BMJ Open

A generalized model for predicting the local COVID-19 outbreak around the world based on meteorological conditions

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041397
Article Type:	Original research
Date Submitted by the Author:	09-Jun-2020
Complete List of Authors:	Chen, Biqing; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Liang, Hao ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Yuan, Xiaomin ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Colorectal Surgery Hu, Yingying ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Xu, Miao; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Zhao, Yating ; Nanjing University Zhang, Binfen ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Tian, Fang ; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Zhu, Xuejun ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology, Research Center of Chinese Medicine
Keywords:	Epidemiology < INFECTIOUS DISEASES, EPIDEMIOLOGY, Public health < INFECTIOUS DISEASES, Infection control < INFECTIOUS DISEASES

Title: A generalized model for predicting the local COVID-19 outbreak around the world based on meteorological conditions

Authors: Biqing Chen[1], Hao Liang[2], Xiaomin Yuan[3], Yingying Hu[2], Miao Xu[2], Yating Zhao[4], Binfen Zhang[2], Fang Tian[1], Xuejun Zhu[1,2*]

Affiliations:

[1] Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[2] Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[3] Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[4] School of Atmospheric Sciences, Nanjing University, Nanjing, China

*Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.

Abstract: Background The COVID-19 has become a pandemic worldwide. **Methods** We collected 382,596 records of weather data with four meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility, and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed cases in total) in over 500 areas worldwide from January 20 to April 9. Epidemic data were modeled against weather data to find a model that could best predict the future outbreak. **Results** Significant correlations of the daily new confirmed case counts with the weather 3~7 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average temperature at 7.9 ℃, relative humidity at 70%~80%, wind speed at 4~10 miles / hour, and visibility less than 10 statute miles. A short-term model with these meteorological variables in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-term model using temperature to predict the pandemic in the next week or month was derived. Taken China as a discovery dataset, it was well validated with worldwide data. According to this model, there are five different viral transmission pattern, "restricted', "controlled", "natural", "tropical", "southern". This model's prediction performance correlates with the actual observations best (over 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not much human interference by epidemic prevention measures. **Conclusion** This model can be used for prediction of the future outbreak, and illustrating the effect of epidemic control for a certain area.

Keywords: COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control

Strengths and limitations of this study

- The number of daily new confirmed cases is significantly correlated with the weather 3~7 days ago.
- Average temperature at 7.9 ℃, relative humidity at 70%~80%, wind speed at 4~10 miles / hour, and visibility less than 10 statute miles are the best weather conditions for the spread of SARS-CoV-2.
- A short-term model consisted of four meteorological factors as a weather coefficient to multiply with the extant confirmed cases could predict the new case count in the following three days very well.

- A long-term model with temperature could be used to predict the new case count in the next week or month for a certain area.
- As the prediction model could illustrate the effect of epidemic control for a certain area, China and other early outbreak countries have effectively reduced more than 50% of the potential outbreak.

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 has spread all over the world and has great social and economic impact worldwide (*1,2*). It exhibits high human-to-human transmissibility compared to other coronavirus like SARS (3). As of April 28 in 2020, the reported cumulative confirmed case count reached over three million and reported death is over 0.21 million globally (4). It would be crucial to predict the future trend of COVID-19 outbreak ahead, in order to make proper prevention and control strategies accordingly in time.

Besides population mobility and human-to-human contact, meteorological conditions have been suggested to be involved in the transmission of droplet-mediated viral diseases (*5,6*). As droplets carrying the coronavirus can travel in gaseous clouds as far as eight metres and stay suspended in the air for hours (7), the suspending time and viability of the coronavirus outside body would be largely affected by the environment. Wind speed could affect the suspending time of droplets, while visibility and humidity reflect the amount of particles in the air, determining the coronavirus payload. Temperature affects virus's viability in the environment. As SARS-CoV-2 is enveloped, it might be more vulnerable to adverse conditions like high temperature.

The impact of weather on epidemiology has been mentioned in human's history. The ancient Chinese had a theory called "Five Movement and Six Weather" to study climate change and its relationship with human health. According to this theory, plague is likely to outbreak in 2020, in consistency with the pandemic. Currently, there are a few studies on preprint servers discussing the relationship of temperature and humidity with the pandemic, but none is systematical investigation or proposes validated practical model for prediction (*8-13*).

Herein, this study intends to investigate the relationship between meteorological factors and epidemic transmission rate on a world scale. Four meteorological variables, i.e., average temperature, relative humidity, wind speed, and visibility, were collected as well as the confirmed case counts daily for 80 days for over 500 areas around the world. Five time delays of the epidemic situation from the exposure day were considered and compared to determine the most reasonable time delay. A multivariate polynomial regression model with meteorological factors as a "weather coefficient" of the extant case count was established in a discovery Chinese dataset, and then validated by worldwide data. Five transmission modes, indicating different levels of epidemic control, were revealed by this model. In this view, this model can not only predict future outbreak, but also be used to evaluate the effect of epidemic prevention measures for a certain area.

Materials and Methods

<u>Epidemiological data</u>

Epidemiological data were collected from the World Health Organization (WHO) (*4*), European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (*14*). The daily new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data were obtained for every Chinese city or district as a discovery dataset, while those for 21 Italian provinces and all the other nations were taken as replication datasets (Supplementary Materials).

<u>Weather data</u>

We obtained hourly values of meteorological observations and geographic factors (latitude and elevation) from the Integrated Surface Database of USA National Centers for Environmental Information (*15*). Temperature, dew point, wind speed, and visibility were collected, and relative humidity was calculated accordingly (Supplementary Materials). Daily data were calculated by averaging the hourly data for each variable in each day.

Statistical modeling

The number of daily new confirmed cases was taken as a dependent variable. Four meteorological variables, namely, average temperature, wind speed, visibility, and relative humidity, and the extant confirmed case counts were taken as independent variables. Considering that there is a latency stage from the day one get infected to the day being confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were collected needs to be taken into consideration. As it is reported that the latency period for COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection were taken into consideration, that is, weather data and extant confirmed cases count data were collected on the day, three days before, seven days before, 3~7 days before, and 14 days before collecting the epidemiological data. A Loess regression interpolation approach was adopted to visually identify the relationship between meteorological variables and confirmed new case counts. Basic statistics and modeling was conducted in R 3.5.1 (*16*).

Model validation and application

The best fitted model was validated in the replication datasets by correlating the observed actual epidemiological data with the predicted values from the model in the datasets. We used these fitted models to calculate a predicted value for case counts for each studied site, and then compared this predicted value with the real observed case counts by calculating a Pearson's correlation coefficient between them.

Patient and Public Involvement

No specific patients were included in the current study.

Results

3~7 days delay of the outbreak from exposure

The average temperature, relative humidity, wind speed, and visibility ranges in the replication datasets were similar to the discovery dataset (see Supplementary Results for detailed datasets description). Regression interpolation showed that the weather 3~7 days ago was correlated with the confirmed new case counts in a most reasonable manner, as well as weather one week ago. The effects of temperature and relative humidity on the new confirmed case count exhibited a bell-shaped trend, while wind speed and visibility were negatively correlated with the new case count (Fig. 1). It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure under certain adverse weather might exhibit its effect after 3~7 days.

<u>Contribution of single meteorological factor to the outbreak</u>

To elucidate the contribution of each meteorological factor to the case counts, we first performed single-factor regression modeling for each meteorological variable in the discovery dataset. Temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility was fitted into a linear model. It was found that visibility was correlated with the outbreak best, followed by temperature, relative humidity, and wind speed. The new case count was significantly negatively correlated with visibility (Spearman's correlation ρ = -0.14, $p < 0.001$), temperature (ρ = -0.14, $p < 0.001$), and wind speed (ρ = -0.07, $p < 0.001$), but positively correlated with relative humidity (ρ = 0.05, $p < 0.001$). For Wuhan data, a model only with temperature as a parameter could already explained 45% of the variance in the epidemic data ($p = 4\times10^{-4}$), while wind speed and visibility could explain over 25% of the variance. According to the fitted single-factor models for Wuhan, SARS-CoV-2 transmission reaches a peak when mean temperature is 7.9 ℃ (Fig. 2A) , relative humidity is 77.6% (Fig. 2B), and wind speed is 5.2 miles / hour (Fig. 2C). The effects of geographic factors such as latitude and elevation, and the pure influence from the extant case count were further investigated (Fig. S1), illustrating that

COVID-19 mainly outbreaks at latitude 30°~50°(Fig. S1A) and elevation < 500 metre (Fig. S1B). New confirmed case count was positively correlated with the extant confirmed case count (Fig. S1C).

Short-term prediction model

We further combined different meteorological variables into a complex short-term model, and took the extant confirmed case count as a base for meteorological factors to multiply. The short-term model fitted was as follows:

$$\text{New Case Count} = \left(-0.13 \times T^2 + 1.45 \times T - 608 \times RH^2 + 974 \times RH - 0.23 \times SPD^2 + 0.89 \times SPD - 7.45 \times VSB - 200\right) \times \alpha \times \text{Extant Case Count}$$

where T is temperature in ℃, RH is relative humidity in percentage, SPD is wind speed in miles per hour, VSB is visibility in statute miles, α is a site-specific constant, with a default of 0.002. All parameters take the means of values 3~7 days before the day new case count is evaluated.

In this model, all the four meteorological variables are added together in their proper forms to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate of SARS-CoV-2, and thus influences the number of people that catch infection from the extant confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a multiplicative factor α in the equation, which seems site-related and determines the strength of the "weather coefficient" on viral transmission. The value of the multiplicative factor α is determined by first substitute the general value 0.002 into the formula, and then plot the observed case count vs. predicted one, to find the extent of underestimation or overestimation.

Substitute data from the past two months, a good prediction performance was obtained for this short-term model, with the predicted values significantly correlated to the observed ones for most areas (Fig. 3). However, only the extant confirmed case count data could not predict the new case count 3~7 days later as well as the weather-combined model did (Fig. S2).

Different modes of viral transmission illustrated by the model

The observed versus predicted data exhibited different correlation patterns for different areas, meaning different viral transmission modes, which may indicate the effect of epidemic control for certain area.

Data from Chinese top-affected cities were not very well predicted and obviously overestimated by this model with the default multiplicative factor α (Pearson's correlation coefficient $r = 0.31$, $p < 0.001$; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention measures at the beginning of the pandemic. This viral transmission mode suggested by the not well correlated prediction pattern is called "restricted".

For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the predicted outbreak was well correlated with the actual observations at the beginning when the extant confirmed cases were not in very large numbers, but the prediction deviates from the observation as the confirmed cases increase, in detail, there's large overestimation of prediction ($r_{\text{Wuhan}} = 0.47$, $p = 0.02$, $r_{\text{Italy}} = 0.86$, $r_{\text{Japan}} = 0.71$, $r_{\text{Iran}} = 0.64$, $p < 0.001$, $r_{\text{Korea}} = 0.06$, $p = 0.68$; Fig. 3B). It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15, the day when shelter hospitals had been put into use for seven days (the average latency period for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention and control taken in these areas is effective (so-called "controlled" mode). The number of cases had been decreased by 82% for Wuhan, over 95% for Korea, Japan, and Italy, and 52% for Iran at most due to epidemic control (the largest gap between prediction and observation).

For most European and American countries, the predicted outbreak was linear correlated with the observed data very well ($r_{\text{France}} = 0.96$, $r_{\text{United States}} = 0.92$, $r_{\text{United Kingdom}} = 0.92$, $r_{\text{Spain}} = 0.84$, $r_{\text{Germany}} = 0.73$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without much man-made epidemic prevention and control measures. Estimation of daily new case counts by this short-term model performed very well for European countries, while this model underestimated the outbreak in the United States.

Although the weather is not suitable for tropical areas, the viral transmitted in natural mode, manifested as good linear correlation between the prediction and the observation ($r_{\text{India}} = 0.95$, $r_{\text{Singapore}} = 0.84$, $r_{\text{Thailand}} = 0.82$, $p < 0.001$; Fig. 3D), with just relatively small daily new case counts compared to temperate regions.

Countries in the southern atmosphere displayed similar pattern as the "controlled" with large overestimation by the model when the confirmed cases increase, leading to not good prediction performance ($r_{\text{Australia}} = 0.28$, $p = 0.04$, $r_{\text{South Africa}} = 0.03$, $p = 0.87$; Fig. 3E). It might be due to the effect of epidemic prevention measures in these countries.

<u>Long-term simplified model</u>

For long-term prediction, another simplified model with average temperature as a weather factor was derived as follows:

$$\text{new case count} = (\,-\,0.14 \times \mathrm{T}^2 + 0.93 \times \mathrm{T} + 100) \times \beta \times \text{Extant Case Count}$$

where T is temperature in ℃, β is a site-related constant, with a default of 0.003. All parameters take values 7 days before the day new case count is evaluated.

With the model, the prediction performance was still good ($r = 0.64$ in the current datasets, $p < 0.001$; Fig. 3F). The long-term simplified prediction model also showed five prediction-observation correlation patterns, indicating different modes of viral transmission, for the studied areas.

Discussion

This research discovers nonlinear dose-response relationship for meteorological factors, in consistency with previous studies (10). Predictions of COVID-19 outbreak scale by the models were well correlated with the observations around the world, suggesting the importance of weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many respiratory infectious diseases, such as influenza, is dependent upon temperature and relative humidity (*5,6*). Recent published papers on preprint servers have reported roles of temperature and absolute humidity in the COVID-19 transmission, but their conclusions are diverse (*8-13*). In contrast to the findings by Cai et al (8), this study suggests significant impact of mean temperature on the daily new case count, indicating a need for sufficient time delay between exposure and confirmation for weather to exhibit its effect. In contrary to other two studies (*9,10*), this research suggests that there is a relatively not wide temperature and humidity ranges for the pandemic. There is an optimal temperature for SARS-CoV-2 at 7.9 ℃, which is colder than that suggested by Bu et al (12) but in consistency with the estimation by Wang et al (10); and most areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al suggested (12). It is of notice that different from other viral respiratory diseases such as influenza, high relative humidity is better for SARS-CoV-2 to spread, suggesting that a sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more important for the spread than the effect of dry air on the human immune system. Different from other studies (13), this study also finds significant involvement of wind speed, in a quadric manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In addition, the current study discovered that visibility was significantly negatively correlated with new case count and played a more important role in viral spread than humidity did. New case count decreases rapidly when visibility is high than 13 statute miles, indicating that caution should be taken if visibility drops below 10 statute miles.

In the prediction model, there is a constant multiplicative factor which determines the strength of the weather coefficient on the epidemic transmission. It seems site-specific, as adjusting it could make the prediction for one site very close to the observation. This constant might reflect the influence of epidemic management and control measures. Various degrees of isolation for various areas around the world lead to different degrees of weather effect. When evaluate the prediction performance by the short-term model and the long-term model, they both exhibit different prediction-observation correlation patterns, suggesting that changes in the degree of epidemic control and isolation policy would lead to deviation from the original prediction and thus different prediction-observation correlation patterns. Therefore, by plotting the predicted versus observed new case counts and adjusting the multiplicative factor (α and β), it would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the observed case counts dropped dramatically from the predictions for Wuhan seven days after their shelter hospitals were put in use, suggesting the importance and necessity of building shelter hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to natural transmission mode. Another thing worth attention is that although the weather in tropical areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore, these tropical areas should still be on the alert against future outbreak of COVID-19.

Although those cases with travel history to China or indicated by the World Health Organization as "imported case only" were excluded in this study, leaving the world data most likely local transmitted, it's difficult to separate the imported cases from local transmission very well in practice, which might explain the not excellent correlations of predictions with observations. Furthermore, the relationship of weather and COVID-19 could be complex, since the human immune system has an innate seasonal rhythm, and the immune system could also be affected by weather *vice versa*, for example, dry air would reduce the amount of mucus on the airway mucosa, and thus increasing the probability of viral invasion, while wet air provides droplets for virus to adhere.

In summary, this study has found significant correlations with the COVID-19 epidemic trend for not only temperature and humidity, but also wind speed and visibility. It proposed a comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version and a long-term version. The short-term version uses the combination of four meteorological factors as a "weather coefficient" of the extant case count in the past week and can be used to predict epidemic situation in the future three days; the short-term version uses average temperature as the "weather coefficient" seven days ago and can predict the outbreak in one month if combined with weather forecast. This model is easy to use for predicting the COVID-19 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case count for the future couple of days or month. This model will be very helpful for local governments to make timely policies on epidemic control, for instance, the allocation of medical equipments such as ventilators and medical resources such as hospitals, beds and health-care workers, according to the prediction results.

Acknowledgments: We thank Dr. Zhisheng Huang for advices on data collecting and processing, Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on modeling.

Funding: the Priority Academic Program Development of Jiangsu Higher Education Institutions – the third period (NO.035062002003c).

Author contributions: BC, YZ and XZ design and interpret the reported analyses and results; HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse data; BC drafted the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical support; XZ supervised the research.

Competing interests: Authors declare no competing interests.

Data and materials availability: Weather data and epidemiological data is all obtained from public databases. Detailed modeling results are available upon request by emailing Biqing Chen, bq_chen@qq.com.

References:

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.

2. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 2020;382(13):1199–207.

3. Chan JFW, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020;395(10223):514–23.

4. World Health Organization. 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (29 March 2020, date last accessed).

5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the United States. Amer J Epidemiol. 2012;176(suppl. 7):S114–22.

6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathol. 2007;3:e151.

7. Bourouiba L. Turbulent gas clouds and respiratory pathogen emissions potential implications for reducing transmission of COVID-19. JAMA. 2020;E1–2.

8. Cai Y, Huang Sr. T, Liu Sr. X, Xu Sr. G. The effects of "Fangcang, Huoshenshan, and Leishenshan" makeshift hospitals and temperature on the mortality of COVID-19. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3 (29 March 2020, date last accessed).

9. Bannister-Tyrrell1 M, Meyer A, Faverjon C, Cameron A. Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.03.18.20036731 (29 March 2020, date last accessed).

10. Wang M, Jiang A, Gong L, et al. Temperature significantly change COVID-19 transmission in 429 cities. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1 (29 March 2020, date last accessed).

11. Luo W, Majumder MS, Liu D, et al. The role of absolute humidity on transmission rates of the COVID-19 outbreak. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.02.12.20022467v1 (29 March 2020, date last accessed).

12. Bu J, Peng D-D, Xiao H, et al. Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.02.13.20022715v1 (29 March 2020, date last accessed).

13. Oliveiros B, Caramelo L, Ferreira NC, Caramelo F. Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. Preprint at Med RXIV. 2020. https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1 (29 March 2020, date last accessed).

14. COVID-19/2019-nCoV Infection Time Series Data Warehouse. 2020. https://github.com/BlankerL/DXY-COVID-19-Data (9 April 2020, date last accessed).

15. Integrated Surface Database of USA National Centers for Environmental Information. 2020. ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/ (9 April 2020, date last accessed).

16. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2018. https://cran.r-project.org/bin/windows/base/old/3.5.1/ (9 January 2020, date last accessed).

FIGURE LEGENDS

Fig. 1. Loess regression interpolation of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for Wuhan city. Five time delay of the confirmation day (when epidemiological data were correlated) from the exposure day (when weather data was correlated) are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, the replication_Italy dataset is provincial data in Italy, the replication_world dataset is national data around the world excluding China, Italy, India, Australia, and South Africa.

Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the

"tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern atmosphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves with 95% confidence intervals in shadow are illustrated for each dataset. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

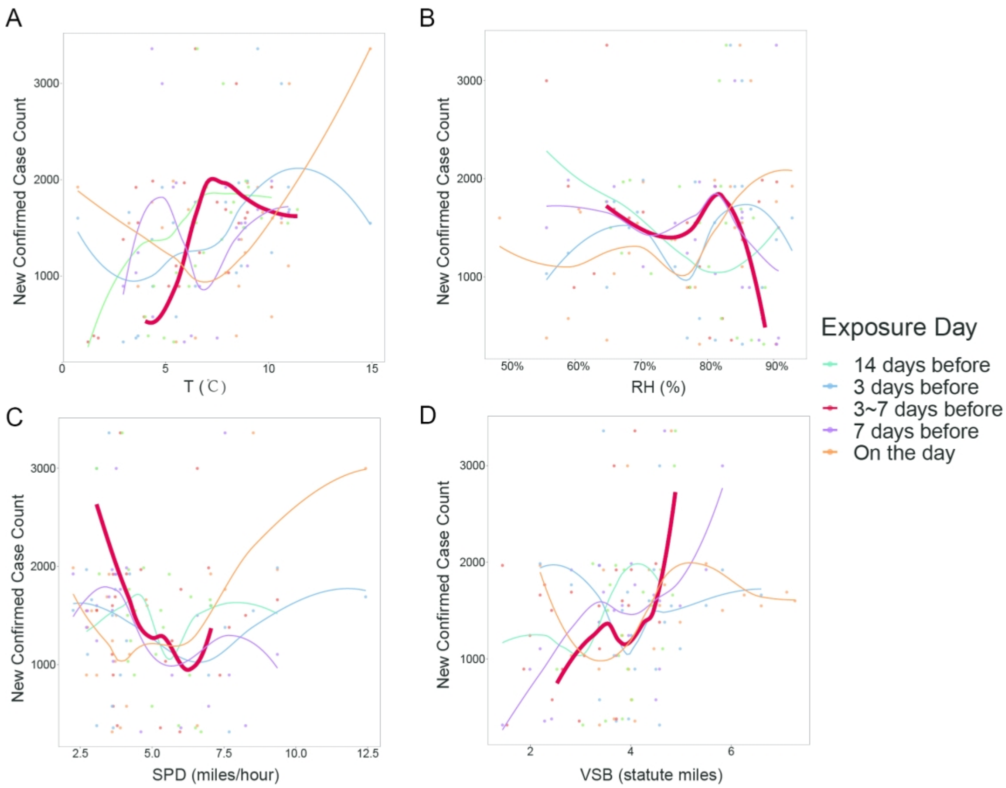


Fig 1. Loess regression interpolation of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for Wuhan city. Five time delay of the confirmation day (when epidemiological data were correlated) from the exposure day (when weather data was correlated) are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

168x132mm (300 x 300 DPI)

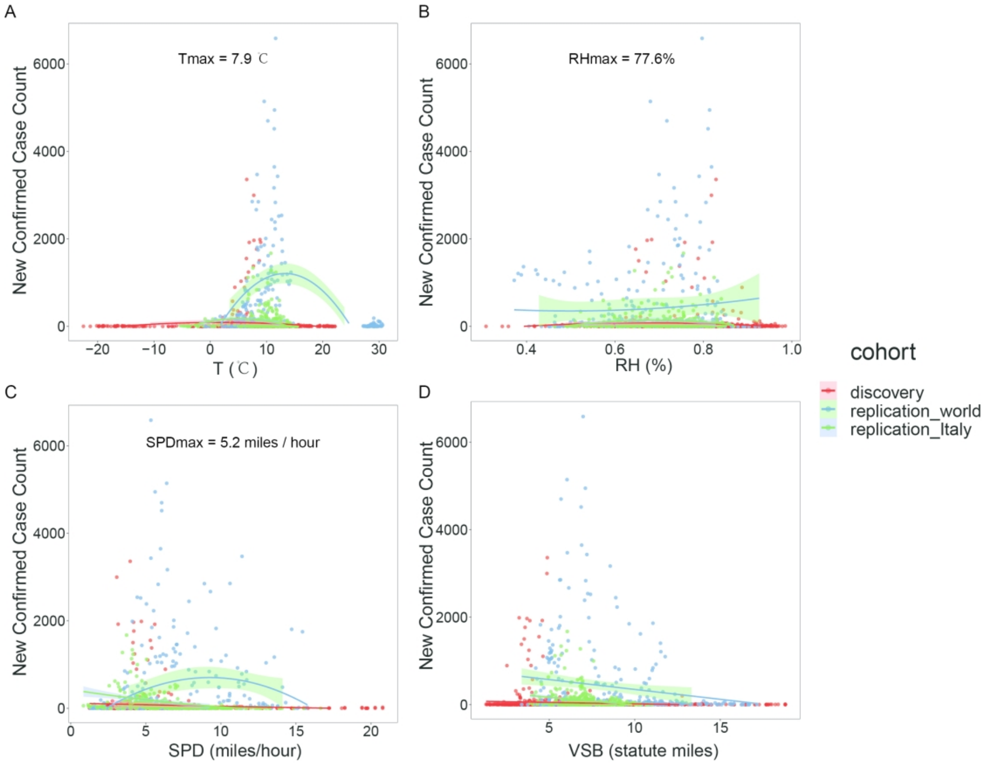


Fig 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, the replication_Italy dataset is provincial data in Italy, the replication_world dataset is national data around the world excluding China, Italy, India, Australia, and South Africa.

241x183mm (300 x 300 DPI)

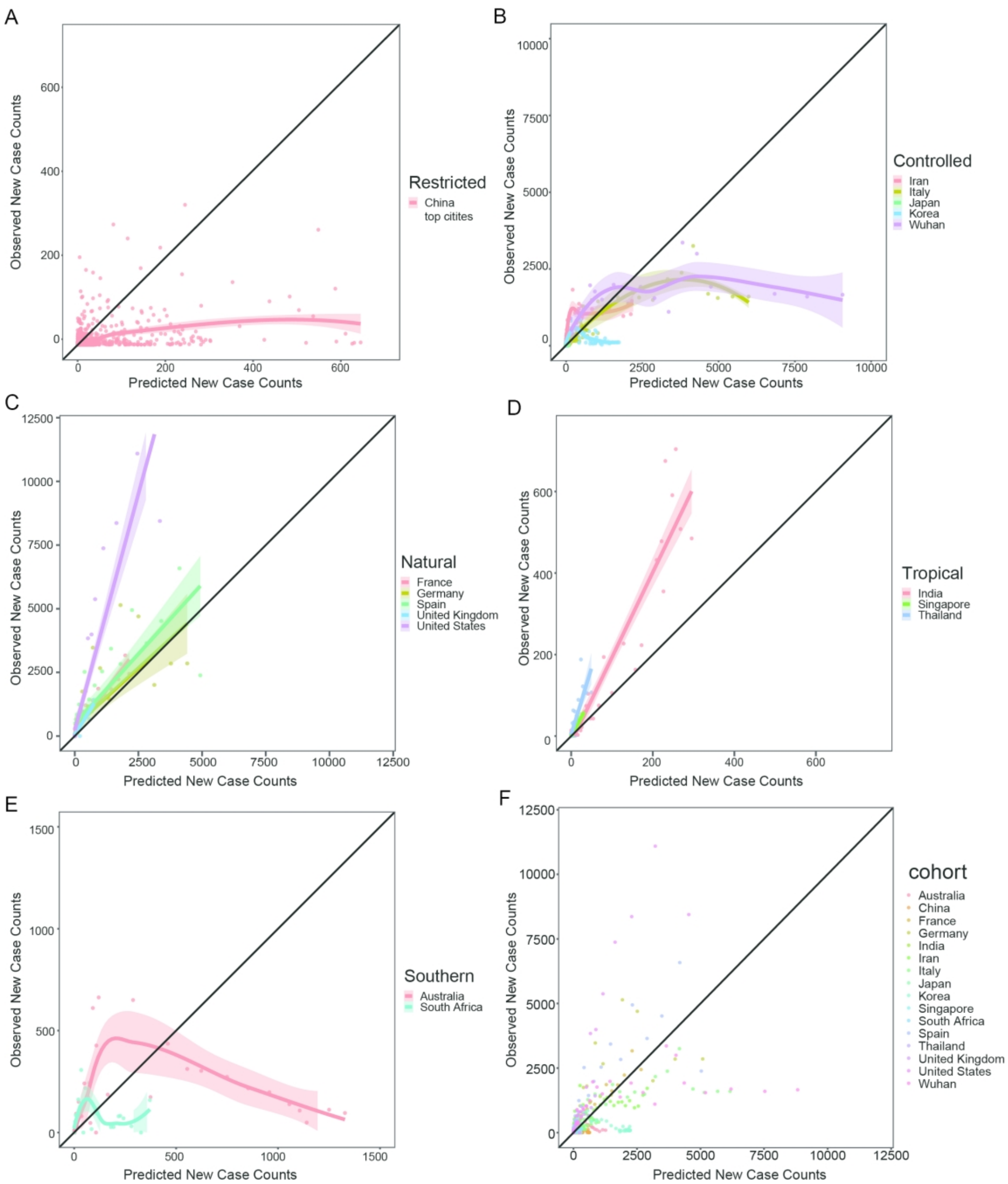


Fig 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern atmosphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves with 95% confidence intervals in shadow are illustrated for each dataset. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

203x240mm (300 x 300 DPI)

Supplementary Materials and Methods

Epidemiological data

Considering the potential confounding effect, only cities with no less than 50 cumulative confirmed cases in one month and without official reports of large imported cases were taken as the discovery dataset. The countries with high COVID-19 incidence except China, namely, United States, United Kingdom, Germany, France, Italy, Spain, Iran, Korea, Japan, and southern atmosphere countries Australia and South Africa, and tropical countries India, Thailand, and Singapore, were selected for replication representing the world's situation.

We scrutinized WHO's situation reports to rule out these countries with only imported cases, and only collected the confirmed cases with possible or confirmed local transmission (i.e., without recent travel history to China).

For Wuhan city, there was a shortage of test kits at the beginning of the pandemic, which would make confirmed case counts much lower than the actual data, thus, we discarded epidemic data before January 28th, the day when domestic test kits have been approved, produced in large quantities, and were available for Wuhan hospitals. As there was a cut down problem for the extant confirmed case count on February 20th for Wuhan, when modeling with the extant confirmed case count, only data before February 20th were used.

Weather data

Temperature and dew point displayed in Fahrenheit were transformed into Celsius forms, and relative humidity was calculated from temperature and dew point using the following formula for each time point:

$$\mathrm{RH} = \begin{cases} e^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T < 0 \\ 10^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T \geq 0 \end{cases}$$

where RH is the relative humidity, D is the dew point in degrees Celsius, T is the temperature in degrees Celsius, and *e* is the base of the natural log.

For each city with epidemiological data, the meteorological station in that city or that was closest to the latitude and longitude coordinates of the city center was chosen. For a city with more than one meteorological stations, the one nearest to the city center was chosen. For a province with epidemiological data, the meteorological station in the capital city of that province was chosen. For a country with only national wide epidemiological data, weather data were averaged across all the meteorological observatories in the cities where outbreak was officially reported. Latitude and elevation for the meteorological observatories were also collected.

<u>Statistical modeling</u>

At first, each meteorological variable was plotted against the confirmed new case counts for the Wuhan dataset. Only one city Wuhan was chosen for illustrating the time delay effect because it is the first city to have an outbreak of COVID-19, there was none reported imported cases for Wuhan, which might obscure the correlation between weather and virus transmission. After choosing the appropriate time delay, data from the discovery dataset were fitted into generalized linear model or non-linear model (basically polynomial models) according to the indentified relationship by Loess regression and knowledge of droplet-mediated viral diseases. Each of the four meteorological variables was fitted into models solely, and then all variables were combined together to compose a comprehensive coefficient that was multiplied with the extant confirmed case counts. All the models were compared with each other to find a best-fitted model with the best fitness. Fitness was evaluated according to log-likelihood, Akaike information criterion, and Bayesian Information Criterion. The final equation supposed that all the meteorological variables composed a coefficient which was multiplied by the extant confirmed case counts on the exposure day, and then derived the new confirmed case counts on the test day.

Supplementary Results

Datasets description

Only Chinese cities with monthly confirmed cases over 50 were included in the discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in all the other Chinese cities. We suppose that it might be due to abrupt large supplement of virus test kits on that day. In order to reduce the potential contamination of modeling by this outlier, we substituted the counts on that day by four, that was 13,436/4=3,359, which was still the largest number but not deviated from the dataset too much. There were also two oddly large new confirmed case counts for Lombardy, which were discarded from the subsequent analysis. Except the outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to 2,997, the average temperature ranged -23.54℃ ~ 22.85℃, the wind speed ranged

1.33 ~ 26 miles per hour, visibility ranged 0.425 ~ 110 statute miles to nearest tenth, and relative humidity ranged 31.4% ~ 100%.

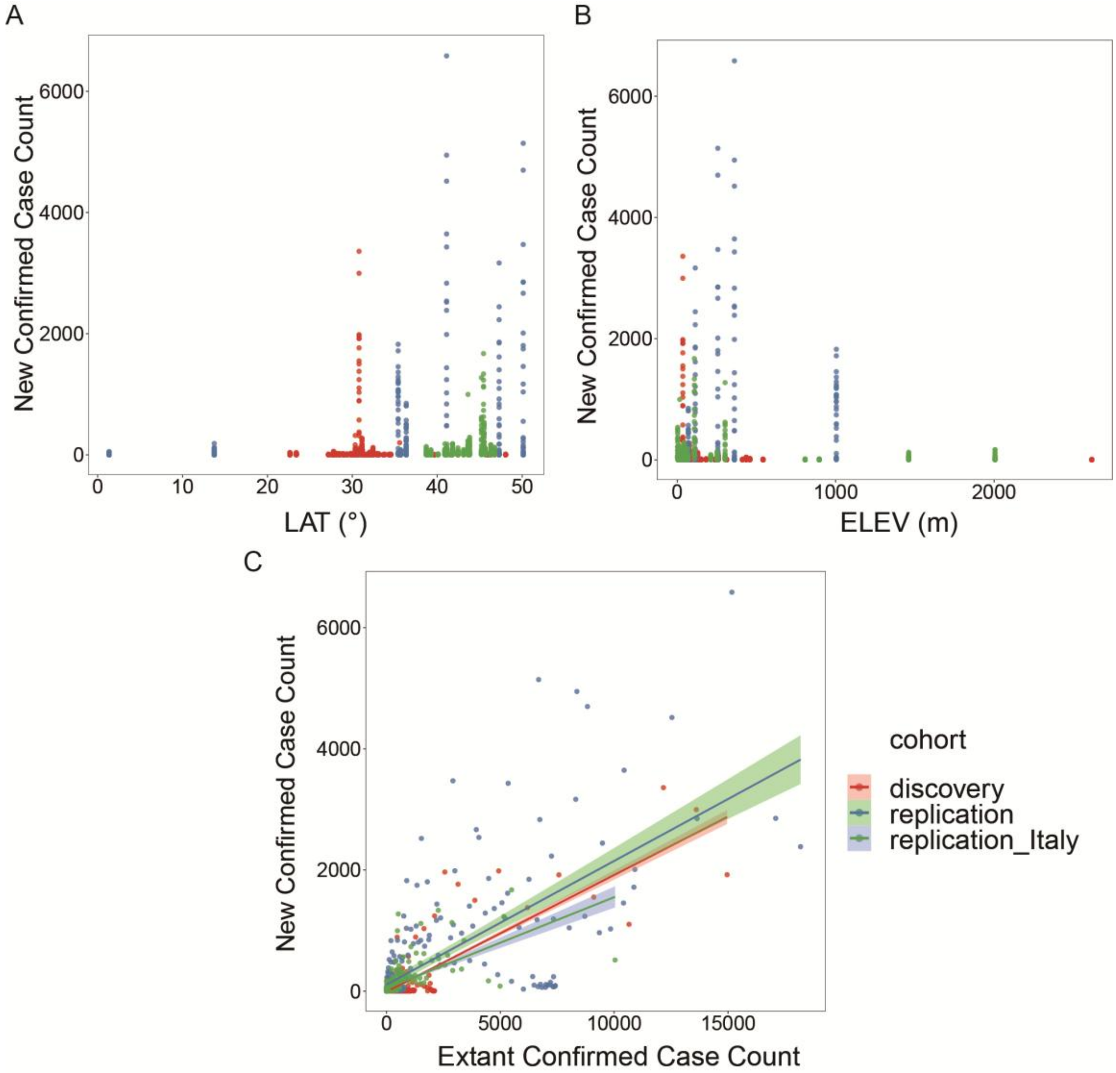


Fig. S1. Scatterplots of new confirmed case count to (A) latitude, (B) elevation, and (C) the extant confirmed case count, for all the studied sites.

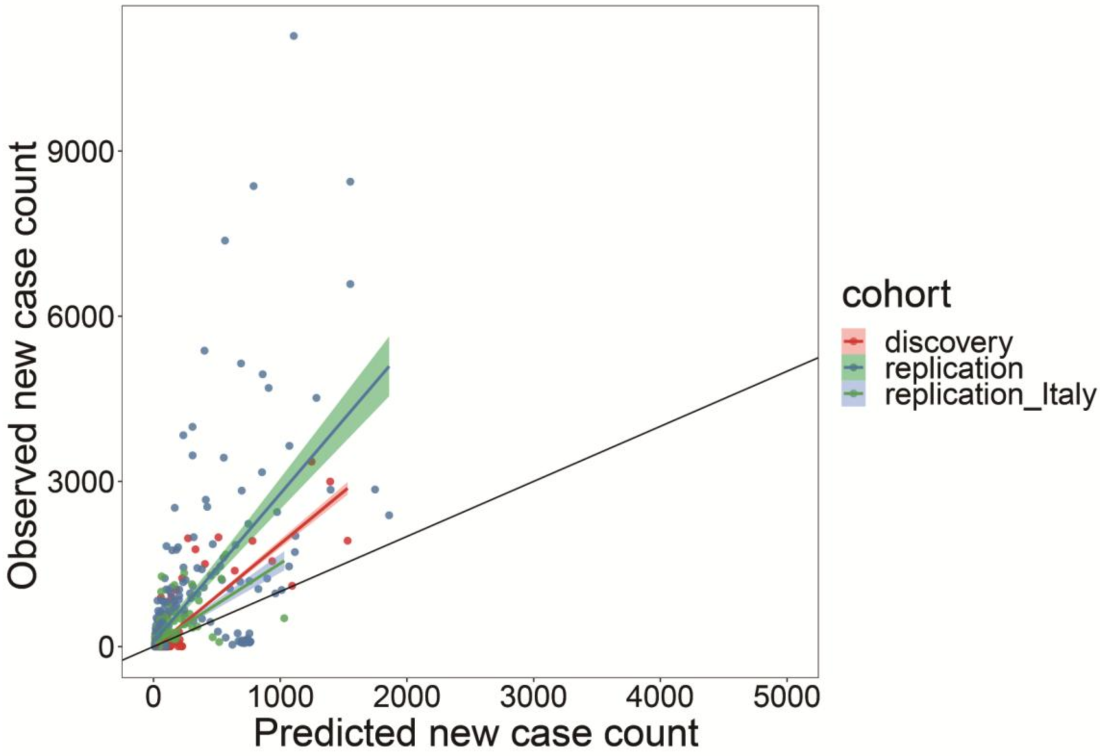


Fig. S2. The observed daily new case counts verse the predicted values by only the extant confirmed case count are illustrated for all cohorts. Linear regression interpolation curves with 95% confidence intervals in shadow are illustrated for each dataset. The black solid line represents that the observed values are equal to the predicted ones.

BMJ Open

Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041397.R1
Article Type:	Original research
Date Submitted by the Author:	14-Aug-2020
Complete List of Authors:	Chen, Biqing; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Liang, Hao ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Yuan, Xiaomin ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Colorectal Surgery Hu, Yingying ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Xu, Miao; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Zhao, Yating ; Nanjing University Zhang, Binfen ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Tian, Fang ; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Zhu, Xuejun ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology, Research Center of Chinese Medicine
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Global health, Infectious diseases, Occupational and environmental medicine, Public health, Qualitative research
Keywords:	Epidemiology < INFECTIOUS DISEASES, EPIDEMIOLOGY, Public health < INFECTIOUS DISEASES, Infection control < INFECTIOUS DISEASES, COVID-19

SCHOLARONE™ Manuscripts

Title: Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Authors: Biqing Chen[1], Hao Liang[2], Xiaomin Yuan[3], Yingying Hu[2], Miao Xu[2], Yating Zhao[4], Binfen Zhang[2], Fang Tian[1], Xuejun Zhu[1,2*]

Affiliations:

[1] Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[2] Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[3] Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[4] School of Atmospheric Sciences, Nanjing University, Nanjing, China

*Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.

Abstract

OBJECTIVE: This study aims to investigate the relationship between daily weather and transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of the COVID-19 spreading rate for a certain area with meteorological factors.

METHODS AND ANALYSIS: We collected 382,596 records of weather data with four meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility, and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed cases in total) in nearly 500 areas worldwide from January 20 to April 9. Epidemic data were modeled against weather data to find a model that could best predict the future outbreak.

RESULTS: Significant correlations of the daily new confirmed case counts with the weather 3~7 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average temperature at 5~15 ℃, relative humidity at 70%~80%, wind speed at 1.5~4.5 meter / second, and visibility less than 10 statute miles. A short-term model with these meteorological variables in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-term model using temperature to predict the pandemic in the next week or month was derived. Taken China as a discovery dataset, it was well validated with worldwide data. According to this model, there are five different viral transmission pattern, "restricted', "controlled", "natural", "tropical", "southern". This model's prediction performance correlates with the actual observations best (over 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not much human interference by epidemic prevention measures.

CONCLUSION: This model can be used for prediction of the future outbreak, and illustrating the effect of epidemic control for a certain area.

Keywords: COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control

Strengths and limitations of this study

- This study investigates the role of daily weather in COVID-19 spread systematically with a comprehensive set of four meteorological factors.
- This research collected a huge amount of data, covering nearly 500 areas worldwide in a long timescale.

- The current study proposes two prediction models on different time scales, a short-term one integrating more detailed meteorological information which is more accurate, and a long-term one with only temperature which is more feasible.
- The influence of weather on virus spread could be confounded by a dozen of manual interventions, such as population mobility and disinfection measures, leading to inaccurate modeling.
- The prediction model (especially the long-term model) might be unsuitable and inaccurate for areas with hot weather.

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 has spread all over the world and has great social and economic impact worldwide (*1,2*). It exhibits high human-to-human transmissibility compared to other coronavirus like SARS (3). As of April 28 in 2020, the reported cumulative confirmed case count reached over three million and reported death is over 0.21 million globally (4). It would be crucial to predict the future trend of COVID-19 outbreak ahead, in order to make proper prevention and control strategies accordingly in time.

Besides population mobility and human-to-human contact, meteorological conditions have been suggested to be involved in the transmission of droplet-mediated viral diseases (*5,6*). As droplets carrying the coronavirus can travel in gaseous clouds as far as eight metres and stay suspended in the air for hours (7), the suspending time and viability of the coronavirus outside body would be largely affected by the environment. Wind speed could affect the suspending time of droplets, while visibility and humidity reflect the amount of particles in the air, determining the coronavirus payload. Temperature affects virus's viability in the environment. As SARS-CoV-2 is enveloped, it might be more vulnerable to adverse conditions like high temperature.

The impact of weather on epidemiology has been mentioned in human's history. The ancient Chinese had a theory called "Five Movement and Six Weather" to study climate change and its relationship with human health. Currently, there are a few studies on preprint servers discussing the relationship of temperature and humidity with the pandemic, but none is systematical investigation or proposes validated practical model for prediction (*8-13*).

Herein, this study intends to investigate the relationship between meteorological factors and epidemic transmission rate on a world scale. Four meteorological variables, i.e., average temperature, relative humidity, wind speed, and visibility, were collected as well as the confirmed case counts daily for 80 days from January 20, 2020 to April 9, 2020 for nearly 500 areas around the world, including 428 Chinese cities and areas, 18 Italian provinces, and 13 other countries. Five time point's delay of virus infection from the exposure day were considered and compared to determine the most reasonable time point's delay. A multivariate polynomial regression model with meteorological factors as a "weather coefficient" of the existing confirmed case count was established in a discovery Chinese dataset, and then validated by worldwide data. Five transmission modes, indicating different levels of epidemic control, were revealed by this model. In this view, this model can not only predict future outbreak, but also be used to evaluate the effect of epidemic prevention measures for a certain area.

Materials and Methods

Epidemiological data

Epidemiological data were collected from the World Health Organization (WHO) (4), European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (10). The daily new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries, namely, United States, United Kingdom, Germany, France, Italy, Spain, Iran, Korea, Japan, Australia, South Africa, India, Thailand, and Singapore. Considering the potential confounding effect, only Chinese cities with no less than 50 cumulative confirmed cases in one month and without official reports of large imported cases (42 in total) were taken as a discovery dataset, while those for Italian provinces and all the other nations were taken as replication datasets (Supplementary Materials).

Weather data

Four meteorological variables were chosen, air temperature, relative humidity, wind speed, and visibility. Temperature could affect virus viability in the environment. Wind speed could affect the suspending time of virus-attached particles. Relative humidity reflects the amount of droplets in the air. Visibility is influenced by the amount of particles such as dust and air pollutants. These two parameters both affect the amount of mediator for the virus to stay in the air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of meteorological observations and geographic factors (latitude and elevation) from the Integrated Surface Database of USA National Centers for Environmental Information (11). Daily data were calculated by averaging the hourly data for each variable in each day.

Statistical modeling

The number of daily new confirmed cases was taken as a dependent variable. Four meteorological variables, namely, average temperature, wind speed, visibility, and relative humidity, and the existing confirmed case counts were taken as independent variables. Considering that there is a latency stage from the day one get infected to the day being confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were collected needs to be taken into consideration. As it is reported that the latency period for COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection were taken into consideration, that is, weather data and existing confirmed cases count data were collected on the day, three days before, seven days before, 3~7 days before, and 14 days before collecting the epidemiological data.

At first, each meteorological variable was fitted into a bunch of single-factor models (either generalized linear model or polynomial model) through non-linear least squares (NLS) modeling using Wuhan dataset under the assumption of 3~7 days exposure delay. The relationship between each meteorological variable and confirmed new case count (linear or quadric) was identified based on model fitness (log-likelihood, Akaike information criterion, Bayesian Information Criterion, etc.) and common knowledge of droplet-mediated viral diseases. Second, the time delay effect was investigated in the Wuhan dataset through Loess regression interpolation and NLS modeling with the previously identified relationship for each meteorological variable. The most possible time delay identified was taken for subsequent analyses.

The contribution of each meteorological factor was investigated with the Wuhan dataset through Spearman's correlation test. Single-factor models were fitted into NLS models again with data from the discovery dataset (all Chinese cities with monthly confirmed cases over 50) under the assumption of previously determined relationship and pre-defined time delay, to determine the exact coefficients accompanied with each meteorological factor and to find out the most suitable environmental condition for SARS-CoV-2. Then, two final prediction models (short-term model and long-term model) were developed using the discovery dataset with the previously determined coefficients. The prediction model supposed that all the meteorological variables added together to compose a coefficient which was multiplied by the existing confirmed case count on the exposure day, and then derived the new confirmed case counts on the test day. The short-term model took all four variables, while the long-term model only considered temperature as it is easy to be forecasted. There was a constant coefficient for the total equation, that was multiplied by the existing confirmed case count. Its default value was obtained by model fitting in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was investigated with all datasets covering the world's top cities and areas. The correlation of existing confirmed case counts with newly confirmed case counts was also investigated. Basic statistics and modeling was conducted in R 3.5.1 (https://cran.***r***-project.org/).

144 Model validation and application

145 The best fitted model was validated in the replication datasets (Italian city-level data and 146 other nation-level data) by correlating the observed actual epidemiological data with the 147 predicted values from the model in the datasets. We used these fitted models to calculate a 148 predicted value for case counts for each studied site, and then compared this predicted value with 149 the real observed case counts by calculating a Spearman's correlation coefficient ρ between them.

150 Patient and Public Involvement

151 No specific patients were included in the current study. Epidemiological data were 152 downloaded from online open-source databases. The public were not involved in the planning 153 and design of the study.

154 **Results**

155 The Weather's influence on SARS-CoV-2 transmission displays 3~7 days time delay

156 The ranges of average temperature, relative humidity, wind speed, and visibility in the replication 157 datasets were similar to those in the discovery dataset (see Supplementary Results for detailed 158 datasets description). To investigate whether the influence of meteorological factors is linear or 159 quadric, both linear and non-linear square modeling were performed under different relationship 160 assumptions to compare model fitness statistics using the Wuhan dataset with a 3~7 days delay of 161 infection. It was suggested that the effect of temperature and wind speed is better depicted as 162 quadric (Table S1), which was also supported by Loess regression interpolation (Fig. 1). The 163 mode for relative humidity and visibility was hard to be determined, as statistics supported both 164 relationships (Table S1). Considering the common knowledge of coronavirus transmission and 165 the trend showed by Loess regression interpolation, relative humidity exerted its impact in a 166 quadric trend while visibility exerted its impact in a linear trend (Fig. 1, Supplementary results).

Furthermore, we investigated the time delay from weather exposure to COVID-19 confirmation with the above determined relationships and NLS modeling using Wuhan dataset. Model fitness statistics showed that the number of confirmed new cases was best correlated with air temperature 3~7 days ago, relative humidity and visibility 7 days ago, and wind speed on the exposure day (Table S2). By comprehensive consideration of all four meteorological variables and the differences between statistics values, the weather 3~7 days ago, as well as weather one week ago, could well predict COVID-19 outbreak. It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure under certain adverse weather might exhibit its effect after 3~7 days.

Contribution of single meteorological factor to the outbreak

In the Wuhan dataset, the new case count was significantly positively correlated with temperature (Spearman's correlation ρ = 0.69, p < 0.001) and visibility (ρ = 0.43, p =0.04), and negatively correlated with wind speed (ρ = -0.45, p = 0.03) and relative humidity (ρ = -0.33, p = 0.12) 3~7 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind speed, visibility, and relative humidity. A model only with temperature as a parameter could already explained 45% of the variance in the epidemic data ($p = 4\times10^{-4}$), while wind speed and visibility could explain over 25% of the variance. To elucidate the contribution of each meteorological factor to the case counts and to determine the exact coefficients, we first performed single-factor regression modeling for each meteorological variable in the discovery dataset with the relationship identified before under the assumption of 3~7 days delay of viral infection. Temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility was fitted into a linear model. According to the fitted single-factor models (Supplementary Results), SARS-CoV-2 transmission reaches a peak when mean temperature is

6.18 ℃ (Fig. 2A), relative humidity is 78.47% (Fig. 2B), and wind speed is 1.88 meter /second (m/s) (Fig. 2C); and its transmission rate decreases with the increase of visibility (Fig. 2D). The effects of geographic factors such as latitude and elevation, and the pure influence from the existing case count were further investigated in the worldwide datasets (Fig. S1), illustrating that COVID-19 mainly outbreaks at latitude 30°~50° (Fig. S1A) and elevation < 500 metre (Fig. S1B). New confirmed case count was positively correlated with the existing confirmed case count (Fig. S1C).

Short-term prediction model

To deduce a practical comprehensive model, all four meteorological variables with their specific coefficients determined by single-factor modeling were added together to form a complex short-term model, and the existing confirmed case count was taken as a base for meteorological factors to multiply (Supplementary Results). This full model was fitted with the discovery dataset to determine the exact values of the constant coefficient in the equation. The best-fitted short-term model was as follows:

$$\begin{aligned}\text{New Case Count} &= \left(-0.11 \times T^2 + 1.40 \times T - 0.058 \times RH^2 + 9.04 \times RH - 1.36 \times SPD^2 + 5.12 \right.\\ &\left. \times SPD - 7.02 \times VSB - 126.66\right) \times \alpha \times \text{Existing Confirmed Case Count}\end{aligned}$$

where T is temperature in ℃, RH is relative humidity in percentage, SPD is wind speed in m/s, VSB is visibility in statute miles, α is a site-specific constant, with a default of 0.001. All parameters take the means of values 3~7 days before the day new case count is evaluated.

In this model, all the four meteorological variables are added together in their proper forms to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate of SARS-CoV-2, and thus influences the number of people that catch infection from the existing confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a multiplicative constant coefficient α in the equation, which seems site-related. This constant coefficient could adjust the strength of the "weather coefficient" on disease transmission. When we substitute replication datasets into this short-term model with the multiplicative constant coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious underestimation of predicted values against real ones was observed although the predicted values correlated with the real ones very well. We supposed it was due to site-specific difference in the multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the pandemic had been controlled early. Thus, we further re-fitted this composed model with all datasets to determine a more accurate value of the multiplicative constant coefficient α, which was 0.001 then. In practical application, we need to first plot the observed case count vs. predicted one with a default α value 0.001, and then examine the extent of underestimation or overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the impact size of "weather coefficient" for a certain site.

Substitute data from the past two months, a good prediction performance was obtained for this short-term model, with the predicted values significantly correlated to the observed ones for most areas (Fig. 3). However, only the existing confirmed case count data could not predict the new case count 3~7 days later as well as the weather-combined model did (Table S3).

Different modes of viral transmission illustrated by the model

The observed versus predicted data exhibited different correlation patterns for different areas, meaning different viral transmission modes, which may indicate the effect of epidemic control for certain area.

Data from Chinese top-affected cities were not very well predicted and obviously overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p <$

0.001; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention measures at the beginning of the pandemic. This viral transmission mode suggested by the not well correlated prediction pattern is called "restricted".

For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the predicted outbreak was well correlated with the actual observations at the beginning when the existing confirmed cases were not in very large numbers, but the prediction deviates from the observation as the confirmed cases increase, in detail, there's large overestimation of prediction ($\rho_{Wuhan} = 0.69$, $\rho_{Italy} = 0.87$, $\rho_{Japan} = 0.80$, $\rho_{Iran} = 0.86$, $p < 0.001$, $\rho_{Korea} = 0.43$, $p = 0.002$; Fig. 3B). It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15, the day when shelter hospitals had been put into use for seven days (the average latency period for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention and control taken in these areas is effective (so-called "controlled" mode). The number of cases had been decreased by 72% for Wuhan, over 95% for Korea, Japan, and Italy, and 37% for Iran at most due to epidemic control (the largest gap between prediction and observation).

For most European and American countries, the predicted outbreak was linear correlated with the observed data very well ($\rho_{France} = 0.96$, $\rho_{United\ States} = 0.93$, $\rho_{United\ Kingdom} = 0.83$, $\rho_{Spain} = 0.97$, $\rho_{Germany} = 0.94$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without much man-made epidemic prevention and control measures. Estimation of daily new case counts by this short-term model performed very well for European countries, while this model underestimated the outbreak in the United States.

Although the weather is not suitable for tropical areas, the viral transmitted in natural mode, manifested as good linear correlation between the prediction and the observation ($\rho_{India} = 0.94$,

$\rho_{Singapore} = 0.66$, $p < 0.001$, $\rho_{Thailand} = 0.56$, $p = 0.001$; Fig. 3D), with just relatively small daily new case counts compared to temperate regions.

Countries in the southern hemisphere displayed similar pattern as the "controlled" with large overestimation by the model when the confirmed cases increase, leading to not good prediction performance ($\rho_{Australia} = 0.79$, $p < 0.001$, $\rho_{South\ Africa} = 0.34$, $p = 0.08$; Fig. 3E). It might be due to the effect of epidemic prevention measures and hot summer weather in these countries.

Long-term simplified model

Long-term prediction depends on weather forecast, which generally reports only average temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as temperature in a short time delay (3~7 days ago), we again performed single-factor regression modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a quadric function (Supplementary Results). This simplified model with average temperature as a weather factor was derived as follows:

$$\text{new case count} = \left(-0.10 \times T^2 + 1.11 \times T + 46.42\right) \times \beta \times \text{Existing Confirmed Case Count}$$

where T is temperature in ℃, β is a site-related multiplicative constant coefficient, with a default of 0.006. All parameters take values 14 days before the day new case count is evaluated.

With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, $p < 0.001$; Fig. 3F). The long-term simplified prediction model also showed five prediction-observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the studied areas. This model could directly predict the newly emerging cases 14 days later, and be used to predict COVID-19 outbreak in the future month by summing up the daily new case count and combining weather forecast (usually available for the future 15 days).

Discussion

This research discovers nonlinear dose-response relationship for meteorological factors, in consistency with previous studies (12). Predictions of COVID-19 outbreak scale by the models were well correlated with the observations around the world, suggesting the importance of weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many respiratory infectious diseases, such as influenza, is dependent upon temperature and relative humidity (*5,6*). Recent published papers on preprint servers have reported roles of temperature and absolute humidity in the COVID-19 transmission, but their conclusions are diverse (*8-13*). In contrast to the findings by Cai et al (8), this study suggests significant impact of mean temperature on the daily new case count, indicating a need for sufficient time delay between exposure and confirmation for weather to exhibit its effect. In contrary to other two studies (*9,10*), this research suggests that there is a relatively not wide temperature and humidity ranges for the pandemic. There is an optimal temperature for SARS-CoV-2 at 6.18 ℃, which is colder than that suggested by Bu et al (14) but in consistency with the estimation by Wang et al (12); and most areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al suggested (14). It is of notice that different from other viral respiratory diseases such as influenza(15)(16), high relative humidity is better for SARS-CoV-2 to spread, suggesting that a sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more important for the spread than the effect of dry air on the human immune system. Different from other studies (17), this study also finds significant involvement of wind speed, in a quadric manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In addition, the current study discovered that visibility was significantly negatively correlated with new case count and played a more important role in viral spread than humidity did (from spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g., dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case count decreases rapidly when visibility is high than 13 statute miles, indicating that caution should be taken if visibility drops below 10 statute miles.

In the prediction model, there is a multiplicative constant coefficient which determines the strength of the weather coefficient on the epidemic transmission. It seems site-specific, as adjusting it could make the prediction for one site very close to the observation. This constant might reflect the influence of a couple of site-specific confounding factors, such as epidemic control measures, sun radiation, and population density. Various degrees of isolation for various areas around the world lead to different degrees of weather effect. When evaluate the prediction performance by the short-term model and the long-term model, they both exhibit different prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of epidemic control and isolation policy would lead to deviation from the original prediction and thus different prediction-observation correlation patterns. Therefore, by plotting the predicted versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the observed case counts dropped dramatically from the predictions for Wuhan seven days after their shelter hospitals were put in use, suggesting the importance and necessity of building shelter hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to natural transmission mode. Another thing worth attention is that although the weather in tropical areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore, these tropical areas should still be on the alert against future outbreak of COVID-19.

Although those cases with travel history to China or indicated by the World Health Organization as "imported case only" were excluded in this study to make the world data most likely local transmitted, it was difficult to separate the imported cases from local transmission very well in practice. It might explain the not excellent correlations of predictions with observations. Furthermore, the relationship of weather and COVID-19 could be complex, since the human immune system has an innate seasonal rhythm, and the immune system could also be affected by weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets for virus to adhere.

There are several limitations of this study. First of all, this prediction model (especially the long-term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter, as the models were derived using Chinese datasets, mainly in the first three months of 2020. The prediction became inaccurate and could be largely deviated from real observations under hot weather, which might explain the obvious bad prediction performance for countries in the southern hemisphere. One explanation for the inaccurate prediction in areas with high temperature could be that SARS-CoV-2 transmission in these areas was mainly not influenced by weather, but in another direct transmission way, such as face-to-face contact or spread in gathering crowd. Second, it seems that the prediction performance drops with the increase in new case count, suggesting that the prediction model might become inaccurate and not suitable for very large new case count. This could be due to that there was less data points with large new case count. Therefore, the model's prediction performance would be better with more data points, especially the large case count points. Third, the short-term prediction model must use all four meteorological factors, while these factors are not always available for any one certain area. Fourth, this study included various areas covering a long period into modeling, thus, there were a bunch of variable confounding factors, such as population mobility and disinfection measures, which were not controlled and thus could impede the model accuracy. Fifth, as we could only obtain country-level epidemiological data, the corresponding meteorological data were obtained for their capital cities, leading to not exact pairing of epidemiological data and meteorological data.

Conclusion

In summary, this study has found significant correlations with the COVID-19 epidemic trend for not only temperature and humidity, but also wind speed and visibility. It proposed a comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version and a long-term version. The short-term version uses the combination of four meteorological factors as a "weather coefficient" of the existing confirmed case count in the past week and can be used to predict epidemic situation in the future three days; the short-term version uses average temperature as the "weather coefficient" seven days ago and can predict the outbreak in one month if combined with weather forecast. This model is easy to use for predicting the COVID-19 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case count for the future couple of days or month. This model will be very helpful for local governments to make timely policies on epidemic control, for instance, the allocation of medical equipments such as ventilators and medical resources such as hospitals, beds and health-care workers, according to the prediction results.

Acknowledgments: We thank Dr. Zhisheng Huang for advices on data collecting and processing, Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on modeling.

Funding: the Priority Academic Program Development of Jiangsu Higher Education Institutions – the third period (NO.035062002003c), the Yizhong Research Fund of Jiangsu Provincial Hospital of Chinese Medicine (Y19066).

Author contributions: BC, YZ and XZ design and interpret the reported analyses and results; HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse the data, drafted, and revised the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical support; XZ supervised the research.

Competing interests: Authors declare no competing interests.

Data and materials availability: Weather data and epidemiological data is all obtained from public databases. Detailed modeling results are available upon request by emailing to Biqing Chen, bq_chen@qq.com.

References:

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.

2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med. 2020;382(13):1199–207.

3. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet [Internet]. Elsevier Ltd; 2020;395(10223):514–23. Available from: http://dx.doi.org/10.1016/S0140-6736(20)30154-9

4. Organization WH. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [Internet]. 2020 [cited 2020 Mar 29]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the United States. Amer J Epidemiol. 2012;176(suppl. 7):S114–22.

6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathol. 2007;3:e151.

7. Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential Implications for Reducing Transmission of COVID-19. JAMA. 2020;E1–2.

8. Cai Y. The Effects of "Fangcang, Huoshenshan [Internet]. Available from: https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3

9. Proverbio AM, Crotti N, Manfredi M, Zani A. Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. medRxiv. 2020;(February).

10. https://github.com/BlankerL/DXY-COVID-19-Data.

11. ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/. 2020.

12. Wang M, Jiang A, Gong L, Lu L, Guo W, Li C, et al. Temperature Significantly Change COVID-19 Transmission in 429 cities. Block Caving – A Viable Altern [Internet]. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1

13. Luo W, Majumder MS, Liu D, Poirier C, Mandl KD, Lipsitch M, et al. The role of absolute humidity on transmission rates of the COVID-19 outbreak. medRxiv [Internet]. 2020;2020.02.12.20022467. Available from: http://medrxiv.org/content/early/2020/02/17/2020.02.12.20022467.abstract%0Ahttps://www.medrxiv.org/content/10.1101/2020.02.12.20022467v1

14. Bu J, Peng D-D, Xiao H, Yue Q, Han Y, Lin Y, et al. Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020. medRxiv [Internet]. 2020;(May):2020.02.13.20022715. Available from: https://www.medrxiv.org/content/10.1101/2020.02.13.20022715v1

15. Peci A, Winter AL, Li Y, Gnaneshan S, Liu J, Mubareka S, et al. Effects of Absolute Humidity, Relative Humidity, Temperature, and Wind Speed on Influenza Activity in Toronto, Ontario, Canada. Appl Env Microbiol. 2019;85(6):e02426-18.

16. Kudo E, Song E, Yockey LJ, Rakib T, Wong PW, Homer RJ, et al. Low ambient humidity impairs barrier function and innate resistance against influenza infection. Proc Natl Acad Sci U S A. 2019;116(22):10905–10.

17. Oliveiros B, Caramelo L, Ferreira NC, Caramelo F. Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. medRxiv [Internet]. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1

FIGURE LEGENDS

Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot

represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

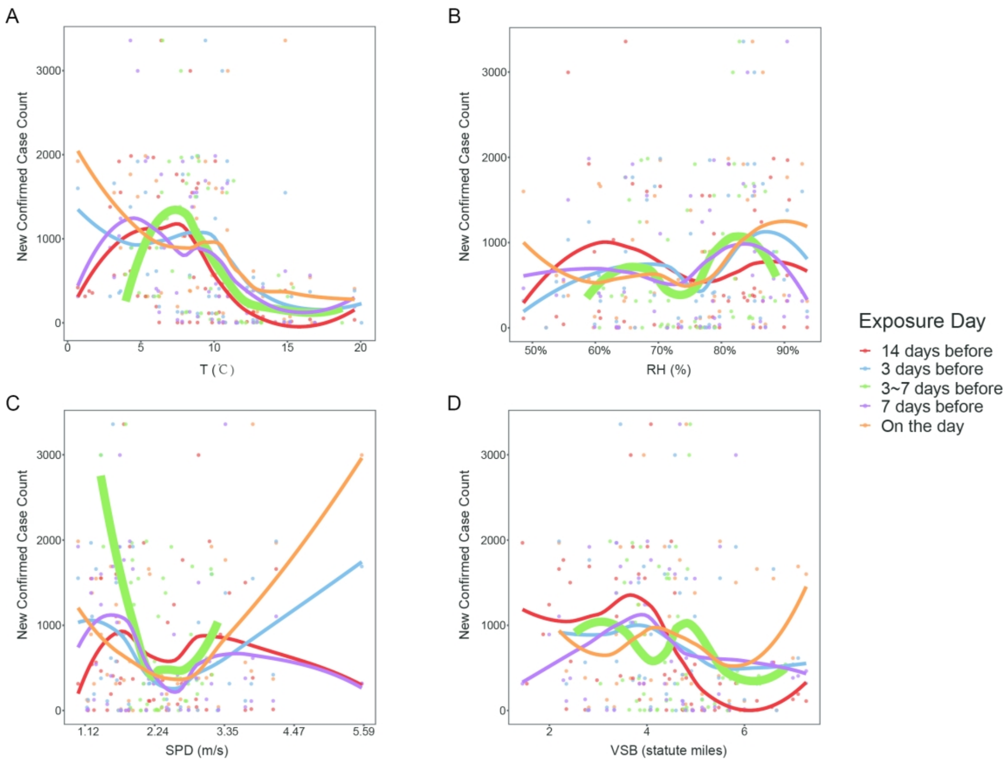


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

207x156mm (300 x 300 DPI)

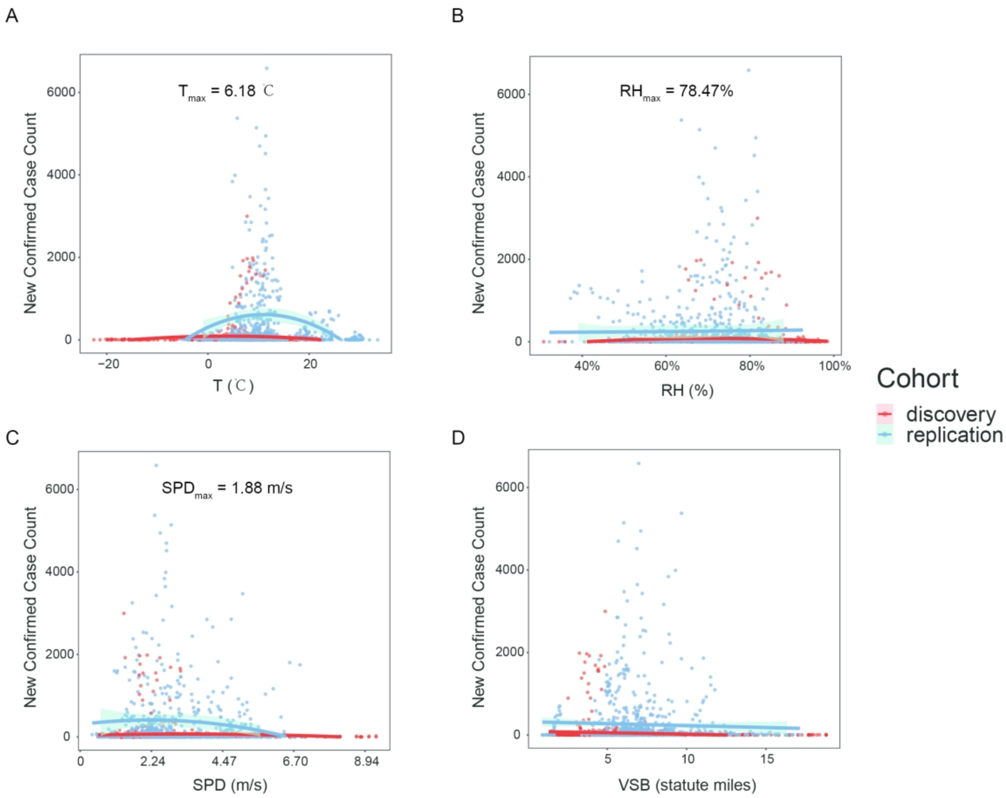


Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

218x172mm (300 x 300 DPI)

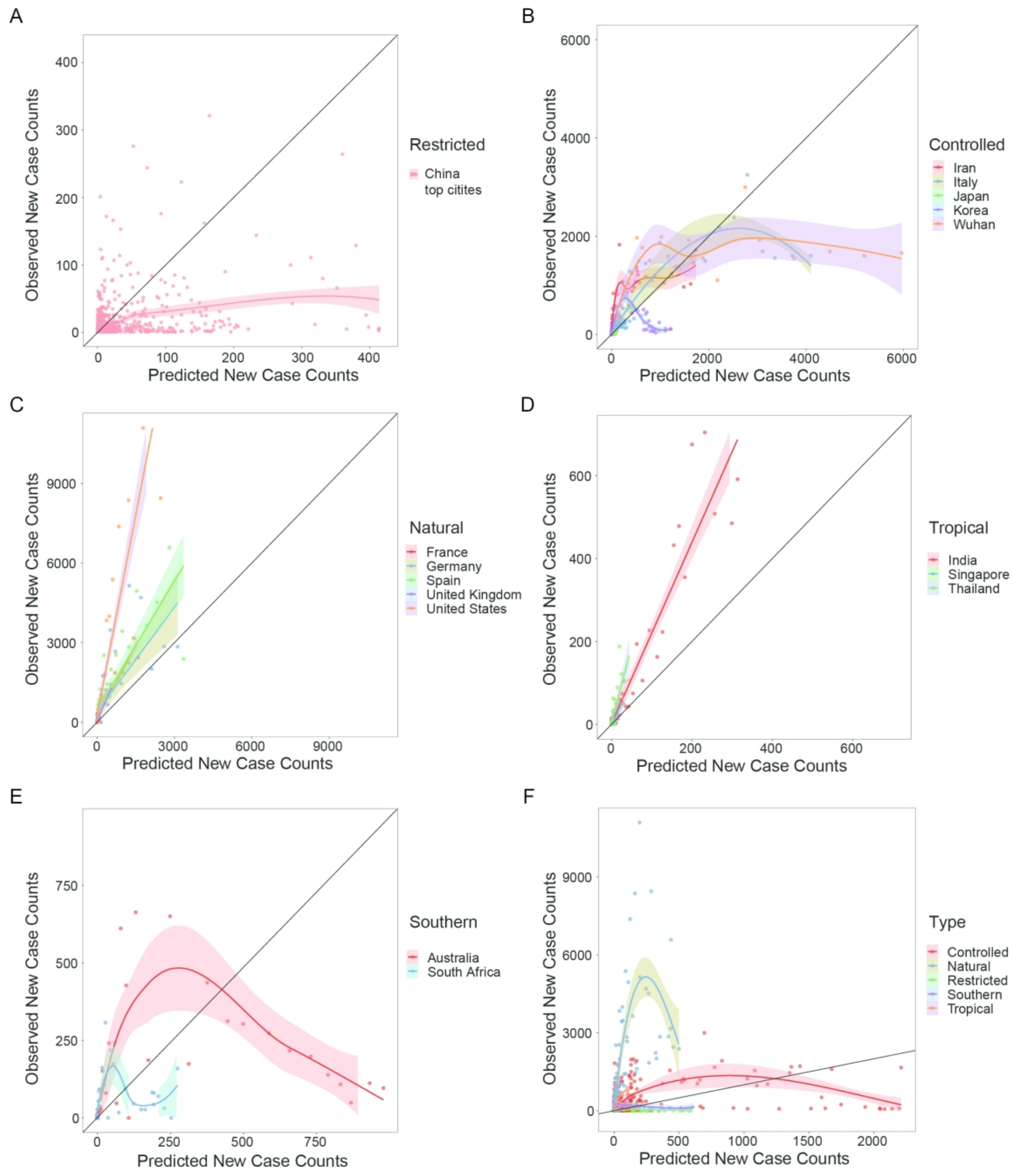


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

209x244mm (300 x 300 DPI)

Supplementary Materials and Methods

Epidemiological data

We scrutinized WHO's situation reports to rule out these countries with only imported cases, and only collected the confirmed cases with possible or confirmed local transmission (i.e., without recent travel history to China).

For Wuhan city, there was a shortage of test kits at the beginning of the pandemic, which would make confirmed case counts much lower than the actual data, thus, we discarded epidemic data before January 28th, the day when domestic test kits have been approved, produced in large quantities, and were available for Wuhan hospitals. As there was a cut down problem for the existing confirmed case count on February 20th for Wuhan, when modeling with the existing confirmed case count, only data before February 20th were used.

Weather data

Temperature and dew point displayed in Fahrenheit were transformed into Celsius forms, and relative humidity was calculated from temperature and dew point using the following formula for each time point:

$$\mathrm{RH} = \begin{cases} e^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T < 0 \\ 10^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T \geq 0 \end{cases}$$

where RH is the relative humidity, D is the dew point in degrees Celsius, T is the temperature in degrees Celsius, and *e* is the base of the natural log.

For each city with epidemiological data, the meteorological station in that city or that was closest to the latitude and longitude coordinates of the city center was chosen.

For a city with more than one meteorological stations, the one nearest to the city center was chosen. For a province with epidemiological data, the meteorological station in the capital city of that province was chosen. For a country with only national wide epidemiological data, weather data were averaged across all the meteorological observatories in the cities where outbreak was officially reported. Latitude and elevation for the meteorological observatories were also collected.

<u>Statistical modeling</u>

Only one city Wuhan was chosen for illustrating the time delay effect because it is the first city to have an outbreak of COVID-19, there was none reported imported cases for Wuhan, which might obscure the correlation between weather and virus transmission.

Supplementary Results

<u>Datasets description</u>

Only Chinese cities with monthly confirmed cases over 50 were included in the discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in all the other Chinese cities. We suppose that it might be due to abrupt large supplement of virus test kits or data correction on that day. In order to reduce the potential contamination of modeling by this outlier, data on that day were discarded from the subsequent analysis. There were also two oddly large new confirmed case counts for Lombardy, which were discarded from the subsequent analysis. Except the

outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to 2,997, the average temperature ranged -22.54℃ ~ 22.16℃, the wind speed ranged 0.56 ~ 9.29 meter per second, visibility ranged 1.3 ~ 18.8 statute miles, and relative humidity ranged 30.84% ~ 98.52%.

Model selection

With the increase of relative humidity, the amount of droplets in the air increases, leading to more virus load. However, as the air gets humid, human's respiratory tract could better defend virus infection. Thus, the relationship of relative humidity could be complex, not pure linear. Giving comprehensive consideration, we defined the effect of relative humidity to be quadric. As for visibility, it only affects the amount of particles in the air, which is positively correlated with virus load. Thus, it is most probably to exert its effect linearly.

Although relative humidity and visibility 7 days ago correlated with the confirmed new case counts best, there was not great loss of model fitting statistics for relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days time delay and 3~7 days time delay for temperature.

Fitted models

The fitted single-factor models were as follows:

$$\text{New Case Count} = -0.11305 \times \mathrm{T}^2 + 1.39819 \times \mathrm{T} + 45.11405$$

where T is temperature in ℃.

The estimate p-value for constant was < 0.001. The extremum was $-1.39819/(2 \times (-0.11305)) = 6.183945$ ℃.

$$\text{New Case Count} = -0.05759 \times \text{RH}^2 + 9.038 \times \text{RH} - 303.0$$

64 where RH is relative humidity in percentage.

65 The extremum was $-9.038/(2 \times (-0.05759)) = 78.46848$ %.

$$\text{New Case Count} = -1.360056 \times \text{SPD}^2 + 5.120123 \times \text{SPD} + 42.1855$$

66 where SPD is wind speed in meter per second (m/s).

67 The extremum was $-5.120123/(2 \times (-1.360056)) = 1.882321$ m/s.

$$\text{New Case Count} = -7.021 \times \text{VSB} + 89.041$$

68 where VSB is visibility in statute miles.

69 The estimate p-value for VSB was < 0.01, constant was < 0.001.

70 Thus, the complex short-term model to be regressed was

$$\begin{aligned}\text{New Case Count} \\ &= (-0.11 \times \text{T}^2 + 1.40 \times \text{T} - 0.058 \times \text{RH}^2 + 9.04 \times \text{RH} - 1.36 \\ &\times \text{SPD}^2 + 5.12 \times \text{SPD} - 7.02 \times \text{VSB} - 126.66) \times \text{a} \\ &\times \text{Existing Confirmed Case Count}\end{aligned}$$

71 where a is a constant to be fitted. All parameters take values 3~7 days before the day

72 new case count is confirmed.

73 Through fitting this full model with the discovery data, a was estimated to be

74 0.0004786 (standard error 0.0000128, *p*-values < 2e-16).

75 For long-term model, the fitted model with temperature 14 days ago was as

76 follows:

$$\text{New Case Count} = -0.10062 \times \mathrm{T}^2 + 1.11189 \times \mathrm{T} + 46.41792$$

77 The estimate p-value for constant was < 0.001. The extremum was $-1.11189/$

78 $(2 \times (-0.10062)) = 5.525194$.

79 Thus, the simplified long-term model to be regressed was:

$$\begin{aligned}\text{New Case Count} &= (-0.10 \times \mathrm{T}^2 + 1.11 \times \mathrm{T} + 46.42) \times \mathrm{b} \\ &\times \text{Existing Confirmed Case Count}\end{aligned}$$

80 where b is a constant to be fitted. All parameters take values 14 days before the day

81 new case count is confirmed.

82 Through fitting this simplified model with the discovery data, b was estimated to

83 be 0.0061382 (standard error 0.0002666, *p*-values < 2e-16).

84 Table S1. Model fitness statistics for comparing and selecting proper fitting
85 relationship

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Linear	493	4.5×10^{-8}	-167	339	342	4860391	0.757
Quadric	421	1.3×10^{-7}	-163	333	337	3370230	0.812
Relative humidity							
Linear	627	9.8×10^{-8}	-172	350	353	7855418	0.401
Quadric	626	8.4×10^{-6}	-171	351	355	7442367	0.358
Wind speed							
Linear	585	3.1×10^{-8}	-170	347	350	6840545	0.380
Quadric	546	2.4×10^{-7}	-168	344	349	5654728	0.423
Visibility							
Linear	594	3.3×10^{-8}	-171	347	351	7059799	0.354
Quadric	598	7.9×10^{-7}	-170	349	353	6799355	0.358

86 Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the
87 log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information
88 Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real
89 values and the predicted values by the predisposed model.
90

91 Table S2. Model fitness statistics for comparing and selecting proper time delay of
92 virus exposure

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Day 0	626	2.6×10^{-8}	-171	351	355	7441513	0.330
Day -3	605	1.3×10^{-8}	-171	349	353	6953553	0.479
Day -7	664	5.4×10^{-8}	-173	353	358	8386957	0.262
Day -14	528	1.1×10^{-7}	-168	343	347	5297229	0.534
Day -3 ~ -7	421	1.3×10^{-7}	-163	333	337	3370230	0.812
Relative humidity							
Day 0	605	5.9×10^{-6}	-171	349	353	6953396	0.389
Day -3	679	4.3×10^{-6}	-173	354	359	8768069	0.065
Day -7	560	5.0×10^{-8}	-169	346	350	5962416	0.524
Day -14	605	9.1×10^{-6}	-171	349	353	6962609	0.326
Day -3 ~ -7	626	8.4×10^{-6}	-171	351	355	7442367	0.358
Wind speed							
Day 0	526	7.4×10^{-8}	-167	343	347	5251026	0.500
Day -3	663	1.4×10^{-8}	-173	353	357	8343427	0.268
Day -7	559	1.1×10^{-8}	-169	346	350	5926891	0.516
Day -14	674	5.2×10^{-8}	-173	354	358.	8643076	0.014
Day -3 ~ -7	546	2.4×10^{-7}	-168	344	349	5654728	0.423
Visibility							

Day 0	646	4.2×10^{-9}	-173	351	354	8343221	0.286
Day -3	663	5.1×10^{-8}	-173	352	355	8804055	0.016
Day -7	514	3.9×10^{-8}	-168	341	344	5290247	0.502
Day -14	635	1.1×10^{-8}	-172	350	354	8052388	0.272
Day -3 ~ -7	594	3.3×10^{-8}	-171	347	351	7059799	0.354

Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real values and the predicted values by the predisposed model.

98 Table S3. Model fitness statistics for weather-combined model and epidemic only

99 model

Model	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Weather-combined	147	1.8×10^{-9}	-6239	12481	12491	21128810	0.171
Epidemic-only	149	2.1×10^{-8}	-6251	12507	12517	21689551	0.152

100 Note: The weather-combined model is the short-term model with multiplicative constant to be fitted. The

101 epidemic-only model is the model only with existing confirmed case count as an independent variable, assuming a

102 linear function.

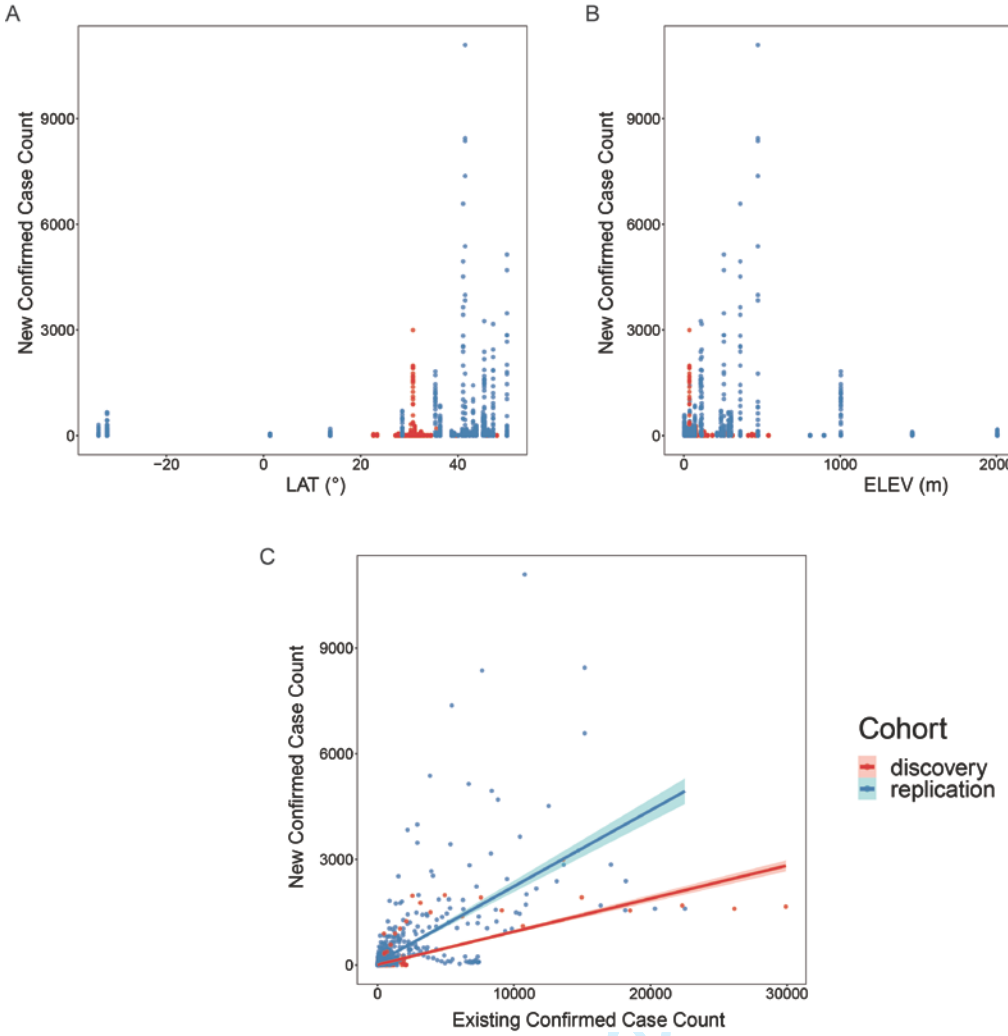


Fig. S1. Scatterplots of new confirmed case count to (A) latitude, (B) elevation, and (C) the existing confirmed case count, for all the studied sites. Linear regression (C) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow.

BMJ Open

Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041397.R2
Article Type:	Original research
Date Submitted by the Author:	22-Sep-2020
Complete List of Authors:	Chen, Biqing; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Liang, Hao ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Yuan, Xiaomin ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Colorectal Surgery Hu, Yingying ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Xu, Miao; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Zhao, Yating ; Nanjing University Zhang, Binfen ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Tian, Fang ; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Zhu, Xuejun ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology, Research Center of Chinese Medicine
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Global health, Infectious diseases, Occupational and environmental medicine, Public health, Qualitative research
Keywords:	Epidemiology < INFECTIOUS DISEASES, EPIDEMIOLOGY, Public health < INFECTIOUS DISEASES, Infection control < INFECTIOUS DISEASES, COVID-19

Title: Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Authors: Biqing Chen[1], Hao Liang[2], Xiaomin Yuan[3], Yingying Hu[2], Miao Xu[2], Yating Zhao[4], Binfen Zhang[2], Fang Tian[1], Xuejun Zhu[1,2*]

Affiliations:

[1] Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[2] Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[3] Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[4] School of Atmospheric Sciences, Nanjing University, Nanjing, China

*Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.

Abstract

OBJECTIVES: This study aims to investigate the relationship between daily weather and transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of the COVID-19 spreading rate for a certain area with meteorological factors.

DESIGN: A retrospective, qualitative study.

METHODS AND ANALYSIS: We collected 382,596 records of weather data with four meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility, and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed cases in total) in nearly 500 areas worldwide from January 20 to April 9. Epidemic data were modeled against weather data to find a model that could best predict the future outbreak.

RESULTS: Significant correlations of the daily new confirmed case counts with the weather 3~7 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average temperature at 5~15 ℃, relative humidity at 70%~80%, wind speed at 1.5~4.5 meter / second, and visibility less than 10 statute miles. A short-term model with these meteorological variables in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-term model using temperature to predict the pandemic in the next week or month was derived. Taken China as a discovery dataset, it was well validated with worldwide data. According to this model, there are five different viral transmission pattern, "restricted', "controlled", "natural", "tropical", "southern". This model's prediction performance correlates with the actual observations best (over 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not much human interference by epidemic prevention measures.

CONCLUSIONS: This model can be used for prediction of the future outbreak, and illustrating the effect of epidemic control for a certain area.

Keywords: COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control

Strengths and limitations of this study

- This study investigates the role of daily weather in COVID-19 spread systematically with a comprehensive set of four meteorological factors.

- This research collected a huge amount of data, covering nearly 500 areas worldwide in a long timescale.
- The current study proposes mathematical models integrating meteorological information for predicting COVID-19 case counts in the future.
- The influence of weather on virus spread could be confounded by a dozen of manual interventions, such as population mobility and disinfection measures, leading to inaccurate modeling.
- The prediction model (especially the long-term model) might be unsuitable and inaccurate for areas with hot weather.

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 has spread all over the world and has great social and economic impact worldwide (*1,2*). It exhibits high human-to-human transmissibility compared to other coronavirus like SARS (3). As of April 28 in 2020, the reported cumulative confirmed case count reached over three million and reported death is over 0.21 million globally (4). It would be crucial to predict the future trend of COVID-19 outbreak ahead, in order to make proper prevention and control strategies accordingly in time.

Besides population mobility and human-to-human contact, meteorological conditions have been suggested to be involved in the transmission of droplet-mediated viral diseases (*5,6*). As droplets carrying the coronavirus can travel in gaseous clouds as far as eight metres and stay suspended in the air for hours (7), the suspending time and viability of the coronavirus outside body would be largely affected by the environment. Wind speed could affect the suspending time of droplets, while visibility and humidity reflect the amount of particles in the air, determining the coronavirus payload. Temperature affects virus's viability in the environment. As SARS-CoV-2 is enveloped, it might be more vulnerable to adverse conditions like high temperature.

The impact of weather on epidemiology has been mentioned in human's history. The ancient Chinese had a theory called "Five Movement and Six Weather" to study climate change and its relationship with human health. Currently, there are a few studies on preprint servers discussing the relationship of temperature and humidity with the pandemic, but none is systematical investigation or proposes validated practical model for prediction (*8-13*).

Herein, this study intends to investigate the relationship between meteorological factors and epidemic transmission rate on a world scale. Four meteorological variables, i.e., average temperature, relative humidity, wind speed, and visibility, were collected as well as the confirmed case counts daily for 80 days from January 20, 2020 to April 9, 2020 for nearly 500 areas around the world, including 428 Chinese cities and areas, 18 Italian provinces, and 13 other countries. Five time point's delay of virus infection from the exposure day were considered and compared to determine the most reasonable time point's delay. A multivariate polynomial regression model with meteorological factors as a "weather coefficient" of the existing confirmed case count was established in a discovery Chinese dataset, and then validated by worldwide data. Five transmission modes, indicating different levels of epidemic control, were revealed by this model. In this view, this model can not only predict future outbreak, but also be used to evaluate the effect of epidemic prevention measures for a certain area.

Materials and Methods

Epidemiological data

Epidemiological data were collected from the World Health Organization (WHO) (4), European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (10). The daily new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries, namely, United States, United Kingdom, Germany, France, Italy, Spain, Iran, Korea, Japan, Australia, South Africa, India, Thailand, and Singapore. Considering the potential confounding effect, only Chinese cities with no less than 50 cumulative confirmed cases in one month and without official reports of large imported cases (42 in total) were taken as a discovery dataset, while those for Italian provinces and all the other nations were taken as replication datasets (Supplementary Materials).

Weather data

97 Four meteorological variables were chosen, air temperature, relative humidity, wind speed, and visibility. Temperature could affect virus viability in the environment. Wind speed could affect the suspending time of virus-attached particles. Relative humidity reflects the amount of droplets in the air. Visibility is influenced by the amount of particles such as dust and air pollutants. These two parameters both affect the amount of mediator for the virus to stay in the air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of meteorological observations and geographic factors (latitude and elevation) from the Integrated Surface Database of USA National Centers for Environmental Information (11). Daily data were calculated by averaging the hourly data for each variable in each day.

107 Statistical modeling

108 The number of daily new confirmed cases was taken as a dependent variable. Four meteorological variables, namely, average temperature, wind speed, visibility, and relative humidity, and the existing confirmed case counts were taken as independent variables. Considering that there is a latency stage from the day one get infected to the day being confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were collected needs to be taken into consideration. As it is reported that the latency period for COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection were taken into consideration, that is, weather data and existing confirmed cases count data were collected on the day, three days before, seven days before, 3~7 days before, and 14 days before collecting the epidemiological data.

118 To investigate whether the influence of meteorological factors is linear or quadric, both linear and non-linear modeling were performed under different relationship assumptions to

compare model fitness statistics. Each meteorological variable was fitted into a bunch of single-factor models (either generalized linear model or polynomial model) through non-linear least squares (NLS) modeling using the Wuhan dataset with a 3~7 days delay of infection. The relationship between each meteorological variable and confirmed new case count (linear or quadric) was identified based on model fitness (log-likelihood, Akaike information criterion, Bayesian Information Criterion, etc.) and common knowledge of droplet-mediated viral diseases.

Second, the proper time delay from weather exposure to COVID-19 confirmation was investigated in the Wuhan dataset through Loess regression interpolation and NLS modeling with the previously identified relationship for each meteorological variable. The most possible time delay identified was taken for subsequent analyses.

The contribution of each meteorological factor to the case counts was first investigated with the Wuhan dataset under the assumption of previously defined time delay through Spearman's correlation test. Then, we performed single-factor NLS regression modeling for each meteorological variable in the discovery dataset (all Chinese cities with monthly confirmed cases over 50) under the assumption of previously determined relationship and pre-defined time delay, to determine the exact coefficients accompanied with each meteorological factor and to find out the most suitable environmental condition for SARS-CoV-2.

Then, two final prediction models (short-term model and long-term model) were developed using the discovery dataset with the previously determined coefficients. The prediction model supposed that all the meteorological variables, with their specific coefficients determined by single-factor modeling, were added together to compose a weather coefficient. The new confirmed case count on the test day is calculated by multiplying the weather coefficient with the existing confirmed case count on the exposure day (the time delay between test day and exposure day is determined in previous analysis), and then multiply by a constant coefficient. The short-term model took all four variables, while the long-term model only considered temperature as it is easy to be forecasted. There was a constant coefficient for the total equation, that was multiplied by the existing confirmed case count. Its exact value was determined by model fitting in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was investigated with all datasets covering the world's top cities and areas. The correlation of existing confirmed case counts with newly confirmed case counts was also investigated. Basic statistics and modeling was conducted in R 3.5.1 (https://cran.***r***-project.org/).

Model validation and application

The best fitted model was validated in the replication datasets (Italian city-level data and other nation-level data) by correlating the observed actual epidemiological data with the predicted values from the model in the datasets. We used these fitted models to calculate a predicted value for case counts for each studied site, and then compared this predicted value with the real observed case counts by calculating a Spearman's correlation coefficient ρ between them.

Patient and Public Involvement

No specific patients were included in the current study. Epidemiological data were downloaded from online open-source databases. The public were not involved in the planning and design of the study.

Results

The Weather's influence on SARS-CoV-2 transmission displays 3~7 days time delay

The ranges of average temperature, relative humidity, wind speed, and visibility in the replication datasets were similar to those in the discovery dataset (see Supplementary Results for detailed datasets description). Non-linear modeling with Wuhan dataset under the assumption of 3~7 days delay of infection confirmation suggested that the effect of temperature and wind speed is better depicted as quadric (Table S1), which was also supported by Loess regression interpolation (Fig. 1). The mode for relative humidity and visibility was hard to be determined, as statistics supported both relationships (Table S1). Considering the common knowledge of coronavirus transmission and the trend showed by Loess regression interpolation, relative humidity exerted its impact in a quadric trend while visibility exerted its impact in a linear trend (Fig. 1, Supplementary results).

Furthermore, investigation of the time delay effect in the Wuhan dataset showed that the number of confirmed new cases was best correlated with air temperature 3~7 days ago, relative humidity and visibility 7 days ago, and wind speed on the exposure day (Table S2). By comprehensive consideration of all four meteorological variables and the differences between statistics values, the weather 3~7 days ago, as well as weather one week ago, could well predict COVID-19 outbreak. It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure under certain adverse weather might exhibit its effect after 3~7 days.

<u>Contribution of single meteorological factor to the outbreak</u>

In the Wuhan dataset, the new case count was significantly positively correlated with temperature (Spearman's correlation ρ = 0.69, p < 0.001) and visibility (ρ = 0.43, p =0.04), and negatively correlated with wind speed (ρ = -0.45, p = 0.03) and relative humidity (ρ = -0.33, p = 0.12) 3~7 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind speed, visibility, and relative humidity. A model only with temperature as a parameter could already explained 45% of the variance in the epidemic data ($p = 4\times10^{-4}$), while wind speed and visibility could explain over 25% of the variance. According to the fitted single-factor models (temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility was fitted into a linear model, see the Supplementary Results for details), SARS-CoV-2 transmission reaches a peak when mean temperature is 6.18 ℃ (Fig. 2A), relative humidity is 78.47% (Fig. 2B), and wind speed is 1.88 meter /second (m/s) (Fig. 2C); and its transmission rate decreases with the increase of visibility (Fig. 2D). The effects of geographic factors such as latitude and elevation, and the pure influence from the existing case count were further investigated in the worldwide datasets (Fig. S1), illustrating that COVID-19 mainly outbreaks at latitude 30°~50° (Fig. S1A) and elevation < 500 metre (Fig. S1B). New confirmed case count was positively correlated with the existing confirmed case count (Fig. S1C).

Short-term prediction model

We further derived a full model combined with all four meteorological variables and fitted this model with the discovery dataset (Supplementary Results). The best-fitted short-term model was as follows:

$$\text{New Case Count} = \left(-0.11 \times T^2 + 1.40 \times T - 0.058 \times RH^2 + 9.04 \times RH - 1.36 \times SPD^2 + 5.12 \times SPD - 7.02 \times VSB - 126.66\right) \times \alpha \times \text{Existing Confirmed Case Count}$$

where T is temperature in ℃, RH is relative humidity in percentage, SPD is wind speed in m/s, VSB is visibility in statute miles, α is a site-specific constant, with a default of 0.001. All parameters take the means of values 3~7 days before the day new case count is evaluated.

In this model, all the four meteorological variables are added together in their proper forms to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate of SARS-CoV-2, and thus influences the number of people that catch infection from the existing confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a multiplicative constant coefficient α in the equation, which seems site-related. This constant coefficient could adjust the strength of the "weather coefficient" on disease transmission. When

211 we substitute replication datasets into this short-term model with the multiplicative constant coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious underestimation of predicted values against real ones was observed although the predicted values correlated with the real ones very well. We supposed it was due to site-specific difference in the multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the pandemic had been controlled early. Thus, we further re-fitted this composed model with all datasets to determine a more accurate value of the multiplicative constant coefficient α, which was 0.001 then. In practical application, we need to first plot the observed case count vs. predicted one with a default α value 0.001, and then examine the extent of underestimation or overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the impact size of "weather coefficient" for a certain site.

Substitute data from the past two months, a good prediction performance was obtained for this short-term model, with the predicted values significantly correlated to the observed ones for most areas (Fig. 3). However, only the existing confirmed case count data could not predict the new case count 3~7 days later as well as the weather-combined model did (Table S3).

Different modes of viral transmission illustrated by the model

The observed versus predicted data exhibited different correlation patterns for different areas, meaning different viral transmission modes, which may indicate the effect of epidemic control for certain area.

Data from Chinese top-affected cities were not very well predicted and obviously overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p < 0.001$; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention measures at the beginning of the pandemic. This viral transmission mode suggested by the not well correlated prediction pattern is called "restricted".

For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the predicted outbreak was well correlated with the actual observations at the beginning when the existing confirmed cases were not in very large numbers, but the prediction deviates from the observation as the confirmed cases increase, in detail, there's large overestimation of prediction ($\rho_{Wuhan} = 0.69$, $\rho_{Italy} = 0.87$, $\rho_{Japan} = 0.80$, $\rho_{Iran} = 0.86$, $p < 0.001$, $\rho_{Korea} = 0.43$, $p = 0.002$; Fig. 3B). It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15, the day when shelter hospitals had been put into use for seven days (the average latency period for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention and control taken in these areas is effective (so-called "controlled" mode). The number of cases had been decreased by 72% for Wuhan, over 95% for Korea, Japan, and Italy, and 37% for Iran at most due to epidemic control (the largest gap between prediction and observation).

For most European and American countries, the predicted outbreak was linear correlated with the observed data very well ($\rho_{France} = 0.96$, $\rho_{United\ States} = 0.93$, $\rho_{United\ Kingdom} = 0.83$, $\rho_{Spain} = 0.97$, $\rho_{Germany} = 0.94$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without much man-made epidemic prevention and control measures. Estimation of daily new case counts by this short-term model performed very well for European countries, while this model underestimated the outbreak in the United States.

Although the weather is not suitable for tropical areas, the viral transmitted in natural mode, manifested as good linear correlation between the prediction and the observation ($\rho_{India} = 0.94$, $\rho_{Singapore} = 0.66$, $p < 0.001$, $\rho_{Thailand} = 0.56$, $p = 0.001$; Fig. 3D), with just relatively small daily new case counts compared to temperate regions.

Countries in the southern hemisphere displayed similar pattern as the "controlled" with large overestimation by the model when the confirmed cases increase, leading to not good prediction performance ($\rho_{Australia} = 0.79$, $p < 0.001$, $\rho_{South\ Africa} = 0.34$, $p = 0.08$; Fig. 3E). It might be due to the effect of epidemic prevention measures and hot summer weather in these countries.

Long-term simplified model

Long-term prediction depends on weather forecast, which generally reports only average temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as temperature in a short time delay (3~7 days ago), we again performed single-factor regression modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a quadric function (Supplementary Results). This simplified model with average temperature as a weather factor was derived as follows:

$$\text{new case count} = \left(-0.10 \times T^2 + 1.11 \times T + 46.42\right) \times \beta \times \text{Existing Confirmed Case Count}$$

where T is temperature in ℃, β is a site-related multiplicative constant coefficient, with a default of 0.006. All parameters take values 14 days before the day new case count is evaluated.

With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, $p < 0.001$; Fig. 3F). The long-term simplified prediction model also showed five prediction-observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the studied areas. This model could directly predict the newly emerging cases 14 days later, and be used to predict COVID-19 outbreak in the future month by summing up the daily new case count and combining weather forecast (usually available for the future 15 days).

Discussion

This research discovers nonlinear dose-response relationship for meteorological factors, in consistency with previous studies (12). Predictions of COVID-19 outbreak scale by the models were well correlated with the observations around the world, suggesting the importance of weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many respiratory infectious diseases, such as influenza, is dependent upon temperature and relative humidity (*5,6*). Recent published papers on preprint servers have reported roles of temperature and absolute humidity in the COVID-19 transmission, but their conclusions are diverse (*8-13*). In contrast to the findings by Cai et al (8), this study suggests significant impact of mean temperature on the daily new case count, indicating a need for sufficient time delay between exposure and confirmation for weather to exhibit its effect. In contrary to other two studies (*9,10*), this research suggests that there is a relatively not wide temperature and humidity ranges for the pandemic. There is an optimal temperature for SARS-CoV-2 at 6.18 ℃, which is colder than that suggested by Bu et al (14) but in consistency with the estimation by Wang et al (12); and most areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al suggested (14). It is of notice that different from other viral respiratory diseases such as influenza(15)(16), high relative humidity is better for SARS-CoV-2 to spread, suggesting that a sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more important for the spread than the effect of dry air on the human immune system. Different from other studies (17), this study also finds significant involvement of wind speed, in a quadric manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In addition, the current study discovered that visibility was significantly negatively correlated with new case count and played a more important role in viral spread than humidity did (from spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g., dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case count decreases rapidly when visibility is high than 13 statute miles, indicating that caution should be taken if visibility drops below 10 statute miles.

In the prediction model, there is a multiplicative constant coefficient which determines the strength of the weather coefficient on the epidemic transmission. It seems site-specific, as adjusting it could make the prediction for one site very close to the observation. This constant might reflect the influence of a couple of site-specific confounding factors, such as epidemic control measures, sun radiation, and population density. Various degrees of isolation for various areas around the world lead to different degrees of weather effect. When evaluate the prediction performance by the short-term model and the long-term model, they both exhibit different prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of epidemic control and isolation policy would lead to deviation from the original prediction and thus different prediction-observation correlation patterns. Therefore, by plotting the predicted versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the observed case counts dropped dramatically from the predictions for Wuhan seven days after their shelter hospitals were put in use, suggesting the importance and necessity of building shelter hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to natural transmission mode. Another thing worth attention is that although the weather in tropical areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore, these tropical areas should still be on the alert against future outbreak of COVID-19.

Although those cases with travel history to China or indicated by the World Health Organization as "imported case only" were excluded in this study to make the world data most likely local transmitted, it was difficult to separate the imported cases from local transmission very well in practice. It might explain the not excellent correlations of predictions with observations. Furthermore, the relationship of weather and COVID-19 could be complex, since the human immune system has an innate seasonal rhythm, and the immune system could also be affected by weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets for virus to adhere.

There are several limitations of this study. First of all, this prediction model (especially the long-term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter, as the models were derived using Chinese datasets, mainly in the first three months of 2020. The prediction became inaccurate and even improper under hot weather (i.e., the predicted values of long-term model become negative when air temperature is higher than 28 ℃), which might explain the obvious bad prediction performance for countries in the southern hemisphere and tropical areas. One explanation for the inaccurate prediction in areas with high temperature could be that SARS-CoV-2 transmission in these areas was mainly not influenced by weather, but in another direct transmission way, such as face-to-face contact or spread in gathering crowd. Second, it seems that the prediction performance drops with the increase in new case count, suggesting that the prediction model might become inaccurate and not suitable for very large new case count. This could be due to (1) the influence of weather on COVID-19 spread might weaken when the number of cases increases, while other factors such as social distance become more important at a later stage; (2) there was less data points with large new case count, which might lead to larger variance. Third, the short-term prediction model must use all four meteorological factors, while these factors are not always available for any one certain area. Fourth, this study included various areas covering a long period into modeling, thus, there were a bunch of variable confounding factors, such as population mobility and disinfection measures, which were not controlled and thus could impede the model accuracy. Fifth, as we could only obtain country-level epidemiological data, the corresponding meteorological data were obtained for their capital cities, leading to not exact pairing of epidemiological data and meteorological data. Sixth, there is a general lack of data and cases in the current study, since we only collected data covering two and a half months while the pandemic has persisted over seven months up to now.

Conclusion

In summary, this study has found significant correlations with the COVID-19 epidemic trend for not only temperature and humidity, but also wind speed and visibility. It proposed a comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version and a long-term version. The short-term version uses the combination of four meteorological factors as a "weather coefficient" of the existing confirmed case count in the past week and can be used to predict epidemic situation in the future three days; the short-term version uses average temperature as the "weather coefficient" seven days ago and can predict the outbreak in one month if combined with weather forecast. This model is easy to use for predicting the COVID-19 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case count for the future couple of days or month. This model will be very helpful for local governments to make timely policies on epidemic control, for instance, the allocation of medical equipments such as ventilators and medical resources such as hospitals, beds and health-care workers, according to the prediction results.

Acknowledgments: We thank Dr. Zhisheng Huang for advices on data collecting and processing, Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on modeling.

Funding: the Priority Academic Program Development of Jiangsu Higher Education Institutions – the third period (NO.035062002003c), the Yizhong Research Fund of Jiangsu Provincial Hospital of Chinese Medicine (Y19066).

Author contributions: BC, YZ and XZ design and interpret the reported analyses and results; HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse the data, drafted, and revised the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical support; XZ supervised the research.

Competing interests: Authors declare no competing interests.

Data and materials availability: Weather data and epidemiological data is all obtained from public databases. Detailed modeling results are available upon request by emailing to Biqing Chen, bq_chen@qq.com.

References:

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.

2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med. 2020;382(13):1199–207.

3. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet [Internet]. Elsevier Ltd; 2020;395(10223):514–23. Available from: http://dx.doi.org/10.1016/S0140-6736(20)30154-9

4. Organization WH. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [Internet]. 2020 [cited 2020 Mar 29]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the United States. Amer J Epidemiol. 2012;176(suppl. 7):S114–22.

6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathol. 2007;3:e151.

7. Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential Implications for Reducing Transmission of COVID-19. JAMA. 2020;E1–2.

8. Cai Y. The Effects of "Fangcang, Huoshenshan [Internet]. Available from: https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3

9. Proverbio AM, Crotti N, Manfredi M, Zani A. Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. medRxiv. 2020;(February).

10. https://github.com/BlankerL/DXY-COVID-19-Data.

11. ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/. 2020.

12. Wang M, Jiang A, Gong L, Lu L, Guo W, Li C, et al. Temperature Significantly Change COVID-19 Transmission in 429 cities. Block Caving – A Viable Altern [Internet]. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1

13. Luo W, Majumder MS, Liu D, Poirier C, Mandl KD, Lipsitch M, et al. The role of absolute humidity on transmission rates of the COVID-19 outbreak. medRxiv [Internet]. 2020;2020.02.12.20022467. Available from: http://medrxiv.org/content/early/2020/02/17/2020.02.12.20022467.abstract%0Ahttps://www.medrxiv.org/content/10.1101/2020.02.12.20022467v1

14. Bu J, Peng D-D, Xiao H, Yue Q, Han Y, Lin Y, et al. Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020. medRxiv [Internet]. 2020;(May):2020.02.13.20022715. Available from: https://www.medrxiv.org/content/10.1101/2020.02.13.20022715v1

15. Peci A, Winter AL, Li Y, Gnaneshan S, Liu J, Mubareka S, et al. Effects of Absolute Humidity, Relative Humidity, Temperature, and Wind Speed on Influenza Activity in Toronto, Ontario, Canada. Appl Env Microbiol. 2019;85(6):e02426-18.

16. Kudo E, Song E, Yockey LJ, Rakib T, Wong PW, Homer RJ, et al. Low ambient humidity impairs barrier function and innate resistance against influenza infection. Proc Natl Acad Sci U S A. 2019;116(22):10905–10.

17. Oliveiros B, Caramelo L, Ferreira NC, Caramelo F. Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. medRxiv [Internet]. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1

436 **FIGURE LEGENDS**

437 **Fig. 1.** Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

442 **Fig. 2.** Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

449 **Fig. 3.** The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot

458 represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

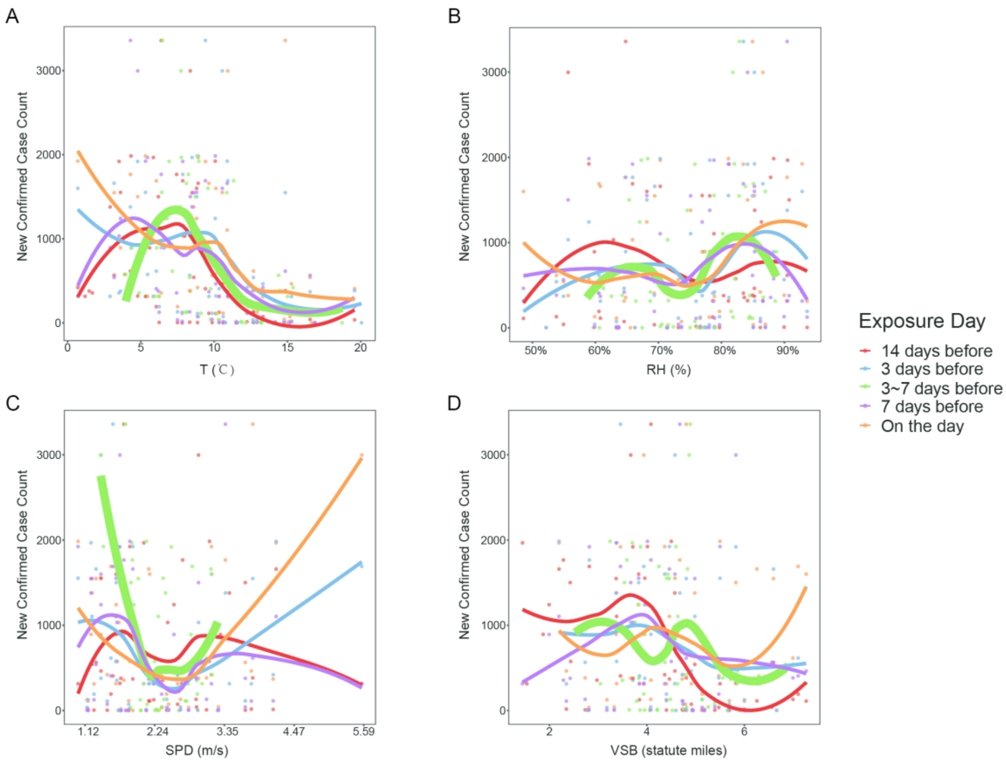


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

207x156mm (300 x 300 DPI)

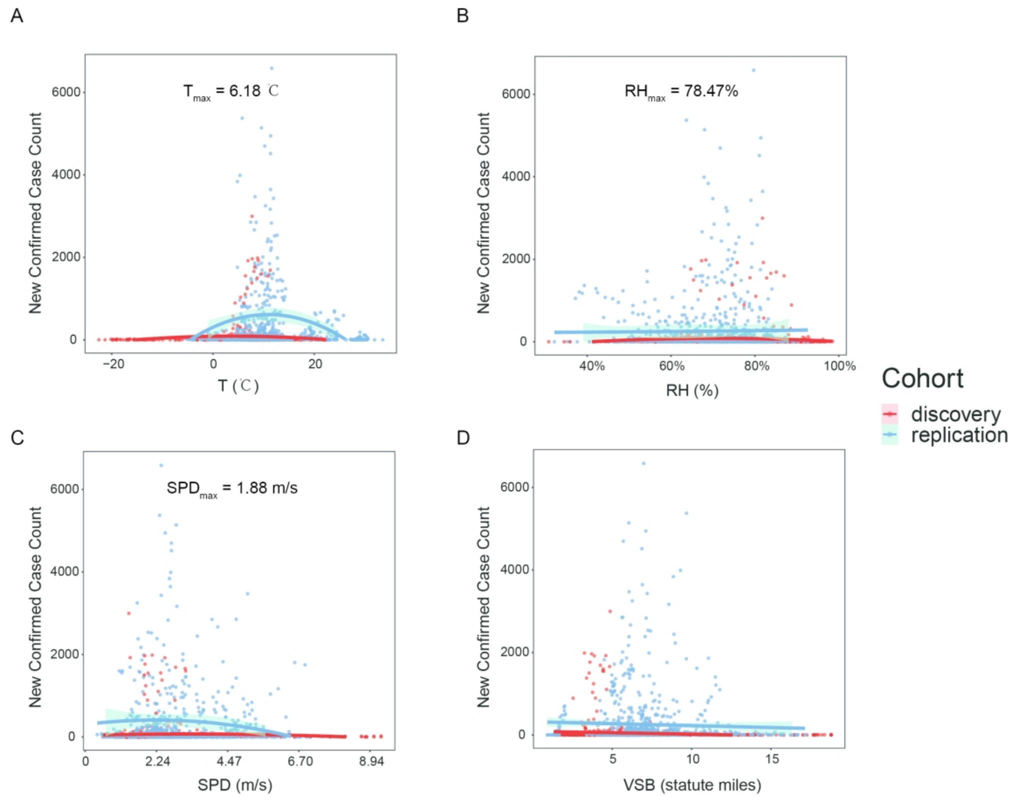


Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

218x172mm (300 x 300 DPI)

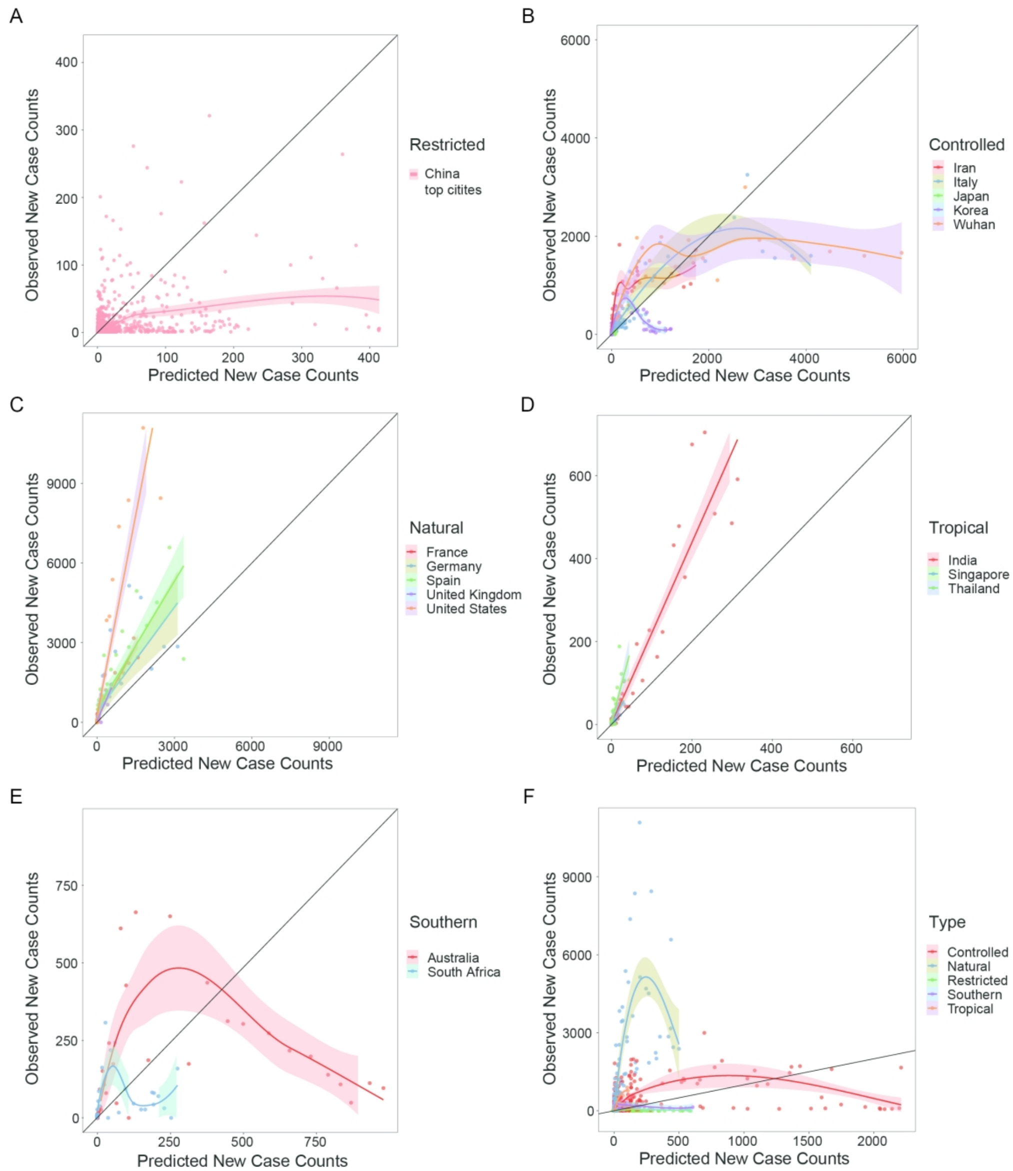


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

209x244mm (300 x 300 DPI)

Supplementary Materials and Methods

Epidemiological data

We scrutinized WHO's situation reports to rule out these countries with only imported cases, and only collected the confirmed cases with possible or confirmed local transmission (i.e., without recent travel history to China).

For Wuhan city, there was a shortage of test kits at the beginning of the pandemic, which would make confirmed case counts much lower than the actual data, thus, we discarded epidemic data before January 28th, the day when domestic test kits have been approved, produced in large quantities, and were available for Wuhan hospitals. As there was a cut down problem for the existing confirmed case count on February 20th for Wuhan, when modeling with the existing confirmed case count, only data before February 20th were used.

Weather data

Temperature and dew point displayed in Fahrenheit were transformed into Celsius forms, and relative humidity was calculated from temperature and dew point using the following formula for each time point:

$$\mathrm{RH} = \begin{cases} e^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}}-\frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T < 0 \\ 10^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}}-\frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T \geq 0 \end{cases}$$

where RH is the relative humidity, D is the dew point in degrees Celsius, T is the temperature in degrees Celsius, and *e* is the base of the natural log.

For each city with epidemiological data, the meteorological station in that city or that was closest to the latitude and longitude coordinates of the city center was chosen.

For a city with more than one meteorological stations, the one nearest to the city center was chosen. For a province with epidemiological data, the meteorological station in the capital city of that province was chosen. For a country with only national wide epidemiological data, weather data were averaged across all the meteorological observatories in the cities where outbreak was officially reported. Latitude and elevation for the meteorological observatories were also collected.

Statistical modeling

Only one city Wuhan was chosen for illustrating the time delay effect because it is the first city to have an outbreak of COVID-19, there was none reported imported cases for Wuhan, which might obscure the correlation between weather and virus transmission.

Supplementary Results

Datasets description

Only Chinese cities with monthly confirmed cases over 50 were included in the discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in all the other Chinese cities. We suppose that it might be due to abrupt large supplement of virus test kits or data correction on that day. In order to reduce the potential contamination of modeling by this outlier, data on that day were discarded from the subsequent analysis. There were also two oddly large new confirmed case counts for Lombardy, which were discarded from the subsequent analysis. Except the

outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to 2,997, the average temperature ranged -22.54℃ ~ 22.16℃, the wind speed ranged 0.56 ~ 9.29 meter per second, visibility ranged 1.3 ~ 18.8 statute miles, and relative humidity ranged 30.84% ~ 98.52%.

Model selection

With the increase of relative humidity, the amount of droplets in the air increases, leading to more virus load. However, as the air gets humid, human's respiratory tract could better defend virus infection. Thus, the relationship of relative humidity could be complex, not pure linear. Giving comprehensive consideration, we defined the effect of relative humidity to be quadric. As for visibility, it only affects the amount of particles in the air, which is positively correlated with virus load. Thus, it is most probably to exert its effect linearly.

Although relative humidity and visibility 7 days ago correlated with the confirmed new case counts best, there was not great loss of model fitting statistics for relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days time delay and 3~7 days time delay for temperature.

Fitted models

The fitted single-factor models were as follows:

$$\text{New Case Count} = -0.11305 \times \mathrm{T}^2 + 1.39819 \times \mathrm{T} + 45.11405$$

where T is temperature in ℃.

The estimate p-value for constant was < 0.001. The extremum was $-1.39819/(2 \times (-0.11305)) = 6.183945$ ℃.

$$\text{New Case Count} = -0.05759 \times \text{RH}^2 + 9.038 \times \text{RH} - 303.0$$

where RH is relative humidity in percentage.

The extremum was $-9.038/(2 \times (-0.05759)) = 78.46848$ %.

$$\text{New Case Count} = -1.360056 \times \text{SPD}^2 + 5.120123 \times \text{SPD} + 42.1855$$

where SPD is wind speed in meter per second (m/s).

The extremum was $-5.120123/(2 \times (-1.360056)) = 1.882321$ m/s.

$$\text{New Case Count} = -7.021 \times \text{VSB} + 89.041$$

where VSB is visibility in statute miles.

The estimate p-value for VSB was < 0.01, constant was < 0.001.

Thus, the complex short-term model to be regressed was

$$\begin{aligned}\text{New Case Count} \\ &= (-0.11 \times \text{T}^2 + 1.40 \times \text{T} - 0.058 \times \text{RH}^2 + 9.04 \times \text{RH} - 1.36 \\ &\times \text{SPD}^2 + 5.12 \times \text{SPD} - 7.02 \times \text{VSB} - 126.66) \times \text{a} \\ &\times \text{Existing Confirmed Case Count}\end{aligned}$$

where a is a constant to be fitted. All parameters take values 3~7 days before the day new case count is confirmed.

Through fitting this full model with the discovery data, a was estimated to be 0.0004786 (standard error 0.0000128, *p*-values < 2e-16).

For long-term model, the fitted model with temperature 14 days ago was as follows:

$$\text{New Case Count} = -0.10062 \times \text{T}^2 + 1.11189 \times \text{T} + 46.41792$$

77 The estimate p-value for constant was < 0.001. The extremum was $-1.11189/$

78 $(2 \times (-0.10062)) = 5.525194$.

79 Thus, the simplified long-term model to be regressed was:

$$\begin{aligned}&\text{New Case Count}\\&\qquad = (-0.10 \times \text{T}^2 + 1.11 \times \text{T} + 46.42) \times \text{b}\\&\qquad \times \text{Existing Confirmed Case Count}\end{aligned}$$

80 where b is a constant to be fitted. All parameters take values 14 days before the day

81 new case count is confirmed.

82 Through fitting this simplified model with the discovery data, b was estimated to

83 be 0.0061382 (standard error 0.0002666, *p*-values < 2e-16).

Table S1. Model fitness statistics for comparing and selecting proper fitting relationship

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Linear	493	4.5×10^{-8}	-167	339	342	4860391	0.757
Quadric	421	1.3×10^{-7}	-163	333	337	3370230	0.812
Relative humidity							
Linear	627	9.8×10^{-8}	-172	350	353	7855418	0.401
Quadric	626	8.4×10^{-6}	-171	351	355	7442367	0.358
Wind speed							
Linear	585	3.1×10^{-8}	-170	347	350	6840545	0.380
Quadric	546	2.4×10^{-7}	-168	344	349	5654728	0.423
Visibility							
Linear	594	3.3×10^{-8}	-171	347	351	7059799	0.354
Quadric	598	7.9×10^{-7}	-170	349	353	6799355	0.358

Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real values and the predicted values by the predisposed model.

Table S2. Model fitness statistics for comparing and selecting proper time delay of virus exposure

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Day 0	626	2.6×10^{-8}	-171	351	355	7441513	0.330
Day -3	605	1.3×10^{-8}	-171	349	353	6953553	0.479
Day -7	664	5.4×10^{-8}	-173	353	358	8386957	0.262
Day -14	528	1.1×10^{-7}	-168	343	347	5297229	0.534
Day -3 ~ -7	421	1.3×10^{-7}	-163	333	337	3370230	0.812
Relative humidity							
Day 0	605	5.9×10^{-6}	-171	349	353	6953396	0.389
Day -3	679	4.3×10^{-6}	-173	354	359	8768069	0.065
Day -7	560	5.0×10^{-8}	-169	346	350	5962416	0.524
Day -14	605	9.1×10^{-6}	-171	349	353	6962609	0.326
Day -3 ~ -7	626	8.4×10^{-6}	-171	351	355	7442367	0.358
Wind speed							
Day 0	526	7.4×10^{-8}	-167	343	347	5251026	0.500
Day -3	663	1.4×10^{-8}	-173	353	357	8343427	0.268
Day -7	559	1.1×10^{-8}	-169	346	350	5926891	0.516
Day -14	674	5.2×10^{-8}	-173	354	358.	8643076	0.014
Day -3 ~ -7	546	2.4×10^{-7}	-168	344	349	5654728	0.423
Visibility							

Day 0	646	4.2×10^{-9}	-173	351	354	8343221	0.286
Day -3	663	5.1×10^{-8}	-173	352	355	8804055	0.016
Day -7	514	3.9×10^{-8}	-168	341	344	5290247	0.502
Day -14	635	1.1×10^{-8}	-172	350	354	8052388	0.272
Day -3 ~ -7	594	3.3×10^{-8}	-171	347	351	7059799	0.354

93 Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the
94 log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information
95 Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real
96 values and the predicted values by the predisposed model.

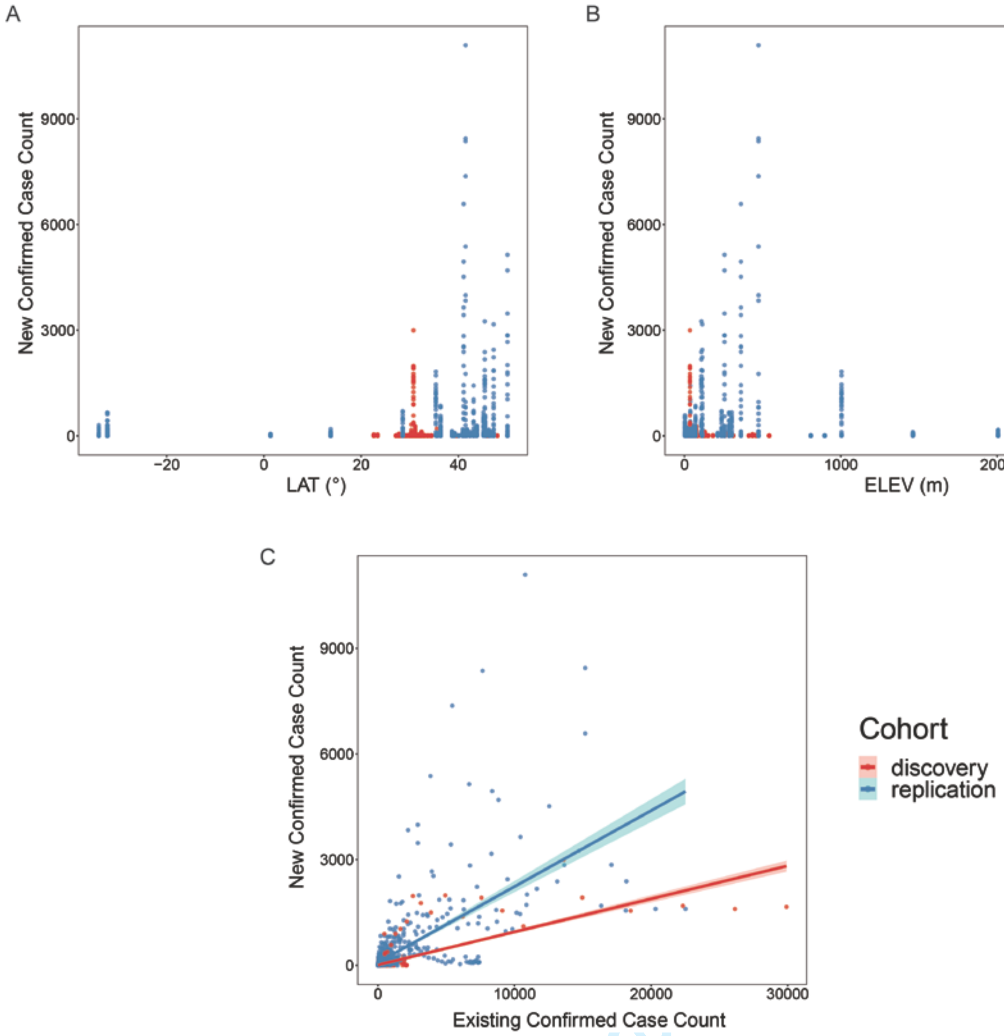
97

98 Table S3. Model fitness statistics for weather-combined model and epidemic only
99 model

Model	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Weather-combined	147	1.8×10^{-9}	-6239	12481	12491	21128810	0.171
Epidemic-only	149	2.1×10^{-8}	-6251	12507	12517	21689551	0.152

100 Note: The weather-combined model is the short-term model with multiplicative constant to be fitted. The
101 epidemic-only model is the model only with existing confirmed case count as an independent variable, assuming a
102 linear function.

103



104 **Fig. S1.** Scatterplots of new confirmed case count to (A) latitude, (B) elevation, and
105 (C) the existing confirmed case count, for all the studied sites. Linear regression (C)
106 interpolation curves are illustrated for each dataset, with 95% confidence intervals
107 showing in shadow.

Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041397.R3
Article Type:	Original research
Date Submitted by the Author:	28-Oct-2020
Complete List of Authors:	Chen, Biqing; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Liang, Hao ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Yuan, Xiaomin ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Colorectal Surgery Hu, Yingying ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Xu, Miao; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Zhao, Yating ; Nanjing University Zhang, Binfen ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology Tian, Fang ; Affiliated Hospital of Nanjing University of Chinese Medicine, Research Center of Chinese Medicine Zhu, Xuejun ; Affiliated Hospital of Nanjing University of Chinese Medicine, Department of Hematology, Research Center of Chinese Medicine
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Global health, Infectious diseases, Occupational and environmental medicine, Public health, Qualitative research
Keywords:	Epidemiology < INFECTIOUS DISEASES, EPIDEMIOLOGY, Public health < INFECTIOUS DISEASES, Infection control < INFECTIOUS DISEASES, COVID-19

Title: Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study

Authors: Biqing Chen[1], Hao Liang[2], Xiaomin Yuan[3], Yingying Hu[2], Miao Xu[2], Yating Zhao[4], Binfen Zhang[2], Fang Tian[1], Xuejun Zhu[1,2*]

Affiliations:

[1] Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[2] Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[3] Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.

[4] School of Atmospheric Sciences, Nanjing University, Nanjing, China

*Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.

Abstract

OBJECTIVES: This study aims to investigate the relationship between daily weather and transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of the COVID-19 spreading rate for a certain area with meteorological factors.

DESIGN: A retrospective, qualitative study.

METHODS AND ANALYSIS: We collected 382,596 records of weather data with four meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility, and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed cases in total) in nearly 500 areas worldwide from January 20 to April 9 in 2020. Epidemic data were modeled against weather data to find a model that could best predict the future outbreak.

RESULTS: Significant correlations of the daily new confirmed case counts with the weather 3~7 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average temperature at 5~15 ℃, relative humidity at 70%~80%, wind speed at 1.5~4.5 meter / second, and visibility less than 10 statute miles. A short-term model with these four meteorological variables in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-term model using temperature to predict the pandemic in the next week or month was derived. Taken China as a discovery dataset, it was well validated with worldwide data. According to this model, there are five viral transmission patterns, "restricted', "controlled", "natural", "tropical", "southern". This model's prediction performance correlates with actual observations best (over 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not much human interference by epidemic prevention measures.

CONCLUSIONS: This model can be used for prediction of the future outbreak, and illustrating the effect of epidemic control for a certain area.

Keywords: COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control

Strengths and limitations of this study

- This study investigates the role of daily weather in COVID-19 spread systematically with a comprehensive set of four meteorological factors.

- This research collected a huge amount of data, covering nearly 500 areas worldwide in a long timescale.
- The current study proposes mathematical models integrating meteorological information for predicting COVID-19 case counts in the future.
- The influence of weather on virus spread could be confounded by a dozen of manual interventions, such as population mobility and disinfection measures, leading to inaccurate modeling.
- The prediction model (especially the long-term model) might be unsuitable and inaccurate for areas with hot weather.

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 has spread all over the world and has unprecedented great social and economic impact worldwide[1,2]. It exhibits high human-to-human transmissibility compared to other coronavirus like SARS[3]. It would be crucial to predict the future trend of COVID-19 outbreak ahead, in order to make proper prevention and control strategies accordingly in time.

Besides population mobility and human-to-human contact, meteorological conditions have been suggested to be involved in the transmission of droplet-mediated viral diseases[4,5]. As droplets carrying the coronavirus can travel in gaseous clouds as far as eight metres and stay suspended in the air for hours[5], the suspending time and viability of the coronavirus outside body would be largely affected by the environment. Wind speed could affect the suspending time of droplets, while visibility and humidity reflect the amount of particles in the air, determining the coronavirus payload. Temperature affects virus's viability in the environment. As SARS-CoV-2 is enveloped, it might be more vulnerable to adverse conditions like high temperature.

The impact of weather on epidemiology has been mentioned in human's history. The ancient Chinese had a theory called "Five Movement and Six Weather" to study climate change and its relationship with human health. Currently, there are a few studies on preprint servers discussing the relationship of temperature and humidity with the pandemic, but none is systematical investigation or proposes validated practical model for prediction[6–10].

Herein, this study intends to investigate the relationship between meteorological factors and epidemic transmission rate on a world scale. Four meteorological variables, i.e., average temperature, relative humidity, wind speed, and visibility, were collected as well as the confirmed case counts daily for 81 days from January 20, 2020 to April 9, 2020 for nearly 500 areas around the world, including over 400 Chinese cities and areas, 18 Italian provinces, and 13 other countries. Five time point's delay of virus infection from exposure were considered and compared to determine the most reasonable time point's delay. A multivariate polynomial regression model with meteorological factors as a "weather coefficient" of the existing confirmed case count was established in a discovery Chinese dataset, and then validated by worldwide data. Five transmission modes, indicating different levels of epidemic control, were revealed by this model. In this view, this model can not only predict future outbreak, but also be used to evaluate the effect of epidemic prevention measures for a certain area.

Materials and Methods

Epidemiological data

Epidemiological data were collected from the World Health Organization (WHO)[11], European Centre for Disease Control and Prevention, and DXY-COVID-19-Data[12]. The daily new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries, namely, United States, United Kingdom, Germany, France, Spain, Iran, Korea, Japan, Australia, South Africa, India, Thailand, and Singapore. Considering the potential confounding effect, only Chinese cities with no less than 50 cumulative confirmed cases in one month and without official reports of large imported cases (42 cities in total) were taken as a discovery dataset, while those for Italian provinces and all the other nations were taken as replication datasets (Supplementary Materials).

Weather data

Four meteorological variables were chosen, air temperature, relative humidity, wind speed, and visibility. Temperature could affect virus viability in the environment. Wind speed could

98 affect the suspending time of virus-attached particles. Relative humidity reflects the amount of droplets in the air. Visibility is influenced by the amount of particles such as dust and air pollutants. These two parameters both affect the amount of mediator for the virus to stay in the air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of meteorological observations and geographic factors (latitude and elevation) from the Integrated Surface Database of USA National Centers for Environmental Information[13]. Daily data were calculated by averaging the hourly data for each variable in each day.

106 Statistical modeling

107 The number of daily new confirmed cases was taken as a dependent variable. Four meteorological variables, namely, average temperature, wind speed, visibility, and relative humidity, and the existing confirmed case counts were taken as independent variables. Considering that there is a latency stage from the day one get infected to the day being confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were collected needs to be taken into consideration. As it is reported that the latency period for COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection were taken into consideration, that is, weather data and existing confirmed cases count data were collected on the day, three days before, seven days before, 3~7 days before, and 14 days before collecting the new confirmed case count data.

117 To investigate whether the influence of meteorological factors is linear or quadric, both linear and non-linear modeling were performed under different relationship assumptions to compare model fitness statistics. Each meteorological variable was fitted into a bunch of single-factor models (either generalized linear model or polynomial model) through non-linear least

121 squares (NLS) modeling using the Wuhan dataset with a 3~7 days delay of infection. The
122 relationship between each meteorological variable and confirmed new case count (linear or
123 quadric) was identified based on model fitness (log-likelihood, Akaike information criterion,
124 Bayesian Information Criterion, etc.) and common knowledge of droplet-mediated viral diseases.

125 Second, the proper time delay from weather exposure to COVID-19 confirmation was
126 investigated in the Wuhan dataset through Loess regression interpolation and NLS modeling with
127 the previously identified relationship for each meteorological variable. The most possible time
128 delay identified was taken for subsequent analyses.

129 To investigate the degree of contribution for each meteorological factor to the COVID-19
130 case counts, Spearman's correlation test (a non-parametric method that measures the strength and
131 direction of associations) was first adopted, with the Wuhan dataset under the assumption of
132 previously defined time delay. Nevertheless, here we assumed monotonic correlations between
133 COVID-19 case count and meteorological variables, while we could not exclude the possibility
134 that the real relationship was not monotonic, which might impede the accuracy of correlation
135 analysis. Then, we performed single-factor NLS regression modeling for each meteorological
136 variable in the discovery dataset under the assumption of previously determined relationship and
137 pre-defined time delay, to determine the exact coefficients accompanied with each
138 meteorological factor and to find out the most suitable environmental condition for SARS-CoV-2.

139 Then, two final prediction models (short-term model and long-term model) were developed
140 using the discovery dataset with the previously determined coefficients. The prediction model
141 supposed that all the meteorological variables, with their specific coefficients determined by
142 single-factor modeling, were added together to compose a weather coefficient. The new
143 confirmed case count on the test day is calculated by multiplying the weather coefficient with the
144 existing confirmed case count on the exposure day (the time delay between test day and exposure

day is determined in previous analysis), and then multiply by a constant coefficient. The short-term model took all four variables, while the long-term model only considered temperature as it is easy to be forecasted. There was a constant coefficient for the total equation, that was multiplied by the existing confirmed case count. Its exact value was determined by model fitting in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was investigated with all datasets covering the world's top cities and areas. The correlation of existing confirmed case counts with newly confirmed case counts was also investigated. Basic statistics and modeling was conducted in R 3.5.1 (https://cran.*r*-project.org/).

Model validation and application

The best fitted model was validated in the replication datasets (Italian city-level data and other nation-level data) by correlating the observed actual epidemiological data with the predicted values from the model in the datasets. We used these fitted models to calculate a predicted value for case counts for each studied site, and then compared this predicted value with the real observed case counts by calculating a Spearman's correlation coefficient ρ between them.

Patient and Public Involvement

No specific patients were included in the current study. Epidemiological data were downloaded from online open-source databases. The public were not involved in the planning and design of the study.

Results

The Weather's influence on SARS-CoV-2 transmission displays 3~7 days time delay

The ranges of average temperature, relative humidity, wind speed, and visibility in the replication datasets were similar to those in the discovery dataset (see Supplementary Results for detailed datasets description). Non-linear modeling with Wuhan dataset under the assumption of 3~7 days delay of infection confirmation suggested that the effect of temperature and wind speed is better depicted as quadric (Table S1), which was also supported by Loess regression interpolation (Fig. 1). The mode for relative humidity and visibility was hard to be determined, as statistics supported both relationships (Table S1). Considering the common knowledge of coronavirus transmission and the trend showed by Loess regression interpolation, relative humidity exerted its impact in a quadric trend while visibility exerted its impact in a linear trend (Fig. 1, Supplementary results).

Furthermore, investigation of the time delay effect in the Wuhan dataset showed that the number of confirmed new cases was best correlated with air temperature 3~7 days ago, relative humidity and visibility 7 days ago, and wind speed on the exposure day (Table S2). By comprehensive consideration of all four meteorological variables and the differences between statistics values, the weather 3~7 days ago, as well as weather one week ago, could well predict COVID-19 outbreak. It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure under certain adverse weather might exhibit its effect after 3~7 days.

Contribution of single meteorological factor to the outbreak

In the Wuhan dataset, the new case count was significantly positively correlated with temperature (Spearman's correlation ρ = 0.69, p < 0.001) and visibility (ρ = 0.43, p =0.04), and negatively correlated with wind speed (ρ = -0.45, p = 0.03) and relative humidity (ρ = -0.33, p = 0.12) 3~7 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind speed, visibility, and relative humidity. A model only with temperature as a parameter could already explained 45% of the variance in the epidemic data ($p = 4\times10^{-4}$), while wind speed and visibility could explain over 25% of the variance. According to the fitted single-factor models (temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility was fitted into a linear model, see the Supplementary Results for details), SARS-CoV-2 transmission reaches a peak when mean temperature is 6.18 ℃ (Fig. 2A), relative humidity is 78.47% (Fig. 2B), and wind speed is 1.88 meter /second (m/s) (Fig. 2C); and its transmission rate decreases with the increase of visibility (Fig. 2D). The effects of geographic factors such as latitude and elevation, and the pure influence from the existing case count were further investigated in the worldwide datasets (Fig. S1), illustrating that COVID-19 mainly outbreaks at latitude 30°~50° (Fig. S1A) and elevation < 500 metre (Fig. S1B). New confirmed case count was positively correlated with the existing confirmed case count (Fig. S1C).

Short-term prediction model

We further derived a full model combined with all four meteorological variables and fitted this model with the discovery dataset (Supplementary Results). The best-fitted short-term model was as follows:

$$\begin{aligned}\text{New Case Count} &= \left(-0.11 \times \mathrm{T}^2 + 1.40 \times \mathrm{T} - 0.058 \times \mathrm{RH}^2 + 9.04 \times \mathrm{RH} - 1.36 \times \mathrm{SPD}^2 + 5.12 \right.\\ &\left. \times \mathrm{SPD} - 7.02 \times \mathrm{VSB} - 126.66\right) \times \alpha \times \text{Existing Confirmed Case Count}\end{aligned}$$

where T is temperature in ℃, RH is relative humidity in percentage (defined as over 15%), SPD is wind speed in m/s, VSB is visibility in statute miles, α is a site-specific constant, with a default of 0.001. All parameters take the means of values 3~7 days before the day new case count is evaluated.

In this model, all the four meteorological variables are added together in their proper forms to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate of SARS-CoV-2, and thus influences the number of people that catch infection from the existing confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a multiplicative constant coefficient α in the equation, which seems site-related. This constant coefficient could adjust the strength of the "weather coefficient" on disease transmission. When we substitute replication datasets into this short-term model with the multiplicative constant coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious underestimation of predicted values against real ones was observed although the predicted values correlated with the real ones very well. We supposed it was due to site-specific difference in the multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the pandemic had been controlled early. Thus, we further re-fitted this composed model with all datasets to determine a more accurate value of the multiplicative constant coefficient α, which was 0.001 then. In practical application, we need to first plot the observed case count vs. predicted one with a default α value 0.001, and then examine the extent of underestimation or overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the impact size of "weather coefficient" for a certain site.

Substitute data from the past two months, a good prediction performance was obtained for this short-term model, with the predicted values significantly correlated to the observed ones for most areas (Fig. 3). However, only the existing confirmed case count data could not predict the new case count 3~7 days later as well as the weather-combined model did (Table S3).

Different modes of viral transmission illustrated by the model

The observed versus predicted data exhibited different correlation patterns for different areas, meaning different viral transmission modes, which may indicate the effect of epidemic control for certain area.

Data from Chinese top-affected cities were not very well predicted and obviously overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p < 0.001$; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention measures at the beginning of the pandemic. This viral transmission mode suggested by the not well correlated prediction pattern is called "restricted".

For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the predicted outbreak was well correlated with the actual observations at the beginning when the existing confirmed cases were not in very large numbers, but the prediction deviates from the observation as the confirmed cases increase, in detail, there's large overestimation of prediction ($\rho_{Wuhan} = 0.69$, $\rho_{Italy} = 0.87$, $\rho_{Japan} = 0.80$, $\rho_{Iran} = 0.86$, $p < 0.001$, $\rho_{Korea} = 0.43$, $p = 0.002$; Fig. 3B). It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15, the day when shelter hospitals had been put into use for seven days (the average latency period for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention and control taken in these areas is effective (so-called "controlled" mode). The number of cases had been decreased by 72% for Wuhan, over 95% for Korea, Japan, and Italy, and 37% for Iran at most due to epidemic control (the largest gap between prediction and observation).

For most European and American countries, the predicted outbreak was linear correlated with the observed data very well ($\rho_{France} = 0.96$, $\rho_{United\ States} = 0.93$, $\rho_{United\ Kingdom} = 0.83$, $\rho_{Spain} = 0.97$, $\rho_{Germany} = 0.94$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without much man-made epidemic prevention and control measures. Estimation of daily new case counts by this short-term model performed very well for European countries, while this model underestimated the outbreak in the United States.

Although the weather is not suitable for tropical areas, the viral transmitted in natural mode, manifested as good linear correlation between the prediction and the observation ($\rho_{India} = 0.94$, $\rho_{Singapore} = 0.66$, $p < 0.001$, $\rho_{Thailand} = 0.56$, $p = 0.001$; Fig. 3D), with just relatively small daily new case counts compared to temperate regions.

Countries in the southern hemisphere displayed similar pattern as the "controlled" with large overestimation by the model when the confirmed cases increase, leading to not good prediction performance ($\rho_{Australia} = 0.79$, $p < 0.001$, $\rho_{South\ Africa} = 0.34$, $p = 0.08$; Fig. 3E). It might be due to the effect of epidemic prevention measures and hot summer weather in these countries.

<u>Long-term simplified model</u>

Long-term prediction depends on weather forecast, which generally reports only average temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as temperature in a short time delay (3~7 days ago), we again performed single-factor regression modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a quadric function (Supplementary Results). This simplified model with average temperature as a weather factor was derived as follows:

$$\text{new case count} = \left(-0.10 \times T^2 + 1.11 \times T + 46.42\right) \times \beta \times \text{Existing Confirmed Case Count}$$

where T is temperature in ℃, β is a site-related multiplicative constant coefficient, with a default of 0.006. All parameters take values 14 days before the day new case count is evaluated.

With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, $p < 0.001$; Fig. 3F). The long-term simplified prediction model also showed five prediction-observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the studied areas. This model could directly predict the newly emerging cases 14 days later, and be used to predict COVID-19 outbreak in the future month by summing up the daily new case count and combining weather forecast (usually available for the future 15 days).

Discussion

This research discovers nonlinear dose-response relationship for meteorological factors, in consistency with previous studies[7]. Predictions of COVID-19 outbreak scale by the models were well correlated with the observations around the world, suggesting the importance of weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many respiratory infectious diseases, such as influenza, is dependent upon temperature and relative humidity[4]. Recent published papers on preprint servers have reported roles of temperature and absolute humidity in the COVID-19 transmission, but their conclusions are diverse[6–10]. In contrast to the findings by Cai et al[10], this study suggests significant impact of mean temperature on the daily new case count, indicating a need for sufficient time delay between exposure and confirmation for weather to exhibit its effect. In contrary to other two studies[6,7], this research suggests that there is a relatively not wide temperature and humidity ranges for the pandemic. There is an optimal temperature for SARS-CoV-2 at 6.18 ℃, which is colder than that suggested by Bu et al[9] but in consistency with the estimation by Wang et al[7]; and most areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al suggested[9]. It is of notice that different from other viral respiratory diseases such as influenza[14,15], high relative humidity is better for SARS-CoV-2 to spread, suggesting that a sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more important for the spread than the effect of dry air on the human immune system. Different from other studies[16], this study also finds significant involvement of wind speed, in a quadric manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In addition, the current study discovered that visibility was significantly negatively correlated with new case count and played a more important role in viral spread than humidity did (from spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g., dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case count decreases rapidly when visibility is high than 13 statute miles, indicating that caution should be taken if visibility drops below 10 statute miles.

In the prediction model, there is a multiplicative constant coefficient which determines the strength of the weather coefficient on the epidemic transmission. It seems site-specific, as adjusting it could make the prediction for one site very close to the observation. This constant might reflect the influence of a couple of site-specific confounding factors, such as epidemic control measures, sun radiation, and population density. Various degrees of isolation for various areas around the world lead to different degrees of weather effect. When evaluate the prediction performance by the short-term model and the long-term model, they both exhibit different prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of epidemic control and isolation policy would lead to deviation from the original prediction and thus different prediction-observation correlation patterns. Therefore, by plotting the predicted versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the observed case counts dropped dramatically from the predictions for Wuhan seven days after their shelter hospitals were put in use, suggesting the importance and necessity of building shelter hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to natural transmission mode. Another thing worth attention is that although the weather in tropical areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore, these tropical areas should still be on the alert against future outbreak of COVID-19.

Although those cases with travel history to China or indicated by the World Health Organization as "imported case only" were excluded in this study to make the world data most likely local transmitted, it was difficult to separate the imported cases from local transmission very well in practice. It might explain the not excellent correlations of predictions with observations. Furthermore, the relationship of weather and COVID-19 could be complex, since the human immune system has an innate seasonal rhythm, and the immune system could also be affected by weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets for virus to adhere.

There are several limitations of this study. First of all, this prediction model (especially the long-term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter, as the models were derived using Chinese datasets, mainly in the first three months of 2020. The prediction became inaccurate and even improper under hot weather (i.e., the predicted values of long-term model become negative when air temperature is higher than 28 ℃), which might explain the obvious bad prediction performance for countries in the southern hemisphere and tropical areas. One explanation for the inaccurate prediction in areas with high temperature could be that SARS-CoV-2 transmission in these areas was mainly not influenced by weather, but in another direct transmission way, such as face-to-face contact or spread in gathering crowd. Second, it seems that the prediction performance drops with the increase in new case count, suggesting that the prediction model might become inaccurate and not suitable for very large new case count. This could be due to (1) the influence of weather on COVID-19 spread might weaken when the number of cases increases, while other factors such as social distance become more important at a later stage; (2) there was less data points with large new case count, which might lead to larger variance. Third, the short-term prediction model must use all four meteorological factors, while these factors are not always available for any one certain area. Fourth, this study included various areas covering a long period into modeling, thus, there were a bunch of variable confounding factors, such as population mobility and disinfection measures, which were not controlled and thus could impede the model accuracy. Fifth, as we could only obtain country-level epidemiological data, the corresponding meteorological data were obtained for their capital cities, leading to not exact pairing of epidemiological data and meteorological data. Sixth, there is a general lack of data and cases in the current study, since we only collected data covering two and a half months while the pandemic has persisted over seven months up to now.

Conclusion

In summary, this study has found significant correlations with the COVID-19 epidemic trend for not only temperature and humidity, but also wind speed and visibility. It proposed a comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version and a long-term version. The short-term version uses the combination of four meteorological factors as a "weather coefficient" of the existing confirmed case count in the past week and can be used to predict epidemic situation in the future three days; the short-term version uses average temperature as the "weather coefficient" seven days ago and can predict the outbreak in one month if combined with weather forecast. This model is easy to use for predicting the COVID-19 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case count for the future couple of days or month. This model will be very helpful for local governments to make timely policies on epidemic control, for instance, the allocation of medical equipments such as ventilators and medical resources such as hospitals, beds and health-care workers, according to the prediction results.

Acknowledgments: We thank Dr. Zhisheng Huang for advices on data collecting and processing, Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on modeling, Dr. Zhongfa Yang for advices on manuscript revision.

Funding: the Priority Academic Program Development of Jiangsu Higher Education Institutions – the third period (NO.035062002003c), the National Natural Science Foundation of China (NO. 82001206), the Yizhong Research Fund of Jiangsu Provincial Hospital of Chinese Medicine (Y19066).

Author contributions: BC, YZ and XZ design and interpret the reported analyses and results; HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse the data, drafted, and revised the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical support; XZ supervised the research and revised the manuscript.

Competing interests: Authors declare no competing interests.

Data and materials availability: Weather data and epidemiological data is all obtained from public databases. Detailed modeling results are available upon request by emailing to Biqing Chen, bq_chen@qq.com.

References:

1 Zhu N, Zhang D, Wang W, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727–33. doi:10.1056/NEJMoa2001017

2 Dewey C, Hingle S, Goelz E, *et al.* I DEAS AND O PINIONS Supporting Clinicians During the COVID-19 Pandemic. 2020;**2019**:2019–21. doi:10.7326/M19

3 Chan JFW, Yuan S, Kok KH, *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 2020;**395**:514–23. doi:10.1016/S0140-6736(20)30154-9

4 Lowen AC, Mubareka S, Steel J, *et al.* Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathol* 2007;**3**:e151.

5 Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential Implications for Reducing Transmission of COVID-19. *JAMA* 2020;:E1–2. doi:10.1001/jama.2020.4756

6 Proverbio AM, Crotti N, Manfredi M, *et al.* Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. *medRxiv* Published Online First: 2020. doi:https://doi.org/10.1101/2020.03.18.20036731

7 Wang M, Jiang A, Gong L, *et al.* Temperature Significantly Change COVID-19 Transmission in 429 cities. *Block Caving – A Viable Altern* Published Online First: 2020. doi:https://doi.org/10.1101/2020.02.22.20025791.

8 Luo W, Majumder MS, Liu D, *et al.* The role of absolute humidity on transmission rates of the COVID-19 outbreak. *medRxiv* 2020;:2020.02.12.20022467. doi:10.1101/2020.02.12.20022467

9 Bu J, Peng D-D, Xiao H, *et al.* Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020. *medRxiv* 2020;:2020.02.13.20022715. doi:10.1101/2020.02.13.20022715

10 Cai Y. The Effects of "Fangcang, Huoshenshan. https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3

11 Organization WH. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. 2020.https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed 29 Mar 2020).

12 https://github.com/BlankerL/DXY-COVID-19-Data.

13 ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/. 2020.

14 Peci A, Winter AL, Li Y, *et al.* Effects of Absolute Humidity, Relative Humidity, Temperature, and Wind Speed on Influenza Activity in Toronto, Ontario, Canada. *Appl Env Microbiol* 2019;**85**:e02426-18. doi:10.1128/AEM.02426-18

15 Kudo E, Song E, Yockey LJ, *et al.* Low ambient humidity impairs barrier function and innate resistance against influenza infection. *Proc Natl Acad Sci U S A* 2019;**116**:10905–10.

16 Oliveiros B, Caramelo L, Ferreira NC, *et al.* Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. *medRxiv* Published Online First: 2020. doi:https://doi.org/10.1101/2020.03.05.20031872

FIGURE LEGENDS

Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time

point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

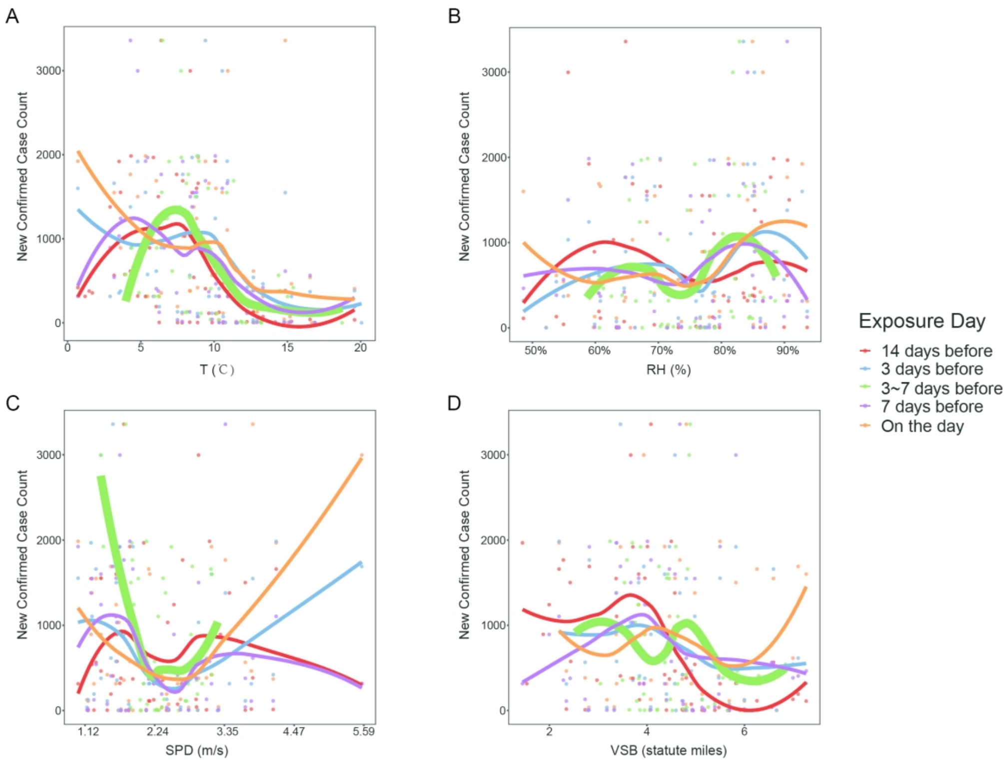


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of confirmation from viral infection are displayed together in one figure, namely, exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.

207x156mm (300 x 300 DPI)

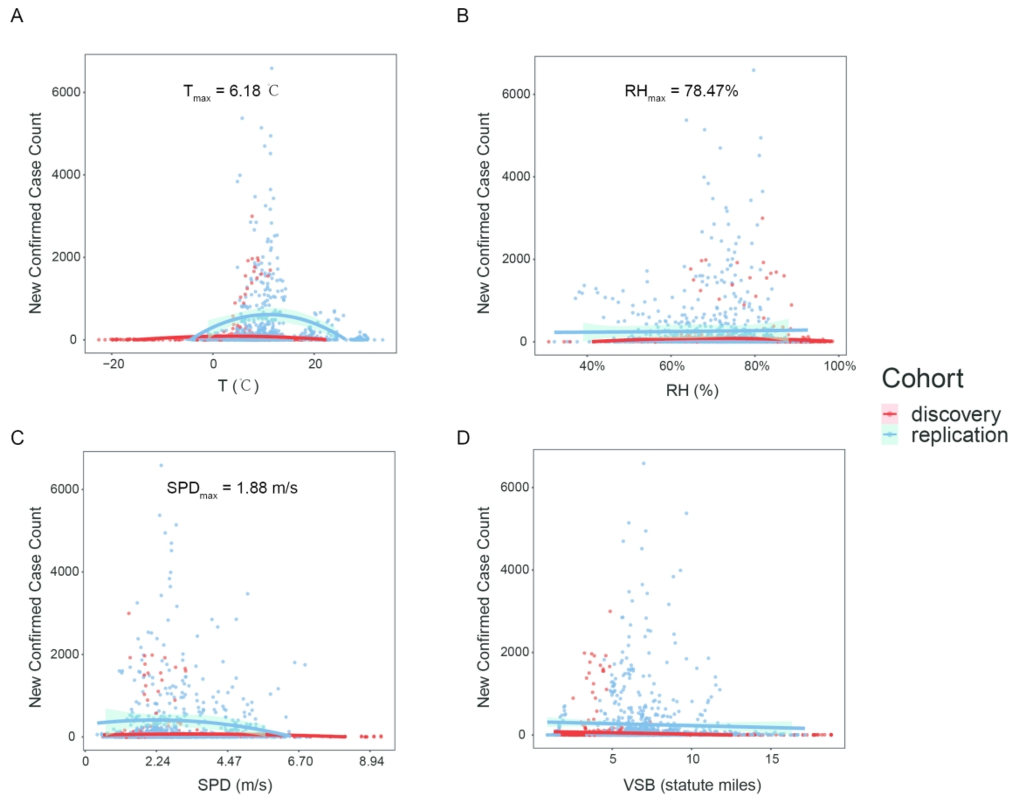


Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in ℃, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset. Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset includes the major outbreak Chinese cities, while the replication datasets included provincial data in Italy, and national data around the world(except China).

218x172mm (300 x 300 DPI)

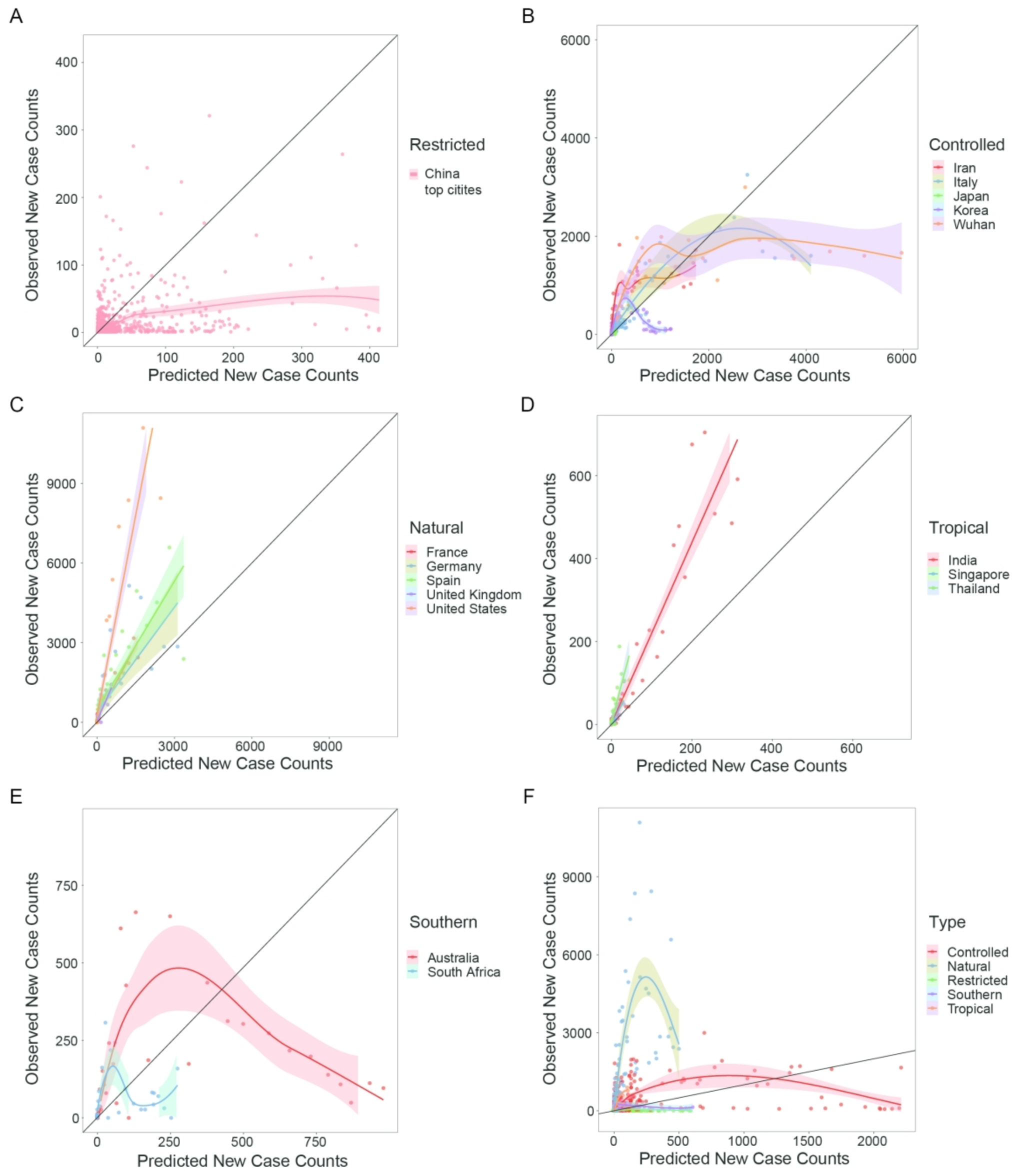


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern" pattern including countries in the southern hemisphere, Australia and South Africa. Each dot represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black solid line represents that the observed values are equal to the predicted ones, and dots closer to this line means better prediction performance.

209x244mm (300 x 300 DPI)

Supplementary Materials and Methods

Epidemiological data

We scrutinized WHO's situation reports to rule out these countries with only imported cases, and only collected the confirmed cases with possible or confirmed local transmission (i.e., without recent travel history to China).

For Wuhan city, there was a shortage of test kits at the beginning of the pandemic, which would make confirmed case counts much lower than the actual data, thus, we discarded epidemic data before January 28th, the day when domestic test kits have been approved, produced in large quantities, and were available for Wuhan hospitals. As there was a cut down problem for the existing confirmed case count on February 20th for Wuhan, when modeling with the existing confirmed case count, only data before February 20th were used.

Weather data

Temperature and dew point displayed in Fahrenheit were transformed into Celsius forms, and relative humidity was calculated from temperature and dew point using the following formula for each time point:

$$\mathrm{RH} = \begin{cases} e^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T < 0 \\ 10^{\frac{7.5\mathrm{D}}{237.3+\mathrm{D}} - \frac{7.5\mathrm{T}}{237.3+\mathrm{T}}} \times 100\%, & T \geq 0 \end{cases}$$

where RH is the relative humidity, D is the dew point in degrees Celsius, T is the temperature in degrees Celsius, and *e* is the base of the natural log.

For each city with epidemiological data, the meteorological station in that city or that was closest to the latitude and longitude coordinates of the city center was chosen.

For a city with more than one meteorological stations, the one nearest to the city center was chosen. For a province with epidemiological data, the meteorological station in the capital city of that province was chosen. For a country with only national wide epidemiological data, weather data were averaged across all the meteorological observatories in the cities where outbreak was officially reported. Latitude and elevation for the meteorological observatories were also collected.

Statistical modeling

Only one city Wuhan was chosen for illustrating the time delay effect because it is the first city to have an outbreak of COVID-19, there was none reported imported cases for Wuhan, which might obscure the correlation between weather and virus transmission.

Supplementary Results

Datasets description

Only Chinese cities with monthly confirmed cases over 50 were included in the discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in all the other Chinese cities. We suppose that it might be due to abrupt large supplement of virus test kits or data correction on that day. In order to reduce the potential contamination of modeling by this outlier, data on that day were discarded from the subsequent analysis. There were also two oddly large new confirmed case counts for Lombardy, which were discarded from the subsequent analysis. Except the

outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to 2,997, the average temperature ranged -22.54℃ ~ 22.16℃, the wind speed ranged 0.56 ~ 9.29 meter per second, visibility ranged 1.3 ~ 18.8 statute miles, and relative humidity ranged 30.84% ~ 98.52%.

Model selection

With the increase of relative humidity, the amount of droplets in the air increases, leading to more virus load. However, as the air gets humid, human's respiratory tract could better defend virus infection. Thus, the relationship of relative humidity could be complex, not pure linear. Giving comprehensive consideration, we defined the effect of relative humidity to be quadric. As for visibility, it only affects the amount of particles in the air, which is positively correlated with virus load. Thus, it is most probably to exert its effect linearly.

Although relative humidity and visibility 7 days ago correlated with the confirmed new case counts best, there was not great loss of model fitting statistics for relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days time delay and 3~7 days time delay for temperature.

Fitted models

The fitted single-factor models were as follows:

$$\text{New Case Count} = -0.11305 \times \mathrm{T}^2 + 1.39819 \times \mathrm{T} + 45.11405$$

where T is temperature in ℃.

The estimate p-value for constant was < 0.001. The extremum was $-1.39819/(2 \times (-0.11305)) = 6.183945$ ℃.

$$\text{New Case Count} = -0.05759 \times \text{RH}^2 + 9.038 \times \text{RH} - 303.0$$

where RH is relative humidity in percentage.

The extremum was $-9.038/(2 \times (-0.05759)) = 78.46848$ %.

$$\text{New Case Count} = -1.360056 \times \text{SPD}^2 + 5.120123 \times \text{SPD} + 42.1855$$

where SPD is wind speed in meter per second (m/s).

The extremum was $-5.120123/(2 \times (-1.360056)) = 1.882321$ m/s.

$$\text{New Case Count} = -7.021 \times \text{VSB} + 89.041$$

where VSB is visibility in statute miles.

The estimate p-value for VSB was < 0.01, constant was < 0.001.

Thus, the complex short-term model to be regressed was

$$\begin{aligned}\text{New Case Count} &\\ &= (-0.11 \times \text{T}^2 + 1.40 \times \text{T} - 0.058 \times \text{RH}^2 + 9.04 \times \text{RH} - 1.36 \\ &\times \text{SPD}^2 + 5.12 \times \text{SPD} - 7.02 \times \text{VSB} - 126.66) \times \text{a} \\ &\times \text{Existing Confirmed Case Count}\end{aligned}$$

where a is a constant to be fitted. All parameters take values 3~7 days before the day new case count is confirmed.

Through fitting this full model with the discovery data, a was estimated to be 0.0004786 (standard error 0.0000128, *p*-values < 2e-16).

For long-term model, the fitted model with temperature 14 days ago was as follows:

$$\text{New Case Count} = -0.10062 \times \mathrm{T}^2 + 1.11189 \times \mathrm{T} + 46.41792$$

77 The estimate p-value for constant was < 0.001. The extremum was $-1.11189/$

78 $(2 \times (-0.10062)) = 5.525194.$

79 Thus, the simplified long-term model to be regressed was:

$$\text{New Case Count} = (-0.10 \times \mathrm{T}^2 + 1.11 \times \mathrm{T} + 46.42) \times \mathrm{b} \times \text{Existing Confirmed Case Count}$$

80 where b is a constant to be fitted. All parameters take values 14 days before the day

81 new case count is confirmed.

82 Through fitting this simplified model with the discovery data, b was estimated to

83 be 0.0061382 (standard error 0.0002666, *p*-values < 2e-16).

Table S1. Model fitness statistics for comparing and selecting proper fitting relationship

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Linear	493	4.5×10^{-8}	-167	339	342	4860391	0.757
Quadric	421	1.3×10^{-7}	-163	333	337	3370230	0.812
Relative humidity							
Linear	627	9.8×10^{-8}	-172	350	353	7855418	0.401
Quadric	626	8.4×10^{-6}	-171	351	355	7442367	0.358
Wind speed							
Linear	585	3.1×10^{-8}	-170	347	350	6840545	0.380
Quadric	546	2.4×10^{-7}	-168	344	349	5654728	0.423
Visibility							
Linear	594	3.3×10^{-8}	-171	347	351	7059799	0.354
Quadric	598	7.9×10^{-7}	-170	349	353	6799355	0.358

Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real values and the predicted values by the predisposed model.

91 Table S2. Model fitness statistics for comparing and selecting proper time delay of

92 virus exposure

	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Temperature							
Day 0	626	$2.6{\times}10^{-8}$	-171	351	355	7441513	0.330
Day -3	605	$1.3{\times}10^{-8}$	-171	349	353	6953553	0.479
Day -7	664	$5.4{\times}10^{-8}$	-173	353	358	8386957	0.262
Day -14	528	$1.1{\times}10^{-7}$	-168	343	347	5297229	0.534
Day -3 ~ -7	421	$1.3{\times}10^{-7}$	-163	333	337	3370230	0.812
Relative humidity							
Day 0	605	$5.9{\times}10^{-6}$	-171	349	353	6953396	0.389
Day -3	679	$4.3{\times}10^{-6}$	-173	354	359	8768069	0.065
Day -7	560	$5.0{\times}10^{-8}$	-169	346	350	5962416	0.524
Day -14	605	$9.1{\times}10^{-6}$	-171	349	353	6962609	0.326
Day -3 ~ -7	626	$8.4{\times}10^{-6}$	-171	351	355	7442367	0.358
Wind speed							
Day 0	526	$7.4{\times}10^{-8}$	-167	343	347	5251026	0.500
Day -3	663	$1.4{\times}10^{-8}$	-173	353	357	8343427	0.268
Day -7	559	$1.1{\times}10^{-8}$	-169	346	350	5926891	0.516
Day -14	674	$5.2{\times}10^{-8}$	-173	354	358.	8643076	0.014
Day -3 ~ -7	546	$2.4{\times}10^{-7}$	-168	344	349	5654728	0.423
Visibility							

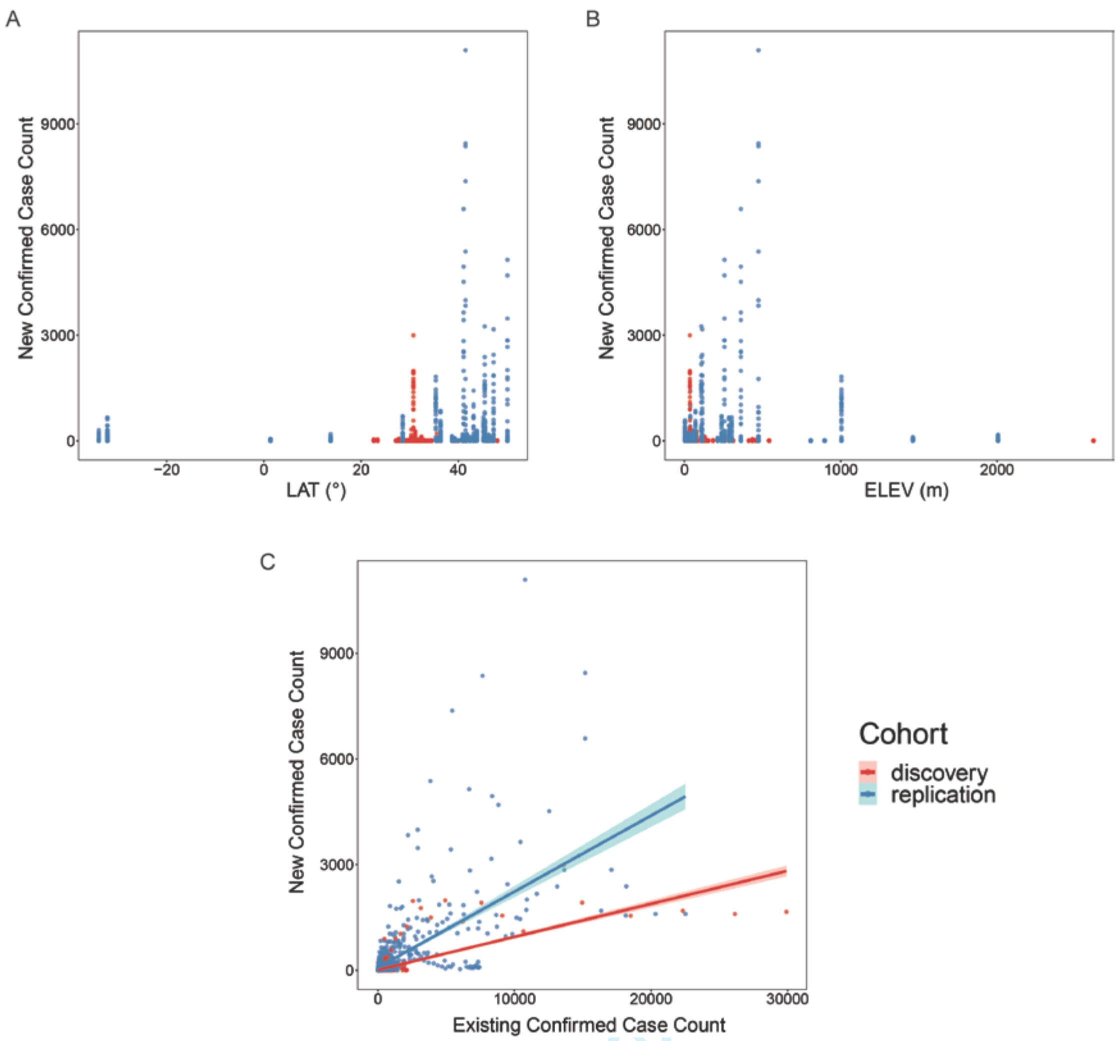
Day 0	646	4.2×10^{-9}	-173	351	354	8343221	0.286
Day -3	663	5.1×10^{-8}	-173	352	355	8804055	0.016
Day -7	514	3.9×10^{-8}	-168	341	344	5290247	0.502
Day -14	635	1.1×10^{-8}	-172	350	354	8052388	0.272
Day -3 ~ -7	594	3.3×10^{-8}	-171	347	351	7059799	0.354

Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information Criterion for the model; deviance, deviance of the model; Corr, Spearman's correlation coefficient between the real values and the predicted values by the predisposed model.

98 Table S3. Model fitness statistics for weather-combined model and epidemic only
99 model

Model	sigma	finTol	logLik	AIC	BIC	deviance	Corr
Weather-combined	147	1.8×10^{-9}	-6239	12481	12491	21128810	0.171
Epidemic-only	149	2.1×10^{-8}	-6251	12507	12517	21689551	0.152

100 Note: The weather-combined model is the short-term model with multiplicative constant to be fitted. The
101 epidemic-only model is the model only with existing confirmed case count as an independent variable, assuming a
102 linear function.



103

104 **Fig. S1.** Scatterplots of new confirmed case count to (A) latitude, (B) elevation, and
105 (C) the existing confirmed case count, for all the studied sites. Linear regression (C)
106 interpolation curves are illustrated for each dataset, with 95% confidence intervals
107 showing in shadow.