencing the confirmed cases, closely followed by wind speed. On the other hand, the dew point and relative humidity have relatively less influence on confirmed case predictability. In a review on the effect of meteorology on COVID-19 transmission, Kerr et al. (2021) pointed out that temperature and humidity have an established role in virus transmission. Further, no study has been conducted so far that has used a complete annual cycle of data (Kerr et al. 2021). Instead, space has been substituted for time

variations. As mentioned earlier, low temperatures are anticipated to cause higher transmissibility of the virus. A negative correlation, however, is not observed to be significant at any of the study locations. Relative humidity, however, shows a significant negative correlation with the number of COVID-19 cases at 18 locations.

The influence of lockdown variables is quite low as compared to meteorological variables and the number of previous day's deaths. Partial lockdown, effective only for a few hours, is seen to be highly effective in confining cases as compared to a complete lockdown and no lockdown. The unlock period is also shown to be the second influencing factor among the lockdown measures governing the number of confirmed cases. This finding is very useful to policymakers and the economic growth point of view. Policymakers can opt for partial lockdowns instead of complete or no lockdown at all to sustain economic activity, viz. a viz. confine the spread of the virus. Moreover, it can be seen from the importance matrix that meteorological variables alone cannot be used as predictors to predict or estimate the number of confirmed cases for any district. Instead, the number of the previous day's deaths and lockdown details are also required. One of the major limitations of the *RF* model is that it does not provide the direction of the relationship like regression models. The negative or positive impact of the input variables on the response variable cannot be ascertained. The correlation of the response variable with the predictors can be used to ascertain the direction of the relationship.

The results of *RF* models are then further improved by fitting an *ARI* model on the residuals of the model. The *ARI* model is observed to have a coefficient $\varphi_0 = -0.000127$ ($p > 0.05$) and $\varphi_1 = 0.46$ (0.44, 0.48) with $p < 0.05$. The estimations obtained by the *ARI* model i.e., $\hat{r}_t$ are added to the estimations of the *RF* model ($\widehat{RF}_t$) as in Eq. (4). The performance of the final results of the hybrid model is given in Table 4. The prediction results for the training and testing sets in terms of box plot are shown in Fig. 3a and b, respectively. R^2 for the training set for both the models is the same but for the testing set, it is slightly

Table 4 Model evaluation for training and testing set

Statistics	RF model		Hybrid model	
	Training	Testing	Training	Testing
R ²	0.96	0.76	0.96	0.79
RMSE	0.000037	0.000286	0.000033	0.000205
NRMSE	0.36	0.55	0.29	0.38

Fig. 3 **a** Observed and predicted number of cases using hybrid model in 18 districts for training set **b** Observed and predicted number of cases using hybrid model in 18 districts for testing set

better for the hybrid model. *NRMSE* has also moderately improved for the hybrid model for both training and testing sets. The predictions provided by the *RF* model are nevertheless undermined as the hybrid model is based on its output. In addition, the predictions of the *RF* model are based on the exogenous predictors and not on the historical lag time series of confirmed cases. The model can reasonably provide a good prediction of the number of confirmed COVID-19 cases based on the exogenous meteorological variables for a particular day along with the previous day's deaths. The *RF* model's performance is quite acceptable, though the hybrid model moderately outperforms the *RF* model.

4 Conclusion

Considering the fierceness of COVID-19, many interventions have been applied in India to contain its spread. In the last one and a half years span, several studies have been conducted to assess the impact of climatic and weather conditions on its spread. The earlier studies on the linkage between meteorological variables and the number of COVID-19 confirmed cases are based on the data obtained at an early stage. Drawing any inferences based on the short-time data may lead to erroneous policy interventions to contain the disease. This study attempts to study the influence of temperature, humidity, wind speed, dew point, previous day's number of deaths due to COVID-19 and the government's interventions on the number of COVID-19 confirmed cases in 18 districts of India with the data observed from April 26, 2020 to June 15, 2021. It is also attempted to identify the important predictors of the number of cases in those districts using the random forest model and the hybrid model, which is obtained by modelling the residuals of the random forest model. When comparing the degrees of restrictions in terms of complete, partial, unlock phase,s and no lockdown, partial lockdown is observed to be more important than complete lockdown and no lockdown in predicting the number of confirmed COVID-19 cases. Since the time span of the data in the study is reasonable large, the information is useful to policy makers. To some extent, the meteorological variables are observed to be useful in predicting the number of COVID-19 cases. The data on the number of previous day's deaths is more important in governing the number of COVID-19 cases. Instead of restricting everything completely to contain the spread of the virus or allowing all the activities, the government can opt for partial lockdowns to minimize the economic losses to individuals and the government.

Acknowledgements This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Authors are thankful to the anonymous reviewers for the constructive comments that helped improve the quality of the manuscript.

References

- Ambade B, Sankar TK, Kumar A, Gautam AS, Gautam S (2021) COVID-19 lockdowns reduce the Black carbon and polycyclic aromatic hydrocarbons of the Asian atmosphere: source apportionment and health hazard evaluation. *Environ Develop Sustain*. <https://doi.org/10.1007/s10668-020-01167-1>
- Araujo, M. B., Naimi, B., 2020. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. <https://doi.org/10.1101/2020.03.12.20034728>.

- BBC, 2021. Covid-19: India in a 'delicate phase' of its coronavirus battle as cases surge. <https://www.bbc.com/news/world-asia-india-56206004>.
- Bisht, L., Gupta, V., Singh, A., Gautam, A.S., Gautam, S., 2022. Heavy metal concentration and its distribution analysis in urban road dust: A case study from most populated city of Indian state of Uttarakhand. *Spatial and Spatio-temporal Epidemiology* 40: 100470. (Accepted - <https://doi.org/10.1016/j.sste.2021.100470>).
- Breiman L (2001) Random forests. *Machine Learn* 45(1):5–32
- Breiman, L., 2002. Manual on setting up, using and understanding Random Forests V3.1. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.
- Chelani A, Gautam S (2021) Lockdown during COVID-19 pandemic: a case study from Indian cities shows insignificant effects on persistent property of urban air quality. *Geosci Front*. <https://doi.org/10.1016/j.gsf.2021.101284>
- Chelani AB, Devotta S (2006) Air quality modeling using a hybrid autoregressive and nonlinear model. *Atm Env* 40:1774–1780
- Chen S, Prettnier K, Kuhn M, Geldsetzer P, Wang C, Barnighausen T, Bloom DE (2021) Climate and the spread of COVID-19. *Sci Rep* 11:9042
- Cole MA, Ozgen C, Strobl E (2020) Air pollution exposure and Covid-19 in Dutch municipalities. *Environ Resour Econ* 76(4):581–610. <https://doi.org/10.1007/s10640-020-00491-4>
- Coticini E, Frediani B, Caro D (2020) Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? *Environ Pollut* 261:114465. <https://doi.org/10.1016/j.envpol.2020.114465>
- COVID-19-India, 2021. Accessed via COVID-19india.org.
- Damette O, Mathonnat C, Goutte S (2021) Meteorological factors against COVID-19 and the role of human mobility. *PLoS ONE* 16(6):e0252405. <https://doi.org/10.1371/journal.pone.0252405>
- Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, Moncada-Herrera JA (2008) A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco. *Chile Atmos Environ* 42(35):8331–8340
- Gautam AS, Dilwaliya N, Srivastava A, Kumar S, Baudh K, Singh D, Gautam S (2020) Temporary reduction in air pollution due to anthropogenic activity switch-off during COVID-19 lockdown in northern parts of India. *Environ Develop Sustain*. <https://doi.org/10.1007/s10668-020-00994-6>
- Gautam, S., Setu, S., Khan, M.G.Q., Khan, B., 2022. Analysis of the health, economic and environmental impacts of COVID-19: The Bangladesh perspective. *Geosystems and Geoenvironment*. (Accepted - <https://doi.org/10.1016/j.geogeo.2021.100011>).
- Gautam S, Sammuell C, Gautam AS, Kumar S (2021a) Strong link between coronavirus count and bad air: a case study of India. *Environ Develop Sustain*. <https://doi.org/10.1007/s10668-021-01366-4>
- Gautam S, Gautam AS, Singh K, James EJ, Brema J (2021b) Investigations on the relationship among lightning, aerosol concentration, and meteorological parameters with specific reference to the wet and hot humid tropical zone of the southern parts of India. *Environ Technol Innov*. <https://doi.org/10.1016/j.eti.2021.101414>
- Gautam S, Sammuell C, Bhardwaj A, Esfandabadi ZC, Santosh M, Gautam AS, Joshi A, Justin A, Wessley JW, James EJ (2021c) Vertical profiling of atmospheric air pollutants in rural India: a case study on particulate matter (PM₁₀/PM_{2.5}/PM₁), carbon dioxide, and formaldehyde. *Measurement* 185:110061. <https://doi.org/10.1016/j.measurement.2021.110061>
- Gautam S (2020a) The influence of COVID – 19 on air quality in India: a boon or inutile. *Bull Environ Contam Toxicol* 104(6):724–726
- Gautam S (2020b) COVID – 19: Air pollution remains low as people stay at home. *Air Qual Atmos Health*. <https://doi.org/10.1007/s11869-020-00842-6>
- Gautam S, Trivedi UK (2020) Global Implication of Bioaerosol in Pandemic. *Environ Develop Sustain* 22:3861–3865
- Golakota ARK, Gautam S, Santosh M, Sudan HA, Gandhi R, Jebadurai VS, Shu CM (2021) Bioaerosols: characterization, pathways, sampling strategies, and challenges to geo-environment and health. *Gondwana Res* 99:178–203. <https://doi.org/10.1016/j.gr.2021.07.003>
- Goyal SK, ChalapatiRao CV (2007) Assessment of atmospheric assimilation potential for industrial development in an urban environment: Kochi (India). *Sci Total Environ* 376(1–3):27–39
- Humbal C, Gautam S, Trivedi U (2018) A review on recent progress in observations, and health effects of Bioaerosols. *Environ Int* 118:189–193
- Humbal C, Gautam S et al (2019) Evaluating the colonization and distribution of fungal and bacterial bioaerosol in Rajkot, western India using multi-proxy approach. *Air Qual Atmos Health* 12(6):693–704
- Izquierdo-Verdiguier E, Zurita-Milla R (2020) An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *Int J Appl Earth Obs Geoinf* 88:102051. <https://doi.org/10.1016/j.jag.2020.102051>
- Kane MJ, Price N, Scotch M, Rabinowitz P (2014) Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatic* 15(1):1–9
- Kerr, G. H., Badr, H.S., Gardner, L.M., Perez-Saez, J., Zaitchik, B.F., 2021. Associations between meteorology and COVID-19 in early studies: Inconsistencies, uncertainties, and recommendations. *One Health*. 12 (100225), ISSN 2352–7714. <https://doi.org/10.1016/j.onehlt.2021.100225>.
- Kontschieder P, Bulò SR, Bischof H, Pelillo M (2011) Structured class-labels in random forests for semantic image labelling. *Inter Conf Comp Vision*. <https://doi.org/10.1109/ICCV.2011.6126496>
- Liaw A, Wiener M (2002) Classification and regression by random-forest. *R News* 2(3):18–22
- Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, He X, Wang B, Fu S, Niu T, Yan J, Shi Y, Ren X, Niu J, Zhu W, Li S, Luo B, Zhang K (2020) Impact of meteorological factors on the COVID-19 transmission: a multi-city study in China. *Sci Total Environ* 726:138513. <https://doi.org/10.1016/j.scitotenv.2020.138513>
- Méndez-Arriaga F (2020) The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase I. *Sci Total Environ* 735:139560. <https://doi.org/10.1016/j.scitotenv.2020.139560>
- Moore PJ, Lyons TJ, Gallacher J (2019) Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *PLoS ONE* 14(2):e0211558
- R Development Core Team (2010) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A (2020) Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. *JAMA. Network Open* 3(6):e2011834 (pmid:32525550)
- Shaman J, Kohn M (2009) Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci USA* 106(9):3243–3248
- Shi P, Dong Y, Yan H, Zhao H, Li X, Liu W, He M, Tang S, Xi S (2020) Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Sci Total Environ* 728:138890. <https://doi.org/10.1016/j.scitotenv.2020.138890>

- Stone M (1974) Cross-validated choice and assessment of statistical predictions. *J R Stat Soc: Ser B (methodol)* 36(2):111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- URL1, 2020. <https://www.tribuneindia.com/news/nation/centre-extends-nationwide-lockdown-till-may-31-new-guidelines-issued-86042>.
- URL2, 2020. <https://indianexpress.com/article/coronavirus/unlock-2-guidelines-july-coronavirus-6482179/>.
- URL3, 2020. <https://indianexpress.com/article/india/unlock-3-0-guidelines-rules-whats-allowed-whats-not-6529596/>.
- URL4, 2020. <https://www.firstpost.com/india/unlock-4-0-schools-to-reopen-for-classes-9-to-12-from-21-sep-on-voluntary-basis-centre-issues-guidelines-8797971.html>.
- URL5, 2020. <https://www.jagran.com/news/national-unlock-5-full-guidelines-know-what-to-open-from-1-october2020-20808875.html>.
- URL6, 2020. <https://www.india.com/hindi-news/india-hindi/unlock-6-0-guidelines-unlock-6-0-starts-from-today-in-india-know-whats-allowed-and-whats-not-unlock-6-full-guidelines-4193462/>.
- URL7, 2020. <https://hindi.careerindia.com/news/unlock-7-0-guidelines-in-hindi-for-december-2020-002841.html>.
- URL8, 2020. <https://www.indiatoday.in/india/story/india-lockdown-pm-narendra-modi-speech-coronavirus-1659266-2020-03-24>.
- Weather Underground, 2021. Accessed via www.wunderground.com, accessed on 28/7/2021.
- Worldometer, 2021. Accessed via <https://www.worldometers.info/coronavirus/country/india/>, accessed on 24/8/2021.
- Wu Y, Jing W, Liu J, Ma Q, Yuan J, Wang Y et al (2020) Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Sci Total Environ* 729:139051
- Xie J, Zhu Y (2020) Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci Total Environ* 724:138201. <https://doi.org/10.1016/j.scitotenv.2020.138201>
- Yao Y, Pan J, Liu Z, Meng X, Wang W, Kan H, Wang W (2020) No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. *Eur Respir* 55(5):2000517. <https://doi.org/10.1183/13993003.00517-2020>
- Yu R, Yang Y, Yang L, Han G, Move OA (2016) RAQ- A random forest approach for predicting air quality in urban sensing systems. *Sensors* 16:86
- Yuan J, Wu Y, Jing W, Liu J, Du M, Wang Y, Liu M (2021) Association between meteorological factors and daily new cases of COVID-19 in 188 countries: a time series analysis. *Sci Total Environ* 780:146538. <https://doi.org/10.1016/j.scitotenv.2021.146538>
- Zhu Y, Xie J, Huang F, Cao L (2020) Association between short-term exposure to air pollution and COVID-19 infection: evidence from China. *Sci Total Environ* 727:138704. <https://doi.org/10.1016/j.scitotenv.2020.138704>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.