

RESEARCH ARTICLE

A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers

Md. Siddikur Rahman *, Arman Hossain Chowdhury 

Department of Statistics, Begum Rokeya University, Rangpur, Rangpur, Bangladesh

* siddikur@brur.ac.bd OPEN ACCESS

Citation: Rahman M.S, Chowdhury AH (2022) A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers. PLoS ONE 17(9): e0273319. <https://doi.org/10.1371/journal.pone.0273319>

Editor: Randeep Singh, Amity University, INDIA

Received: May 14, 2022

Accepted: August 6, 2022

Published: September 13, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0273319>

Copyright: © 2022 Rahman, Chowdhury. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its [Supporting information files](#).

Funding: The author(s) received no specific funding for this work.

Abstract

COVID-19 pandemic has become a global major public health concern. Examining the meteorological risk factors and accurately predicting the incidence of the COVID-19 pandemic is an extremely important challenge. Therefore, in this study, we analyzed the relationship between meteorological factors and COVID-19 transmission in SAARC countries. We also compared the predictive accuracy of Autoregressive Integrated Moving Average (ARIMAX) and eXtreme Gradient Boosting (XGBoost) methods for precise modelling of COVID-19 incidence. We compiled a daily dataset including confirmed COVID-19 case counts, minimum and maximum temperature ($^{\circ}\text{C}$), relative humidity (%), surface pressure (kPa), precipitation (mm/day) and maximum wind speed (m/s) from the onset of the disease to January 29, 2022, in each country. The data were divided into training and test sets. The training data were used to fit ARIMAX model for examining significant meteorological risk factors. All significant factors were then used as covariates in ARIMAX and XGBoost models to predict the COVID-19 confirmed cases. We found that maximum temperature had a positive impact on the COVID-19 transmission in Afghanistan ($\beta = 11.91$, 95% CI: 4.77, 19.05) and India ($\beta = 0.18$, 95% CI: 0.01, 0.35). Surface pressure had a positive influence in Pakistan ($\beta = 25.77$, 95% CI: 7.85, 43.69) and Sri Lanka ($\beta = 411.63$, 95% CI: 49.04, 774.23). We also found that the XGBoost model can help improve prediction of COVID-19 cases in SAARC countries over the ARIMAX model. The study findings will help the scientific communities and policymakers to establish a more accurate early warning system to control the spread of the pandemic.

Introduction

The novel coronavirus disease 2019 (COVID-19), induced by severe acute respiratory syndrome 2 [1, 2], has become a serious public health threat globally. The disease has quickly spread over the world because of its extremely human-to-human transmission characteristics [3–5]. As of July 02, 2022, more than 553.87 million confirmed cases and over 6.36 million deaths have been reported globally [6]. It has already been studied that meteorological factors like temperature, relative humidity and wind speed have been linked to the development of the

Competing interests: The authors have declared that no competing interests exist.

transmission of recognized coronavirus infections such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [7, 8]. According to laboratory tests, SARS-CoV-2 is very stable in cold conditions but vulnerable to rising temperatures [9]. Different previous studies also investigated that meteorological factors such as temperature [10–12], humidity [13, 14], and wind speed [15] might affect COVID-19 transmission [16–18]. The transmission of the COVID-19 pandemic is reduced as temperature rises in China as well as other regions of the world [13, 16, 18]. It was also found that wind speed had lagged correlations with COVID-19 incidence in various Turkish cities [19]. Humidity was also a major meteorological factor in reducing COVID-19 viral transmission in China, Pakistan, Sri Lanka and other countries [11, 14, 20]. However, the humidity was also negatively associated with the COVID-19 epidemic in Indonesia and New York [21, 22].

Different studies widely used different types of statistical approaches including correlation, regression analysis, generalized additive model, and generalized linear model to analyze the influence of environmental variables on COVID-19 transmission [5, 13, 19, 21–24]. Besides these, several studies have used Autoregressive Integrated Moving Average with exogeneous variables model to determine the association of climate variables with COVID-19 transmission and forecasting [23, 25, 26]. Time-series modelling is a popular forecasting method for understanding the dynamic association of important variables. However, the transmission of COVID-19 disease is often influenced by several factors which exhibit nonlinear influences which cause problems [27]. This problem can be easily solved by machine learning techniques [28, 29]. Given the uncertainty around decisions on the accurate time of the emergence and disappearance of the disease, short-term forecasting is crucial to create better plans and more appropriate responses. The eXtreme Gradient Boosting (XGBoost) is an uptrend machine learning technique in time series modelling. The XGBoost model can generate a high precision result for its self-learning characteristics. This study contributes to the advancement of the time-series prediction of COVID-19. Consequently, an initial benchmarking is given to demonstrate the potential of machine learning for future research. The study further suggests that a genuine novelty in COVID-19 prediction can be realized by a data-driven XGBoost machine learning model. Currently, no study used this technique for determining the association between meteorological factors and COVID-19 transmission and prediction. Therefore, the study aimed to: (a) identify the meteorological risk factors; (b) compare the predictive accuracy of the ARIMAX and XGBoost for precise modelling of COVID-19 incidence in the South Asian Association for Regional Cooperation (SAARC) countries.

In this study, our proposed methodology (Fig 1) and results are useful to select a suitable model for COVID-19 prediction. The findings from this study will help the countries' policy-makers for taking effective strategies to establish a more accurate early warning system to control the spread of the pandemic.

Materials and methods

Data source

The daily COVID-19 confirmed cases data of the SAARC countries (Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka) were collected from the Johns Hopkins Coronavirus Resource Center [30]. The meteorological data of each country were obtained based on hourly meteorological observations from the NASA Langley Research Center (LaRC) website [31], including minimum and maximum temperatures (°C), relative humidity (%), maximum wind speed (m/s), surface pressure (kPa) and precipitation (mm/day). The study period was from the onset of COVID-19 to January 29, 2022, for each SAARC country.

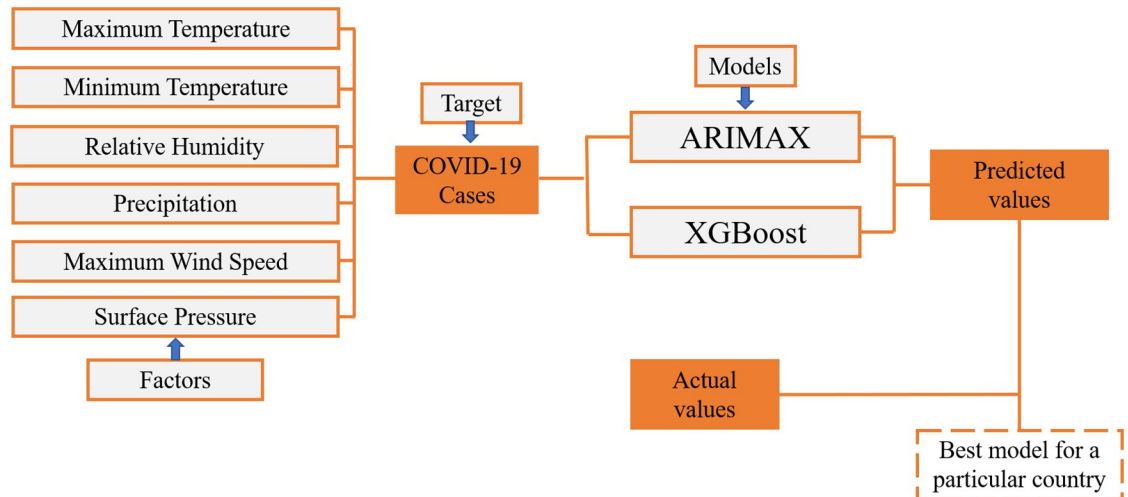


Fig 1. Schematic of the proposed methodology.

<https://doi.org/10.1371/journal.pone.0273319.g001>

Model building, prediction, and performance evaluation

Predictive modeling and statistical analyses were conducted using RStudio (Version 4.1.0) [32]. The 'tseries' and stats packages were used to process the time series. The ARIMAX models were built with the 'forecast' package using auto.arima function for choosing the best model based on the Corrected Akaike Information Criteria (AICc) values [33]. The 'forecastxgb' package was used for building the XGBoost model. Details data and necessary codes for predictive modeling and statistical analysis are provided in supplements (S1 Table and S1 Text).

In this study, predictive accuracy of ARIMAX and XGBoost models was compared to determine which was more suitable for predicting COVID-19 confirmed cases in SAARC countries based on meteorological risk factors. The data were divided into training and test sets. All the significant meteorological factors were used as covariates in ARIMAX and XGBoost models for predicting the COVID-19 confirmed cases. The ensemble machine learning technique XGBoost were built using the lagged meteorological variables as covariates by frequently changing several parameters. The adjusted parameters for the model of each SAARC country are nrounds, nrounds_method = 'cv', nfold, seas_method, trend_method = 'none'.

Predictive accuracy refers to the capacity of the model to predict COVID-19 incidence. There are several metrics for computing the model's accuracy [34]. However, in this study, we used four prominent performance metrics such as the mean absolute percentage error (MAPE), mean percentage error (MPE), mean absolute error (MAE) and root mean square error (RMSE). The mathematical form of the error measures are as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \tag{1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \tag{3}$$

$$MPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i - y_i}{y_i} \right) \times 100\% \tag{4}$$

Where n represents the number of observations, $\hat{y}_i - y_i$ represents the error between the predicted and actual value.

ARIMAX model

The Autoregressive Integrated Moving Average (ARIMA) model introduced by Box and Jenkins (2013) is widely used for predicting time series data because of its capacity to handle non-stationary data [35]. ARIMA(p, d, q) combines the Autoregressive (AR) and Moving Average (MA) models, with the 'I' indicating integration; where p stands for autoregressive order, d for differencing order, and q stands for moving average order [36]. The AR(p) in ARIMA stands for a linear combination of p prior observations with a random error factor that determines a variable's future value which can be mathematically expressed as

$$Y_t = C + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} \dots \phi_p Y_{t-p} + \epsilon_t \tag{5}$$

Where, Y_t and ϵ_t are the actual value and random error terms at time t, ϕ_i ($i = 1, 2, 3, 4, \dots$) represents model parameters, and c is a constant. The order of the model is a positive integer p. The MA(q) model incorporates a dependent variable for previous errors which can be expressed as

$$Y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \theta_4 \epsilon_{t-4} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{6}$$

Where μ indicates the series mean, θ_j ($j = 1, 2, 3 \dots q$) indicates model parameters, and q is the model's order [37].

The ARIMA model may be stated in its basic form as

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \dots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{7}$$

where y'_t represents differenced series (it can be more than one); $\phi_1, \phi_2, \dots, \phi_p$ are the coefficients of AR terms and $\theta_1, \theta_2 \dots \theta_q$ are the coefficients of moving average term.

The ARIMAX model is the generalization of the ARIMA model. It enhances the ARIMA model's capabilities by including several meteorological information such as temperature, humidity, precipitation and other meteorological conditions in time series modelling. An ARIMAX model is formed as follows:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t} + \eta_t \tag{8}$$

where, y_t represents the response variable for the given time series; $x_{1,t} \dots x_{k,t}$ are the features or exogenous variables of the time series that potentially explain y_t ; η_t is the regression model error that describes the ARIMA model (Eq 3) [38].

XGBoost model

The XGBoost model is a supervised machine learning technique and an emerging machine learning method for time series forecasting in recent years [39, 40]. It uses an improved generalized gradient boosting library that can rapidly assess the value of all input attributes [41–43]. Boosting is a technique that combines hundreds of low-accuracy prediction models into a single high-accuracy model by frequently integrating the models under tolerable parameter values

[44–46]. The objective function of the model is as follows:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \quad (9)$$

Where y_i stands for the observed values, x_i stands for the feature vector, n stands for the sample size, m stands for the number of iterations, and f_m stands for the error in m iterations. l stands for the loss function, which computes the deviation between the label and the forecasting in the previous phase as well as the output of the new tree, and Ω stands for the regularization term, which reduces the new tree's output variation [37, 39, 47].

Result

As of 29 January 2022, India had reported the highest total of 41.1 million COVID-19 confirmed cases, resulting in 0.5 million fatalities, whereas Bhutan had reported the lowest COVID-19 confirmed cases and fatalities among SAARC countries (Table 1).

The maximum temperature among the SAARC countries varies from -3.38°C (Nepal) to 47.01°C (India) and the minimum temperature varies from -26.17°C (Afghanistan) to 31.67°C (India). The highest average maximum temperature was observed in India (32.54°C). The highest level of humidity was observed in Maldives (100%), but the lowest level of humidity was observed in Afghanistan (5.06%). Bangladesh had the highest maximum wind speed at 10M (15.68 m/s), but Bhutan had the lowest (1.62 m/s). The highest surface pressure was observed in Bangladesh (101.88 kPa) and the lowest surface pressure was observed in Nepal (66.08 kPa) as illustrated in Fig 2.

The time series figure depicts the trend of COVID-19 confirmed cases from the onset of the disease to January 29, 2022, in each SAARC country. Daily confirmed cases in Bangladesh, Nepal and Pakistan fluctuated at different periods including a highly upward trend. The pattern in Afghanistan and Sri Lanka was remarkably similar, indicating a downward tendency. Overall, Bhutan and Maldives had a comparatively lower rate of COVID-19 transmission than other SAARC countries (Fig 3). The cross-correlation between COVID-19 confirmed cases and meteorological variables was formed at 0 to 30 lags. Only positive lags were considered to explore the influence of meteorological factors on the COVID-19 transmission in a certain period [48]. In Afghanistan, the maximum and minimum temperature at lag 0 showed a significant relationship with COVID-19 confirmed cases. The only maximum temperature at lag 4 showed a significant relationship in India. Maximum wind speed showed a significant relationship in Bangladesh at lag 9 and Maldives at lag 13 days. Relative humidity at a lag of 26

Table 1. Summary statistics of COVID-19 confirmed cases and deaths for SAARC countries till January 29, 2022.

Countries	Daily confirmed cases				Daily Deaths			
	Min	Max	Mean \pm SD	Total	Min	Max	Mean \pm SD	Total
Afghanistan	0	3243	228.50 \pm 397.09	161,306	0	159	10.49 \pm 18.95	7405
Bangladesh	0	16,230	2559 \pm 3163.93	1,773,149	0	264	40.88 \pm 53.22	28,329
Bhutan	0	205	6.57 \pm 19.39	4566	0	1	0.005 \pm 0.08	4
India	0	533,035	56,214 \pm 85,201.04	41,092,522	0	4529	665.80 \pm 904.03	486,718
Maldives	0	2813	192.30 \pm 390.21	133,288	0	10	0.40 \pm 0.98	274
Nepal	0	10,052	1287.20 \pm 1937.83	947,394	0	619	15.90 \pm 39.42	11,703
Pakistan	0	12,073	2011 \pm 1758.87	1,417,991	0	313	41.49 \pm 38.87	29,248
Sri Lanka	0	11,366	829.80 \pm 1237.04	609,047	0	334	20.98 \pm 42.64	15,400

Min: Minimum; Max: Maximum; SD: Standard Deviation.

<https://doi.org/10.1371/journal.pone.0273319.t001>

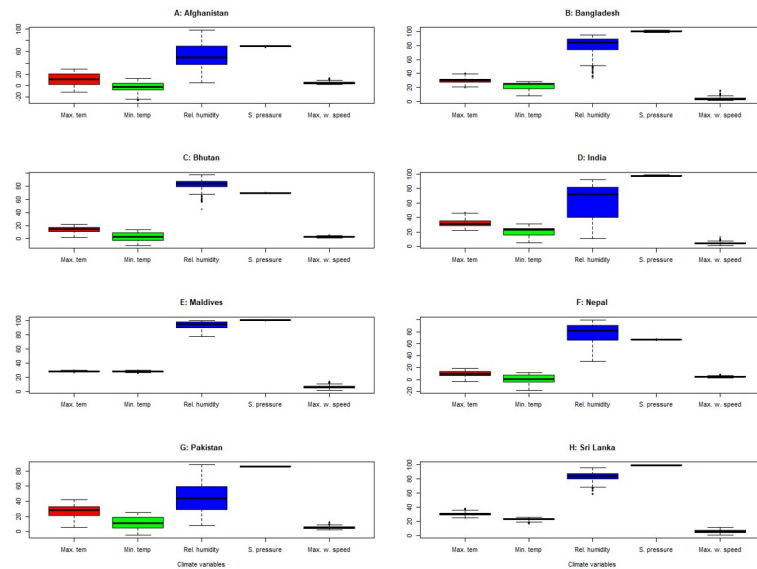


Fig 2. Boxplot of meteorological variables for SAARC countries. Max. tem: Maximum temperature; Min. temp: Minimum temperature; Rel. humidity: Relative humidity; S. pressure: Surface pressure; Max. w. speed: Maximum wind speed.

<https://doi.org/10.1371/journal.pone.0273319.g002>

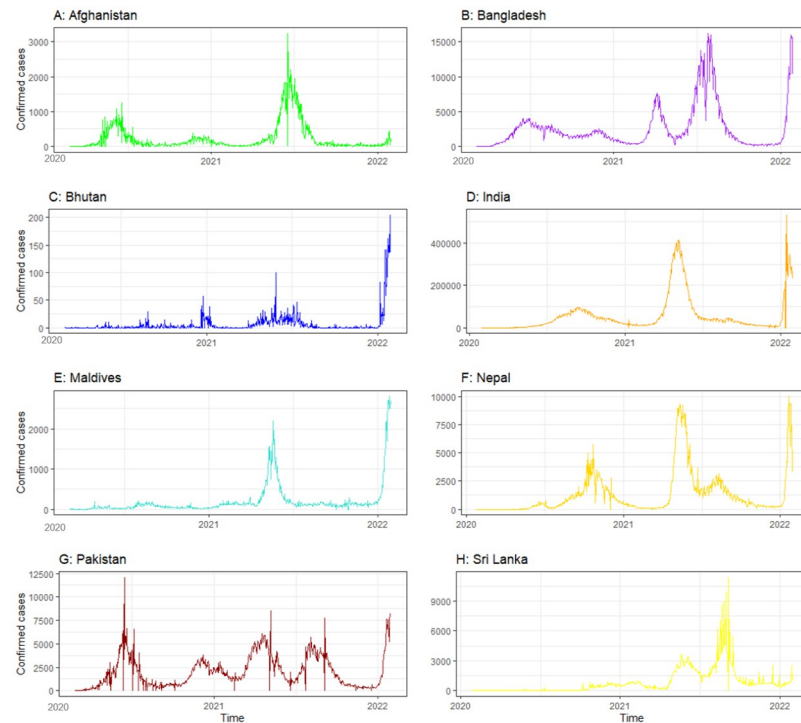


Fig 3. Time series plot showing the trend of COVID-19 confirmed cases for SAARC countries.

<https://doi.org/10.1371/journal.pone.0273319.g003>

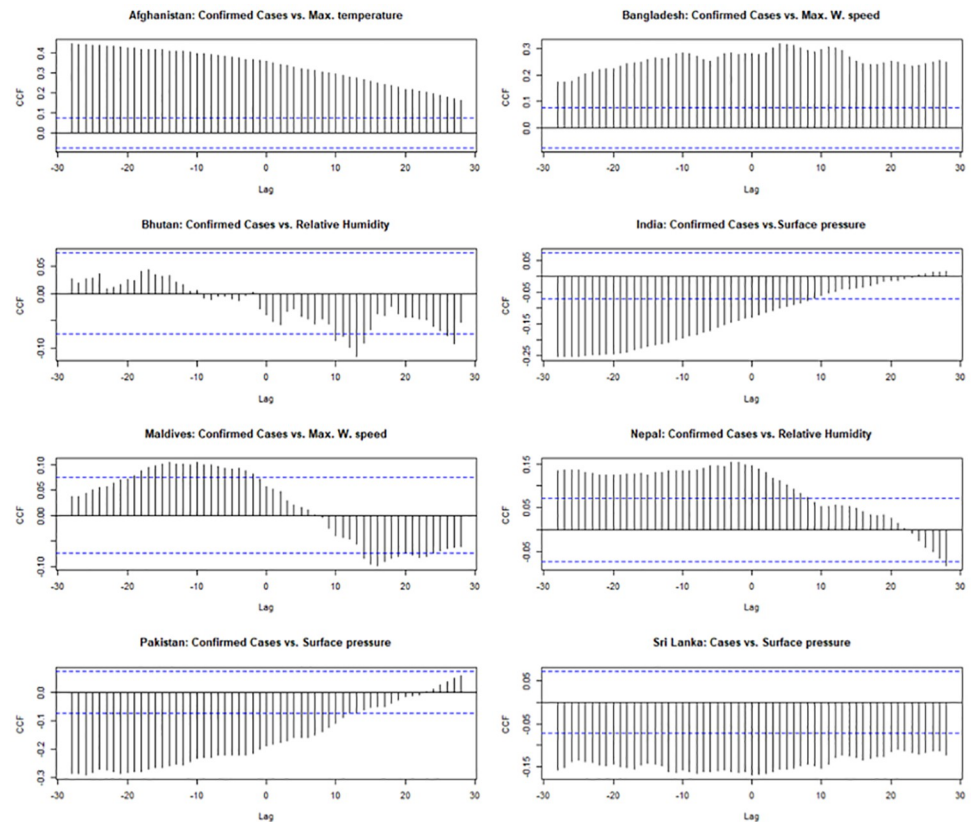


Fig 4. Cross-correlation between COVID-19 confirmed cases and meteorological variables in SAARC countries. Max. temperature: Maximum temperature; Max. W. speed: Maximum Wind speed; CCF: Cross-Correlation Function. <https://doi.org/10.1371/journal.pone.0273319.g004>

days in Bhutan and lag of 10 days in Nepal showed a significant correlation with COVID-19 confirmed cases. Surface pressure showed a significant correlation with COVID-19 confirmed cases in India at lag of 9 days, in Sri Lanka at lag of 13 days and in Pakistan at lag of 28 days (Fig 4).

The aforementioned meteorological factors were used as covariates in ARIMAX model at different lags to determine their influence on COVID-19 confirmed cases. For example, in Afghanistan, the maximum and minimum temperature at lag 0 was used as covariates for building the ARIMAX model. Similarly for Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka, the lagged variables were used as covariates and the influence of those variables on the disease was shown in Table 2.

Table 2 depicts the minimum temperature with a lag of 0 (i.e., same day) in Afghanistan ($\beta = -8.93$, 95% CI: -14.30, -3.56) negatively impact the transmission of COVID-19 cases. The maximum temperature with a lag of 4 days in India ($\beta = 0.18$, 95% CI: 0.01, 0.35) and with a lag of 0 (i.e., same day) in Afghanistan ($\beta = 11.91$, 95% CI: 4.77, 19.05) had a positive influence on the transmission of COVID-19 confirmed cases. Maximum wind speed with a lag of 9 days in Bangladesh ($\beta = -53.89$, 95% CI: -93.45, -14.32) and a lag of 13 days in Maldives ($\beta = -4.24$, 95% CI: -8.31, -0.18) negatively impacts the transmission of COVID-19 confirmed cases. Relative humidity with a lag of 10 days in Nepal ($\beta = -4.84$, 95% CI: -9.20, -0.48) and at a lag of 26 days in Bhutan ($\beta = -0.12$, 95% CI: -0.22, -0.02) negatively impacts COVID-19 confirmed cases. Surface pressure positively impacts COVID-19 confirmed cases in Pakistan ($\beta = 25.77$,

Table 2. Estimated parameters with 95% confidence intervals of significant meteorological factors of ARIMAX models.

Factors	Afghanistan	Bangladesh	Bhutan	India	Maldives	Nepal	Pakistan	Sri Lanka
	ARIMAX (3,1,0)	ARIMAX (0,1,0)	ARIMAX (5,1,0)	ARIMAX (2,1,0)	ARIMAX (1,1,0)	ARIMAX (1,1,0)	ARIMAX (4,1,0)	ARIMAX (5,1,0)
Min. temperature (0)	-8.93* (-14.30, -3.56)							
Max. temperature (0)	11.91* (4.77, 19.05)							
Max. temperature (4)				0.18* (0.01, 0.35)				
Max. W. speed (9)		-53.89* (-93.45, -14.32)						
Max. W. speed (13)					-4.24* (-8.31, -0.18)			
Rel. humidity (10)						-4.84* (-9.20, -0.48)		
Rel. humidity (26)			-0.12* (-0.22, -0.02)					
Surface pressure (9)				-1.91* (-3.75, -0.06)				
Surface pressure (13)								411.63* (49.04, 774.23)
Surface pressure (28)							25.77* (7.85, 43.69)	

Max. temperature: Maximum temperature; Min. temperature: Minimum temperature; Rel. humidity: Relative humidity; Max. W. speed: Maximum Wind speed; * indicates significance at 5% level.

<https://doi.org/10.1371/journal.pone.0273319.t002>

95% CI: 7.85, 43.69) with a lag of 28 days and Sri Lanka ($\beta = 411.63$, 95% CI: 49.04, 774.23) with a lag of 13 days. Moreover, surface pressure with a lag of 9 days in India ($\beta = -1.91$, 95% CI: -3.75, -0.06) negatively impacts the transmission of COVID-19 confirmed cases. The detailed result about the influence of meteorological factors on COVID-19 transmission is presented in [Table 2](#).

The average value of the error measures in the XGBoost model is lower than the ARIMAX models for all the SAARC countries ([Fig 5](#)). Hence, in our study, it was found that XGBoost performs better in predicting COVID-19 confirmed cases in most of the SAARC countries. The detailed procedure of ARIMAX and XGBoost model fitting for COVID-19 confirmed case prediction is presented in [S1 File](#).

Discussion

This study predicted the effect of meteorological factors on the transmission of COVID-19 confirmed cases in SAARC countries. In South Asia, Bangladesh experiences subtropical monsoon weather, with annual average temperatures hovering from 26 to 36°C [49]. Afghanistan experiences hot, dry summer and chilly winter. In summer, the highest temperature in the country reaches up to 50°C, whereas in winter it is -25°C [50]. India has two distinct climate conditions: tropical monsoon weather and tropical wet and dry weather [51]. In Pakistan, there has a wide range of typical temperatures, from 2°C to 38°C [52]. There are distinct rainy and dry seasons in Sri Lanka’s tropical climate. The coastal regions of Sri Lanka get year-round temperatures of 28°C whereas the highland regions experience lower, more moderate

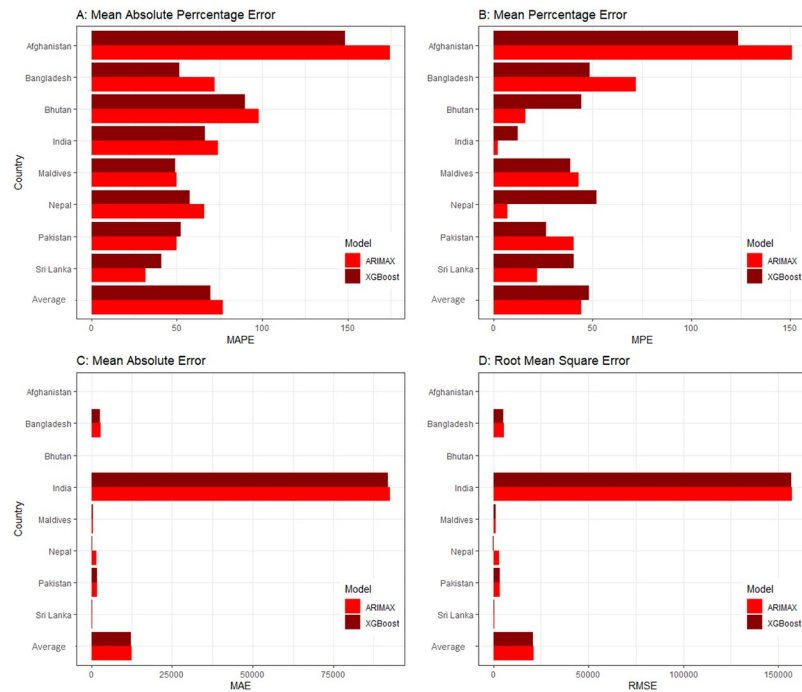


Fig 5. Performance metrics of the ARIMAX and XGBoost models for predicting COVID-19 confirmed cases in SAARC countries.

<https://doi.org/10.1371/journal.pone.0273319.g005>

temperatures of 16°C to 20°C [53]. The vast elevational differences in Bhutan result in a diversified climate [54]. The climate of Nepal varies according to altitude: subtropical with a rainy season in the southern flat strip, moderate in the low mountains, and chilly in the Himalayan peaks [55].

The study found that the meteorological factors have both positive and negative influences on the transmission of COVID-19 confirmed cases. For instance, the maximum temperature had a positive influence on the transmission of COVID-19 confirmed cases in Afghanistan and India which is similar to some previous studies in the EU [56]. This study also found a negative impact of minimum temperature on COVID-19 transmission which is in line with some previous studies conducted in China and the USA [10, 16, 57]. But some previous studies conducted in Spain and China claimed that temperature had no impact on COVID-19 transmission [58, 59]. We also found that relative humidity had a negative influence on the transmission of COVID-19 cases in Bhutan and Nepal which is in line with some previous studies [10, 21, 22, 60]. We found that surface pressure had a positive influence in Pakistan and Sri Lanka as well as negative impact on the COVID-19 confirmed cases in India which is also in line with some previous studies [61, 62]. It was also stated by some studies that surface pressure had no impact on COVID-19 transmission [25]. This study also found a statistically significant association of maximum wind speed with COVID-19 confirmed cases in Bangladesh and Maldives which is similar to a previous study [63]. This study didn't find any statistically significant association of precipitation with COVID-19 confirmed cases while previous studies examined that this is associated with the transmission of COVID-19 confirmed cases [64].

This paper evaluated the applicability of two popular models ARIMAX and XGBoost for predicting the COVID-19 incidence in SAARC countries. The models showed promising results in terms of predicting the time series without the assumptions that traditional

epidemiological models require. Machine learning models, as an alternative to epidemiological models, showed potential for COVID-19 prediction. Considering the availability of only a small amount of training data, it is expected that machine learning will be developed further as the basis for, or a component of, future COVID-19 outbreak prediction models. The XGBoost model is an uptrend machine learning technique in time series modelling. The novelty of our study is that we predicted COVID-19 confirmed cases with ARIMAX model and a data-driven eXtreme Gradient Boosting algorithm using the significant meteorological factors as covariates. The XGBoost technique offers several benefits in terms of model forecasting, including the non-requirement of data preprocessing, complete feature extraction and high prediction accuracy. This study used this technique to predict COVID-19 confirmed cases using the significant meteorological variables. Because it features a higher late trimming penalty than a standard Gradient boosting decision tree, which reduces the likelihood of overfitting [65]. The XGBoost model was developed by adjusting its different parameters. We selected the most traditional ARIMAX as a baseline for our study. The study found that the XGBoost model performs better in predicting the COVID-19 confirmed cases in most of the SAARC countries. In this study, we used these models as a case study to find the significant relationship between meteorological factors and the COVID-19 transmission and compared the prediction accuracy of those models to determine the best model. The findings of this study are also useful for all other COVID-19-affected countries similar to SAARC countries.

Limitations

This study used ARIMAX and XGBoost predictive models to investigate the impact of meteorological factors on COVID-19 transmission in SAARC countries. Therefore, a limitation of the study is that, for example, socioeconomic, demographic, healthcare facilities, human mobilities and population density covariates were not incorporated in this study. These covariates might be correlated with the COVID-19 transmission and should be investigated further based on the data availability.

Conclusion

This study shows the machine learning-based XGBoost model performs better than the ARIMAX model in predicting the COVID-19 incidence in SAARC countries. In the absence of effective COVID-19 prevention strategies, our proposed predictive model is useful for government authorities, researchers and planners to put forward strategic plans to control the spread of COVID-19. It is, therefore, possible for other nations to adopt the suggested frameworks and prevention measures. By exploring the influence of meteorological risk factors on COVID-19 transmission, we can help people to establish a more accurate early warning system and recommend developing appropriate environmental policies to control the spread of the pandemic.

Supporting information

S1 Table. Time series COVID-19 data of SAARC countries along with meteorological variables from the onset of COVID-19 incidence to January 29, 2022.

(XLSX)

S1 Text. R codes.

(TXT)

S1 File.

(DOCX)

Acknowledgments

We are grateful to research student Md. Selim Mondol for his generous support and help.

Author Contributions

Conceptualization: Md. Siddikur Rahman.

Data curation: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Formal analysis: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Funding acquisition: Arman Hossain Chowdhury.

Investigation: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Methodology: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Project administration: Md. Siddikur Rahman.

Resources: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Software: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Supervision: Md. Siddikur Rahman.

Validation: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Visualization: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Writing – original draft: Md. Siddikur Rahman, Arman Hossain Chowdhury.

Writing – review & editing: Md. Siddikur Rahman, Arman Hossain Chowdhury.

References

1. Acter T, Uddin N, Das J, Akhter A, Choudhury TR, Kim S. Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. *Sci Total Environ.* 2020; 730: 138996. <https://doi.org/10.1016/j.scitotenv.2020.138996> PMID: 32371230
2. Rahman MS, Karamehic-Muratovic A, Amrin M, Chowdhury AH, Selim Mondol M, Haque U, et al. COVID-19 Epidemic in Bangladesh among Rural and Urban Residents: An Online Cross-Sectional Survey of Knowledge, Attitudes, and Practices. 2020.
3. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet.* 2020; 395: 470–473. [https://doi.org/10.1016/S0140-6736\(20\)30185-9](https://doi.org/10.1016/S0140-6736(20)30185-9) PMID: 31986257
4. Bogoch II, Watts A, Thomas-Bachli A, Huber C, Kraemer MUG, Khan K. Potential for global spread of a novel coronavirus from China. *J Travel Med.* 2020; 27: 1–3. <https://doi.org/10.1093/jtm/taaa011> PMID: 31985790
5. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nat* 2020 5797798. 2020; 579: 265–269. <https://doi.org/10.1038/s41586-020-2008-3> PMID: 32015508
6. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. [cited 2 Jul 2022]. <https://covid19.who.int/>
7. Nair H, Brooks WA, Katz M, Roca A, Berkley JA, Madhi SA, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *Lancet.* 1917; 378: 1917–1930. [https://doi.org/10.1016/S0140-6736\(11\)61051-9](https://doi.org/10.1016/S0140-6736(11)61051-9)
8. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. 2003; 361.
9. Chin AWH, Chu JTS, Perera MRA, Hui KPY, Yen HL, Chan MCW, et al. Stability of SARS-CoV-2 in different environmental conditions. *The Lancet Microbe.* 2020; 1: e10. [https://doi.org/10.1016/S2666-5247\(20\)30003-3](https://doi.org/10.1016/S2666-5247(20)30003-3) PMID: 32835322

10. Wang J, Tang K, Feng K, Lin X, Lv W, Chen K, et al. Impact of temperature and relative humidity on the transmission of COVID-19: A modelling study in China and the United States. *BMJ Open*. 2021; 11: 1–16. <https://doi.org/10.1136/bmjopen-2020-043863> PMID: 33597143
11. Wu Y, Jing W, Liu J, Ma Q, Yuan J, Wang Y, et al. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Sci Total Environ*. 2020; 729: 139051. <https://doi.org/10.1016/j.scitotenv.2020.139051> PMID: 32361460
12. Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, et al. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Sci Total Environ*. 2020; 724: 138226. <https://doi.org/10.1016/j.scitotenv.2020.138226> PMID: 32408453
13. Qi H, Xiao S, Shi R, Ward MP, Chen Y, Tu W, et al. COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci Total Environ*. 2020; 728: 138778. <https://doi.org/10.1016/j.scitotenv.2020.138778> PMID: 32335405
14. Haque SE, Rahman M. Association between temperature, humidity, and COVID-19 outbreaks in Bangladesh. *Environ Sci Policy*. 2020; 114: 253–255. <https://doi.org/10.1016/j.envsci.2020.08.012> PMID: 32863760
15. Ali Q, Raza A, Saghir S, Khan MTI. Impact of wind speed and air pollution on COVID-19 transmission in Pakistan. *Int J Environ Sci Technol*. 2021; 18: 1287–1298. <https://doi.org/10.1007/s13762-021-03219-z> PMID: 33747099
16. Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, et al. Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. *Sci Total Environ*. 2020; 726: 138513. <https://doi.org/10.1016/j.scitotenv.2020.138513> PMID: 32304942
17. Pal SK, Masum MH. Effects of meteorological parameters on COVID-19 transmission trends in Bangladesh. *Environ Sustain* 2021 43. 2021; 4: 559–568. <https://doi.org/10.1007/S42398-021-00195-5>
18. Sobral MFF, Duarte GB, da Penha Sobral AIG, Marinho MLM, de Souza Melo A. Association between climate variables and global transmission of SARS-CoV-2. *Sci Total Environ*. 2020; 729: 138997. <https://doi.org/10.1016/j.scitotenv.2020.138997> PMID: 32353724
19. Şahin M. Impact of weather on COVID-19 pandemic in Turkey. *Sci Total Environ*. 2020; 728: 138810. <https://doi.org/10.1016/j.scitotenv.2020.138810> PMID: 32334158
20. Guo C, Bo Y, Lin C, Li HB, Zeng Y, Zhang Y, et al. Meteorological factors and COVID-19 incidence in 190 countries: An observational study. *Sci Total Environ*. 2021; 757: 143783. <https://doi.org/10.1016/j.scitotenv.2020.143783> PMID: 33257056
21. Bashir MF, Ma B, Bilal, Komal B, Bashir MA, Tan D, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci Total Environ*. 2020; 728: 138835. <https://doi.org/10.1016/j.scitotenv.2020.138835> PMID: 32334162
22. Tosepu R, Gunawan J, Effendy DS, Ahmad LOAI, Lestari H, Bahar H, et al. Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Sci Total Environ*. 2020; 725: 138436. <https://doi.org/10.1016/j.scitotenv.2020.138436> PMID: 32298883
23. Guo X-J, Zhang H, Zeng Y-P. Transmissibility of COVID-19 and its association with temperature and humidity. 2020.
24. Xie J, Zhu Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci Total Environ*. 2020; 724: 138201. <https://doi.org/10.1016/j.scitotenv.2020.138201> PMID: 32408450
25. Hossain MS, Ahmed S, Uddin MJ. Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model. *Sci Total Environ*. 2021; 761: 143315. <https://doi.org/10.1016/j.scitotenv.2020.143315> PMID: 33162141
26. Li Y, Chu S, Zhao H, Rong F, Liu C, Zhao S, et al. Rapid Prediction and Evaluation of COVID-19 Epidemic in the United States Based on Feature Selection and Improved ARIMAX Model. *ACM Int Conf Proceeding Ser*. 2021.
27. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, et al. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. *Infect Drug Resist*. 2019; 12: 1011–1020. <https://doi.org/10.2147/IDR.S190418> PMID: 31118707
28. Fox R. Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J Theor Biol*. 2005; 234: 187–199. <https://doi.org/10.1016/j.jtbi.2004.11.031> PMID: 15757678
29. Rahman MS, Safa NT, Sultana S, Salam S, Karamehic-Muratovic A, Overgaard HJ. Role of artificial intelligence-internet of things (AI-IoT) based emerging technologies in the public health response to infectious diseases in Bangladesh. *Parasite Epidemiol Control*. 2022; 18: e00266. <https://doi.org/10.1016/j.parepi.2022.e00266> PMID: 35975103
30. Johns Hopkins Coronavirus Resource Center. [cited 29 Jan 2022]. <https://coronavirus.jhu.edu/region>

31. Stackhouse P: NASA Langley Research Center (LaRC) POWER Project. Hampton, VA: NASA [National Aeronautics and Space Administration] and LaRC. 2021 [cited 20 Jun 2022]. <https://power.larc.nasa.gov/data-access-viewer/>
32. RStudio: Integrated Development Environment for R RStudio Team. In: RStudio, PBC, Boston, MA (2021) [Internet]. [cited 14 Dec 2021]. <https://www.rstudio.com/>
33. Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. *J Stat Softw.* 2008; 27: 1–22. <https://doi.org/10.18637/JSS.V027.I03>
34. Prajapati S, Swaraj A, Lalwani R, Narwal A, Verma K, Singh G. Comparison of Traditional and Hybrid Time Series Models for Forecasting COVID-19 Cases. 2019; 8.
35. Box G. Box and Jenkins: Time Series Analysis, Forecasting and Control. *A Very Br Aff.* 2013; 161–215. https://doi.org/10.1057/9781137291264_6
36. Kumar N, Susan S. COVID-19 Pandemic Prediction using Time Series Forecasting Models. 2020 11th Int Conf Comput Commun Netw Technol ICCCNT 2020. 2020.
37. Rahman MS, Chowdhury AH, Amrin M. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. *Plos Glob Public Heal.* 2022; 2019: 1–13. <https://doi.org/10.1371/journal.pgph.0000495>
38. Hyndman RJ. AG. Forecasting: principles and practice, 2nd edition. In: OTexts. 2018.
39. Lv CX, An SY, Qiao BJ, Wu W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infect Dis.* 2021; 21: 1–13.
40. Alim M, Ye GH, Guan P, Huang DS, Zhou B Sen, Wu W. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: A time-series study. *BMJ Open.* 2020; 10: 1–8. <https://doi.org/10.1136/bmjopen-2020-039676> PMID: 33293308
41. Zheng Y, Zhu Y, Ji M, Wang R, Liu X, Zhang M, et al. A Learning-Based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics. *Patterns.* 2020; 1: 100092. <https://doi.org/10.1016/j.patter.2020.100092> PMID: 32838344
42. Hu CA, Chen CM, Fang YC, Liang SJ, Wang HC, Fang WF, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. *BMJ Open.* 2020; 10: e033898. <https://doi.org/10.1136/bmjopen-2019-033898> PMID: 32102816
43. Rahman MS, Pientong C, Zafar S, Ekalaksananan T, Paul RE, Haque U, et al. Mapping the spatial distribution of the dengue vector *Aedes aegypti* and predicting its abundance in northeastern Thailand using machine-learning approach. *One Heal.* 2021; 13: 100358. <https://doi.org/10.1016/j.onehlt.2021.100358> PMID: 34934797
44. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet.* 2019; 10: 1–7. <https://doi.org/10.3389/fgene.2019.01077> PMID: 31781160
45. Shrivastav LK, Jha SK. A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. *Appl Intell.* 2021; 51: 2727–2739. <https://doi.org/10.1007/s10489-020-01997-6> PMID: 34764559
46. Babajide Mustapha I, Saeed F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules.* 2016; 21: 1–11. <https://doi.org/10.3390/molecules21080983> PMID: 27483216
47. Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS One.* 2018; 13: 1–13. <https://doi.org/10.1371/journal.pone.0195875> PMID: 29672639
48. Yan L, Wang H, Zhang X, Li MY, He J. Impact of meteorological factors on the incidence of bacillary dysentery in Beijing, China: A time series analysis (1970–2012). *PLoS One.* 2017; 12: e0182937. <https://doi.org/10.1371/journal.pone.0182937> PMID: 28796834
49. Rahman MM, Avtar R, Yunus AP, Dou J, Misra P, Takeuchi W, et al. Monitoring effect of spatial growth on land surface temperature in Dhaka. *Remote Sens.* 2020; 12. <https://doi.org/10.3390/rs12071191>
50. Afghanistan's climate—SCA. [cited 6 Jul 2022]. <https://swedishcommittee.org/afghanistan/climate/>
51. Ravindranath NH, Sukumar R. Climate Change and Tropical Forests in India. *Potential Impacts Clim Chang Trop For Ecosyst.* 1998; 423–441. https://doi.org/10.1007/978-94-017-2730-3_21
52. Pakistan—Climatology | Climate Change Knowledge Portal. [cited 6 Jul 2022]. <https://climateknowledgeportal.worldbank.org/country/pakistan/climate-data-historical>
53. Discover the climate and geography of Sri Lanka. [cited 6 Jul 2022]. <https://www.worldtravelguide.net/guides/asia/sri-lanka/weather-climate-geography/>
54. Bhutan—Climatology | Climate Change Knowledge Portal. [cited 6 Jul 2022]. <https://climateknowledgeportal.worldbank.org/country/bhutan/climate-data-historical>

55. Nepal climate: average weather, temperature, precipitation, when to go. [cited 6 Jul 2022]. <https://www.climatestotravel.com/climate/nepal>
56. Ozyigit A. Understanding Covid-19 transmission: The effect of temperature and health behavior on transmission rates. *Infect Dis Heal.* 2020; 25: 233–238. <https://doi.org/10.1016/j.idh.2020.07.001> PMID: 32674969
57. Mandal CC, Panwar MS. Can the summer temperatures reduce COVID-19 cases? *Public Health.* 2020; 185: 72–79. <https://doi.org/10.1016/j.puhe.2020.05.065> PMID: 32574871
58. Briz-Redón Á, Serrano-Aroca Á. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Sci Total Environ.* 2020; 728: 138811. <https://doi.org/10.1016/j.scitotenv.2020.138811> PMID: 32361118
59. Yao Y, Pan J, Liu Z, Meng X, Wang W, Kan H, et al. No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. *Eur Respir J.* 2020; 55.
60. Ganegoda NC, Wijaya KP, Amadi M, Erandi KKWH, Aldila D. Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany. *Sci Rep.* 2021; 11: 1–16. <https://doi.org/10.1038/s41598-021-90873-5> PMID: 34050241
61. Kumar G, Kumar RR. A correlation study between meteorological parameters and COVID-19 pandemic in Mumbai, India. *Diabetes Metab Syndr Clin Res Rev.* 2020; 14: 1735–1742. <https://doi.org/10.1016/j.dsx.2020.09.002> PMID: 32919321
62. Sarkodie SA, Owusu PA. Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. *Environ Res.* 2020; 191: 110101. <https://doi.org/10.1016/j.envres.2020.110101> PMID: 32835681
63. Islam N, Bukhari Q, Jameel Y, Shabnam S, Erzurumluoglu AM, Siddique MA, et al. COVID-19 and climatic factors: A global analysis. *Environ Res.* 2021; 193: 110355. <https://doi.org/10.1016/j.envres.2020.110355> PMID: 33127399
64. Menebo MM. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. *Sci Total Environ.* 2020; 737: 139659. <https://doi.org/10.1016/j.scitotenv.2020.139659> PMID: 32492607
65. Wu W, Guo J, An S, Guan P, Ren Y, Xia L, et al. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu Province, China. *PLoS One.* 2015; 10: 1–13. <https://doi.org/10.1371/journal.pone.0135492> PMID: 26270814