

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 **Title: A generalized model for predicting the local COVID-19 outbreak**
5
6 **around the world based on meteorological conditions**
7
8

9 **Authors:** Biqing Chen¹, Hao Liang², Xiaomin Yuan³, Yingying Hu², Miao Xu², Yating Zhao⁴,
10
11 Binfen Zhang², Fang Tian¹, Xuejun Zhu^{1,2*}
12
13

14
15
16 5 **Affiliations:**
17

18 ¹ Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the
19
20 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
21
22

23 ² Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated
24
25 Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
26
27

28 ³ Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the
29 10
30 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
31
32

33 ⁴ School of Atmospheric Sciences, Nanjing University, Nanjing, China
34
35

36 *Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.
37
38

39
40 15 **Abstract: Background** The COVID-19 has become a pandemic worldwide. **Methods** We
41
42 collected 382,596 records of weather data with four meteorological factors, i.e., average
43
44 temperature, relative humidity, wind speed, and visibility, and 15,192 records of epidemic data
45
46 with daily new confirmed case counts (1,587,209 confirmed cases in total) in over 500 areas
47
48 worldwide from January 20 to April 9. Epidemic data were modeled against weather data to find
49
50 a model that could best predict the future outbreak. **Results** Significant correlations of the daily
51 20
52 new confirmed case counts with the weather 3~7 days ago were found. SARS-CoV-2 is easy to
53
54 spread under weather conditions of average temperature at 7.9 °C, relative humidity at 70%~80%,
55
56
57
58
59
60

1
2
3 wind speed at 4~10 miles / hour, and visibility less than 10 statute miles. A short-term model
4
5 with these meteorological variables in the past 3~7 days was derived to predict the daily increase
6
7 in COVID-19; and a long-term model using temperature to predict the pandemic in the next week
8
9 or month was derived. Taken China as a discovery dataset, it was well validated with worldwide
10
11 data. According to this model, there are five different viral transmission pattern, "restricted",
12 5 "controlled", "natural", "tropical", "southern". This model's prediction performance correlates
13
14 with the actual observations best (over 0.9 correlation coefficient) under natural spread mode of
15
16 SARS-CoV-2 when there is not much human interference by epidemic prevention measures.
17
18 **Conclusion** This model can be used for prediction of the future outbreak, and illustrating the
19
20 effect of epidemic control for a certain area.
21
22
23
24 10

25
26
27
28
29 **Keywords:** COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control
30
31
32
33
34

35 **Strengths and limitations of this study**

- 36
37
38 15 ● The number of daily new confirmed cases is significantly correlated with the weather 3~7
39
40 days ago.
41
42
43 ● Average temperature at 7.9 °C, relative humidity at 70%~80%, wind speed at 4~10 miles /
44
45 hour, and visibility less than 10 statute miles are the best weather conditions for the spread of
46
47 SARS-CoV-2.
48
49
50
51 20 ● A short-term model consisted of four meteorological factors as a weather coefficient to
52
53 multiply with the extant confirmed cases could predict the new case count in the following
54
55 three days very well.
56
57
58
59
60

- 1
2
3 ● A long-term model with temperature could be used to predict the new case count in the next
4 week or month for a certain area.
5
6
7
8 ● As the prediction model could illustrate the effect of epidemic control for a certain area,
9
10 China and other early outbreak countries have effectively reduced more than 50% of the
11 potential outbreak.
12
13 5
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- 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.

For peer review only

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

1
2
3 temperature, relative humidity, wind speed, and visibility, were collected as well as the confirmed
4 case counts daily for 80 days for over 500 areas around the world. Five time delays of the
5 epidemic situation from the exposure day were considered and compared to determine the most
6 reasonable time delay. A multivariate polynomial regression model with meteorological factors
7 as a "weather coefficient" of the extant case count was established in a discovery Chinese dataset,
8 and then validated by worldwide data. Five transmission modes, indicating different levels of
9 epidemic control, were revealed by this model. In this view, this model can not only predict
10 future outbreak, but also be used to evaluate the effect of epidemic prevention measures for a
11 certain area.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

27 **Materials and Methods**

28 Epidemiological data

29
30
31
32 Epidemiological data were collected from the World Health Organization (WHO) (4),
33 European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (14). The daily
34 new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data
35 were obtained for every Chinese city or district as a discovery dataset, while those for 21 Italian
36 provinces and all the other nations were taken as replication datasets (Supplementary Materials).
37
38
39
40
41
42
43

44 Weather data

45
46 We obtained hourly values of meteorological observations and geographic factors (latitude
47 and elevation) from the Integrated Surface Database of USA National Centers for Environmental
48 Information (15). Temperature, dew point, wind speed, and visibility were collected, and relative
49 humidity was calculated accordingly (Supplementary Materials). Daily data were calculated by
50 averaging the hourly data for each variable in each day.
51
52
53
54
55
56
57
58
59
60

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

1
2
3 The average temperature, relative humidity, wind speed, and visibility ranges in the replication
4 datasets were similar to the discovery dataset (see Supplementary Results for detailed datasets
5 description). Regression interpolation showed that the weather 3~7 days ago was correlated with
6 the confirmed new case counts in a most reasonable manner, as well as weather one week ago.
7
8 The effects of temperature and relative humidity on the new confirmed case count exhibited a
9 bell-shaped trend, while wind speed and visibility were negatively correlated with the new case
10 count (Fig. 1). It coincided with the latency period of 3~7 days for SARS-CoV-2, that is,
11 exposure under certain adverse weather might exhibit its effect after 3~7 days.
12
13
14
15
16
17
18
19
20
21

22 Contribution of single meteorological factor to the outbreak

23
24
25 10 To elucidate the contribution of each meteorological factor to the case counts, we first performed
26 single-factor regression modeling for each meteorological variable in the discovery dataset.
27 Temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility
28 was fitted into a linear model. It was found that visibility was correlated with the outbreak best,
29 followed by temperature, relative humidity, and wind speed. The new case count was
30 significantly negatively correlated with visibility (Spearman's correlation $\rho = -0.14$, $p < 0.001$),
31 temperature ($\rho = -0.14$, $p < 0.001$), and wind speed ($\rho = -0.07$, $p < 0.001$), but positively
32 correlated with relative humidity ($\rho = 0.05$, $p < 0.001$). For Wuhan data, a model only with
33 temperature as a parameter could already explained 45% of the variance in the epidemic data ($p =$
34 4×10^{-4}), while wind speed and visibility could explain over 25% of the variance. According to the
35 fitted single-factor models for Wuhan, SARS-CoV-2 transmission reaches a peak when mean
36 temperature is 7.9 °C (Fig. 2A), relative humidity is 77.6% (Fig. 2B), and wind speed is 5.2
37 miles / hour (Fig. 2C). The effects of geographic factors such as latitude and elevation, and the
38 pure influence from the extant case count were further investigated (Fig. S1), illustrating that
39
40
41
42
43
44
45
46
47
48 20
49
50
51
52
53
54
55
56
57
58
59
60

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:

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

where T is temperature in °C, 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

1
2
3 most areas (Fig. 3). However, only the extant confirmed case count data could not predict the
4
5 new case count 3~7 days later as well as the weather-combined model did (Fig. S2).
6

7 Different modes of viral transmission illustrated by the model

8
9
10 The observed versus predicted data exhibited different correlation patterns for different areas,
11
12 meaning different viral transmission modes, which may indicate the effect of epidemic control
13
14 for certain area.
15

16
17 Data from Chinese top-affected cities were not very well predicted and obviously
18
19 overestimated by this model with the default multiplicative factor α (Pearson's correlation
20
21 coefficient $r = 0.31$, $p < 0.001$; Fig. 3A). It might be due to the reason that most Chinese cities
22
23 took actions quickly after the outbreak in Wuhan was reported, thus, these cities were under strict
24
25 epidemic prevention measures at the beginning of the pandemic. This viral transmission mode
26
27 suggested by the not well correlated prediction pattern is called "restricted".
28
29

30
31 For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the
32
33 predicted outbreak was well correlated with the actual observations at the beginning when the
34
35 extant confirmed cases were not in very large numbers, but the prediction deviates from the
36
37 observation as the confirmed cases increase, in detail, there's large overestimation of prediction
38
39 ($r_{\text{Wuhan}} = 0.47$, $p = 0.02$, $r_{\text{Italy}} = 0.86$, $r_{\text{Japan}} = 0.71$, $r_{\text{Iran}} = 0.64$, $p < 0.001$, $r_{\text{Korea}} = 0.06$, $p = 0.68$;
40
41 Fig. 3B). It is of notice that the dramatic deviation of predictions for Wuhan occurred after
42
43 February 15, the day when shelter hospitals had been put into use for seven days (the average
44
45 latency period for COVID-19). Therefore, the deviated prediction pattern indicates that the
46
47 outbreak prevention and control taken in these areas is effective (so-called "controlled" mode).
48
49 The number of cases had been decreased by 82% for Wuhan, over 95% for Korea, Japan, and
50
51 Italy, and 52% for Iran at most due to epidemic control (the largest gap between prediction and
52
53 observation).
54
55
56
57
58
59
60

1
2
3 For most European and American countries, the predicted outbreak was linear correlated
4 with the observed data very well ($r_{\text{France}} = 0.96$, $r_{\text{United States}} = 0.92$, $r_{\text{United Kingdom}} = 0.92$, $r_{\text{Spain}} =$
5 0.84 , $r_{\text{Germany}} = 0.73$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without
6 much man-made epidemic prevention and control measures. Estimation of daily new case counts
7 by this short-term model performed very well for European countries, while this model
8 underestimated the outbreak in the United States.
9

10
11 Although the weather is not suitable for tropical areas, the viral transmitted in natural mode,
12 manifested as good linear correlation between the prediction and the observation ($r_{\text{India}} = 0.95$,
13 $r_{\text{Singapore}} = 0.84$, $r_{\text{Thailand}} = 0.82$, $p < 0.001$; Fig. 3D), with just relatively small daily new case
14 counts compared to temperate regions.
15

16
17 Countries in the southern atmosphere displayed similar pattern as the "controlled" with large
18 overestimation by the model when the confirmed cases increase, leading to not good prediction
19 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
20 effect of epidemic prevention measures in these countries.
21

22 Long-term simplified model

23
24 For long-term prediction, another simplified model with average temperature as a weather factor
25 was derived as follows:
26

$$27 \text{ new case count} = (-0.14 \times T^2 + 0.93 \times T + 100) \times \beta \times \text{Extant Case Count}$$

28 where T is temperature in °C, β is a site-related constant, with a default of 0.003. All parameters
29 take values 7 days before the day new case count is evaluated.
30

31 With the model, the prediction performance was still good ($r = 0.64$ in the current datasets, $p <$
32 0.001 ; Fig. 3F). The long-term simplified prediction model also showed five prediction-
33

1
2
3 observation correlation patterns, indicating different modes of viral transmission, for the studied
4
5 areas.
6
7

8 **Discussion**

9

10
11 This research discovers nonlinear dose-response relationship for meteorological factors, in
12
13 consistency with previous studies (10). Predictions of COVID-19 outbreak scale by the models
14 5
15 were well correlated with the observations around the world, suggesting the importance of
16
17 weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many
18
19 respiratory infectious diseases, such as influenza, is dependent upon temperature and relative
20
21 humidity (5,6). Recent published papers on preprint servers have reported roles of temperature
22
23 and absolute humidity in the COVID-19 transmission, but their conclusions are diverse (8-13). In
24
25 10
26 contrast to the findings by Cai et al (8), this study suggests significant impact of mean
27
28 temperature on the daily new case count, indicating a need for sufficient time delay between
29
30 exposure and confirmation for weather to exhibit its effect. In contrary to other two studies (9,10),
31
32 this research suggests that there is a relatively not wide temperature and humidity ranges for the
33
34 pandemic. There is an optimal temperature for SARS-CoV-2 at 7.9 °C, which is colder than that
35
36 suggested by Bu et al (12) but in consistency with the estimation by Wang et al (10); and most
37 15
38 areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al
39
40 suggested (12). It is of notice that different from other viral respiratory diseases such as influenza,
41
42 high relative humidity is better for SARS-CoV-2 to spread, suggesting that a sufficient amount of
43
44 droplets in the air to support the suspension of SARS-CoV-2 is more important for the spread
45
46 than the effect of dry air on the human immune system. Different from other studies (13), this
47
48 20
49 study also finds significant involvement of wind speed, in a quadric manner, indicating that mild
50
51 wind might be more suitable for the virus to suspend in the air. In addition, the current study
52
53
54
55
56
57
58
59
60

1
2
3 discovered that visibility was significantly negatively correlated with new case count and played
4 a more important role in viral spread than humidity did. New case count decreases rapidly when
5 visibility is high than 13 statute miles, indicating that caution should be taken if visibility drops
6 below 10 statute miles.
7
8
9
10

11
12
13 5 In the prediction model, there is a constant multiplicative factor which determines the strength of
14 the weather coefficient on the epidemic transmission. It seems site-specific, as adjusting it could
15 make the prediction for one site very close to the observation. This constant might reflect the
16 influence of epidemic management and control measures. Various degrees of isolation for various
17 areas around the world lead to different degrees of weather effect. When evaluate the prediction
18 performance by the short-term model and the long-term model, they both exhibit different
19 prediction-observation correlation patterns, suggesting that changes in the degree of epidemic
20 control and isolation policy would lead to deviation from the original prediction and thus
21 different prediction-observation correlation patterns. Therefore, by plotting the predicted versus
22 observed new case counts and adjusting the multiplicative factor (α and β), it would be easy to
23 evaluate the effect of epidemic prevention measures. It is of notice that the observed case counts
24 dropped dramatically from the predictions for Wuhan seven days after their shelter hospitals were
25 put in use, suggesting the importance and necessity of building shelter hospitals for strict
26 isolation rather than just home isolation. With the use of shelter hospitals and very strict isolation
27 measures, the outbreak in one area could be reduced by 52~99% compared to natural
28 transmission mode. Another thing worth attention is that although the weather in tropical areas
29 like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on
30 spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore,
31 these tropical areas should still be on the alert against future outbreak of COVID-19.
32
33
34
35
36 15
37
38
39
40
41
42
43
44
45
46
47 20
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Although those cases with travel history to China or indicated by the World Health Organization
4 as "imported case only" were excluded in this study, leaving the world data most likely local
5 transmitted, it's difficult to separate the imported cases from local transmission very well in
6 practice, which might explain the not excellent correlations of predictions with observations.
7
8 Furthermore, the relationship of weather and COVID-19 could be complex, since the human
9 immune system has an innate seasonal rhythm, and the immune system could also be affected by
10 weather *vice versa*, for example, dry air would reduce the amount of mucus on the airway
11 mucosa, and thus increasing the probability of viral invasion, while wet air provides droplets for
12 virus to adhere.
13
14

15
16
17
18
19
20
21
22
23
24 In summary, this study has found significant correlations with the COVID-19 epidemic trend for
25 not only temperature and humidity, but also wind speed and visibility. It proposed a
26 comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version
27 and a long-term version. The short-term version uses the combination of four meteorological
28 factors as a "weather coefficient" of the extant case count in the past week and can be used to
29 predict epidemic situation in the future three days; the short-term version uses average
30 temperature as the "weather coefficient" seven days ago and can predict the outbreak in one
31 month if combined with weather forecast. This model is easy to use for predicting the COVID-19
32 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case
33 count for the future couple of days or month. This model will be very helpful for local
34 governments to make timely policies on epidemic control, for instance, the allocation of medical
35 equipments such as ventilators and medical resources such as hospitals, beds and health-care
36 workers, according to the prediction results.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

- 1
2
3 4. World Health Organization. 2020. [https://www.who.int/emergencies/diseases/novel-](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports)
4 coronavirus-2019/situation-reports (29 March 2020, date last accessed).
5
6
7
- 8 5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30
9 years of county-level evidence from the United States. *Amer J Epidemiol.* 2012;176(suppl.
10 7):S114–22.
11
12
- 13 6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on
14 relative humidity and temperature. *PLoS Pathol.* 2007;3:e151.
15
16
17
- 18 7. Bourouiba L. Turbulent gas clouds and respiratory pathogen emissions potential
19 implications for reducing transmission of COVID-19. *JAMA.* 2020;E1–2.
20
21
22
- 23 8. Cai Y, Huang Sr. T, Liu Sr. X, Xu Sr. G. The effects of "Fangcang, Huoshenshan, and
24 Leishenshan" makeshift hospitals and temperature on the mortality of COVID-19. Preprint
25 at Med RXIV. 2020. <https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3>
26 10 (29 March 2020, date last accessed).
27
28
29
30
31
32
33
34
- 35 9. Bannister-Tyrrell M, Meyer A, Faverjon C, Cameron A. Preliminary evidence that higher
36 temperatures are associated with lower incidence of COVID-19, for cases reported
37 globally up to 29th February 2020. Preprint at Med RXIV. 2020.
38 15 <https://www.medrxiv.org/content/10.1101/2020.03.18.20036731> (29 March 2020, date last
39 accessed).
40
41
42
43
44
45
46
- 47 10. Wang M, Jiang A, Gong L, et al. Temperature significantly change COVID-19
48 transmission in 429 cities. Preprint at Med RXIV. 2020.
49 <https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1> (29 March 2020, date
50 20 last accessed).
51
52
53
54
55
56
57
58
59
60

- 1
2
3 11. Luo W, Majumder MS, Liu D, et al. The role of absolute humidity on transmission rates of
4 the COVID-19 outbreak. Preprint at Med RXIV. 2020.
5
6 <https://www.medrxiv.org/content/10.1101/2020.02.12.20022467v1> (29 March 2020, date
7 last accessed).
8
9
10
11
12 5 12. Bu J, Peng D-D, Xiao H, et al. Analysis of meteorological conditions and prediction of
13 epidemic trend of 2019-nCoV infection in 2020. Preprint at Med RXIV. 2020.
14
15 <https://www.medrxiv.org/content/10.1101/2020.02.13.20022715v1> (29 March 2020, date
16 last accessed).
17
18
19
20
21
22 13. Oliveiros B, Caramelo L, Ferreira NC, Caramelo F. Role of temperature and humidity in
23 the modulation of the doubling time of COVID-19 cases. Preprint at Med RXIV. 2020.
24
25 10 <https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1> (29 March 2020, date
26 last accessed).
27
28
29
30
31
32 14. COVID-19/2019-nCoV Infection Time Series Data Warehouse. 2020.
33
34 <https://github.com/BlankerL/DXY-COVID-19-Data> (9 April 2020, date last accessed).
35
36
37 15. Integrated Surface Database of USA National Centers for Environmental Information.
38 15 2020. <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/> (9 April 2020, date last accessed).
39
40
41
42 16. R Core Team, R: A language and environment for statistical computing. R Foundation for
43 Statistical Computing. 2018. <https://cran.r-project.org/bin/windows/base/old/3.5.1/> (9
44 January 2020, date last accessed).
45
46
47
48
49
50 20
51
52
53
54
55
56
57
58
59
60

FIGURE LEGENDS

Fig. 1. Loess regression interpolation of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in °C, (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 °C, (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

1
2
3 "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern"
4
5 pattern including countries in the southern atmosphere, Australia and South Africa. Each dot
6
7 represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves
8
9 with 95% confidence intervals in shadow are illustrated for each dataset. The black solid line
10
11 represents that the observed values are equal to the predicted ones, and dots closer to this line
12 5
13 means better prediction performance.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

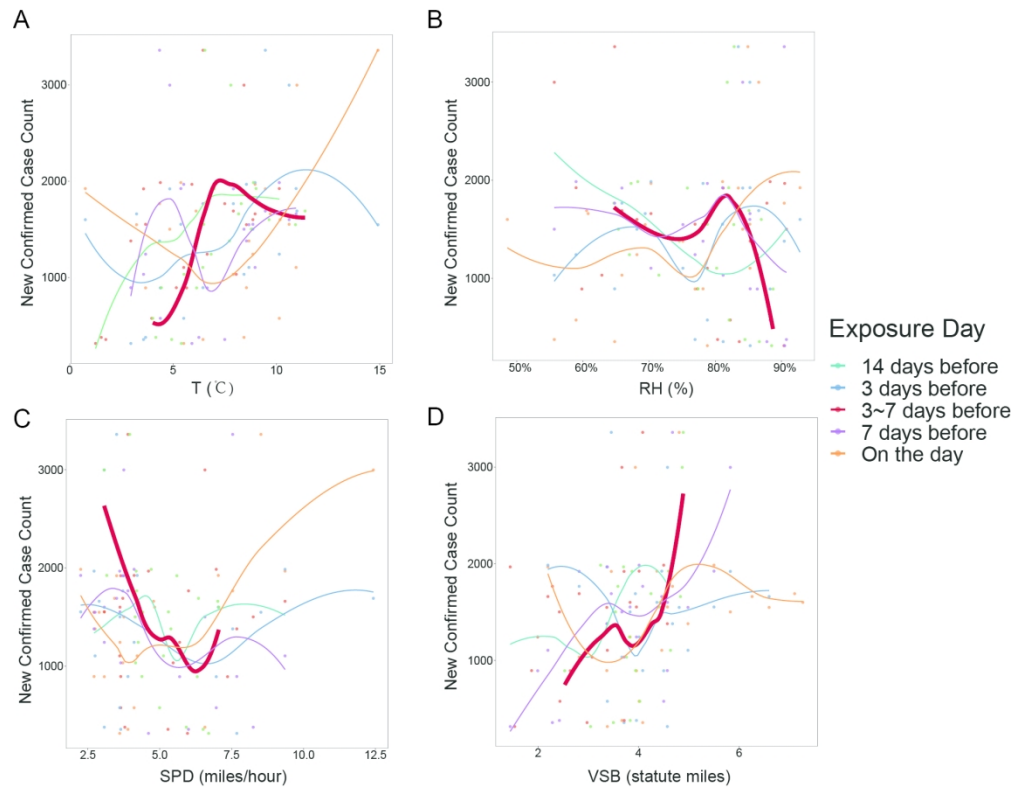


Fig 1. Loess regression interpolation of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for Wuhan city. Five time delay of the confirmation 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)

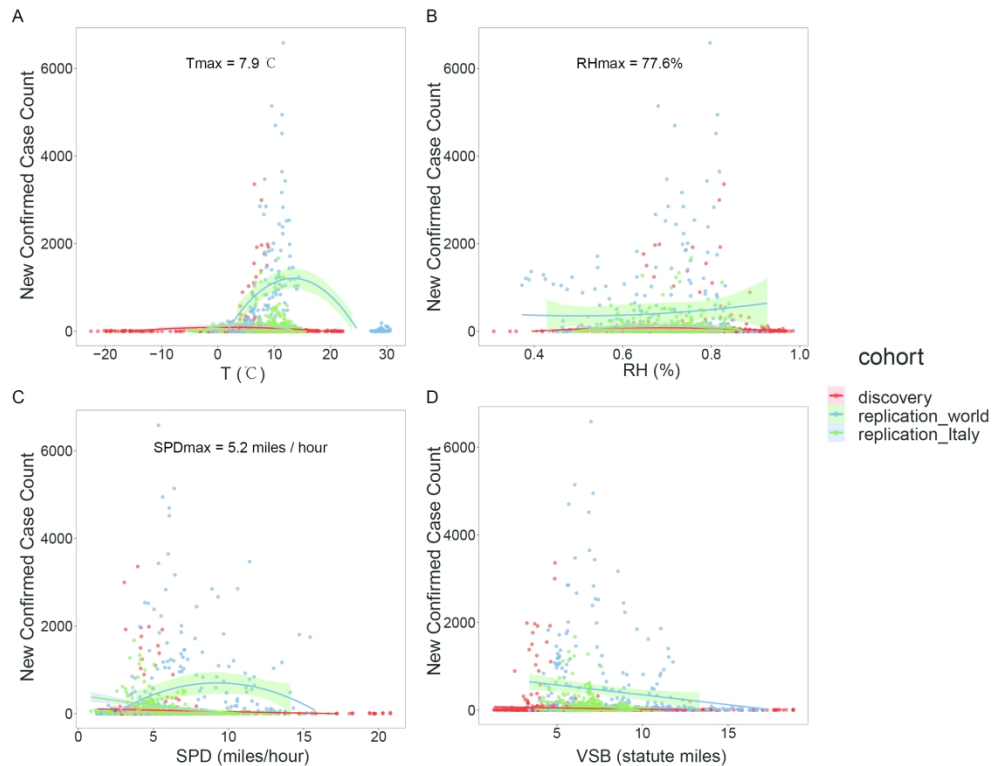


Fig 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in miles per hour, (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, and SPD, 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)

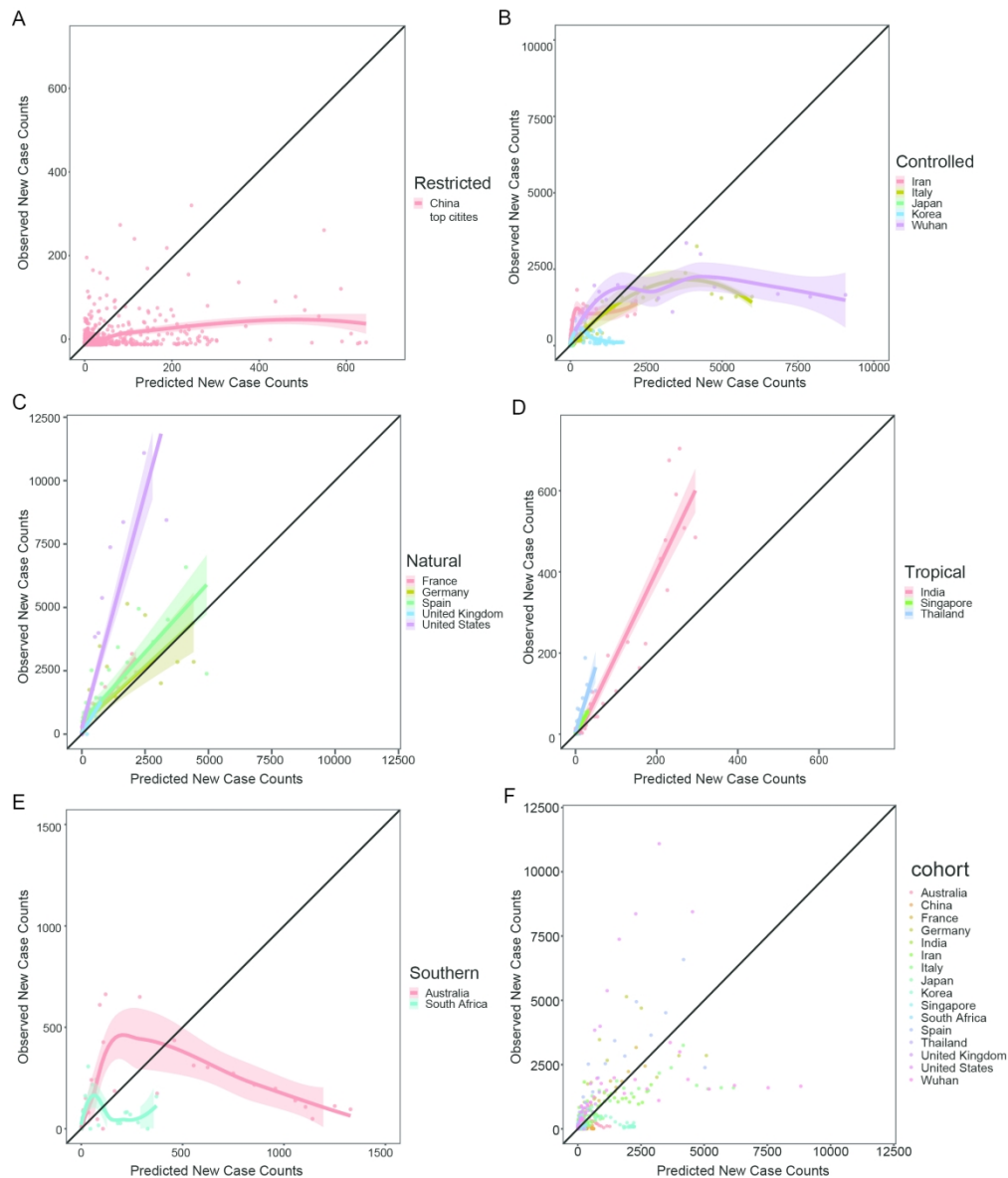


Fig 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" 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)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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:

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

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

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 identified relationship by Loess regression and knowledge of droplet-mediated viral diseases. Each of the four

1
2
3
4 meteorological variables was fitted into models solely, and then all variables were
5
6 combined together to compose a comprehensive coefficient that was multiplied with
7
8 the extant confirmed case counts. All the models were compared with each other to
9
10 find a best-fitted model with the best fitness. Fitness was evaluated according to
11
12 log-likelihood, Akaike information criterion, and Bayesian Information Criterion. The
13
14 final equation supposed that all the meteorological variables composed a coefficient
15
16 which was multiplied by the extant confirmed case counts on the exposure day, and
17
18 then derived the new confirmed case counts on the test day.
19
20
21
22
23

24 **Supplementary Results**

25 Datasets description

26
27
28
29
30 Only Chinese cities with monthly confirmed cases over 50 were included in the
31
32 discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in
33
34 Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily
35
36 confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in
37
38 all the other Chinese cities. We suppose that it might be due to abrupt large
39
40 supplement of virus test kits on that day. In order to reduce the potential
41
42 contamination of modeling by this outlier, we substituted the counts on that day by
43
44 four, that was $13,436/4=3,359$, which was still the largest number but not deviated
45
46 from the dataset too much. There were also two oddly large new confirmed case
47
48 counts for Lombardy, which were discarded from the subsequent analysis. Except the
49
50 outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to
51
52 2,997, the average temperature ranged $-23.54^{\circ}\text{C} \sim 22.85^{\circ}\text{C}$, the wind speed ranged
53
54
55
56
57
58
59
60

1
2
3
4 1.33 ~ 26 miles per hour, visibility ranged 0.425 ~ 110 statute miles to nearest tenth,
5
6 and relative humidity ranged 31.4% ~ 100%.
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

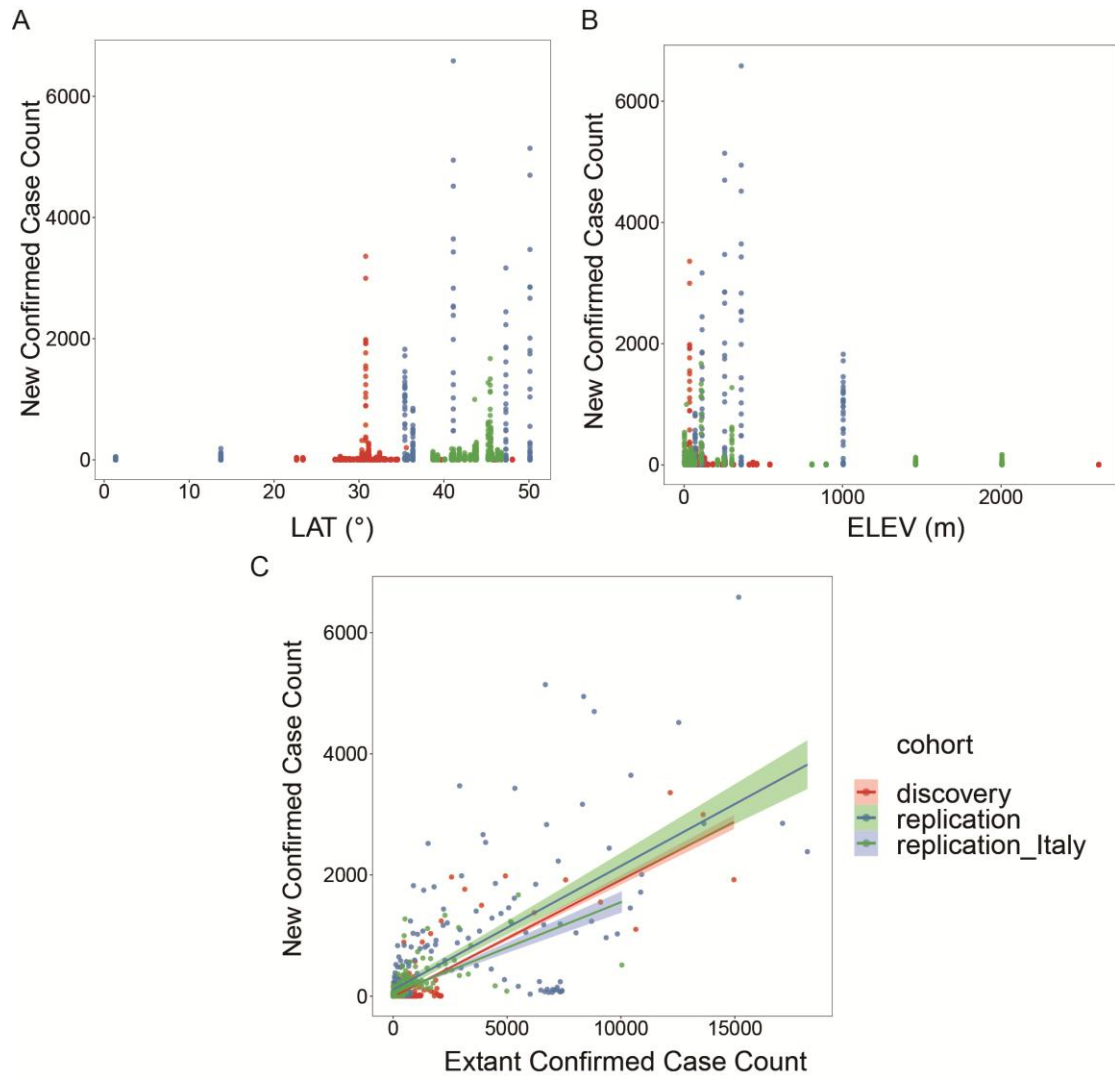


Fig. S1. Scatterplots of new confirmed case count to (A) latitude, (B) elevation, and (C) the extant confirmed case count, for all the studied sites.

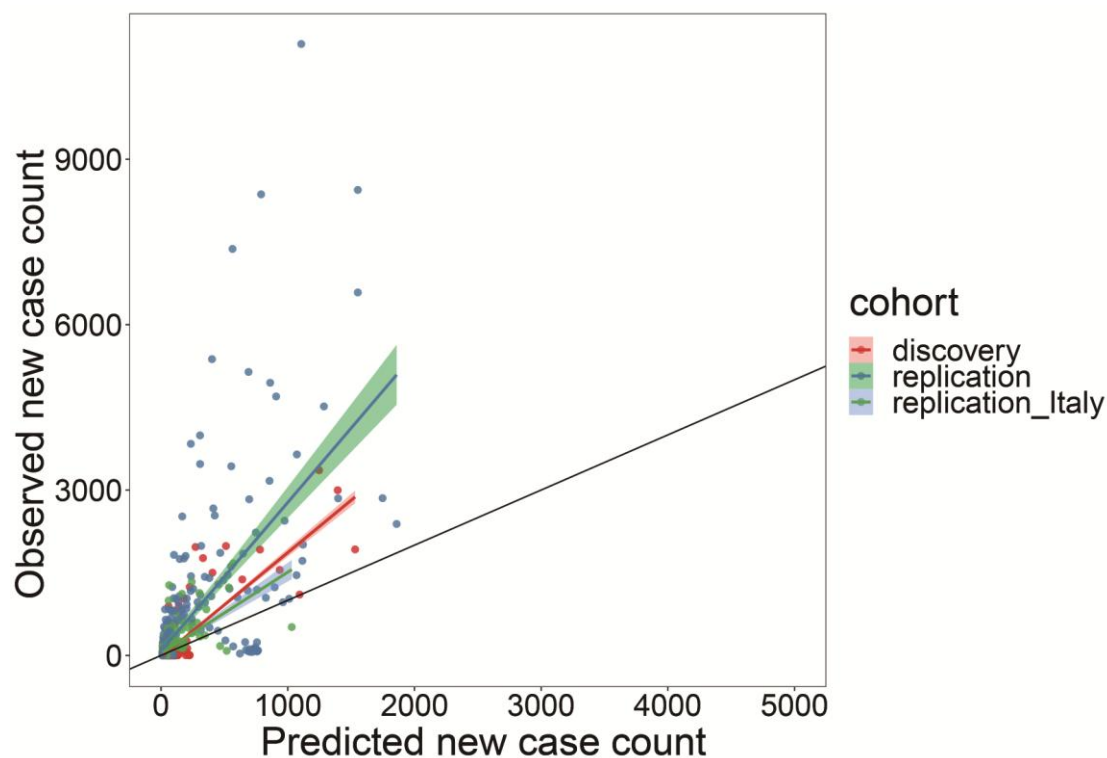


Fig. S2. The observed daily new case counts verse the predicted values by only the extant confirmed case count are illustrated for all cohorts. Linear regression interpolation curves with 95% confidence intervals in shadow are illustrated for each dataset. The black solid line represents that the observed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



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 1 **Title: Predicting the local COVID-19 outbreak around the world with**
5
6 2 **meteorological conditions: a model-based qualitative study**
7
8

9 3 **Authors:** Biqing Chen¹, Hao Liang², Xiaomin Yuan³, Yingying Hu², Miao Xu², Yating Zhao⁴,
10 4 Binfen Zhang², Fang Tian¹, Xuejun Zhu^{1,2*}
11
12
13

14
15 5 **Affiliations:**
16
17

18 6 ¹ Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the
19 7 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
20
21

22 8 ² Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated
23 9 Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
24
25

26 10 ³ Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the
27 11 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
28
29

30 12 ⁴ School of Atmospheric Sciences, Nanjing University, Nanjing, China
31
32

33 13 *Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.
34
35

36 14
37 15 **Abstract**
38
39

40 16 **OBJECTIVE:** This study aims to investigate the relationship between daily weather and
41 17 transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of
42 18 the COVID-19 spreading rate for a certain area with meteorological factors.
43
44

45 19 **METHODS AND ANALYSIS:** We collected 382,596 records of weather data with four
46 20 meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility,
47 21 and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed
48
49
50
51
52
53
54
55
56
57
58
59
60

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 °C, 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.

- 1
2
3 44 ● The current study proposes two prediction models on different time scales, a short-term one
4
5 45 integrating more detailed meteorological information which is more accurate, and a long-
6
7 term one with only temperature which is more feasible.
8 46
9
10 47 ● The influence of weather on virus spread could be confounded by a dozen of manual
11
12 interventions, such as population mobility and disinfection measures, leading to inaccurate
13 48
14 modeling.
15 49
16
17 ● The prediction model (especially the long-term model) might be unsuitable and inaccurate
18 50
19 for areas with hot weather.
20 51
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

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

1
2
3 75 case counts daily for 80 days from January 20, 2020 to April 9, 2020 for nearly 500 areas around
4
5 76 the world, including 428 Chinese cities and areas, 18 Italian provinces, and 13 other countries.
6
7
8 77 Five time point's delay of virus infection from the exposure day were considered and compared to
9
10 78 determine the most reasonable time point's delay. A multivariate polynomial regression model
11
12 79 with meteorological factors as a "weather coefficient" of the existing confirmed case count was
13
14 80 established in a discovery Chinese dataset, and then validated by worldwide data. Five
15
16 81 transmission modes, indicating different levels of epidemic control, were revealed by this model.
17
18
19 82 In this view, this model can not only predict future outbreak, but also be used to evaluate the
20
21 83 effect of epidemic prevention measures for a certain area.
22
23

24 84 **Materials and Methods**

27 85 Epidemiological data

28
29
30 86 Epidemiological data were collected from the World Health Organization (WHO) (4),
31
32 87 European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (10). The daily
33
34 88 new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data
35
36 89 were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries,
37
38
39 90 namely, United States, United Kingdom, Germany, France, Italy, Spain, Iran, Korea, Japan,
40
41 91 Australia, South Africa, India, Thailand, and Singapore. Considering the potential confounding
42
43 92 effect, only Chinese cities with no less than 50 cumulative confirmed cases in one month and
44
45 93 without official reports of large imported cases (42 in total) were taken as a discovery dataset,
46
47 94 while those for Italian provinces and all the other nations were taken as replication datasets
48
49
50 95 (Supplementary Materials).
51

52 96 Weather data

1
2
3 97 Four meteorological variables were chosen, air temperature, relative humidity, wind speed,
4
5 98 and visibility. Temperature could affect virus viability in the environment. Wind speed could
6
7 affect the suspending time of virus-attached particles. Relative humidity reflects the amount of
8 99 droplets in the air. Visibility is influenced by the amount of particles such as dust and air
9
10 100 pollutants. These two parameters both affect the amount of mediator for the virus to stay in the
11
12 101 air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative
13
14 102 humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of
15
16 103 meteorological observations and geographic factors (latitude and elevation) from the Integrated
17
18 104 Surface Database of USA National Centers for Environmental Information (11). Daily data were
19
20 105 calculated by averaging the hourly data for each variable in each day.
21
22
23
24 106
25
26

27 107 Statistical modeling

28
29 108 The number of daily new confirmed cases was taken as a dependent variable. Four
30
31 109 meteorological variables, namely, average temperature, wind speed, visibility, and relative
32
33 110 humidity, and the existing confirmed case counts were taken as independent variables.
34
35 111 Considering that there is a latency stage from the day one get infected to the day being
36
37 112 confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were
38
39 113 collected needs to be taken into consideration. As it is reported that the latency period for
40
41 114 COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection
42
43 115 were taken into consideration, that is, weather data and existing confirmed cases count data were
44
45 116 collected on the day, three days before, seven days before, 3~7 days before, and 14 days before
46
47 117 collecting the epidemiological data.
48
49
50
51

52 118 At first, each meteorological variable was fitted into a bunch of single-factor models (either
53
54 119 generalized linear model or polynomial model) through non-linear least squares (NLS) modeling
55
56
57
58
59
60

1
2
3 120 using Wuhan dataset under the assumption of 3~7 days exposure delay. The relationship between
4
5
6 121 each meteorological variable and confirmed new case count (linear or quadric) was identified
7
8 122 based on model fitness (log-likelihood, Akaike information criterion, Bayesian Information
9
10 123 Criterion, etc.) and common knowledge of droplet-mediated viral diseases. Second, the time
11
12 124 delay effect was investigated in the Wuhan dataset through Loess regression interpolation and
13
14
15 125 NLS modeling with the previously identified relationship for each meteorological variable. The
16
17 126 most possible time delay identified was taken for subsequent analyses.
18

19 127 The contribution of each meteorological factor was investigated with the Wuhan dataset
20
21
22 128 through Spearman's correlation test. Single-factor models were fitted into NLS models again with
23
24 129 data from the discovery dataset (all Chinese cities with monthly confirmed cases over 50) under
25
26 130 the assumption of previously determined relationship and pre-defined time delay, to determine
27
28
29 131 the exact coefficients accompanied with each meteorological factor and to find out the most
30
31 132 suitable environmental condition for SARS-CoV-2. Then, two final prediction models (short-
32
33 133 term model and long-term model) were developed using the discovery dataset with the previously
34
35 134 determined coefficients. The prediction model supposed that all the meteorological variables
36
37
38 135 added together to compose a coefficient which was multiplied by the existing confirmed case
39
40 136 count on the exposure day, and then derived the new confirmed case counts on the test day. The
41
42 137 short-term model took all four variables, while the long-term model only considered temperature
43
44
45 138 as it is easy to be forecasted. There was a constant coefficient for the total equation, that was
46
47 139 multiplied by the existing confirmed case count. Its default value was obtained by model fitting
48
49 140 in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was
50
51
52 141 investigated with all datasets covering the world's top cities and areas. The correlation of existing
53
54 142 confirmed case counts with newly confirmed case counts was also investigated. Basic statistics
55
56 143 and modeling was conducted in R 3.5.1 (<https://cran.r-project.org/>).
57
58
59
60

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). To investigate whether the influence of meteorological factors is linear or quadric, both linear and non-linear square modeling were performed under different relationship assumptions to compare model fitness statistics using the Wuhan dataset with a 3~7 days delay of infection. It was 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).

1
2
3 167 Furthermore, we investigated the time delay from weather exposure to COVID-19 confirmation
4
5
6 168 with the above determined relationships and NLS modeling using Wuhan dataset. Model fitness
7
8 169 statistics showed that the number of confirmed new cases was best correlated with air
9
10 170 temperature 3~7 days ago, relative humidity and visibility 7 days ago, and wind speed on the
11
12 171 exposure day (Table S2). By comprehensive consideration of all four meteorological variables
13
14
15 172 and the differences between statistics values, the weather 3~7 days ago, as well as weather one
16
17 173 week ago, could well predict COVID-19 outbreak. It coincided with the latency period of 3~7
18
19 174 days for SARS-CoV-2, that is, exposure under certain adverse weather might exhibit its effect
20
21
22 175 after 3~7 days.

24 176 Contribution of single meteorological factor to the outbreak

26
27 177 In the Wuhan dataset, the new case count was significantly positively correlated with temperature
28
29
30 178 (Spearman's correlation $\rho = 0.69$, $p < 0.001$) and visibility ($\rho = 0.43$, $p = 0.04$), and negatively
31
32 179 correlated with wind speed ($\rho = -0.45$, $p = 0.03$) and relative humidity ($\rho = -0.33$, $p = 0.12$) 3~7
33
34 180 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind
35
36
37 181 speed, visibility, and relative humidity. A model only with temperature as a parameter could
38
39 182 already explained 45% of the variance in the epidemic data ($p = 4 \times 10^{-4}$), while wind speed and
40
41 183 visibility could explain over 25% of the variance. To elucidate the contribution of each
42
43 184 meteorological factor to the case counts and to determine the exact coefficients, we first
44
45
46 185 performed single-factor regression modeling for each meteorological variable in the discovery
47
48 186 dataset with the relationship identified before under the assumption of 3~7 days delay of viral
49
50 187 infection. Temperature, relative humidity, and wind speed were fitted into quadratic models; and
51
52
53 188 visibility was fitted into a linear model. According to the fitted single-factor models
54
55 189 (Supplementary Results), SARS-CoV-2 transmission reaches a peak when mean temperature is

1
2
3 190 6.18 °C (Fig. 2A), relative humidity is 78.47% (Fig. 2B), and wind speed is 1.88 meter /second
4
5
6 191 (m/s) (Fig. 2C); and its transmission rate decreases with the increase of visibility (Fig. 2D). The
7
8 192 effects of geographic factors such as latitude and elevation, and the pure influence from the
9
10 193 existing case count were further investigated in the worldwide datasets (Fig. S1), illustrating that
11
12 COVID-19 mainly outbreaks at latitude 30°~50° (Fig. S1A) and elevation < 500 metre (Fig. S1B).
13 194
14
15 195 New confirmed case count was positively correlated with the existing confirmed case count (Fig.
16
17 196 S1C).
18
19

20 197 Short-term prediction model

21
22
23 198 To deduce a practical comprehensive model, all four meteorological variables with their specific
24
25 199 coefficients determined by single-factor modeling were added together to form a complex short-
26
27 200 term model, and the existing confirmed case count was taken as a base for meteorological factors
28
29 to multiply (Supplementary Results). This full model was fitted with the discovery dataset to
30 201 determine the exact values of the constant coefficient in the equation. The best-fitted short-term
31
32 202 model was as follows:
33
34 203

35
36
37 New Case Count

$$38 204 = (-0.11 \times T^2 + 1.40 \times T - 0.058 \times RH^2 + 9.04 \times RH - 1.36 \times SPD^2 + 5.12$$

$$39 \times SPD - 7.02 \times VSB - 126.66) \times \alpha \times \text{Existing Confirmed Case Count}$$

40
41
42 205 where T is temperature in °C, RH is relative humidity in percentage, SPD is wind speed in m/s,
43
44 VSB is visibility in statute miles, α is a site-specific constant, with a default of 0.001. All
45 206 parameters take the means of values 3~7 days before the day new case count is evaluated.
46
47 207
48
49

50 208 In this model, all the four meteorological variables are added together in their proper forms
51
52 209 to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate
53
54 210 of SARS-CoV-2, and thus influences the number of people that catch infection from the existing
55
56
57
58
59
60

1
2
3 211 confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a
4
5 212 multiplicative constant coefficient α in the equation, which seems site-related. This constant
6
7
8 213 coefficient could adjust the strength of the "weather coefficient" on disease transmission. When
9
10 214 we substitute replication datasets into this short-term model with the multiplicative constant
11
12 215 coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious
13
14
15 216 underestimation of predicted values against real ones was observed although the predicted values
16
17 217 correlated with the real ones very well. We supposed it was due to site-specific difference in the
18
19 218 multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the
20
21
22 219 pandemic had been controlled early. Thus, we further re-fitted this composed model with all
23
24 220 datasets to determine a more accurate value of the multiplicative constant coefficient α , which
25
26 221 was 0.001 then. In practical application, we need to first plot the observed case count vs.
27
28 222 predicted one with a default α value 0.001, and then examine the extent of underestimation or
29
30
31 223 overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the
32
33 224 impact size of "weather coefficient" for a certain site.

35 225 Substitute data from the past two months, a good prediction performance was obtained for
36
37
38 226 this short-term model, with the predicted values significantly correlated to the observed ones for
39
40 227 most areas (Fig. 3). However, only the existing confirmed case count data could not predict the
41
42 228 new case count 3~7 days later as well as the weather-combined model did (Table S3).

44 45 229 Different modes of viral transmission illustrated by the model

46
47 230 The observed versus predicted data exhibited different correlation patterns for different areas,
48
49 231 meaning different viral transmission modes, which may indicate the effect of epidemic control
50
51 232 for certain area.

53
54 233 Data from Chinese top-affected cities were not very well predicted and obviously
55
56 234 overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p <$

1
2
3 235 0.001; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after
4
5
6 236 the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention
7
8 237 measures at the beginning of the pandemic. This viral transmission mode suggested by the not
9
10 238 well correlated prediction pattern is called "restricted".

11
12 239 For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the
13
14
15 240 predicted outbreak was well correlated with the actual observations at the beginning when the
16
17 241 existing confirmed cases were not in very large numbers, but the prediction deviates from the
18
19 242 observation as the confirmed cases increase, in detail, there's large overestimation of prediction
20
21
22 243 ($\rho_{\text{Wuhan}} = 0.69$, $\rho_{\text{Italy}} = 0.87$, $\rho_{\text{Japan}} = 0.80$, $\rho_{\text{Iran}} = 0.86$, $p < 0.001$, $\rho_{\text{Korea}} = 0.43$, $p = 0.002$; Fig. 3B).
23
24 244 It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15,
25
26 245 the day when shelter hospitals had been put into use for seven days (the average latency period
27
28 246 for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention
29
30
31 247 and control taken in these areas is effective (so-called "controlled" mode). The number of cases
32
33 248 had been decreased by 72% for Wuhan, over 95% for Korea, Japan, and Italy, and 37% for Iran
34
35 249 at most due to epidemic control (the largest gap between prediction and observation).

36
37
38 250 For most European and American countries, the predicted outbreak was linear correlated
39
40 251 with the observed data very well ($\rho_{\text{France}} = 0.96$, $\rho_{\text{United States}} = 0.93$, $\rho_{\text{United Kingdom}} = 0.83$, $\rho_{\text{Spain}} =$
41
42 252 0.97 , $\rho_{\text{Germany}} = 0.94$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without
43
44
45 253 much man-made epidemic prevention and control measures. Estimation of daily new case counts
46
47 254 by this short-term model performed very well for European countries, while this model
48
49 255 underestimated the outbreak in the United States.

50
51 256 Although the weather is not suitable for tropical areas, the viral transmitted in natural mode,
52
53
54 257 manifested as good linear correlation between the prediction and the observation ($\rho_{\text{India}} = 0.94$,

1
2
3 258 $\rho_{\text{Singapore}} = 0.66, p < 0.001, \rho_{\text{Thailand}} = 0.56, p = 0.001$; Fig. 3D), with just relatively small daily
4
5
6 259 new case counts compared to temperate regions.

7
8 260 Countries in the southern hemisphere displayed similar pattern as the "controlled" with large
9
10 261 overestimation by the model when the confirmed cases increase, leading to not good prediction
11
12 262 performance ($\rho_{\text{Australia}} = 0.79, p < 0.001, \rho_{\text{South Africa}} = 0.34, p = 0.08$; Fig. 3E). It might be due to
13
14
15 263 the effect of epidemic prevention measures and hot summer weather in these countries.

16 17 264 Long-term simplified model

18
19
20 265 Long-term prediction depends on weather forecast, which generally reports only average
21
22 266 temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as
23
24 267 temperature in a short time delay (3~7 days ago), we again performed single-factor regression
25
26
27 268 modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a
28
29 269 quadric function (Supplementary Results). This simplified model with average temperature as a
30
31 270 weather factor was derived as follows:

$$32
33
34 271 \text{ new case count} = (-0.10 \times T^2 + 1.11 \times T + 46.42) \times \beta \times \text{Existing Confirmed Case Count}$$

35
36
37 272 where T is temperature in °C, β is a site-related multiplicative constant coefficient, with a default
38
39
40 273 of 0.006. All parameters take values 14 days before the day new case count is evaluated.

41
42 274 With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, p
43
44
45 275 < 0.001 ; Fig. 3F). The long-term simplified prediction model also showed five prediction-
46
47 276 observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the
48
49 277 studied areas. This model could directly predict the newly emerging cases 14 days later, and be
50
51
52 278 used to predict COVID-19 outbreak in the future month by summing up the daily new case count
53
54 279 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 °C, 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

1
2
3 303 spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g.,
4
5 304 dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may
6
7
8 305 indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case
9
10 306 count decreases rapidly when visibility is high than 13 statute miles, indicating that caution
11
12 307 should be taken if visibility drops below 10 statute miles.

14
15 308 In the prediction model, there is a multiplicative constant coefficient which determines the
16
17
18 309 strength of the weather coefficient on the epidemic transmission. It seems site-specific, as
19
20 310 adjusting it could make the prediction for one site very close to the observation. This constant
21
22 311 might reflect the influence of a couple of site-specific confounding factors, such as epidemic
23
24 312 control measures, sun radiation, and population density. Various degrees of isolation for various
25
26
27 313 areas around the world lead to different degrees of weather effect. When evaluate the prediction
28
29 314 performance by the short-term model and the long-term model, they both exhibit different
30
31 315 prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of
32
33
34 316 epidemic control and isolation policy would lead to deviation from the original prediction and
35
36 317 thus different prediction-observation correlation patterns. Therefore, by plotting the predicted
37
38 318 versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it
39
40
41 319 would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the
42
43 320 observed case counts dropped dramatically from the predictions for Wuhan seven days after their
44
45 321 shelter hospitals were put in use, suggesting the importance and necessity of building shelter
46
47 322 hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and
48
49
50 323 very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to
51
52 324 natural transmission mode. Another thing worth attention is that although the weather in tropical
53
54 325 areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on

1
2
3 326 spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore,
4
5 327 these tropical areas should still be on the alert against future outbreak of COVID-19.
6
7

8 328 Although those cases with travel history to China or indicated by the World Health Organization
9
10 329 as "imported case only" were excluded in this study to make the world data most likely local
11
12 transmitted, it was difficult to separate the imported cases from local transmission very well in
13 330
14 practice. It might explain the not excellent correlations of predictions with observations.
15 331
16 Furthermore, the relationship of weather and COVID-19 could be complex, since the human
17 332
18 immune system has an innate seasonal rhythm, and the immune system could also be affected by
19
20 333 weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway
21
22 334
23 mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets
24 335
25 for virus to adhere.
26
27 336
28

29 337 There are several limitations of this study. First of all, this prediction model (especially the long-
30
31 term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter,
32 338
33 as the models were derived using Chinese datasets, mainly in the first three months of 2020. The
34 339
35 prediction became inaccurate and could be largely deviated from real observations under hot
36 340
37 weather, which might explain the obvious bad prediction performance for countries in the
38
39 341 southern hemisphere. One explanation for the inaccurate prediction in areas with high
40
41 342
42 temperature could be that SARS-CoV-2 transmission in these areas was mainly not influenced by
43 343
44 weather, but in another direct transmission way, such as face-to-face contact or spread in
45
46 344
47 gathering crowd. Second, it seems that the prediction performance drops with the increase in new
48 345
49 case count, suggesting that the prediction model might become inaccurate and not suitable for
50 346
51 very large new case count. This could be due to that there was less data points with large new
52 347
53 case count. Therefore, the model's prediction performance would be better with more data points,
54
55 348
56
57
58
59
60

1
2
3 349 especially the large case count points. Third, the short-term prediction model must use all four
4
5 meteorological factors, while these factors are not always available for any one certain area.
6 350
7
8 351 Fourth, this study included various areas covering a long period into modeling, thus, there were a
9
10 352 bunch of variable confounding factors, such as population mobility and disinfection measures,
11
12 353 which were not controlled and thus could impede the model accuracy. Fifth, as we could only
13
14 obtain country-level epidemiological data, the corresponding meteorological data were obtained
15 354
16
17 355 for their capital cities, leading to not exact pairing of epidemiological data and meteorological
18
19 356 data.

22 357 **Conclusion**

23
24
25 358 In summary, this study has found significant correlations with the COVID-19 epidemic trend for
26
27 not only temperature and humidity, but also wind speed and visibility. It proposed a
28 359
29 comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version
30 360
31 and a long-term version. The short-term version uses the combination of four meteorological
32 361
33 factors as a "weather coefficient" of the existing confirmed case count in the past week and can
34 362
35 be used to predict epidemic situation in the future three days; the short-term version uses average
36
37 363 temperature as the "weather coefficient" seven days ago and can predict the outbreak in one
38
39 364 month if combined with weather forecast. This model is easy to use for predicting the COVID-19
40
41 365 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case
42
43 count for the future couple of days or month. This model will be very helpful for local
44 366
45 governments to make timely policies on epidemic control, for instance, the allocation of medical
46 367
47 equipments such as ventilators and medical resources such as hospitals, beds and health-care
48 368
49 workers, according to the prediction results.
50
51 369
52
53 370
54
55
56 371

1
2
3 372 **Acknowledgments:** We thank Dr. Zhisheng Huang for advices on data collecting and processing,
4
5 373 Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on
6
7 modeling.
8 374
9

10
11 375 **Funding:** the Priority Academic Program Development of Jiangsu Higher Education Institutions
12
13 376 – the third period (NO.035062002003c), the Yizhong Research Fund of Jiangsu Provincial
14
15 377 Hospital of Chinese Medicine (Y19066).
16
17

18 378 **Author contributions:** BC, YZ and XZ design and interpret the reported analyses and results;
19
20 379 HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse the data, drafted,
21
22 and revised the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical
23 380 support; XZ supervised the research.
24
25 381
26
27

28 382 **Competing interests:** Authors declare no competing interests.
29

30
31 383 **Data and materials availability:** Weather data and epidemiological data is all obtained from
32
33 384 public databases. Detailed modeling results are available upon request by emailing to Biqing
34
35 385 Chen, bq_chen@qq.com.
36
37

38 386 39 40 41 387 **References:**

- 42
43 388 1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients
44
45 with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.
46 389
47
- 48
49 390 2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in
50
51 391 Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med.
52
53 392 2020;382(13):1199–207.
54
55
56
57
58
59
60

- 1
2
3 393 3. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of
4 pneumonia associated with the 2019 novel coronavirus indicating person-to-person
5 394 transmission: a study of a family cluster. *Lancet* [Internet]. Elsevier Ltd;
6 395 2020;395(10223):514–23. Available from: [http://dx.doi.org/10.1016-S0140-](http://dx.doi.org/10.1016/S0140-)
7
8 396 [6736\(20\)30154-9](http://dx.doi.org/10.1016/S0140-6736(20)30154-9)
9
10 397
11
12
13
14
- 15 398 4. Organization WH. [https://www.who.int/emergencies/diseases/novel-coronavirus-](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports)
16
17 399 [2019/situation-reports](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports) [Internet]. 2020 [cited 2020 Mar 29]. Available from:
18
19 <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
20 400
21
22
- 23 401 5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30
24 years of county-level evidence from the United States. *Amer J Epidemiol*. 2012;176(suppl.
25 402 7):S114–22.
26
27 403
28
29
- 30 404 6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on
31 relative humidity and temperature. *PLoS Pathol*. 2007;3:e151.
32 405
33
34
- 35 406 7. Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential
36 Implications for Reducing Transmission of COVID-19. *JAMA*. 2020;E1–2.
37 407
38
39
- 40 408 8. Cai Y. The Effects of "Fangcang, Huoshenshan [Internet]. Available from:
41 <https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3>
42 409
43
44
- 45 410 9. Proverbio AM, Crotti N, Manfredi M, Zani A. Preliminary evidence that higher
46 temperatures are associated with lower incidence of COVID-19, for cases reported
47 411 globally up to 29th February 2020. *medRxiv*. 2020;(February).
48
49
50 412
51
- 52 10. <https://github.com/BlankerL/DXY-COVID-19-Data>.
53 413
54
- 55 11. <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/>. 2020.
56 414
57

1

2

3 415

4

5 416

6

7 417

8

9

10

11 418

12

13 419

14

15 420

16

17 421

18

19 422

20

21

22 423

23

24 424

25

26 425

27

28 426

29

30 427

31

32 428

33

34 429

35

36 430

37

38 431

39

40 432

41

42 433

43

44 434

45

46 435

47

48 436

49

50 437

51

52 438

53

54 439

55

56 440

57

58 441

59

60

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 °C, (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 °C, (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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

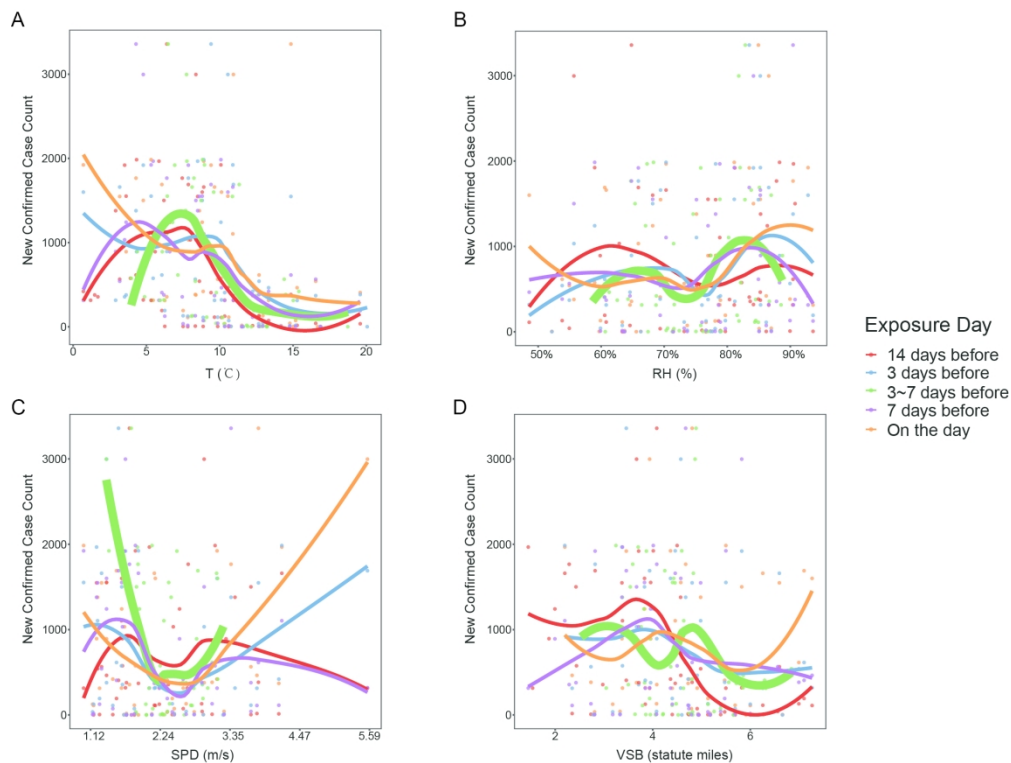
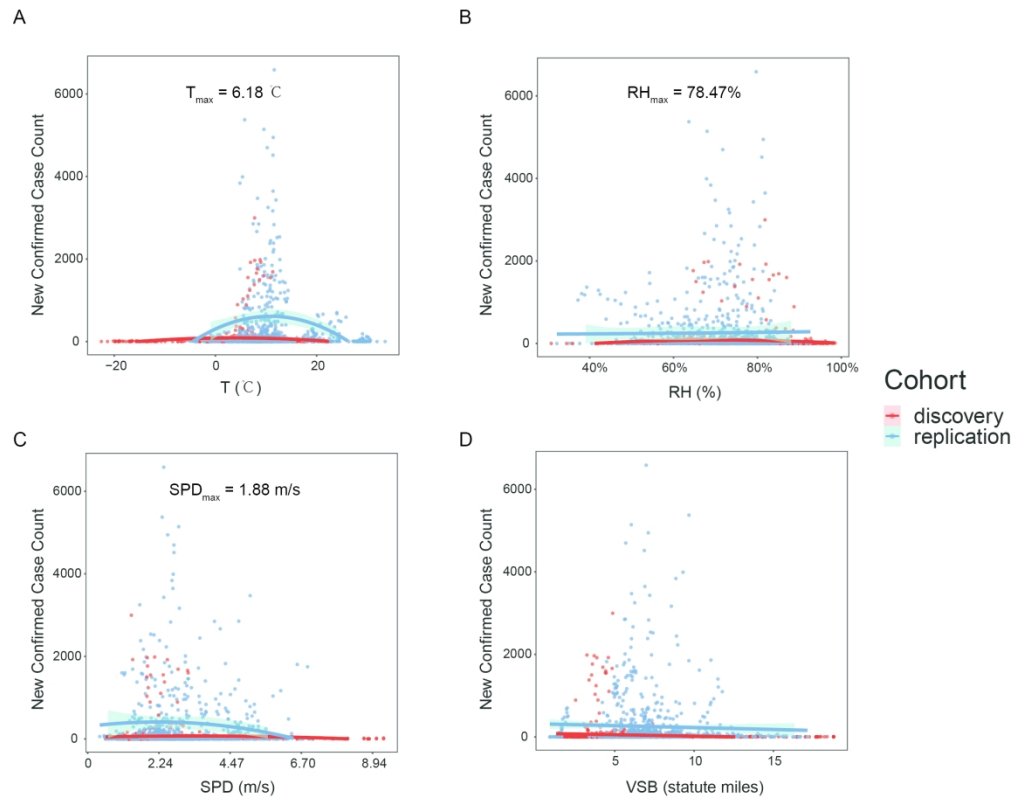


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of 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)



32
33
34
35
36
37

Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in °C, (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).

38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

218x172mm (300 x 300 DPI)

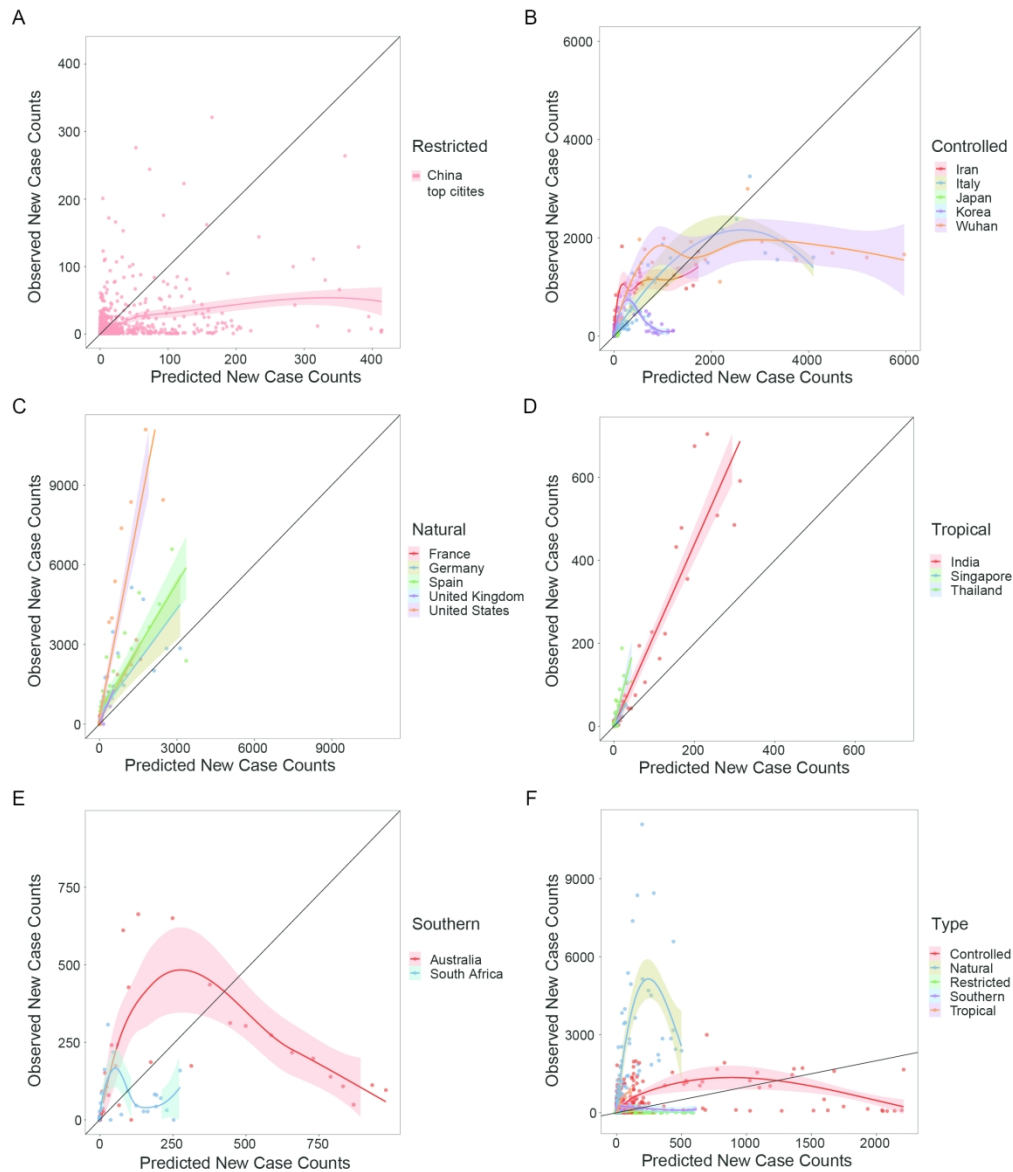


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" 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)

1 **Supplementary Materials and Methods**

2 Epidemiological data

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

6 For Wuhan city, there was a shortage of test kits at the beginning of the pandemic,
7 which would make confirmed case counts much lower than the actual data, thus, we
8 discarded epidemic data before January 28th, the day when domestic test kits have
9 been approved, produced in large quantities, and were available for Wuhan hospitals.
10 As there was a cut down problem for the existing confirmed case count on February
11 20th for Wuhan, when modeling with the existing confirmed case count, only data
12 before February 20th were used.

13 Weather data

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

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

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

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

1
2
3
4 21 For a city with more than one meteorological stations, the one nearest to the city
5
6 22 center was chosen. For a province with epidemiological data, the meteorological
7
8
9 23 station in the capital city of that province was chosen. For a country with only
10
11
12 24 national wide epidemiological data, weather data were averaged across all the
13
14
15 25 meteorological observatories in the cities where outbreak was officially reported.
16
17 26 Latitude and elevation for the meteorological observatories were also collected.

18 19 27 Statistical modeling

20
21
22 28 Only one city Wuhan was chosen for illustrating the time delay effect because it
23
24 29 is the first city to have an outbreak of COVID-19, there was none reported imported
25
26
27 30 cases for Wuhan, which might obscure the correlation between weather and virus
28
29
30 31 transmission.

31 32 32 **Supplementary Results**

33 33 Datasets description

34
35 34 Only Chinese cities with monthly confirmed cases over 50 were included in the
36
37 35 discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in
38
39
40 36 Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily
41
42
43 37 confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in
44
45
46 38 all the other Chinese cities. We suppose that it might be due to abrupt large
47
48
49 39 supplement of virus test kits or data correction on that day. In order to reduce the
50
51
52 40 potential contamination of modeling by this outlier, data on that day were discarded
53
54
55 41 from the subsequent analysis. There were also two oddly large new confirmed case
56
57
58 42 counts for Lombardy, which were discarded from the subsequent analysis. Except the
59
60

1
2
3
4 43 outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to
5
6 44 2,997, the average temperature ranged $-22.54^{\circ}\text{C} \sim 22.16^{\circ}\text{C}$, the wind speed ranged
7
8
9 45 $0.56 \sim 9.29$ meter per second, visibility ranged $1.3 \sim 18.8$ statute miles, and relative
10
11
12 46 humidity ranged $30.84\% \sim 98.52\%$.

13 14 47 Model selection

15
16
17 48 With the increase of relative humidity, the amount of droplets in the air increases,
18
19 49 leading to more virus load. However, as the air gets humid, human's respiratory tract
20
21
22 50 could better defend virus infection. Thus, the relationship of relative humidity could
23
24
25 51 be complex, not pure linear. Giving comprehensive consideration, we defined the
26
27 52 effect of relative humidity to be quadric. As for visibility, it only affects the amount of
28
29
30 53 particles in the air, which is positively correlated with virus load. Thus, it is most
31
32
33 54 probably to exert its effect linearly.

34
35 55 Although relative humidity and visibility 7 days ago correlated with the
36
37 56 confirmed new case counts best, there was not great loss of model fitting statistics for
38
39
40 57 relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days
41
42
43 58 time delay and 3~7 days time delay for temperature.

44 45 59 Fitted models

46
47
48 60 The fitted single-factor models were as follows:

$$49 \quad \text{New Case Count} = -0.11305 \times T^2 + 1.39819 \times T + 45.11405$$

50
51
52 61 where T is temperature in $^{\circ}\text{C}$.

53
54
55 62 The estimate p-value for constant was < 0.001 . The extremum was $-1.39819/$
56
57
58 63 $(2 \times (-0.11305)) = 6.183945 \text{ }^{\circ}\text{C}$.

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

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 $-5.120123/(2 \times (-1.360056)) = 1.882321 \text{ m/s}$.

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

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

New Case Count

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

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:

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

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:

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

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

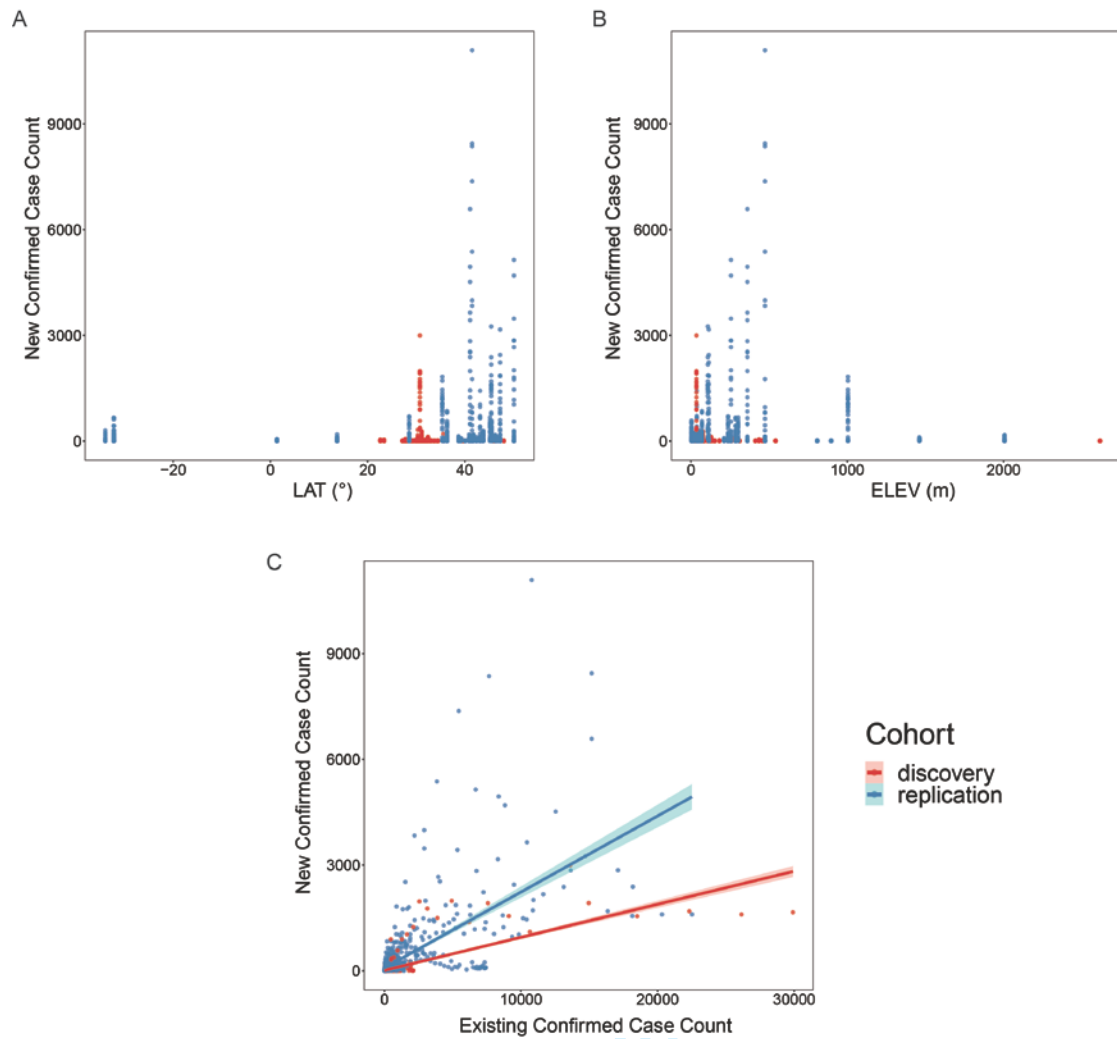
93 Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the
 94 log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information
 95 Criterion for the model; deviance, deviance of the model; Corr, Spearman's 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.

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 1 **Title: Predicting the local COVID-19 outbreak around the world with**
5
6 2 **meteorological conditions: a model-based qualitative study**
7
8

9 3 **Authors:** Biqing Chen¹, Hao Liang², Xiaomin Yuan³, Yingying Hu², Miao Xu², Yating Zhao⁴,
10 4 Binfen Zhang², Fang Tian¹, Xuejun Zhu^{1,2*}
11
12
13

14
15 5 **Affiliations:**
16
17

18 6 ¹ Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the
19 7 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
20
21

22 8 ² Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated
23 9 Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
24
25

26 10 ³ Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the
27 11 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
28
29

30 12 ⁴ School of Atmospheric Sciences, Nanjing University, Nanjing, China
31
32

33 13 *Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.
34
35

36 14
37 15 **Abstract**
38
39

40 16 **OBJECTIVES:** This study aims to investigate the relationship between daily weather and
41 17 transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of
42 18 the COVID-19 spreading rate for a certain area with meteorological factors.
43
44

45 19 **DESIGN:** A retrospective, qualitative study.
46
47

48 20 **METHODS AND ANALYSIS:** We collected 382,596 records of weather data with four
49 21 meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility,
50
51
52

1
2
3 22 and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed
4
5 23 cases in total) in nearly 500 areas worldwide from January 20 to April 9. Epidemic data were
6
7
8 24 modeled against weather data to find a model that could best predict the future outbreak.
9

10
11 25 **RESULTS:** Significant correlations of the daily new confirmed case counts with the weather 3~7
12
13 26 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average
14
15 27 temperature at 5~15 °C, relative humidity at 70%~80%, wind speed at 1.5~4.5 meter / second,
16
17
18 28 and visibility less than 10 statute miles. A short-term model with these meteorological variables
19
20 29 in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-term
21
22 30 model using temperature to predict the pandemic in the next week or month was derived. Taken
23
24
25 31 China as a discovery dataset, it was well validated with worldwide data. According to this model,
26
27 32 there are five different viral transmission pattern, "restricted", "controlled", "natural", "tropical",
28
29 33 "southern". This model's prediction performance correlates with the actual observations best
30
31
32 34 (over 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not
33
34 35 much human interference by epidemic prevention measures.
35

36
37 36 **CONCLUSIONS:** This model can be used for prediction of the future outbreak, and illustrating
38
39 37 the effect of epidemic control for a certain area.
40

41
42 38 **Keywords:** COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control
43
44
45 39
46

47 40 **Strengths and limitations of this study**

- 48
49
50 41 ● This study investigates the role of daily weather in COVID-19 spread systematically with a
51
52 42 comprehensive set of four meteorological factors.
53
54
55
56
57
58
59
60

- 1
2
3 43 ● This research collected a huge amount of data, covering nearly 500 areas worldwide in a long
4
5 44 timescale.
6
7
8 45 ● The current study proposes mathematical models integrating meteorological information for
9
10 46 predicting COVID-19 case counts in the future.
11
12
13 47 ● The influence of weather on virus spread could be confounded by a dozen of manual
14
15 48 interventions, such as population mobility and disinfection measures, leading to inaccurate
16
17 modeling.
18 49
19
20
21 50 ● The prediction model (especially the long-term model) might be unsuitable and inaccurate
22
23 51 for areas with hot weather.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

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

1
2
3 75 case counts daily for 80 days from January 20, 2020 to April 9, 2020 for nearly 500 areas around
4
5 76 the world, including 428 Chinese cities and areas, 18 Italian provinces, and 13 other countries.
6
7
8 77 Five time point's delay of virus infection from the exposure day were considered and compared to
9
10 78 determine the most reasonable time point's delay. A multivariate polynomial regression model
11
12 79 with meteorological factors as a "weather coefficient" of the existing confirmed case count was
13
14
15 80 established in a discovery Chinese dataset, and then validated by worldwide data. Five
16
17 81 transmission modes, indicating different levels of epidemic control, were revealed by this model.
18
19 82 In this view, this model can not only predict future outbreak, but also be used to evaluate the
20
21
22 83 effect of epidemic prevention measures for a certain area.
23

24 84 **Materials and Methods**

27 85 Epidemiological data

30 86 Epidemiological data were collected from the World Health Organization (WHO) (4),
31
32 87 European Centre for Disease Control and Prevention, and DXY-COVID-19-Data (10). The daily
33
34 88 new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data
35
36
37 89 were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries,
38
39 90 namely, United States, United Kingdom, Germany, France, Italy, Spain, Iran, Korea, Japan,
40
41 91 Australia, South Africa, India, Thailand, and Singapore. Considering the potential confounding
42
43
44 92 effect, only Chinese cities with no less than 50 cumulative confirmed cases in one month and
45
46 93 without official reports of large imported cases (42 in total) were taken as a discovery dataset,
47
48 94 while those for Italian provinces and all the other nations were taken as replication datasets
49
50
51 95 (Supplementary Materials).

53 96 Weather data

1
2
3 97 Four meteorological variables were chosen, air temperature, relative humidity, wind speed,
4
5 98 and visibility. Temperature could affect virus viability in the environment. Wind speed could
6
7 affect the suspending time of virus-attached particles. Relative humidity reflects the amount of
8 99 droplets in the air. Visibility is influenced by the amount of particles such as dust and air
9
10 100 pollutants. These two parameters both affect the amount of mediator for the virus to stay in the
11
12 101 air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative
13
14 102 humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of
15
16 103 meteorological observations and geographic factors (latitude and elevation) from the Integrated
17
18 104 Surface Database of USA National Centers for Environmental Information (11). Daily data were
19
20 105 calculated by averaging the hourly data for each variable in each day.
21
22
23
24 106
25
26

27 107 Statistical modeling

28
29 108 The number of daily new confirmed cases was taken as a dependent variable. Four
30
31 109 meteorological variables, namely, average temperature, wind speed, visibility, and relative
32
33 110 humidity, and the existing confirmed case counts were taken as independent variables.
34
35 111 Considering that there is a latency stage from the day one get infected to the day being
36
37 112 confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were
38
39 113 collected needs to be taken into consideration. As it is reported that the latency period for
40
41 114 COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection
42
43 115 were taken into consideration, that is, weather data and existing confirmed cases count data were
44
45 116 collected on the day, three days before, seven days before, 3~7 days before, and 14 days before
46
47 117 collecting the epidemiological data.
48
49
50
51

52 118 To investigate whether the influence of meteorological factors is linear or quadric, both
53
54 119 linear and non-linear modeling were performed under different relationship assumptions to
55
56
57
58
59
60

1
2
3 120 compare model fitness statistics. Each meteorological variable was fitted into a bunch of single-
4
5 121 factor models (either generalized linear model or polynomial model) through non-linear least
6
7
8 122 squares (NLS) modeling using the Wuhan dataset with a 3~7 days delay of infection. The
9
10 123 relationship between each meteorological variable and confirmed new case count (linear or
11
12 124 quadric) was identified based on model fitness (log-likelihood, Akaike information criterion,
13
14
15 125 Bayesian Information Criterion, etc.) and common knowledge of droplet-mediated viral diseases.

16
17 126 Second, the proper time delay from weather exposure to COVID-19 confirmation was
18
19 127 investigated in the Wuhan dataset through Loess regression interpolation and NLS modeling with
20
21
22 128 the previously identified relationship for each meteorological variable. The most possible time
23
24 129 delay identified was taken for subsequent analyses.

25
26 130 The contribution of each meteorological factor to the case counts was first investigated with
27
28
29 131 the Wuhan dataset under the assumption of previously defined time delay through Spearman's
30
31 132 correlation test. Then, we performed single-factor NLS regression modeling for each
32
33 133 meteorological variable in the discovery dataset (all Chinese cities with monthly confirmed cases
34
35 134 over 50) under the assumption of previously determined relationship and pre-defined time delay,
36
37
38 135 to determine the exact coefficients accompanied with each meteorological factor and to find out
39
40 136 the most suitable environmental condition for SARS-CoV-2.

41
42 137 Then, two final prediction models (short-term model and long-term model) were developed
43
44
45 138 using the discovery dataset with the previously determined coefficients. The prediction model
46
47 139 supposed that all the meteorological variables, with their specific coefficients determined by
48
49 140 single-factor modeling, were added together to compose a weather coefficient. The new
50
51
52 141 confirmed case count on the test day is calculated by multiplying the weather coefficient with the
53
54 142 existing confirmed case count on the exposure day (the time delay between test day and exposure
55
56 143 day is determined in previous analysis), and then multiply by a constant coefficient. The short-

1
2
3 144 term model took all four variables, while the long-term model only considered temperature as it
4
5
6 145 is easy to be forecasted. There was a constant coefficient for the total equation, that was
7
8 146 multiplied by the existing confirmed case count. Its exact value was determined by model fitting
9
10 147 in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was
11
12 148 investigated with all datasets covering the world's top cities and areas. The correlation of existing
13
14
15 149 confirmed case counts with newly confirmed case counts was also investigated. Basic statistics
16
17 150 and modeling was conducted in R 3.5.1 (<https://cran.r-project.org/>).
18

19 151 Model validation and application

20
21
22 152 The best fitted model was validated in the replication datasets (Italian city-level data and
23
24 153 other nation-level data) by correlating the observed actual epidemiological data with the
25
26 154 predicted values from the model in the datasets. We used these fitted models to calculate a
27
28
29 155 predicted value for case counts for each studied site, and then compared this predicted value with
30
31 156 the real observed case counts by calculating a Spearman's correlation coefficient ρ between them.
32

33 157 Patient and Public Involvement

34
35 158 No specific patients were included in the current study. Epidemiological data were
36
37
38 159 downloaded from online open-source databases. The public were not involved in the planning
39
40 160 and design of the study.
41

42 161 **Results**

43 44 45 162 The Weather's influence on SARS-CoV-2 transmission displays 3~7 days time delay

46
47
48 163 The ranges of average temperature, relative humidity, wind speed, and visibility in the replication
49
50 164 datasets were similar to those in the discovery dataset (see Supplementary Results for detailed
51
52 165 datasets description). Non-linear modeling with Wuhan dataset under the assumption of 3~7 days
53
54
55 166 delay of infection confirmation suggested that the effect of temperature and wind speed is better
56
57
58
59
60

1
2
3 167 depicted as quadric (Table S1), which was also supported by Loess regression interpolation (Fig.
4
5 168 1). The mode for relative humidity and visibility was hard to be determined, as statistics
6
7 supported both relationships (Table S1). Considering the common knowledge of coronavirus
8 169
9 transmission and the trend showed by Loess regression interpolation, relative humidity exerted its
10 170
11 impact in a quadric trend while visibility exerted its impact in a linear trend (Fig. 1,
12 171
13 Supplementary results).
14
15 172

16
17 173 Furthermore, investigation of the time delay effect in the Wuhan dataset showed that the number
18
19 of confirmed new cases was best correlated with air temperature 3~7 days ago, relative humidity
20 174
21 and visibility 7 days ago, and wind speed on the exposure day (Table S2). By comprehensive
22 175
23 consideration of all four meteorological variables and the differences between statistics values,
24 176
25 the weather 3~7 days ago, as well as weather one week ago, could well predict COVID-19
26
27 177 outbreak. It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure
28
29 178 under certain adverse weather might exhibit its effect after 3~7 days.
30
31 179

34 180 Contribution of single meteorological factor to the outbreak

35
36

37 181 In the Wuhan dataset, the new case count was significantly positively correlated with temperature
38
39 182 (Spearman's correlation $\rho = 0.69$, $p < 0.001$) and visibility ($\rho = 0.43$, $p = 0.04$), and negatively
40
41 183 correlated with wind speed ($\rho = -0.45$, $p = 0.03$) and relative humidity ($\rho = -0.33$, $p = 0.12$) 3~7
42
43 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind
44 184
45 speed, visibility, and relative humidity. A model only with temperature as a parameter could
46 185
47 already explained 45% of the variance in the epidemic data ($p = 4 \times 10^{-4}$), while wind speed and
48 186
49 visibility could explain over 25% of the variance. According to the fitted single-factor models
50
51 187 (temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility
52
53 188 was fitted into a linear model, see the Supplementary Results for details), SARS-CoV-2
54
55 189

transmission reaches a peak when mean temperature is 6.18 °C (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:

New Case Count

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

where T is temperature in °C, 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

1
2
3 211 we substitute replication datasets into this short-term model with the multiplicative constant
4
5 212 coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious
6
7
8 213 underestimation of predicted values against real ones was observed although the predicted values
9
10 214 correlated with the real ones very well. We supposed it was due to site-specific difference in the
11
12 215 multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the
13
14
15 216 pandemic had been controlled early. Thus, we further re-fitted this composed model with all
16
17 217 datasets to determine a more accurate value of the multiplicative constant coefficient α , which
18
19 218 was 0.001 then. In practical application, we need to first plot the observed case count vs.
20
21
22 219 predicted one with a default α value 0.001, and then examine the extent of underestimation or
23
24 220 overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the
25
26 221 impact size of "weather coefficient" for a certain site.
27

28
29 222 Substitute data from the past two months, a good prediction performance was obtained for
30
31 223 this short-term model, with the predicted values significantly correlated to the observed ones for
32
33 224 most areas (Fig. 3). However, only the existing confirmed case count data could not predict the
34
35 225 new case count 3~7 days later as well as the weather-combined model did (Table S3).
36

37 38 226 Different modes of viral transmission illustrated by the model 39

40 227 The observed versus predicted data exhibited different correlation patterns for different areas,
41
42 228 meaning different viral transmission modes, which may indicate the effect of epidemic control
43
44
45 229 for certain area.

46
47 230 Data from Chinese top-affected cities were not very well predicted and obviously
48
49 231 overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p <$
50
51 232 0.001 ; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after
52
53
54 233 the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention
55
56
57
58
59

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 ($\rho_{\text{Wuhan}} = 0.69$, $\rho_{\text{Italy}} = 0.87$, $\rho_{\text{Japan}} = 0.80$, $\rho_{\text{Iran}} = 0.86$, $p < 0.001$, $\rho_{\text{Korea}} = 0.43$, $p = 0.002$; 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 ($\rho_{\text{France}} = 0.96$, $\rho_{\text{United States}} = 0.93$, $\rho_{\text{United Kingdom}} = 0.83$, $\rho_{\text{Spain}} = 0.97$, $\rho_{\text{Germany}} = 0.94$, $p < 0.001$; 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 ($\rho_{\text{India}} = 0.94$, $\rho_{\text{Singapore}} = 0.66$, $p < 0.001$, $\rho_{\text{Thailand}} = 0.56$, $p = 0.001$; Fig. 3D), with just relatively small daily new case counts compared to temperate regions.

1
2
3 257 Countries in the southern hemisphere displayed similar pattern as the "controlled" with large
4
5
6 258 overestimation by the model when the confirmed cases increase, leading to not good prediction
7
8 259 performance ($\rho_{\text{Australia}} = 0.79$, $p < 0.001$, $\rho_{\text{South Africa}} = 0.34$, $p = 0.08$; Fig. 3E). It might be due to
9
10 260 the effect of epidemic prevention measures and hot summer weather in these countries.

11 12 261 Long-term simplified model

13
14
15 262 Long-term prediction depends on weather forecast, which generally reports only average
16
17
18 263 temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as
19
20 264 temperature in a short time delay (3~7 days ago), we again performed single-factor regression
21
22 265 modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a
23
24 266 quadric function (Supplementary Results). This simplified model with average temperature as a
25
26
27 267 weather factor was derived as follows:

$$28$$
$$29 \text{ new case count} = (-0.10 \times T^2 + 1.11 \times T + 46.42) \times \beta \times \text{Existing Confirmed Case Count}$$
$$30$$

31
32 269 where T is temperature in °C, β is a site-related multiplicative constant coefficient, with a default
33
34
35 270 of 0.006. All parameters take values 14 days before the day new case count is evaluated.

36
37
38 271 With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, p
39
40 272 < 0.001 ; Fig. 3F). The long-term simplified prediction model also showed five prediction-
41
42 273 observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the
43
44
45 274 studied areas. This model could directly predict the newly emerging cases 14 days later, and be
46
47 275 used to predict COVID-19 outbreak in the future month by summing up the daily new case count
48
49 276 and combining weather forecast (usually available for the future 15 days).

50 51 52 277 **Discussion**

1
2
3 278 This research discovers nonlinear dose-response relationship for meteorological factors, in
4
5
6 279 consistency with previous studies (12). Predictions of COVID-19 outbreak scale by the models
7
8 280 were well correlated with the observations around the world, suggesting the importance of
9
10 281 weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many
11
12 282 respiratory infectious diseases, such as influenza, is dependent upon temperature and relative
13
14
15 283 humidity (5,6). Recent published papers on preprint servers have reported roles of temperature
16
17 284 and absolute humidity in the COVID-19 transmission, but their conclusions are diverse (8-13). In
18
19 285 contrast to the findings by Cai et al (8), this study suggests significant impact of mean
20
21
22 286 temperature on the daily new case count, indicating a need for sufficient time delay between
23
24 287 exposure and confirmation for weather to exhibit its effect. In contrary to other two studies (9,10),
25
26 288 this research suggests that there is a relatively not wide temperature and humidity ranges for the
27
28
29 289 pandemic. There is an optimal temperature for SARS-CoV-2 at 6.18 °C, which is colder than that
30
31 290 suggested by Bu et al (14) but in consistency with the estimation by Wang et al (12); and most
32
33 291 areas with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al
34
35 292 suggested (14). It is of notice that different from other viral respiratory diseases such as
36
37
38 293 influenza(15)(16), high relative humidity is better for SARS-CoV-2 to spread, suggesting that a
39
40 294 sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more
41
42 295 important for the spread than the effect of dry air on the human immune system. Different from
43
44
45 296 other studies (17), this study also finds significant involvement of wind speed, in a quadric
46
47 297 manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In
48
49 298 addition, the current study discovered that visibility was significantly negatively correlated with
50
51
52 299 new case count and played a more important role in viral spread than humidity did (from
53
54 300 spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g.,
55
56 301 dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may

1
2
3 302 indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case
4
5 303 count decreases rapidly when visibility is high than 13 statute miles, indicating that caution
6
7
8 304 should be taken if visibility drops below 10 statute miles.
9

10
11 305 In the prediction model, there is a multiplicative constant coefficient which determines the
12
13 306 strength of the weather coefficient on the epidemic transmission. It seems site-specific, as
14
15 307 adjusting it could make the prediction for one site very close to the observation. This constant
16
17 308 might reflect the influence of a couple of site-specific confounding factors, such as epidemic
18
19
20 309 control measures, sun radiation, and population density. Various degrees of isolation for various
21
22 310 areas around the world lead to different degrees of weather effect. When evaluate the prediction
23
24 311 performance by the short-term model and the long-term model, they both exhibit different
25
26
27 312 prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of
28
29 313 epidemic control and isolation policy would lead to deviation from the original prediction and
30
31 314 thus different prediction-observation correlation patterns. Therefore, by plotting the predicted
32
33 315 versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it
34
35
36 316 would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the
37
38 317 observed case counts dropped dramatically from the predictions for Wuhan seven days after their
39
40 318 shelter hospitals were put in use, suggesting the importance and necessity of building shelter
41
42
43 319 hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and
44
45 320 very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to
46
47 321 natural transmission mode. Another thing worth attention is that although the weather in tropical
48
49
50 322 areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on
51
52 323 spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore,
53
54 324 these tropical areas should still be on the alert against future outbreak of COVID-19.
55
56
57
58
59

1
2
3 325 Although those cases with travel history to China or indicated by the World Health Organization
4
5 326 as "imported case only" were excluded in this study to make the world data most likely local
6
7
8 327 transmitted, it was difficult to separate the imported cases from local transmission very well in
9
10 328 practice. It might explain the not excellent correlations of predictions with observations.
11
12 329 Furthermore, the relationship of weather and COVID-19 could be complex, since the human
13
14
15 330 immune system has an innate seasonal rhythm, and the immune system could also be affected by
16
17 331 weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway
18
19 332 mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets
20
21
22 333 for virus to adhere.

23
24 334 There are several limitations of this study. First of all, this prediction model (especially the long-
25
26
27 335 term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter,
28
29 336 as the models were derived using Chinese datasets, mainly in the first three months of 2020. The
30
31 337 prediction became inaccurate and even improper under hot weather (i.e., the predicted values of
32
33
34 338 long-term model become negative when air temperature is higher than 28 °C), which might
35
36 339 explain the obvious bad prediction performance for countries in the southern hemisphere and
37
38
39 340 tropical areas. One explanation for the inaccurate prediction in areas with high temperature could
40
41 341 be that SARS-CoV-2 transmission in these areas was mainly not influenced by weather, but in
42
43 342 another direct transmission way, such as face-to-face contact or spread in gathering crowd.
44
45 343 Second, it seems that the prediction performance drops with the increase in new case count,
46
47
48 344 suggesting that the prediction model might become inaccurate and not suitable for very large new
49
50 345 case count. This could be due to (1) the influence of weather on COVID-19 spread might weaken
51
52 346 when the number of cases increases, while other factors such as social distance become more
53
54
55 347 important at a later stage; (2) there was less data points with large new case count, which might

1
2
3 348 lead to larger variance. Third, the short-term prediction model must use all four meteorological
4
5 349 factors, while these factors are not always available for any one certain area. Fourth, this study
6
7 included various areas covering a long period into modeling, thus, there were a bunch of variable
8 350
9 confounding factors, such as population mobility and disinfection measures, which were not
10 351
11 controlled and thus could impede the model accuracy. Fifth, as we could only obtain country-
12 352
13 level epidemiological data, the corresponding meteorological data were obtained for their capital
14
15 353 cities, leading to not exact pairing of epidemiological data and meteorological data. Sixth, there is
16
17 354 a general lack of data and cases in the current study, since we only collected data covering two
18
19 355 and a half months while the pandemic has persisted over seven months up to now.
20
21
22 356

23 24 25 357 **Conclusion**

26
27
28 358 In summary, this study has found significant correlations with the COVID-19 epidemic trend for
29
30 359 not only temperature and humidity, but also wind speed and visibility. It proposed a
31
32 360 comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version
33
34 361 and a long-term version. The short-term version uses the combination of four meteorological
35
36 factors as a "weather coefficient" of the existing confirmed case count in the past week and can
37 362
38 be used to predict epidemic situation in the future three days; the short-term version uses average
39 363
40 temperature as the "weather coefficient" seven days ago and can predict the outbreak in one
41 364
42 month if combined with weather forecast. This model is easy to use for predicting the COVID-19
43
44 365 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case
45
46 366 count for the future couple of days or month. This model will be very helpful for local
47
48 367 governments to make timely policies on epidemic control, for instance, the allocation of medical
49
50 368 equipments such as ventilators and medical resources such as hospitals, beds and health-care
51
52 workers, according to the prediction results.
53 369
54
55 370

1
2
3 371
4
5
6 372 **Acknowledgments:** We thank Dr. Zhisheng Huang for advices on data collecting and processing,
7
8 373 Dr. Siyuan Tan for technical support on analysis, Dr. Yonggao Chen for mathematical support on
9
10 modeling.
11 374

12
13 375 **Funding:** the Priority Academic Program Development of Jiangsu Higher Education Institutions
14
15 – the third period (NO.035062002003c), the Yizhong Research Fund of Jiangsu Provincial
16 376 Hospital of Chinese Medicine (Y19066).
17
18 377

19
20
21 378 **Author contributions:** BC, YZ and XZ design and interpret the reported analyses and results;
22
23 379 HL, XY, YH, BZ, and MX participated in the acquisition of data; BC analyse the data, drafted,
24
25 380 and revised the manuscript; HL and XZ revised the manuscript; YZ and FT provided technical
26
27 support; XZ supervised the research.
28 381

29
30 382 **Competing interests:** Authors declare no competing interests.
31
32

33 383 **Data and materials availability:** Weather data and epidemiological data is all obtained from
34
35 public databases. Detailed modeling results are available upon request by emailing to Biqing
36 384 Chen, bq_chen@qq.com.
37
38 385

39 40 41 386 42 43 387 **References:**

- 46 388 1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients
47
48 with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.
49 389
- 51 390 2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in
52
53 Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med.
54 391
55
56 392 2020;382(13):1199–207.
57
58
59
60

- 1
2
3 393 3. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of
4 pneumonia associated with the 2019 novel coronavirus indicating person-to-person
5 394 pneumonia associated with the 2019 novel coronavirus indicating person-to-person
6 transmission: a study of a family cluster. *Lancet* [Internet]. Elsevier Ltd;
7 395 2020;395(10223):514–23. Available from: [http://dx.doi.org/10.1016-](http://dx.doi.org/10.1016/S0140-)
8 396 [6736\(20\)30154-9](http://dx.doi.org/10.1016/S0140-6736(20)30154-9)
9
10
11
12 397
13
14
- 15 398 4. Organization WH. [https://www.who.int/emergencies/diseases/novel-coronavirus-](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports)
16 2019/situation-reports [Internet]. 2020 [cited 2020 Mar 29]. Available from:
17 399 <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
18
19
20 400
21
22
- 23 401 5. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30
24 years of county-level evidence from the United States. *Amer J Epidemiol*. 2012;176(suppl.
25 402 7):S114–22.
26
27 403
28
29
- 30 404 6. Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on
31 relative humidity and temperature. *PLoS Pathol*. 2007;3:e151.
32 405
33
34
- 35 406 7. Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential
36 Implications for Reducing Transmission of COVID-19. *JAMA*. 2020;E1–2.
37 407
38
39
- 40 408 8. Cai Y. The Effects of "Fangcang, Huoshenshan [Internet]. Available from:
41 <https://www.medrxiv.org/content/10.1101/2020.02.26.20028472v3>
42 409
43
44
- 45 410 9. Proverbio AM, Crotti N, Manfredi M, Zani A. Preliminary evidence that higher
46 temperatures are associated with lower incidence of COVID-19, for cases reported
47 411 globally up to 29th February 2020. *medRxiv*. 2020;(February).
48
49
50 412
51
- 52 10. <https://github.com/BlankerL/DXY-COVID-19-Data>.
53 413
54
- 55 11. <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/2020/>. 2020.
56 414
57

1

2

3 415

4

5 416

6

7 417

8

9

10

11 418

12

13 419

14

15 420

16

17 421

18

19 422

20

21

22 423

23

24 424

25

26 425

27

28 426

29

30 427

31

32 428

33

34 429

35

36 430

37

38 431

39

40 432

41

42 433

43

44 434

45

46 435

47

48 436

49

50 437

51

52 438

53

54

55

56

57

58

59

60

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 °C, (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 °C, (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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

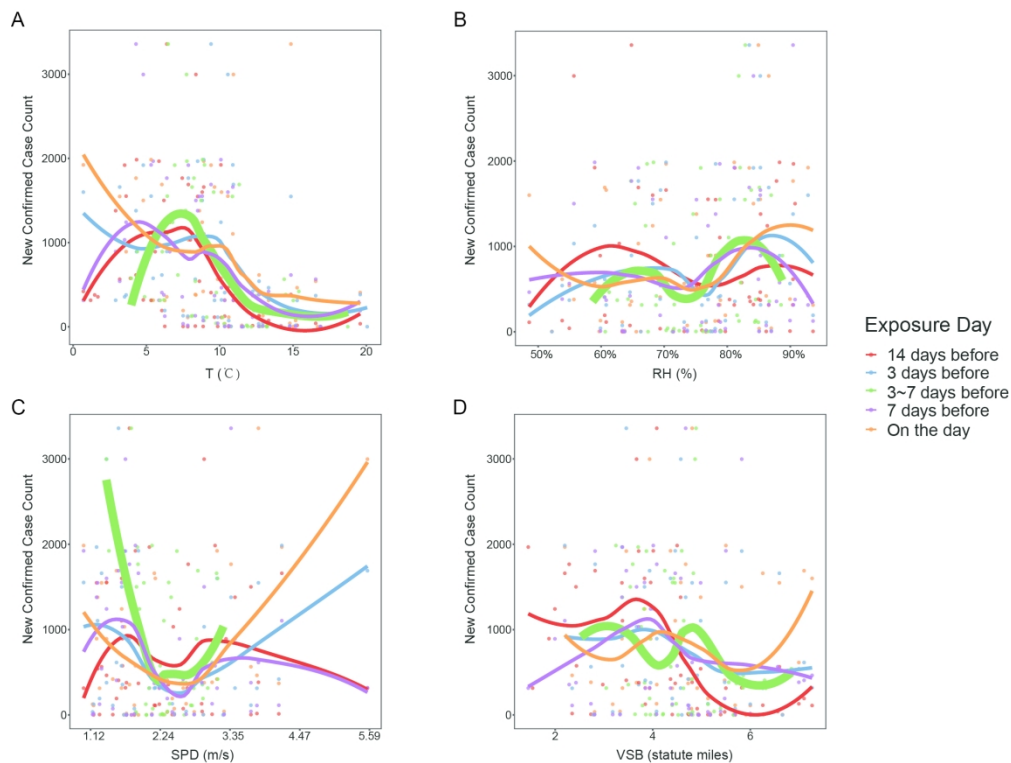
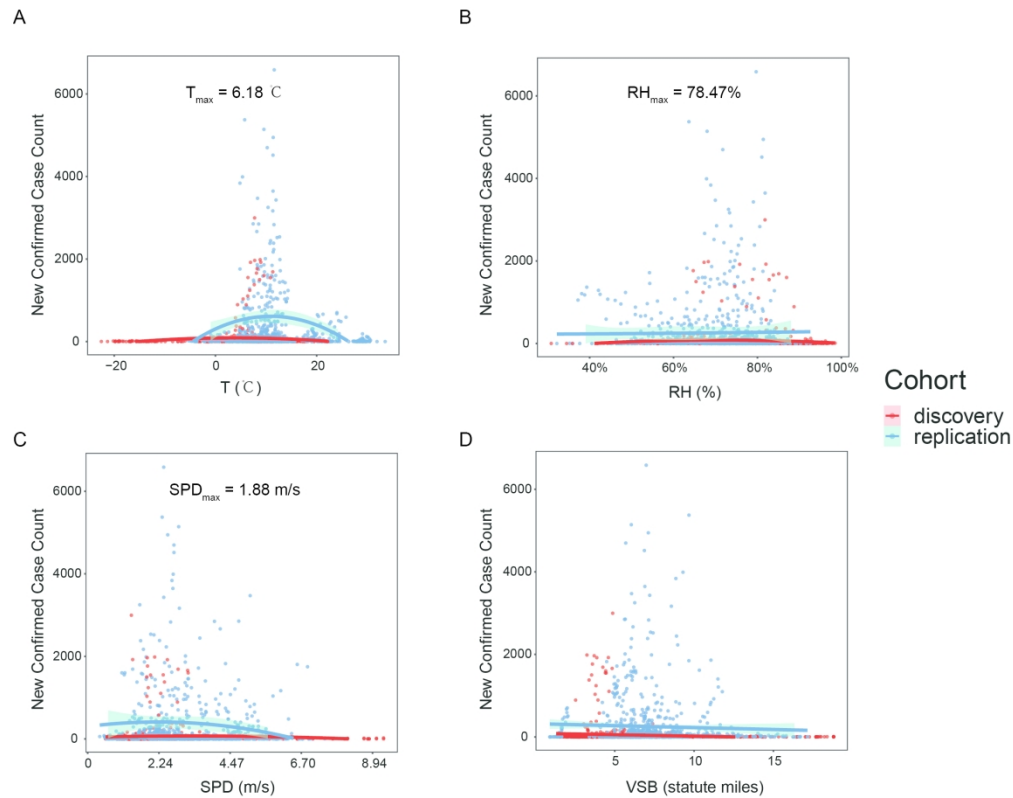


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of 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)



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

38 218x172mm (300 x 300 DPI)

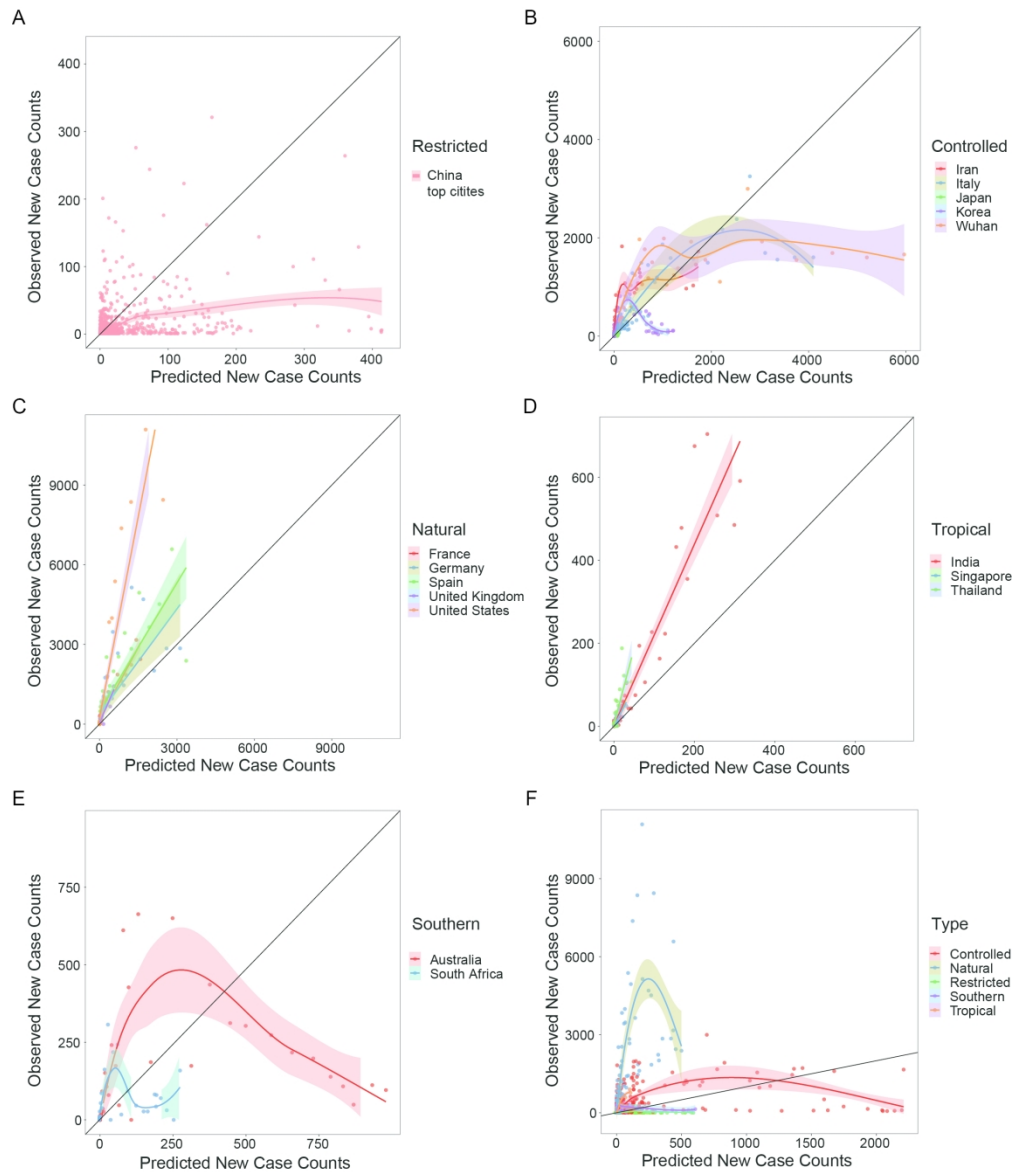


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" 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)

1 **Supplementary Materials and Methods**

2 Epidemiological data

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

6 For Wuhan city, there was a shortage of test kits at the beginning of the pandemic,
7 which would make confirmed case counts much lower than the actual data, thus, we
8 discarded epidemic data before January 28th, the day when domestic test kits have
9 been approved, produced in large quantities, and were available for Wuhan hospitals.
10 As there was a cut down problem for the existing confirmed case count on February
11 20th for Wuhan, when modeling with the existing confirmed case count, only data
12 before February 20th were used.

13 Weather data

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

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

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

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

1
2
3
4 21 For a city with more than one meteorological stations, the one nearest to the city
5
6 22 center was chosen. For a province with epidemiological data, the meteorological
7
8
9 23 station in the capital city of that province was chosen. For a country with only
10
11
12 24 national wide epidemiological data, weather data were averaged across all the
13
14
15 25 meteorological observatories in the cities where outbreak was officially reported.
16
17 26 Latitude and elevation for the meteorological observatories were also collected.

18 19 27 Statistical modeling

20
21
22 28 Only one city Wuhan was chosen for illustrating the time delay effect because it
23
24 29 is the first city to have an outbreak of COVID-19, there was none reported imported
25
26
27 30 cases for Wuhan, which might obscure the correlation between weather and virus
28
29
30 31 transmission.

31 32 32 **Supplementary Results**

33 33 Datasets description

34
35 34 Only Chinese cities with monthly confirmed cases over 50 were included in the
36
37 35 discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in
38
39
40 36 Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily
41
42
43 37 confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in
44
45
46 38 all the other Chinese cities. We suppose that it might be due to abrupt large
47
48
49 39 supplement of virus test kits or data correction on that day. In order to reduce the
50
51
52 40 potential contamination of modeling by this outlier, data on that day were discarded
53
54
55 41 from the subsequent analysis. There were also two oddly large new confirmed case
56
57
58 42 counts for Lombardy, which were discarded from the subsequent analysis. Except the
59
60

1
2
3
4 43 outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to
5
6 44 2,997, the average temperature ranged $-22.54^{\circ}\text{C} \sim 22.16^{\circ}\text{C}$, the wind speed ranged
7
8
9 45 $0.56 \sim 9.29$ meter per second, visibility ranged $1.3 \sim 18.8$ statute miles, and relative
10
11
12 46 humidity ranged $30.84\% \sim 98.52\%$.

13 14 47 Model selection

15
16
17 48 With the increase of relative humidity, the amount of droplets in the air increases,
18
19 49 leading to more virus load. However, as the air gets humid, human's respiratory tract
20
21
22 50 could better defend virus infection. Thus, the relationship of relative humidity could
23
24
25 51 be complex, not pure linear. Giving comprehensive consideration, we defined the
26
27 52 effect of relative humidity to be quadric. As for visibility, it only affects the amount of
28
29
30 53 particles in the air, which is positively correlated with virus load. Thus, it is most
31
32
33 54 probably to exert its effect linearly.

34
35 55 Although relative humidity and visibility 7 days ago correlated with the
36
37 56 confirmed new case counts best, there was not great loss of model fitting statistics for
38
39
40 57 relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days
41
42
43 58 time delay and 3~7 days time delay for temperature.

44 45 59 Fitted models

46
47
48 60 The fitted single-factor models were as follows:

$$49 \quad \text{New Case Count} = -0.11305 \times T^2 + 1.39819 \times T + 45.11405$$

50
51
52 61 where T is temperature in $^{\circ}\text{C}$.

53
54
55 62 The estimate p-value for constant was < 0.001 . The extremum was $-1.39819/$
56
57
58 63 $(2 \times (-0.11305)) = 6.183945 \text{ }^{\circ}\text{C}$.

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

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 $-5.120123/(2 \times (-1.360056)) = 1.882321 \text{ m/s}$.

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

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

New Case Count

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

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:

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

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:

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

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

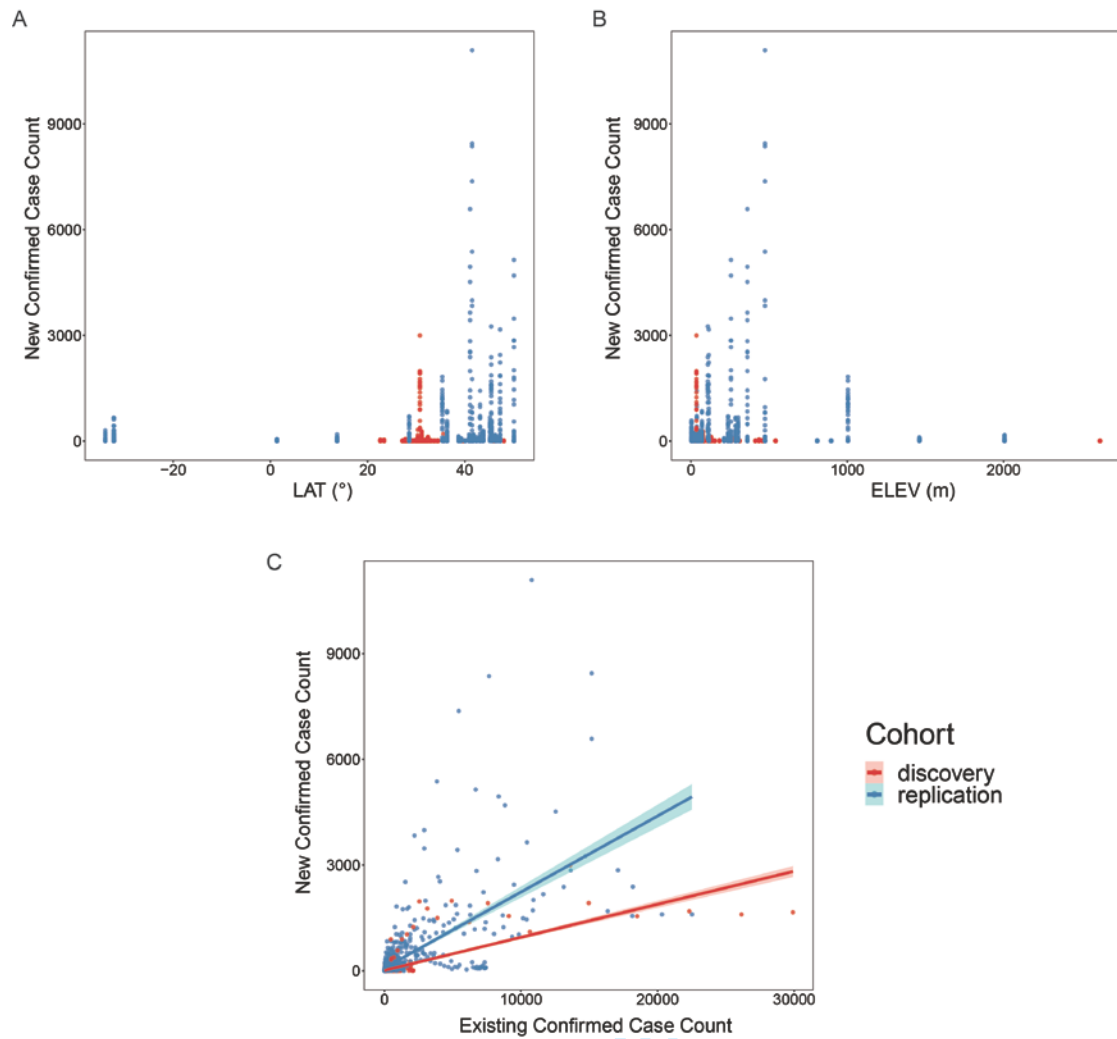
93 Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the
 94 log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information
 95 Criterion for the model; deviance, deviance of the model; Corr, Spearman's 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.

BMJ Open

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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3
4 1 **Title: Predicting the local COVID-19 outbreak around the world with**
5
6 2 **meteorological conditions: a model-based qualitative study**
7
8

9 3 **Authors:** Biqing Chen¹, Hao Liang², Xiaomin Yuan³, Yingying Hu², Miao Xu², Yating Zhao⁴,
10 4 Binfen Zhang², Fang Tian¹, Xuejun Zhu^{1,2*}
11
12
13

14
15 5 **Affiliations:**
16
17

18 6 ¹ Research Center of Chinese Medicine, Jiangsu Province Hospital of Chinese Medicine, the
19 7 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
20
21

22 8 ² Department of Hematology, Jiangsu Province Hospital of Chinese Medicine, the Affiliated
23 9 Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
24
25

26 10 ³ Department of Colorectal Surgery, Jiangsu Province Hospital of Chinese Medicine, the
27 11 Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
28
29

30 12 ⁴ School of Atmospheric Sciences, Nanjing University, Nanjing, China
31
32

33
34 13 *Correspondence to: Xuejun Zhu, zhuxuejun@njucm.edu.cn.
35
36

37 14
38 15 **Abstract**
39
40

41
42 16 **OBJECTIVES:** This study aims to investigate the relationship between daily weather and
43 17 transmission rate of SARS-CoV-2, and to develop a generalized model for future prediction of
44 18 the COVID-19 spreading rate for a certain area with meteorological factors.
45
46
47

48 19 **DESIGN:** A retrospective, qualitative study.
49
50

51
52 20 **METHODS AND ANALYSIS:** We collected 382,596 records of weather data with four
53 21 meteorological factors, i.e., average temperature, relative humidity, wind speed, and visibility,
54
55
56
57

1
2
3 22 and 15,192 records of epidemic data with daily new confirmed case counts (1,587,209 confirmed
4
5 23 cases in total) in nearly 500 areas worldwide from January 20 to April 9 in 2020. Epidemic data
6
7
8 24 were modeled against weather data to find a model that could best predict the future outbreak.
9

10
11 25 **RESULTS:** Significant correlations of the daily new confirmed case counts with the weather 3~7
12
13 26 days ago were found. SARS-CoV-2 is easy to spread under weather conditions of average
14
15 27 temperature at 5~15 °C, relative humidity at 70%~80%, wind speed at 1.5~4.5 meter / second,
16
17
18 28 and visibility less than 10 statute miles. A short-term model with these four meteorological
19
20 29 variables in the past 3~7 days was derived to predict the daily increase in COVID-19; and a long-
21
22 30 term model using temperature to predict the pandemic in the next week or month was derived.
23
24
25 31 Taken China as a discovery dataset, it was well validated with worldwide data. According to this
26
27 32 model, there are five viral transmission patterns, "restricted", "controlled", "natural", "tropical",
28
29 33 "southern". This model's prediction performance correlates with actual observations best (over
30
31
32 34 0.9 correlation coefficient) under natural spread mode of SARS-CoV-2 when there is not much
33
34 35 human interference by epidemic prevention measures.
35

36
37 36 **CONCLUSIONS:** This model can be used for prediction of the future outbreak, and illustrating
38
39 37 the effect of epidemic control for a certain area.
40

41
42 38 **Keywords:** COVID-19, SARS-CoV-2, weather, temperature, prediction model, epidemic control
43
44
45 39
46

47 40 **Strengths and limitations of this study**

- 48
49
50 41 ● This study investigates the role of daily weather in COVID-19 spread systematically with a
51
52 42 comprehensive set of four meteorological factors.
53
54
55
56
57
58
59
60

- 1
2
3 43 ● This research collected a huge amount of data, covering nearly 500 areas worldwide in a long
4
5 44 timescale.
6
7
8 45 ● The current study proposes mathematical models integrating meteorological information for
9
10 46 predicting COVID-19 case counts in the future.
11
12
13 47 ● The influence of weather on virus spread could be confounded by a dozen of manual
14
15 48 interventions, such as population mobility and disinfection measures, leading to inaccurate
16
17 modeling.
18 49
19
20
21 50 ● The prediction model (especially the long-term model) might be unsuitable and inaccurate
22
23 51 for areas with hot weather.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

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

1
2
3 75 the world, including over 400 Chinese cities and areas, 18 Italian provinces, and 13 other
4
5 76 countries. Five time point's delay of virus infection from exposure were considered and compared
6
7
8 77 to determine the most reasonable time point's delay. A multivariate polynomial regression model
9
10 78 with meteorological factors as a "weather coefficient" of the existing confirmed case count was
11
12 79 established in a discovery Chinese dataset, and then validated by worldwide data. Five
13
14 80 transmission modes, indicating different levels of epidemic control, were revealed by this model.
15
16
17 81 In this view, this model can not only predict future outbreak, but also be used to evaluate the
18
19 82 effect of epidemic prevention measures for a certain area.
20
21

22 83 **Materials and Methods**

23 24 25 84 Epidemiological data

26
27 85 Epidemiological data were collected from the World Health Organization (WHO)[11],
28
29
30 86 European Centre for Disease Control and Prevention, and DXY-COVID-19-Data[12]. The daily
31
32 87 new confirmed case counts were collected from January 20, 2020 to April 9, 2020. Incidence data
33
34 88 were obtained for 428 Chinese cities and districts, 18 Italian provinces, and 13 other countries,
35
36
37 89 namely, United States, United Kingdom, Germany, France, Spain, Iran, Korea, Japan, Australia,
38
39 90 South Africa, India, Thailand, and Singapore. Considering the potential confounding effect, only
40
41 91 Chinese cities with no less than 50 cumulative confirmed cases in one month and without official
42
43
44 92 reports of large imported cases (42 cities in total) were taken as a discovery dataset, while those
45
46 93 for Italian provinces and all the other nations were taken as replication datasets (Supplementary
47
48 94 Materials).

49 50 95 Weather data

51
52
53 96 Four meteorological variables were chosen, air temperature, relative humidity, wind speed,
54
55 97 and visibility. Temperature could affect virus viability in the environment. Wind speed could
56
57
58
59

1
2
3 98 affect the suspending time of virus-attached particles. Relative humidity reflects the amount of
4
5 99 droplets in the air. Visibility is influenced by the amount of particles such as dust and air
6
7
8 100 pollutants. These two parameters both affect the amount of mediator for the virus to stay in the
9
10 101 air. Therefore, temperature, dew point, wind speed, and visibility were collected, and relative
11
12 102 humidity was calculated accordingly (Supplementary Materials). We obtained hourly values of
13
14
15 103 meteorological observations and geographic factors (latitude and elevation) from the Integrated
16
17 104 Surface Database of USA National Centers for Environmental Information[13]. Daily data were
18
19 105 calculated by averaging the hourly data for each variable in each day.
20
21

22 106 Statistical modeling

23
24 107 The number of daily new confirmed cases was taken as a dependent variable. Four
25
26
27 108 meteorological variables, namely, average temperature, wind speed, visibility, and relative
28
29 109 humidity, and the existing confirmed case counts were taken as independent variables.
30
31 110 Considering that there is a latency stage from the day one get infected to the day being
32
33
34 111 confirmed, a time delay of the day COVID-19 was confirmed from the day weather data were
35
36 112 collected needs to be taken into consideration. As it is reported that the latency period for
37
38 113 COVID-19 is 3~7 days on average and 14 days at most, five time points delay of virus infection
39
40
41 114 were taken into consideration, that is, weather data and existing confirmed cases count data were
42
43 115 collected on the day, three days before, seven days before, 3~7 days before, and 14 days before
44
45 116 collecting the new confirmed case count data.
46

47 117 To investigate whether the influence of meteorological factors is linear or quadric, both
48
49
50 118 linear and non-linear modeling were performed under different relationship assumptions to
51
52 119 compare model fitness statistics. Each meteorological variable was fitted into a bunch of single-
53
54 120 factor models (either generalized linear model or polynomial model) through non-linear least
55
56
57
58
59
60

1
2
3 121 squares (NLS) modeling using the Wuhan dataset with a 3~7 days delay of infection. The
4
5 122 relationship between each meteorological variable and confirmed new case count (linear or
6
7
8 123 quadric) was identified based on model fitness (log-likelihood, Akaike information criterion,
9
10 124 Bayesian Information Criterion, etc.) and common knowledge of droplet-mediated viral diseases.
11

12 125 Second, the proper time delay from weather exposure to COVID-19 confirmation was
13
14 investigated in the Wuhan dataset through Loess regression interpolation and NLS modeling with
15 126
16
17 127 the previously identified relationship for each meteorological variable. The most possible time
18
19 128 delay identified was taken for subsequent analyses.
20

21
22 129 To investigate the degree of contribution for each meteorological factor to the COVID-19
23
24 130 case counts, Spearman's correlation test (a non-parametric method that measures the strength and
25
26 131 direction of associations) was first adopted, with the Wuhan dataset under the assumption of
27
28 132 previously defined time delay. Nevertheless, here we assumed monotonic correlations between
29
30
31 133 COVID-19 case count and meteorological variables, while we could not exclude the possibility
32
33 134 that the real relationship was not monotonic, which might impede the accuracy of correlation
34
35 135 analysis. Then, we performed single-factor NLS regression modeling for each meteorological
36
37
38 136 variable in the discovery dataset under the assumption of previously determined relationship and
39
40 137 pre-defined time delay, to determine the exact coefficients accompanied with each
41
42 138 meteorological factor and to find out the most suitable environmental condition for SARS-CoV-2.
43

44
45 139 Then, two final prediction models (short-term model and long-term model) were developed
46
47 140 using the discovery dataset with the previously determined coefficients. The prediction model
48
49 141 supposed that all the meteorological variables, with their specific coefficients determined by
50
51 142 single-factor modeling, were added together to compose a weather coefficient. The new
52
53
54 143 confirmed case count on the test day is calculated by multiplying the weather coefficient with the
55
56 144 existing confirmed case count on the exposure day (the time delay between test day and exposure
57
58
59
60

1
2
3 145 day is determined in previous analysis), and then multiply by a constant coefficient. The short-
4
5
6 146 term model took all four variables, while the long-term model only considered temperature as it
7
8 147 is easy to be forecasted. There was a constant coefficient for the total equation, that was
9
10 148 multiplied by the existing confirmed case count. Its exact value was determined by model fitting
11
12 149 in the discovery dataset. The influence of geographic factors, i.e., latitude and elevation, was
13
14
15 150 investigated with all datasets covering the world's top cities and areas. The correlation of existing
16
17 151 confirmed case counts with newly confirmed case counts was also investigated. Basic statistics
18
19 152 and modeling was conducted in R 3.5.1 (<https://cran.r-project.org/>).
20

21 22 153 Model validation and application

23
24 154 The best fitted model was validated in the replication datasets (Italian city-level data and
25
26 155 other nation-level data) by correlating the observed actual epidemiological data with the
27
28 156 predicted values from the model in the datasets. We used these fitted models to calculate a
29
30
31 157 predicted value for case counts for each studied site, and then compared this predicted value with
32
33 158 the real observed case counts by calculating a Spearman's correlation coefficient ρ between them.
34

35 159 Patient and Public Involvement

36
37
38 160 No specific patients were included in the current study. Epidemiological data were
39
40 161 downloaded from online open-source databases. The public were not involved in the planning
41
42 162 and design of the study.
43

44 45 163 **Results**

46 47 48 164 The Weather's influence on SARS-CoV-2 transmission displays 3~7 days time delay

49
50 165 The ranges of average temperature, relative humidity, wind speed, and visibility in the replication
51
52 166 datasets were similar to those in the discovery dataset (see Supplementary Results for detailed
53
54
55 167 datasets description). Non-linear modeling with Wuhan dataset under the assumption of 3~7 days
56
57
58
59
60

1
2
3 168 delay of infection confirmation suggested that the effect of temperature and wind speed is better
4
5 depicted as quadric (Table S1), which was also supported by Loess regression interpolation (Fig.
6 169
7
8 170 1). The mode for relative humidity and visibility was hard to be determined, as statistics
9
10 171 supported both relationships (Table S1). Considering the common knowledge of coronavirus
11
12 172 transmission and the trend showed by Loess regression interpolation, relative humidity exerted its
13
14 impact in a quadric trend while visibility exerted its impact in a linear trend (Fig. 1,
15 173
16
17 174 Supplementary results).

18
19
20 175 Furthermore, investigation of the time delay effect in the Wuhan dataset showed that the number
21
22 176 of confirmed new cases was best correlated with air temperature 3~7 days ago, relative humidity
23
24 177 and visibility 7 days ago, and wind speed on the exposure day (Table S2). By comprehensive
25
26 consideration of all four meteorological variables and the differences between statistics values,
27 178
28
29 179 the weather 3~7 days ago, as well as weather one week ago, could well predict COVID-19
30
31 180 outbreak. It coincided with the latency period of 3~7 days for SARS-CoV-2, that is, exposure
32
33 under certain adverse weather might exhibit its effect after 3~7 days.
34 181
35

36 182 Contribution of single meteorological factor to the outbreak

37
38

39 183 In the Wuhan dataset, the new case count was significantly positively correlated with temperature
40
41 184 (Spearman's correlation $\rho = 0.69$, $p < 0.001$) and visibility ($\rho = 0.43$, $p = 0.04$), and negatively
42
43 correlated with wind speed ($\rho = -0.45$, $p = 0.03$) and relative humidity ($\rho = -0.33$, $p = 0.12$) 3~7
44 185
45 days ago. It suggested that temperature was correlated with the outbreak best, followed by wind
46 186
47 speed, visibility, and relative humidity. A model only with temperature as a parameter could
48 187
49 already explained 45% of the variance in the epidemic data ($p = 4 \times 10^{-4}$), while wind speed and
50
51 188 visibility could explain over 25% of the variance. According to the fitted single-factor models
52
53 189 (temperature, relative humidity, and wind speed were fitted into quadratic models; and visibility
54
55 190
56
57
58
59
60

1
2
3 191 was fitted into a linear model, see the Supplementary Results for details), SARS-CoV-2
4
5
6 192 transmission reaches a peak when mean temperature is 6.18 °C (Fig. 2A), relative humidity is
7
8 193 78.47% (Fig. 2B), and wind speed is 1.88 meter /second (m/s) (Fig. 2C); and its transmission rate
9
10 194 decreases with the increase of visibility (Fig. 2D). The effects of geographic factors such as
11
12
13 195 latitude and elevation, and the pure influence from the existing case count were further
14
15 196 investigated in the worldwide datasets (Fig. S1), illustrating that COVID-19 mainly outbreaks at
16
17 197 latitude 30°~50° (Fig. S1A) and elevation < 500 metre (Fig. S1B). New confirmed case count
18
19
20 198 was positively correlated with the existing confirmed case count (Fig. S1C).

21 22 199 Short-term prediction model

23
24
25 200 We further derived a full model combined with all four meteorological variables and fitted this
26
27
28 201 model with the discovery dataset (Supplementary Results). The best-fitted short-term model was
29
30 202 as follows:

31
32
33
34 203 **New Case Count**

$$35 \quad = (-0.11 \times T^2 + 1.40 \times T - 0.058 \times RH^2 + 9.04 \times RH - 1.36 \times SPD^2 + 5.12 \\ 36 \quad \times SPD - 7.02 \times VSB - 126.66) \times \alpha \times \text{Existing Confirmed Case Count}$$

37
38 204 where T is temperature in °C, RH is relative humidity in percentage (defined as over 15%), SPD
39
40 205 is wind speed in m/s, VSB is visibility in statute miles, α is a site-specific constant, with a default
41
42
43 206 of 0.001. All parameters take the means of values 3~7 days before the day new case count is
44
45 207 evaluated.

46
47
48 208 In this model, all the four meteorological variables are added together in their proper forms
49
50 209 to compose a "weather coefficient" (the equation in brackets), which affects the transmission rate
51
52 210 of SARS-CoV-2, and thus influences the number of people that catch infection from the existing
53
54
55 211 confirmed cases, which then determines the new confirmed case count 3~7 days later. There is a

1
2
3 212 multiplicative constant coefficient α in the equation, which seems site-related. This constant
4
5 213 coefficient could adjust the strength of the "weather coefficient" on disease transmission. When
6
7
8 214 we substitute replication datasets into this short-term model with the multiplicative constant
9
10 215 coefficient α originally determined by the discovery dataset (which was 0.00048), an obvious
11
12 216 underestimation of predicted values against real ones was observed although the predicted values
13
14
15 217 correlated with the real ones very well. We supposed it was due to site-specific difference in the
16
17 218 multiplicative constant coefficient α since the discovery dataset was all Chinese areas where the
18
19 219 pandemic had been controlled early. Thus, we further re-fitted this composed model with all
20
21
22 220 datasets to determine a more accurate value of the multiplicative constant coefficient α , which
23
24 221 was 0.001 then. In practical application, we need to first plot the observed case count vs.
25
26 222 predicted one with a default α value 0.001, and then examine the extent of underestimation or
27
28
29 223 overestimation, to finally determine a proper multiplicative constant coefficient α to adjust the
30
31 224 impact size of "weather coefficient" for a certain site.

32
33 225 Substitute data from the past two months, a good prediction performance was obtained for
34
35 226 this short-term model, with the predicted values significantly correlated to the observed ones for
36
37
38 227 most areas (Fig. 3). However, only the existing confirmed case count data could not predict the
39
40 228 new case count 3~7 days later as well as the weather-combined model did (Table S3).

41 42 229 Different modes of viral transmission illustrated by the model

43
44
45 230 The observed versus predicted data exhibited different correlation patterns for different areas,
46
47 231 meaning different viral transmission modes, which may indicate the effect of epidemic control
48
49 232 for certain area.

50
51 233 Data from Chinese top-affected cities were not very well predicted and obviously
52
53
54 234 overestimated by this model with the default multiplicative constant coefficient α ($\rho = 0.11$, $p <$
55
56 235 0.001 ; Fig. 3A). It might be due to the reason that most Chinese cities took actions quickly after
57

1
2
3 236 the outbreak in Wuhan was reported, thus, these cities were under strict epidemic prevention
4
5 237 measures at the beginning of the pandemic. This viral transmission mode suggested by the not
6
7
8 238 well correlated prediction pattern is called "restricted".
9

10 239 For Wuhan city and some early outbreak countries (Japan, Korea, Iran, and Italy), the
11
12 240 predicted outbreak was well correlated with the actual observations at the beginning when the
13
14
15 241 existing confirmed cases were not in very large numbers, but the prediction deviates from the
16
17 242 observation as the confirmed cases increase, in detail, there's large overestimation of prediction
18
19 243 ($\rho_{\text{Wuhan}} = 0.69$, $\rho_{\text{Italy}} = 0.87$, $\rho_{\text{Japan}} = 0.80$, $\rho_{\text{Iran}} = 0.86$, $p < 0.001$, $\rho_{\text{Korea}} = 0.43$, $p = 0.002$; Fig. 3B).
20
21 It is of notice that the dramatic deviation of predictions for Wuhan occurred after February 15,
22 244
23
24 245 the day when shelter hospitals had been put into use for seven days (the average latency period
25
26 246 for COVID-19). Therefore, the deviated prediction pattern indicates that the outbreak prevention
27
28 247 and control taken in these areas is effective (so-called "controlled" mode). The number of cases
29
30
31 248 had been decreased by 72% for Wuhan, over 95% for Korea, Japan, and Italy, and 37% for Iran
32
33 249 at most due to epidemic control (the largest gap between prediction and observation).
34

35 250 For most European and American countries, the predicted outbreak was linear correlated
36
37
38 251 with the observed data very well ($\rho_{\text{France}} = 0.96$, $\rho_{\text{United States}} = 0.93$, $\rho_{\text{United Kingdom}} = 0.83$, $\rho_{\text{Spain}} =$
39
40 252 0.97 , $\rho_{\text{Germany}} = 0.94$, $p < 0.001$; Fig. 3C), suggesting a natural viral transmission mode without
41
42 253 much man-made epidemic prevention and control measures. Estimation of daily new case counts
43
44
45 254 by this short-term model performed very well for European countries, while this model
46
47 255 underestimated the outbreak in the United States.
48

49 256 Although the weather is not suitable for tropical areas, the viral transmitted in natural mode,
50
51
52 257 manifested as good linear correlation between the prediction and the observation ($\rho_{\text{India}} = 0.94$,
53
54 258 $\rho_{\text{Singapore}} = 0.66$, $p < 0.001$, $\rho_{\text{Thailand}} = 0.56$, $p = 0.001$; Fig. 3D), with just relatively small daily
55
56 259 new case counts compared to temperate regions.
57
58
59
60

1
2
3 260 Countries in the southern hemisphere displayed similar pattern as the "controlled" with large
4
5
6 261 overestimation by the model when the confirmed cases increase, leading to not good prediction
7
8 262 performance ($\rho_{\text{Australia}} = 0.79$, $p < 0.001$, $\rho_{\text{South Africa}} = 0.34$, $p = 0.08$; Fig. 3E). It might be due to
9
10 263 the effect of epidemic prevention measures and hot summer weather in these countries.

11 12 264 Long-term simplified model

13
14
15 265 Long-term prediction depends on weather forecast, which generally reports only average
16
17
18 266 temperature. As temperature 14 days ago could predict COVID-19 outbreak as well as
19
20 267 temperature in a short time delay (3~7 days ago), we again performed single-factor regression
21
22 268 modeling in the discovery dataset, taking temperature 14 days ago as an input, assuming a
23
24 269 quadric function (Supplementary Results). This simplified model with average temperature as a
25
26
27 270 weather factor was derived as follows:

$$28$$
$$29$$
$$30$$
$$31$$
$$\text{new case count} = (-0.10 \times T^2 + 1.11 \times T + 46.42) \times \beta \times \text{Existing Confirmed Case Count}$$

32
33 272 where T is temperature in °C, β is a site-related multiplicative constant coefficient, with a default
34
35 273 of 0.006. All parameters take values 14 days before the day new case count is evaluated.

36
37
38 274 With the model, the prediction performance was still good ($\rho = 0.66$ in the replication datasets, p
39
40 275 < 0.001 ; Fig. 3F). The long-term simplified prediction model also showed five prediction-
41
42 276 observation correlation patterns (Fig. 3F), indicating different modes of viral transmission, for the
43
44
45 277 studied areas. This model could directly predict the newly emerging cases 14 days later, and be
46
47 278 used to predict COVID-19 outbreak in the future month by summing up the daily new case count
48
49 279 and combining weather forecast (usually available for the future 15 days).

50 51 52 280 **Discussion**

1
2
3 281 This research discovers nonlinear dose-response relationship for meteorological factors, in
4
5 282 consistency with previous studies[7]. Predictions of COVID-19 outbreak scale by the models
6
7
8 283 were well correlated with the observations around the world, suggesting the importance of
9
10 284 weather in SARS-CoV-2 transmission. Previous studies have implied the spread of many
11
12 285 respiratory infectious diseases, such as influenza, is dependent upon temperature and relative
13
14
15 286 humidity[4]. Recent published papers on preprint servers have reported roles of temperature and
16
17 287 absolute humidity in the COVID-19 transmission, but their conclusions are diverse[6–10]. In
18
19 288 contrast to the findings by Cai et al[10], this study suggests significant impact of mean
20
21
22 289 temperature on the daily new case count, indicating a need for sufficient time delay between
23
24 290 exposure and confirmation for weather to exhibit its effect. In contrary to other two studies[6,7],
25
26 291 this research suggests that there is a relatively not wide temperature and humidity ranges for the
27
28
29 292 pandemic. There is an optimal temperature for SARS-CoV-2 at 6.18 °C, which is colder than that
30
31 293 suggested by Bu et al[9] but in consistency with the estimation by Wang et al[7]; and most areas
32
33 294 with large spread locate in the humidity range of 60% ~ 90%, more humid than Bu et al
34
35 295 suggested[9]. It is of notice that different from other viral respiratory diseases such as
36
37
38 296 influenza[14,15], high relative humidity is better for SARS-CoV-2 to spread, suggesting that a
39
40 297 sufficient amount of droplets in the air to support the suspension of SARS-CoV-2 is more
41
42 298 important for the spread than the effect of dry air on the human immune system. Different from
43
44
45 299 other studies[16], this study also finds significant involvement of wind speed, in a quadric
46
47 300 manner, indicating that mild wind might be more suitable for the virus to suspend in the air. In
48
49 301 addition, the current study discovered that visibility was significantly negatively correlated with
50
51
52 302 new case count and played a more important role in viral spread than humidity did (from
53
54 303 spearman's correlation coefficient comparison). As visibility reflects the amount of particles (e.g.,
55
56 304 dust and air pollutants) in the air while humidity reflects the amount of water in the air, it may

1
2
3 305 indicates that SARS-CoV-2 is more likely to cling to solid particles than droplets. New case
4
5 306 count decreases rapidly when visibility is high than 13 statute miles, indicating that caution
6
7
8 307 should be taken if visibility drops below 10 statute miles.
9

10
11 308 In the prediction model, there is a multiplicative constant coefficient which determines the
12
13 309 strength of the weather coefficient on the epidemic transmission. It seems site-specific, as
14
15 310 adjusting it could make the prediction for one site very close to the observation. This constant
16
17 311 might reflect the influence of a couple of site-specific confounding factors, such as epidemic
18
19
20 312 control measures, sun radiation, and population density. Various degrees of isolation for various
21
22 313 areas around the world lead to different degrees of weather effect. When evaluate the prediction
23
24 314 performance by the short-term model and the long-term model, they both exhibit different
25
26
27 315 prediction-observation correlation patterns (Fig. 3), suggesting that changes in the degree of
28
29 316 epidemic control and isolation policy would lead to deviation from the original prediction and
30
31 317 thus different prediction-observation correlation patterns. Therefore, by plotting the predicted
32
33 318 versus observed new case counts and adjusting the multiplicative constant coefficient (α and β), it
34
35
36 319 would be easy to evaluate the effect of epidemic prevention measures. It is of notice that the
37
38 320 observed case counts dropped dramatically from the predictions for Wuhan seven days after their
39
40 321 shelter hospitals were put in use, suggesting the importance and necessity of building shelter
41
42
43 322 hospitals for strict isolation rather than just home isolation. With the use of shelter hospitals and
44
45 323 very strict isolation measures, the outbreak in one area could be reduced by 52~99% compared to
46
47 324 natural transmission mode. Another thing worth attention is that although the weather in tropical
48
49
50 325 areas like India is not suitable for viral survival and transmission, SARS-CoV-2 still keeps on
51
52 326 spreading in a linear fashion in these areas, with just low growth rate of the outbreak. Therefore,
53
54 327 these tropical areas should still be on the alert against future outbreak of COVID-19.
55
56
57
58
59
60

1
2
3 328 Although those cases with travel history to China or indicated by the World Health Organization
4
5 329 as "imported case only" were excluded in this study to make the world data most likely local
6
7
8 330 transmitted, it was difficult to separate the imported cases from local transmission very well in
9
10 331 practice. It might explain the not excellent correlations of predictions with observations.
11
12 332 Furthermore, the relationship of weather and COVID-19 could be complex, since the human
13
14
15 333 immune system has an innate seasonal rhythm, and the immune system could also be affected by
16
17 334 weather *vice versa*. For example, dry air would reduce the amount of mucus on the airway
18
19 335 mucosa, and thus increase the probability of viral invasion, while wet air would provide droplets
20
21
22 336 for virus to adhere.

23
24 337 There are several limitations of this study. First of all, this prediction model (especially the long-
25
26
27 338 term model) might be more suitable and accurate for temporal areas in spring, autumn, and winter,
28
29 339 as the models were derived using Chinese datasets, mainly in the first three months of 2020. The
30
31 340 prediction became inaccurate and even improper under hot weather (i.e., the predicted values of
32
33
34 341 long-term model become negative when air temperature is higher than 28 °C), which might
35
36 342 explain the obvious bad prediction performance for countries in the southern hemisphere and
37
38
39 343 tropical areas. One explanation for the inaccurate prediction in areas with high temperature could
40
41 344 be that SARS-CoV-2 transmission in these areas was mainly not influenced by weather, but in
42
43 345 another direct transmission way, such as face-to-face contact or spread in gathering crowd.
44
45 346 Second, it seems that the prediction performance drops with the increase in new case count,
46
47
48 347 suggesting that the prediction model might become inaccurate and not suitable for very large new
49
50 348 case count. This could be due to (1) the influence of weather on COVID-19 spread might weaken
51
52 349 when the number of cases increases, while other factors such as social distance become more
53
54
55 350 important at a later stage; (2) there was less data points with large new case count, which might

1
2
3 351 lead to larger variance. Third, the short-term prediction model must use all four meteorological
4
5 352 factors, while these factors are not always available for any one certain area. Fourth, this study
6
7
8 353 included various areas covering a long period into modeling, thus, there were a bunch of variable
9
10 354 confounding factors, such as population mobility and disinfection measures, which were not
11
12 355 controlled and thus could impede the model accuracy. Fifth, as we could only obtain country-
13
14
15 356 level epidemiological data, the corresponding meteorological data were obtained for their capital
16
17 357 cities, leading to not exact pairing of epidemiological data and meteorological data. Sixth, there is
18
19 358 a general lack of data and cases in the current study, since we only collected data covering two
20
21
22 359 and a half months while the pandemic has persisted over seven months up to now.

23 24 25 360 **Conclusion**

26
27
28 361 In summary, this study has found significant correlations with the COVID-19 epidemic trend for
29
30 362 not only temperature and humidity, but also wind speed and visibility. It proposed a
31
32 363 comprehensive model for prediction of COVID-19 outbreak, composed of a short-term version
33
34
35 364 and a long-term version. The short-term version uses the combination of four meteorological
36
37 365 factors as a "weather coefficient" of the existing confirmed case count in the past week and can
38
39 366 be used to predict epidemic situation in the future three days; the short-term version uses average
40
41
42 367 temperature as the "weather coefficient" seven days ago and can predict the outbreak in one
43
44 368 month if combined with weather forecast. This model is easy to use for predicting the COVID-19
45
46 369 outbreak, by substituting weather data in the recent past week and obtaining an estimate of case
47
48 370 count for the future couple of days or month. This model will be very helpful for local
49
50
51 371 governments to make timely policies on epidemic control, for instance, the allocation of medical
52
53 372 equipments such as ventilators and medical resources such as hospitals, beds and health-care
54
55 373 workers, according to the prediction results.

1
2
3 374
4
5
6 375
7
8 376
9
10
11 377
12
13 378
14
15
16 379
17
18 380
19
20 381
21
22
23 382
24
25 383
26
27
28 384
29
30 385
31
32
33 386
34
35
36 387
37
38 388
39
40 389
41
42
43 390
44
45
46 391
47
48
49 392
50
51 393
52
53
54 394
55
56 395
57
58
59
60

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

- 1
2
3 396 3 Chan JFW, Yuan S, Kok KH, *et al.* A familial cluster of pneumonia associated with the
4
5 397 2019 novel coronavirus indicating person-to-person transmission: a study of a family
6
7 cluster. *Lancet* 2020;**395**:514–23. doi:10.1016/S0140-6736(20)30154-9
8 398
9
10
11 399 4 Lowen AC, Mubareka S, Steel J, *et al.* Influenza virus transmission is dependent on
12
13 400 relative humidity and temperature. *PLoS Pathol* 2007;**3**:e151.
14
15
16 401 5 Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions Potential
17
18 402 Implications for Reducing Transmission of COVID-19. *JAMA* 2020;:E1–2.
19
20 403 doi:10.1001/jama.2020.4756
21
22
23 404 6 Proverbio AM, Crotti N, Manfredi M, *et al.* Preliminary evidence that higher temperatures
24
25 405 are associated with lower incidence of COVID-19, for cases reported globally up to 29th
26
27 February 2020. *medRxiv* Published Online First: 2020.
28 406
29
30 407 doi:https://doi.org/10.1101/2020.03.18.20036731
31
32
33 408 7 Wang M, Jiang A, Gong L, *et al.* Temperature Significantly Change COVID-19
34
35 409 Transmission in 429 cities. *Block Caving – A Viable Altern* Published Online First: 2020.
36
37 410 doi:https://doi.org/10.1101/2020.02.22.20025791.
38
39
40 411 8 Luo W, Majumder MS, Liu D, *et al.* The role of absolute humidity on transmission rates of
41
42 412 the COVID-19 outbreak. *medRxiv* 2020;:2020.02.12.20022467.
43
44
45 413 doi:10.1101/2020.02.12.20022467
46
47
48 414 9 Bu J, Peng D-D, Xiao H, *et al.* Analysis of meteorological conditions and prediction of
49
50 415 epidemic trend of 2019-nCoV infection in 2020. *medRxiv* 2020;:2020.02.13.20022715.
51
52 416 doi:10.1101/2020.02.13.20022715
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 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 °C, (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

1
2
3 437 point's delay of confirmation from viral infection are displayed together in one figure, namely,
4
5 438 exposure on the day, three days before, seven days before, 3~7 days before, and 14 days before.
6
7

8
9 439 **Fig. 2.** Scatterplots of confirmed new case counts to the four meteorological variables, (A)
10
11 440 average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter
12
13
14 441 per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric
15
16 442 regression for T, RH, and SPD, and linear regression for VSB are illustrated for each dataset.
17
18 443 Interpolation curves with 95% confidence intervals are shown in shadow. The discovery dataset
19
20
21 444 includes the major outbreak Chinese cities, while the replication datasets included provincial data
22
23 445 in Italy, and national data around the world(except China).
24
25

26 446 **Fig. 3.** The observed daily new case counts versus the predicted values by the short-term model
27
28
29 447 (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five
30
31 448 prediction-observation correlation patterns, which indicates five viral transmission modes: (A)
32
33 449 the "restricted" pattern including the Chinese top affected cities excluding Wuhan; (B) the
34
35 450 "controlled" pattern including early outbreak areas, namely, Iran, Italy, Japan, and Korea, and
36
37
38 451 Chinese Wuhan city; (C) the "natural" pattern including late outbreak European and American
39
40 452 countries, namely, France, Germany, Spain, United Kingdom, and United States; (D) the
41
42 453 "tropical" pattern including tropical countries India, Singapore, and Thailand; (E) the "southern"
43
44
45 454 pattern including countries in the southern hemisphere, Australia and South Africa. Each dot
46
47 455 represents one day. Loess regression (A, B, E) and linear regression (C, D) interpolation curves
48
49 456 are illustrated for each dataset, with 95% confidence intervals showing in shadow. The black
50
51
52 457 solid line represents that the observed values are equal to the predicted ones, and dots closer to
53
54 458 this line means better prediction performance.
55
56
57
58
59

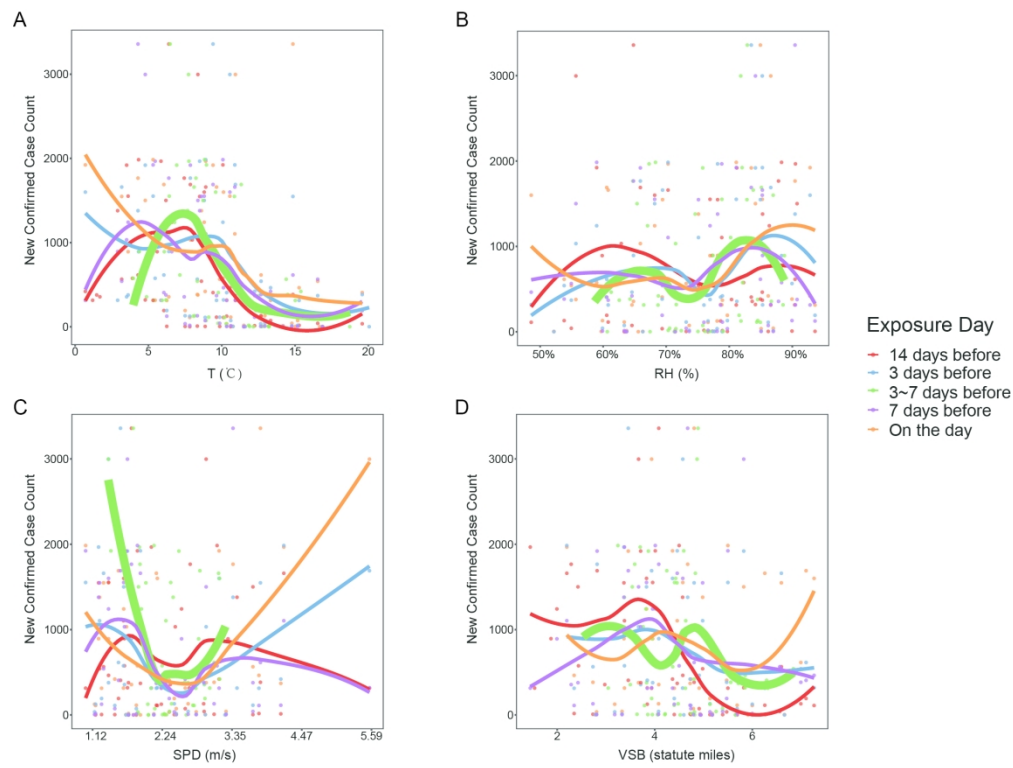


Fig. 1. Loess regression interpolation of confirmed new case count to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for Wuhan city. Five time point's delay of 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)

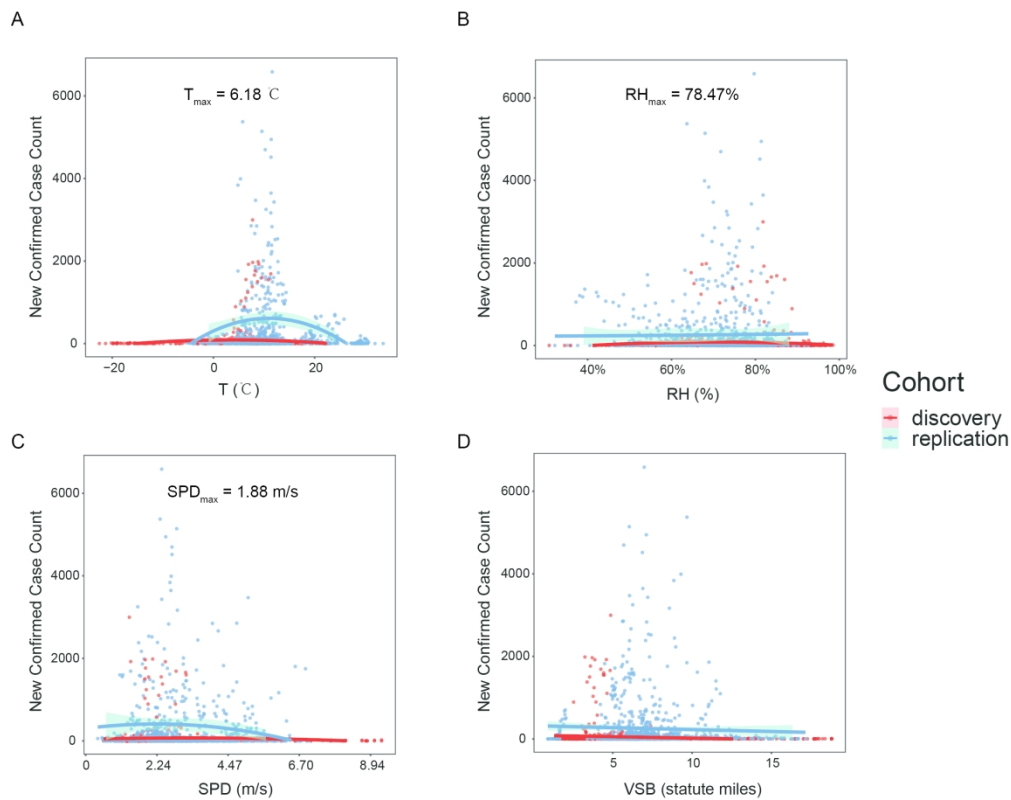


Fig. 2. Scatterplots of confirmed new case counts to the four meteorological variables, (A) average temperature (T) in °C, (B) relative humidity (RH) in %, (C) wind speed (SPD) in meter per second (m/s), (D) visibility (VSB) in statute miles, for all the studied datasets. Quadric regression for T, RH, 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)

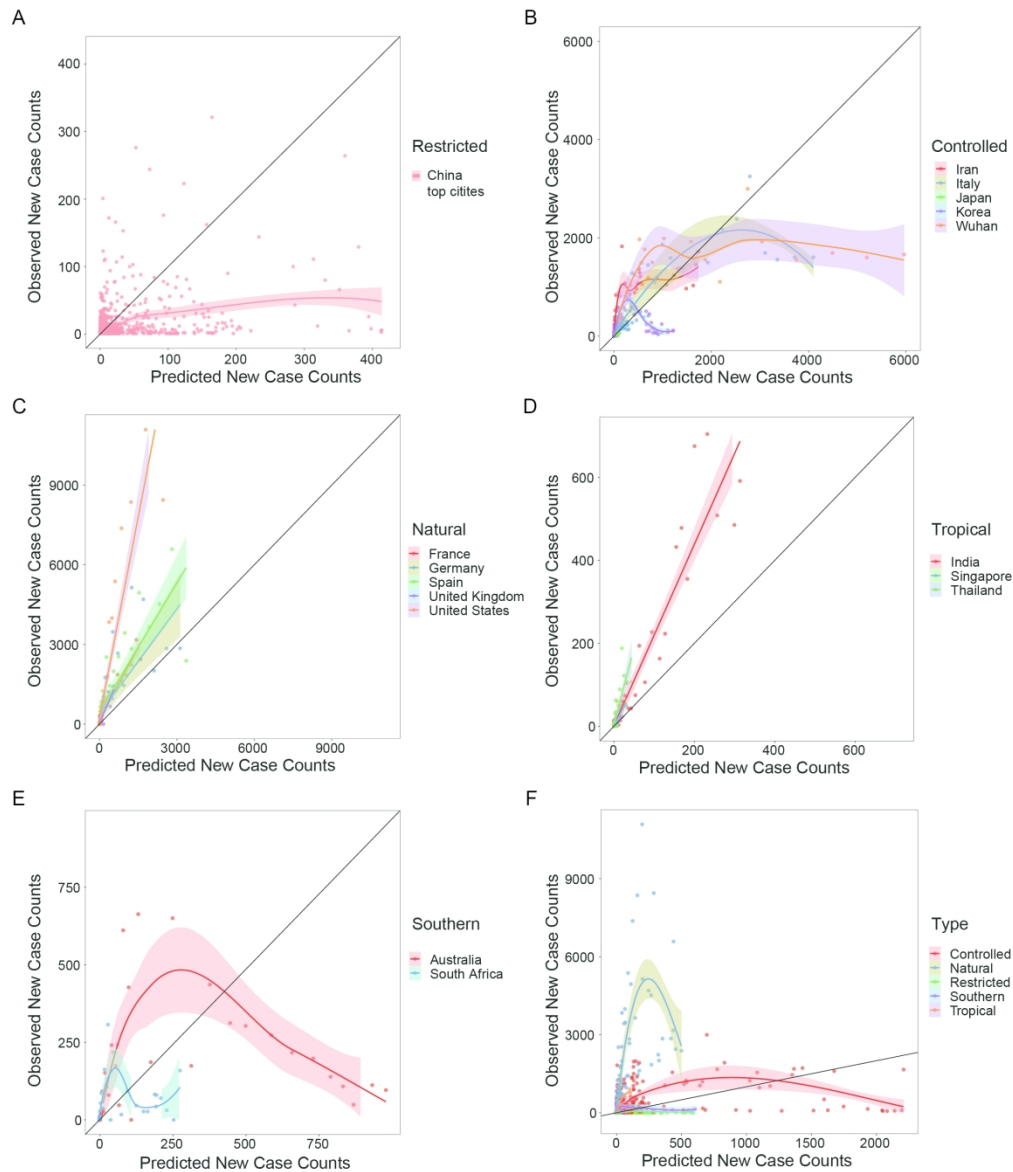


Fig. 3. The observed daily new case counts versus the predicted values by the short-term model (A-E) and the long-term model (F) are illustrated for all the studied areas. The plots exhibit five prediction-observation correlation patterns, which indicates five viral transmission modes: (A) the "restricted" 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)

1 **Supplementary Materials and Methods**

2 Epidemiological data

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

6 For Wuhan city, there was a shortage of test kits at the beginning of the pandemic,
7 which would make confirmed case counts much lower than the actual data, thus, we
8 discarded epidemic data before January 28th, the day when domestic test kits have
9 been approved, produced in large quantities, and were available for Wuhan hospitals.
10 As there was a cut down problem for the existing confirmed case count on February
11 20th for Wuhan, when modeling with the existing confirmed case count, only data
12 before February 20th were used.

13 Weather data

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

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

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

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

1
2
3
4 21 For a city with more than one meteorological stations, the one nearest to the city
5
6 22 center was chosen. For a province with epidemiological data, the meteorological
7
8
9 23 station in the capital city of that province was chosen. For a country with only
10
11
12 24 national wide epidemiological data, weather data were averaged across all the
13
14
15 25 meteorological observatories in the cities where outbreak was officially reported.
16
17 26 Latitude and elevation for the meteorological observatories were also collected.

18 19 20 27 Statistical modeling

21
22 28 Only one city Wuhan was chosen for illustrating the time delay effect because it
23
24
25 29 is the first city to have an outbreak of COVID-19, there was none reported imported
26
27
28 30 cases for Wuhan, which might obscure the correlation between weather and virus
29
30 31 transmission.

32 32 **Supplementary Results**

33 33 Datasets description

34
35 34 Only Chinese cities with monthly confirmed cases over 50 were included in the
36
37
38 35 discovery dataset, which was 60 cities including Wuhan. The confirmed new cases in
39
40
41 36 Wuhan on February 13, 2020, reached 13,436, which was oddly high as the daily
42
43
44 37 confirmed new cases were no larger than 3,000 on all the other dates in Wuhan or in
45
46
47 38 all the other Chinese cities. We suppose that it might be due to abrupt large
48
49
50 39 supplement of virus test kits or data correction on that day. In order to reduce the
51
52
53 40 potential contamination of modeling by this outlier, data on that day were discarded
54
55
56 41 from the subsequent analysis. There were also two oddly large new confirmed case
57
58
59 42 counts for Lombardy, which were discarded from the subsequent analysis. Except the
60

1
2
3
4 43 outliers, the daily confirmed new cases in the discovery dataset ranged from 1 to
5
6 44 2,997, the average temperature ranged $-22.54^{\circ}\text{C} \sim 22.16^{\circ}\text{C}$, the wind speed ranged
7
8
9 45 $0.56 \sim 9.29$ meter per second, visibility ranged $1.3 \sim 18.8$ statute miles, and relative
10
11
12 46 humidity ranged $30.84\% \sim 98.52\%$.

13 14 47 Model selection

15
16
17 48 With the increase of relative humidity, the amount of droplets in the air increases,
18
19 49 leading to more virus load. However, as the air gets humid, human's respiratory tract
20
21
22 50 could better defend virus infection. Thus, the relationship of relative humidity could
23
24
25 51 be complex, not pure linear. Giving comprehensive consideration, we defined the
26
27 52 effect of relative humidity to be quadric. As for visibility, it only affects the amount of
28
29
30 53 particles in the air, which is positively correlated with virus load. Thus, it is most
31
32
33 54 probably to exert its effect linearly.

34
35 55 Although relative humidity and visibility 7 days ago correlated with the
36
37 56 confirmed new case counts best, there was not great loss of model fitting statistics for
38
39
40 57 relative humidity and visibility 3~7 days ago, as compared to the loss between 7 days
41
42
43 58 time delay and 3~7 days time delay for temperature.

44 45 59 Fitted models

46
47
48 60 The fitted single-factor models were as follows:

$$49 \quad \text{New Case Count} = -0.11305 \times T^2 + 1.39819 \times T + 45.11405$$

50
51
52 61 where T is temperature in $^{\circ}\text{C}$.

53
54
55 62 The estimate p-value for constant was < 0.001 . The extremum was $-1.39819/$
56
57
58 63 $(2 \times (-0.11305)) = 6.183945 \text{ }^{\circ}\text{C}$.

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

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

New Case Count

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

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:

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

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:

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

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

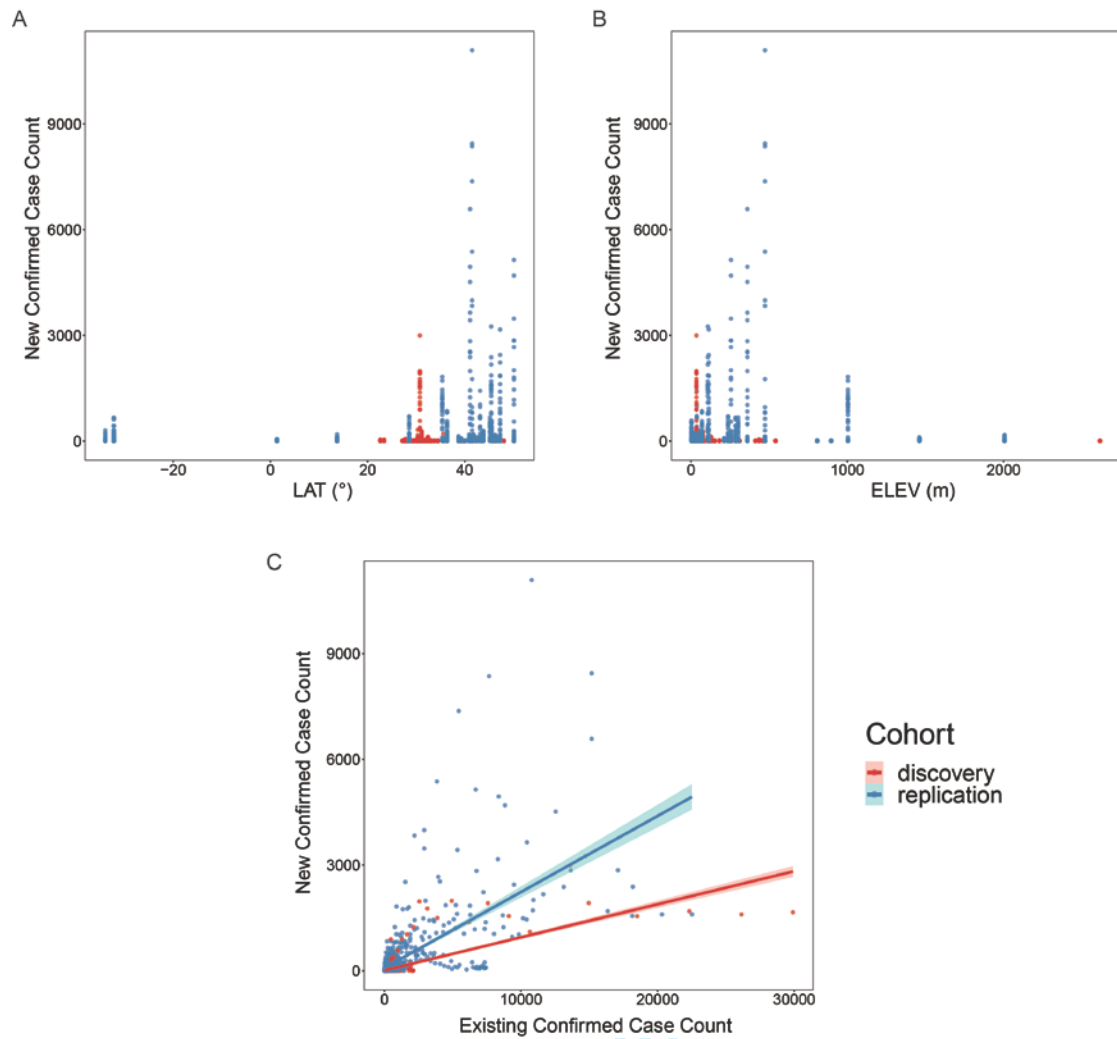
93 Note: sigma, estimated standard error of the residuals; finTol, the achieved convergence tolerance; logLik, the
 94 log-likelihood of the model; AIC, Akaike's Information Criterion for the model; BIC, Bayesian Information
 95 Criterion for the model; deviance, deviance of the model; Corr, Spearman's 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.