Title: The Challenge of Using Epidemiological Case Count Data: The Example of Confirmed COVID-19 Cases and the Weather

Short title: Challenges of using COVID-19 case count data

One Sentence Summary: Measurement issues in the currently available data on confirmed COVID-19 cases undermine the analyses of the drivers of the spread of the disease.

Authors: Francois Cohen[i], Moritz Schwarz[i,ii], Sihan Li[iii], Yangsiyu Lu[i] and Anant Jani[iv]

Abstract: The publicly available data on COVID-19 cases provides an opportunity to better understand this new disease. However, strong attention needs to be paid to the limitations of the data to avoid making inaccurate conclusions. This article, which focuses on the relationship between the weather and COVID-19, raises the concern that the same factors influencing the spread of the disease might also affect the number of tests performed and who gets tested. For example, weather conditions impact the prevalence of respiratory diseases with symptoms similar to COVID-19, and this will likely influence the number of tests performed. This general limitation could severely undermine any similar analysis using existing COVID-19 data or similar epidemiological data, which could, therefore, mislead decision-makers on questions of great policy relevance.

[i] Smith School of Enterprise and the Environment, University of Oxford; and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford.

[ii] Climate Econometrics, Nuffield College, Oxford.

[iii] Environmental Change Institute, University of Oxford.

[iv] Nuffield Department of Primary Care Health Sciences, University of Oxford.

Main text: The publicly available datasets on confirmed COVID-19[v] cases and deaths provide a key opportunity to better understand the drivers of the pandemic. Research using these datasets has been growing at a very fast pace (see an indicative list of references in supplementary material 1). However, little attention has been paid to the reliability of this type of epidemiological data to make statistical inferences.

Our initial aim was to produce a detailed statistical analysis of the relationship between weather conditions and the spread of COVID-19. This question has attracted significant attention from the media (e.g. 1, 2) and the research community (e.g. 3, 4; see a wider list in supplementary material 1) due to the possibility that summer weather might slow the spread of the virus. After going through all the steps of such an analysis, we reached the unexpected conclusion that the limitations of the available COVID-19 data are so severe that we would not be able to make any reliable statistical inference. This applies, for example, to the data provided by the John Hopkins University (5) and the data collated by Xu et al. (2020) (6).

This is a concerning yet very important finding considering that such data is being widely used to make crucial policy decisions on a wide range of topics. Since invalid causal inferences could be made with the publicly available COVID-19 data, and then enter policy-making discourse, there is an urgent need to raise awareness among the scientific community and decision-makers regarding the limitations of the information at their disposal. The elements discussed in this paper are also likely to be applicable to other epidemiological datasets obtained with insufficient testing and monitoring, either during exceptional epidemics or seasonal outbreaks.

***

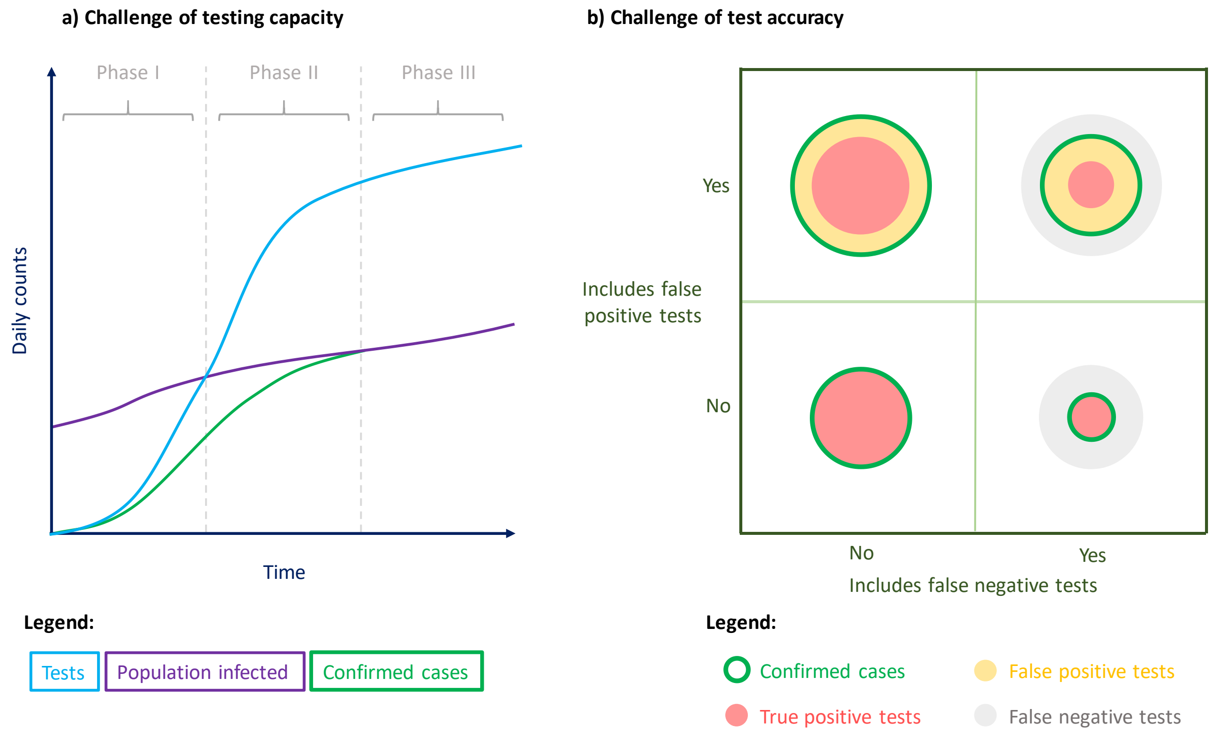
Several challenges could undermine any causal statistical analysis of the influence of a potential determinant, such as the weather, on the spread of COVID-19. To start, confounding variables are likely

[v] In this article we follow Xu et al. (2020) who define COVID-19 cases as individuals for whom SARS-CoV-2 has been detected using rt-PCR.

to pose a significant problem: many factors (e.g. changes in policy or social interactions) are simultaneously influencing how the disease spreads.

In addition, significant challenges come from the limitations of the COVID-19 case count data itself. Firstly, testing capacity has been a major issue in most countries. Before March 1st, 2020, very few countries had sufficient testing capacity. By April 30th, 2020, high-income countries had significantly increased their testing capacity, but testing remained critically infrequent in most low- and middle-income countries.[vi] **Figure 1, panel a** illustrates the effect that insufficient testing capacity has on the number of confirmed cases. It distinguishes between three phases of limited (I), intermediate (II) and widespread (III) testing. In Phases I and II, there is a risk that the number of confirmed cases depends more on the number of tests available than on the actual number of people who have COVID-19, questioning the validity of any analysis relying too heavily on this data.

Figure 1: Difference between actual COVID-19 cases in population and reported confirmed COVID-19 cases. Confirmed COVID-19 cases (green) represent the number of people tested with a positive test result. They include false positive and exclude false negative tests. The circles of panel b represent the size of the populations with true positive, false negative or false positive tests. Quantities in the y-axis of panel a, as well as the size of the circles in panel b, do not represent any true value or proportion.



[vi] Figures on testing are available in Our World In Data (accessed on May 1st, 2020): https://ourworldindata.org/coronavirus

Moreover, there have been numerous concerns regarding the accuracy of the COVID-19 tests performed so far (7, 8, 9, 10). **Figure 1, panel b** illustrates the effects of both false-negative and false-positive test results on the number of confirmed cases. False-negative results would imply that the number of confirmed COVID-19 cases is underestimated. False-positive results would imply that people who do not have COVID-19 are included in the number of confirmed COVID-19 cases. Concerns regarding test accuracy create an additional problem of measurement that might affect statistical analyses.

The two above-mentioned challenges are inherent to all current datasets of COVID-19 confirmed case count and mortality. In addition, specific datasets may have imperfect geographical or time coverage.

To look at the impact of the weather on the spread of COVID-19, we initially used a well-established approach, similar to the ones used previously to look at the impact of the weather on other diseases (e.g. 11, 12) (see details in supplementary material 2). However, the fundamental measurement issues associated with the COVID-19 case count data cannot be corrected by statistical techniques, as we outline below.

The main problem is that the weather could be influencing the number of tests carried out and the segment of the population tested. For example, other respiratory diseases are often similar to COVID-19 in their symptoms (e.g. 13) and are more common during cold weather (e.g. 11, 12), which could influence the number of tests performed on people displaying symptoms of respiratory infection. Therefore, even if the model correctly identified the impact of the weather on COVID-19 case counts, it could not distinguish between the impact of the weather on the spread of the disease and its impact on testing. **Table 1** provides a non-exhaustive list of elements that could undermine any analysis of the impact of the weather on the spread of COVID-19 using data on confirmed cases. The evidence suggests that the weather may correlate with the number of tests conducted and who gets tested. We have not been able to find any specific COVID-19 related evidence that the weather would correlate with test accuracy (e.g. the weather affecting the nasopharyngeal or oropharyngeal swabs used in the PCR analysis), even though this could be possible.

Other points of concern include: the fact that there may be indirect effects of weather conditions on other factors that could have an impact on the spread of COVID-19 (such as social interactions or air pollution); the heterogeneity of impacts across populations and subgroups within a population; or the fact that some people may have travelled and therefore been infected in a different place from where the cases are reported.

Table 1: Non-exhaustive list of reasons why weather conditions could affect the number of COVID-19 tests carried out and who gets tested.

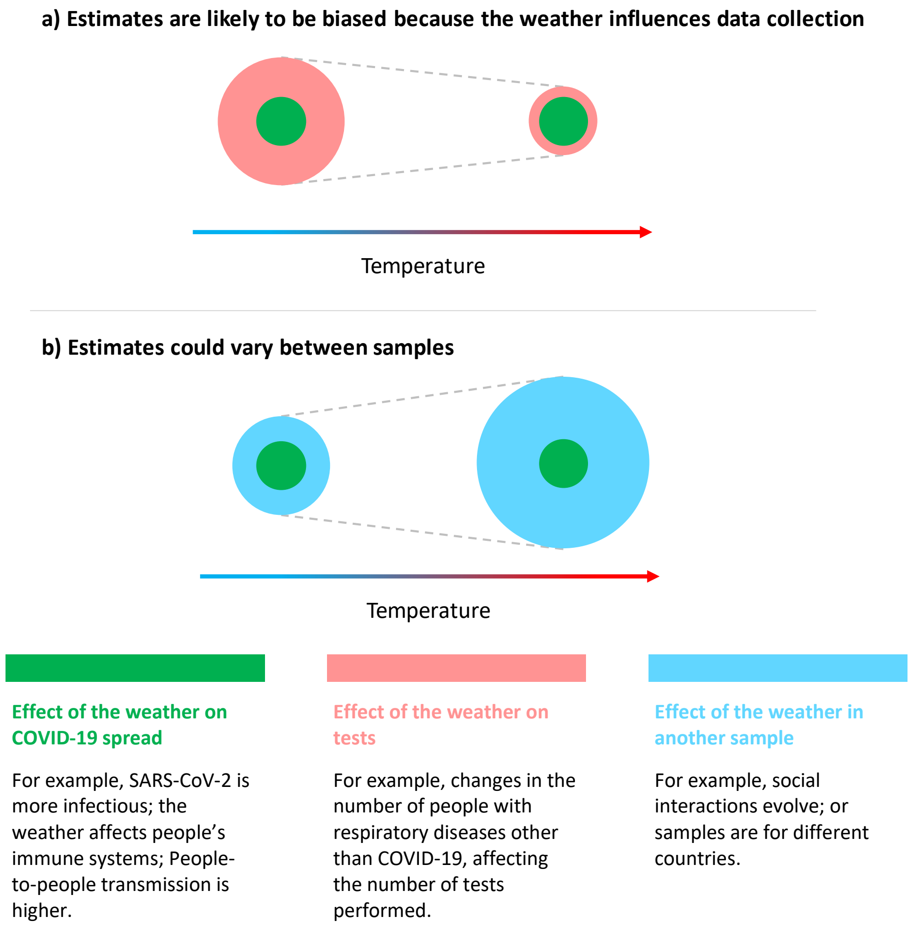
Potential reason	Potential implication
Unrelated respiratory diseases are weather sensitive (e.g. 11, 12) and can be confused with COVID-19 (e.g. 7, 14).	- More patients with symptoms of unrelated respiratory diseases could be tested during cold weather. - The prevalence of other weather-sensitive respiratory diseases might make false-positive results more likely, especially if only radiographic imaging is used, since it is possible to confuse these diseases for COVID-19 (e.g. 7, 14).
The incidence of other pathologies (e.g. cardiovascular diseases) is influenced by the weather (e.g. 11, 12).	- Hospital capacity, and the workload of medical staff and testing structures is affected by weather conditions, with potential implications on the number of tests conducted. - At-risk individuals suffering from unrelated conditions are more likely to be tested for COVID-19, even if they only have mild symptoms for COVID-19.
People may be more inclined to seek medical attention depending on the weather (e.g. 15).	- Due to weather conditions, people may or may not decide to seek medical attention, affecting the number of patients going to the hospital with COVID-19, and the workload of medical staff.

We ran our model (as detailed in the supplementary material 2) and provide results and robustness checks in supplementary material 3. The model would technically suggest a negative correlation (e.g. colder days would be associated with more confirmed COVID-19 cases, and hotter days with fewer cases). Yet, these results could be highly misleading since these estimates are likely to be substantially biased because of the aforementioned reasons.

Figure 2, panel a, provides an illustration of how we could have obtained a negative correlation even if temperature had no impact or a positive impact on the spread of COVID-19 in our sample. The total number of estimated cases is given by the size of the circles as a function of temperature (x-axis). The circles in green correspond to the effects we are interested in – those that explain the influence of temperature on the spread of COVID-19. If temperature has no effect on the spread of COVID-19, then the green circles should be the same size at low and high temperatures. The pink circles represent the possible effect of temperature on testing (as reported in **Table 1**) under the illustrative assumption that high temperatures reduce testing frequency. In this case, the overall result is a negative correlation between temperature and confirmed COVID-19 cases, even if temperature has no effect on the spread of the disease. In practice, we naturally do not know the direction of the bias caused by the effect of temperature on testing when using standard statistical methods. There is also no way for us to evaluate the contribution of each of these effects (green or pink) in our estimate. We arrive at the final size of the circles and cannot be sure if the association that we are interested in is either negative, null or positive.

Figure 2, panel b, focuses on the risk that effects could be different across different samples. The circles in blue capture other underlying factors that are influenced by temperature (such as acclimatisation or the level of social interactions in the population), as well as other socioeconomic factors (such as the demographic characteristics of a population). These factors could be radically different in different regions but may also evolve over time (e.g. between winter and summer seasons).

Figure 2: Effects potentially captured by our estimate. The size of the circles represents the estimated number of cases at different temperatures. These are examples that do not correspond to actual data. In these examples, we assume no correlation between temperature and the effects in green (see legend below), a negative correlation with the effects in pink (example 1) and a positive correlation with those in blue.



a) Estimates are likely to be biased because the weather influences data collection

b) Estimates could vary between samples

🟩 **Effect of the weather on COVID-19 spread**

For example, SARS-CoV-2 is more infectious; the weather affects people's immune systems; People-to-people transmission is higher.

🟥 **Effect of the weather on tests**

For example, changes in the number of people with respiratory diseases other than COVID-19, affecting the number of tests performed.

🟦 **Effect of the weather in another sample**

For example, social interactions evolve; or samples are for different countries.

There are strong reasons to be concerned with the scenario illustrated in **Figure 2, panel b**. In our sample, for example, we only have data from the start of the pandemic until end of April 2020; some countries (e.g. China) may be over-represented in the dataset; and the average daily temperature is relatively low at 10.5°C. Furthermore, many countries have implemented a stringent containment policy during the period covered by the sample. Containment policies may have heightened (or lowered) the sensitivity of the spread of the disease to the weather because social interactions are limited. We are not able to observe how the impact of the weather on COVID-19 might change at different gradients of social interaction. Finally, our estimate is based on small, observed changes in temperatures, and not on radical increases or reductions in temperatures. The spread of COVID-19 may respond differently to large variations in temperature, e.g. by 5°C or 10°C across seasons, making seasonal predictions even more unreliable.

Strong precautions need to be taken before using COVID-19 case count datasets for inference. The results of our model using existing COVID-19 data would seemingly imply a negative association between temperature and confirmed COVID-19 cases. Any projection of COVID-19 cases with such estimates could conclude that, during the upcoming months of June to September 2020, Southern Hemisphere countries would be exposed to higher risks of COVID-19 spread, and Northern Hemisphere countries to lower risks.[vii] These types of unsubstantiated results could be used as a misinformed justification for an early relaxation of effective social distancing measures in the Northern Hemisphere.

These findings have equally strong implications for statistical analyses focusing on other questions that rely on COVID-19 confirmed case count and/or mortality count data. Even though the exact nature of the effects may change, such studies are also at risk of capturing the effect that their parameters of interest have on tests and test results. For example, studies interested in the effect of containment policies may have to consider that these policies substantially affect testing because they change the awareness of the disease in the population, political demands for more testing or the risk of contracting other respiratory diseases. Other studies may also produce estimates that are very specific to the current circumstances in the development of the pandemic and are, therefore, not suitable to use for forecasts of what could happen in the coming months.

In the medium term, more reliable data needs to be gathered, for example through experimental studies that randomly test a sample of the population for COVID-19. In the short term, we are in a situation of

[vii] We performed such a projection to confirm this point (see **supplementary material 4)**.

fundamental uncertainty about how different factors affect or are affected by the widespread societal changes we see with the COVID-19 pandemic. Therefore, scientists, policymakers, journalists and the general public need to be very cautious when discussing how the spread of COVID-19 correlates with the weather or any other factor.

In the long term, this paper suggests that more attention should be given to how epidemiological data is recorded and used during exceptional epidemics and seasonal outbreaks, since insufficient testing and monitoring can undermine essential statistical analyses. This article calls for the complementary use of different methods for data collection, such as random testing in samples of the population.

REFERENCES

[1] Ravilious, K. (2020). Will spring slow spread of coronavirus in northern hemisphere? *The Guardian*, 11th March 2020.

[2] Clive Cookson. (2020). Scientists hopeful warmer weather can slow spread of coronavirus. *The Financial Times*, 25th March 2020.

[3] Araujo, M. B., & Naimi, B. (2020). Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. medRxiv.

[4] Carleton, Tamma et al. “Ultraviolet radiation decreases COVID-19 growth rates: Global causal estimates and seasonal implications” (2020). Unpublished draft available at: http://www.kylemeng.com/ [Accessed 29th April, 2020]

[5] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." The Lancet infectious diseases (2020).

[6] Xu, Bo, et al. "Epidemiological data from the COVID-19 outbreak, real-time case information." *Scientific Data* 7.1 (2020): 1-6.

[7] Ai, Tao, et al. "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." *Radiology* (2020): 200642.

[8] Apostolopoulos, Ioannis D., and Tzani A. Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks." *Physical and Engineering Sciences in Medicine* (2020): 1.

[9] Hu, Emily. "COVID-19 Testing: Challenges, Limitations and Suggestions for Improvement." (2020).

[10] Hall, Lawrence O., et al. "Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset." *arXiv preprint arXiv:2004.02060* (2020).

[11] Deschenes, Olivier, and Enrico Moretti. "Extreme weather events, mortality, and migration." The Review of Economics and Statistics 91.4 (2009): 659-681.

[12] Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., ... & Leone, M. (2015). Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet, 386(9991), 369-375.

[13] WHO. “Q&A: Similarities and Differences – COVID-19 and Influenza”. (2020) Consulted on 29th April, 2020: https://www.who.int/news-room/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza

[14] Chen, Sin-Guang, et al. "Use of Radiographic Features in COVID-19 Diagnosis: Challenges and Perspectives." Journal of the Chinese Medical Association: JCMA (2020).

[15] Norris, John B., et al. "An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics." Decision Support Systems 57 (2014): 428-443.

Acknowledgements. For useful comments, we thank, among others, Eric Beinhocker, Cristina Castro, Antoine Dechezlepretre, Doyne Farmer, Matthieu Glachant, Jeongmin Thomas Han, David Hendry, Cameron Hepburn, José García Quevedo, Daniel Montolio, James Parkhurst, Rafael Perera, Felix Pretis, Andra Tartakowsky, Elisabet Viladecans, Judit Vall Castelo and the team of the Future of Cooling programme at the Oxford Martin School. For technical IT assistance, we thank David Ford. For copyediting, we thank Jack Smith. **Authors contributions:** Cohen is the first author. He had the original idea, wrote most of the paper and the code to produce the econometric analysis. He also coordinated the team. Jani ensured the material was consistent with the epidemiological evidence. Li produced the required climate data for the statistical analysis. Lu helped on literature review, on coding the econometric analysis and on producing the tables. Schwarz helped on data coding and matching and created the projections. All authors contributed to the text. **Funding sources**: Oxford Martin School (Cohen and Jani), The Nature Conservancy (Cohen, Li and Lu), Newton Fund (Li), Clarendon Fund and Robertson Foundation (Schwarz), China Scholarship Council and University of Oxford (Lu). **Competing interests:** the authors have no conflict of interest. **Data and materials availability:** all data and software are publicly available at https://github.com/moritzpschwarz/COVID-19-weather-Oxford.

Supplementary Materials

1. References from the emerging literature on the spread of COVID-19

2. Data and methods to correlate weather conditions to confirmed COVID-19 cases

3. Main results and robustness checks

4. Information on projections

Table A1 – A6

Fig A1 – A2

The Challenge of Using Epidemiological Case Count Data: The Example of Confirmed COVID-19 Cases and the Weather

Francois Cohen[i], Moritz Schwarz[i,ii], Sihan Li[iii], Yangsiyu Lu[i] and Anant Jani[iv]

This version: May 9th, 2020; first version: April 1st, 2020; data update: April 30th, 2020.

Supplementary material

1. References from the emerging literature on the spread of COVID-19

The publicly available datasets on confirmed COVID-19 cases and deaths have been used extensively by the scientific community to understand the spread of the new disease. For example, by May 8th, only two months and a half after its release, the paper by Dong et al. (2020) (16), which presents the publicly available data from the John Hopkins University, was already cited by 383 working papers and published articles (according to Google Scholar).

Below, we provide a concise list of working papers and articles that have recently emerged to look at the spread of COVID-19 and understand its determinants, relying on case count data from the John Hopkins University or similar data sources, or simply commenting on these data sources. This list does not intend to be exhaustive, especially since new work is being produced every day. We chose to include unpublished preprints in this list to demonstrate the extent to which publicly available COVID-19 data is being used by researchers. The sole purpose of this list is to show how important it is to discuss the limitations of the existing data on confirmed COVID-19 cases and deaths, considering its widespread use and policy relevance.

The list includes five sections:

A. Impact of the weather on COVID-19

[i] Smith School of Enterprise and the Environment, University of Oxford; and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford.
[ii] Climate Econometrics, Nuffield College, Oxford
[iii] Environmental Change Institute, University of Oxford.
[iv] Nuffield Department of Primary Care Health Sciences, University of Oxford.

B. Impact of air pollution on COVID-19

C. Impact of non-pharmaceutical interventions

D. Other drivers or impacts of COVID-19

E. Forecasts of the spread of COVID-19 and analyses of mortality rates

A- Impact of the weather on COVID-19

Ma, Yueling, et al. "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China." *Science of The Total Environment* (2020): 138226.

Araujo, M. B., & Naimi, B. (2020). Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. medRxiv.

Bannister-Tyrrell, M., Meyer, A., Faverjon, C., & Cameron, A. (2020). Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020. medRxiv.

Bukhari, Q., & Jameel, Y. (2020). Will Coronavirus Pandemic Diminish by Summer? Available at SSRN 3556998.

Carleton, Tamma et al. “Ultraviolet radiation decreases COVID-19 growth rates: Global causal estimates and seasonal implications” (2020). Unpublished draft available at: http://www.kylemeng.com/ [Accessed 29th April, 2020]

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., ... & Zhu, X. (2020). Roles of meteorological conditions in COVID-19 transmission on a worldwide scale. medRxiv.

Chiyomaru, Katsumi, and Kazuhiro Takemoto. "Global COVID-19 transmission rate is influenced by precipitation seasonality and the speed of climate temperature warming." *medRxiv* (2020).

Ficetola, Gentile Francesco, and Diego Rubolini. "Climate affects global patterns of COVID-19 early outbreak dynamics." *medRxiv* (2020).

Gupta, D. (2020). Effect of Ambient Temperature on COVID-19 Infection Rate. Available at SSRN 3558470.

Luo, Chao, et al. "Possible Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in a Public Bath Center in Huai’an, Jiangsu Province, China." JAMA Network Open 3.3 (2020): e204583-e204583.

Oliveiros, B., Caramelo, L., Ferreira, N. C., & Caramelo, F. (2020). Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. medRxiv.

Poirier, C., Luo, W., Majumder, M. S., Liu, D., Mandl, K., Mooring, T., & Santillana, M. (2020). The Role of Environmental Factors on Transmission Rates of the COVID-19 Outbreak: An Initial Assessment in Two Spatial Scales. Available at SSRN 3552677.

Qu, Guangbo, et al. "An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19)." (2020).

Wang, J., Tang, K., Feng, K., & Lv, W. (2020). High Temperature and High Humidity Reduce the Transmission of COVID-19. Available at SSRN 3551767.

B- Impact of air pollution on COVID-19

Coccia, Mario. "Diffusion of COVID-19 Outbreaks: The Interaction between Air Pollution-to-Human and Human-to-Human Transmission Dynamics in Hinterland Regions with Cold Weather and Low Average Wind Speed." (2020).

Conticini, Edoardo, Bruno Frediani, and Dario Caro. "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?" Environmental Pollution (2020): 114465.

Hale, T., Petherick, A., Phillips, T., & Webster, S. (2020). Variation in government responses to COVID-19. Blavatnik School of Government Working Paper, 31.

Han, Yang, et al. "The Effects of Outdoor Air Pollution Concentrations and Lockdowns on Covid-19 Infections in Wuhan and Other Provincial Capitals in China." (2020).

Pansini, Riccardo, and Davide Fornacca. "COVID-19 higher induced mortality in Chinese regions with lower air quality."

Wu, Xiao, et al. "Exposure to air pollution and COVID-19 mortality in the United States." medRxiv (2020).

C- Impact of non-pharmaceutical interventions

Fang, Yaqing, Yiting Nie, and Marshare Penny. "Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis." *Journal of medical virology* 92.6 (2020): 645-659.

Maier, Benjamin F., and Dirk Brockmann. "Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China." *Science* (2020).

Prem, Kiesha, et al. "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study." *The Lancet Public Health* (2020).

Wilder-Smith, Annelies, Calvin J. Chiew, and Vernon J. Lee. "Can we contain the COVID-19 outbreak with the same measures as for SARS?." *The Lancet Infectious Diseases* (2020).

Banholzer, Nicolas, et al. "Estimating the impact of non-pharmaceutical interventions on documented infections with COVID-19: A cross-country analysis." medRxiv (2020).

Bianconi, Antonio, et al. "Efficiency of Covid-19 Containment by Measuring Time Dependent Doubling Time." arXiv preprint arXiv:2004.04604 (2020).

Chen, You, et al. "Modeling COVID-19 Growing Trends to Reveal the Differences in the Effectiveness of Non-Pharmaceutical Interventions among Countries in the World." medRxiv (2020).

Han, Yang, et al. "The Effects of Outdoor Air Pollution Concentrations and Lockdowns on Covid-19 Infections in Wuhan and Other Provincial Capitals in China." (2020).

Haushofer, J., & Metcalf, C. J. E. (2020). Evaluation of non-pharmaceutical interventions is needed to mitigate the COVID-19 pandemic.

Haushofer, Johannes, and C. Jessica E. Metcalf. "Evaluation of non-pharmaceutical interventions is needed to mitigate the COVID-19 pandemic." (2020).

Pan, Sheng, et al. "Early-Stage Government Control of COVID-19 Limits Spreading and Reduces Fatality by Narrowing the Outbreak-Phase Duration: An Epidemiological Study." Available at SSRN 3564425 (2020).

Piguillem, Facundo, and Liyan Shi. *The optimal covid-19 quarantine and testing policies*. No. 2004. Einaudi Institute for Economics and Finance (EIEF), 2020.

Qiu, Yun, Xi Chen, and Wei Shi. "Impacts of Social and Economic Factors on the Transmission of Coronavirus Disease 2019 (COVID-19) in China." (2020).

Shet, Anita, et al. "Differential COVID-19-attributable mortality and BCG vaccine use in countries." medRxiv (2020).

Tarrataca, L., et al. "Flattening the curves: on-off lock-down strategies for COVID-19 with an application to Brazi." arXiv preprint arXiv:2004.06916 (2020).

Wells, Chad R., et al. "Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak." Proceedings of the National Academy of Sciences 117.13 (2020): 7504-7509.

Yuan, Hsiang-Yu, et al. "Effectiveness of quarantine measure on transmission dynamics of COVID-19 in Hong Kong." *medRxiv* (2020).

D- Other drivers or impacts of COVID-19

Ajzenman, N., Cavalcanti, T., & Da Mata, D. (2020). More than Words: Leaders' Speech and Risky Behavior During a Pandemic. Available at SSRN 3582908.

Bursztyn, Leonardo, et al. "Misinformation during a pandemic." University of Chicago, Becker Friedman Institute for Economics Working Paper 2020-44 (2020).

Fetzer, Thiemo, et al. "Perceptions of Coronavirus Mortality and Contagiousness Weaken Economic Sentiment." arXiv preprint arXiv:2003.03848 (2020).

Kuchler, Theresa, Dominic Russel, and Johannes Stroebel. The geographic spread of COVID-19 correlates with structure of social networks as measured by Facebook. No. w26990. National Bureau of Economic Research, 2020.

E- Forecasts of the spread of COVID-19 and analyses of mortality rates

Onder, Graziano, Giovanni Rezza, and Silvio Brusaferro. "Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy." *Jama* (2020).

Yuan, Jing, et al. "Monitoring transmissibility and mortality of COVID-19 in Europe." International Journal of Infectious Diseases (2020).

Anastassopoulou, Cleo, et al. "Data-based analysis, modelling and forecasting of the COVID-19 outbreak." PloS one 15.3 (2020): e0230405.

Bertozzi, Andrea L., et al. "The challenges of modeling and forecasting the spread of COVID-19." arXiv preprint arXiv:2004.04741 (2020).

Chang, Sheryl L., et al. "Modelling transmission and control of the COVID-19 pandemic in Australia." arXiv preprint arXiv:2003.10218 (2020).

Donsimoni, Jean Roch, et al. "Projecting the Spread of COVID-19 for Germany." Wirtschaftsdienst 100 (2020): 272-276.

Fanelli, Duccio, and Francesco Piazza. "Analysis and forecast of COVID-19 spreading in China, Italy and France." Chaos, Solitons & Fractals 134 (2020): 109761.

Kumar, Santosh. "A Normalized Mortality Rate Showed the Diverse Severity of COVID-19 in the World." (2020).

Ng, Kok Yew, and Meei Mei Gui. "COVID-19: Development of A Robust Mathematical Model and Simulation Package with Consideration for Ageing Population and Time Delay for Control Action and Resusceptibility." arXiv preprint arXiv:2004.01974 (2020).

Pandey, Gaurav, et al. "SEIR and Regression Model based COVID-19 outbreak predictions in India." arXiv preprint arXiv:2004.00958 (2020).

Paul, Abhijit, Samrat Chatterjee, and Nandadulal Bairagi. "Prediction on Covid-19 epidemic for different countries: Focusing on South Asia under various precautionary measures." medRxiv (2020).

Perc, Matjaž, et al. "Forecasting COVID-19." Frontiers in Physics 8 (2020): 127.

Razzak, Junaid A., et al. "Estimating COVID-19 infections in hospital workers in the United States." medRxiv (2020).

Roques, Lionel, et al. "Using early data to estimate the actual infection fatality ratio from COVID-19 in France." medRxiv (2020).

2. Data and methods to correlate weather conditions to confirmed COVID-19 cases

Data

We linked the real-time geo-referenced epidemiological data from Xu et al. (2020) (17) to meteorological data from the 5th generation of European Centre for Medium-Range Weather Forecasts atmospheric reanalyses over the globe (ECMWF-ERA5)[v]. The weather data provides consistent data with high spatial (~0.25 degrees) and temporal (hourly) resolutions. We use daily averages and consider mean, maximum, and minimum temperatures as well as total precipitation and relative humidity (calculated using temperature and dewpoint temperature). We provide the summary statistics for the meteorological data in **Appendix Table A1**.

The dataset from Xu et al. (2020) constitutes a rigorous, multinational effort to provide statistics on COVID-19 at subnational level. It provides information on the location (longitude and the latitude) of confirmed COVID-19 cases with the highest resolution available globally, allowing researchers to produce analyses at sub-national level (18). In our case, this data source allows us to control for local climates and seasonality. This dataset includes data from January onwards and is updated on a regular basis.

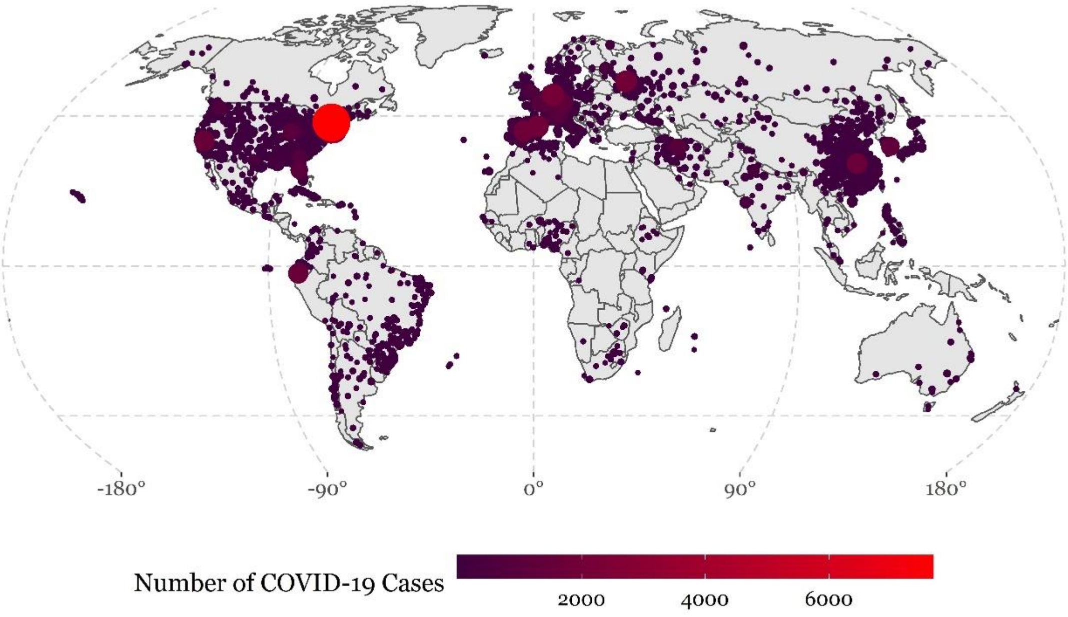
Appendix Figure A1 provides the distribution of the confirmed cases in the dataset. We observe confirmed COVID-19 cases in both Hemispheres, in cold and hot weather. New York (the red dot in **Figure 1**) records the highest number of confirmed cases in an area.

Appendix Table A1: Summary statistics of the meteorological data after it is matched to COVID-19 data

Variable	Mean	Std. Dev.	Min	Max
Avg. temperature (°C)	10.5	9.5	-31.6	38.6
Min. temperature (°C)	5.9	10.0	-36.2	32.6
Max. temperature (°C)	14.7	9.3	-29.0	43.6
Total Precipitation (mm)	2.8	6.9	0.0	293.7
Relative humidity (%)	69.5	15.8	7.4	100.0

[v] Downloadable from Copernicus Climate Change Service: https://cds.climate.copernicus.eu/

Figure A1: Confirmed cases in Xu et al. (2020). Data retrieved on April 30th, 2020. Cases reported with geographical information at country level only were excluded.



Method

We use the following econometric model to look at the correlation between the weather and confirmed COVID-19 cases:

(1) $$\ln(C_{i,t}) - \ln(C_{i,t-1}) = \sum_{x=0}^{X} a_x . W_{i,t-x} + \mu_{c,t} + n_{i,w} + \epsilon_{i,t}$$

In Eq. (1), $\ln(C_{i,t})$ is the logarithm of the total number of confirmed cases of COVID-19 observed in area i on day t. The dependent variable is therefore the first difference of this logarithm. This transformation allows us to scale any change in confirmed cases in relative terms based on the level of confirmed cases the day before. This is to account for the fact that infections can only be proportional to the number of people already infected in an area.

$W_{i,t-x}$ is a matrix of weather-related variables that includes information on the weather at time $t-x$. We include the lagged values of these weather variables (until $t-X$) to capture the correlation between the weather of the previous days and confirmed cases. $W_{i,t-x}$ is modulable. We run our main specifications with average temperature, but we also separate average temperatures into minimum and maximum temperatures and use humidity and precipitation as controls in robustness checks.

We use different values for the total number of lags (X). Our main models use 15 lags ($X = 15$) and therefore covers 16 days. This should cover most cases for the maximum time reported for the incubation of the disease (two weeks maximum) and its detection through testing. We report alternative models with less lags, and more lags, in **supplementary material 3**.

$\mu_{c,t}$ are country by day fixed effects (e.g. the UK on March 25th, 2020). They therefore control for national factors which may vary from day to day and influence the spread of the disease. $n_{i,w}$ is an area-specific (e.g. regions or cities) fixed effect that is assumed to be different every week w and $\varepsilon_{i,t}$ is the error term. The parameters a_k are the vectors of interest to be estimated. The joint use of these fixed effects ($\mu_{c,t}$ and $n_{i,w}$) implies that we only use within-week, within area variations of the weather, expressed as deviations to the national daily average. The model is estimated using the estimator developed by Correia (2018) (19). We cluster standard errors at the country level.

Our identification strategy relies on the assumption that the distribution of these deviations over a week are as good as random, such that their correlation with confirmed cases will identify the effect of the weather on confirmed COVID-19 cases. In the past, many studies (e.g. 20, 21, 22, 23) have used relatively similar statistical frameworks to look at the impact of the weather on diseases that are very well-known to be sensitive to the weather, for example respiratory diseases such as influenza, metabolic diseases like diabetes, or cardiovascular diseases (e.g. stokes).

To match the COVID-19 data with the meteorological data, we had to handle the following imprecisions in the COVID-19 datasets. For 0.7% of confirmed cases, the COVID-19 data does not provide us with an exact date, but with a period when the testing happened. This period is between 2 and 14 days, and most of the time lower than 4 days. We chose to include these observations and add them to each possible day of case confirmation with a weight reflecting that the observation is included for several days. For example, if the date of confirmed cases is a period of 2 days, we add this observation to the case count for each of these 2 days with a weight of 1/2. If the period is 3 days, the observation is added to the case count of these 3 days with a weight of 1/3, and so on. We found that including these observations in the analysis has no impact on the results.

In addition, the COVID-19 dataset does not always provide detailed georeferenced information. The information is provided either at national (7% of observations), regional (43%), city (39%), postcode level (1%) or specific location (10%). We drop the observations that only report national level geographical information and match the other ones with the weather information corresponding to the longitude and latitude reported in the COVID-19 dataset.

We bound the analysis based on the dates when new cases are observed. The model does not include observations before at least one case is observed in an area, and we have dropped the observations at the end of the sample, once the last case has been recorded in an area. This is to avoid measurement errors in case some areas are no longer reported in the COVID-19 dataset we use for the estimation.

3. Main results and robustness checks

The estimation results are reported below in **Appendix Table A2** for the sum of all the lags ($\sum_{x=0}^{X=15} a_x$) and in **Appendix Figure A2** for each individual coefficient separately (the a_x). We find a negative correlation between the outcome variable and temperature. The association is strong for China, while it is only statistically significant at 10% outside China. This suggests that the negative association may not systematically hold. We do not find statistically significant results when looking separately at the United States or the European Union. This could be due to lack of data but also, as explained in the core of the paper, associations are likely to be dependent on the sample used. There are many reasons why this could be the case, starting with different methods used by countries to collect their case data or because social interactions or other parameters influencing disease spread correlate differently with temperature.

Appendix Table A2: Linearized results. The dependent variable is ln $(C_{i,t}) - \ln(C_{i,t-1})$. Standard errors are in brackets and clustered at country level. *** is for statistical significance at 1%. The results displayed for the average temperatures are for the lagged temperatures combined ($\sum_{x=0}^{X} a_x$). The model includes country-by-day fixed effects (e.g. UK, 6th April, 2020) and area-by-week fixed effects (London, 5th–11th April, 2020).

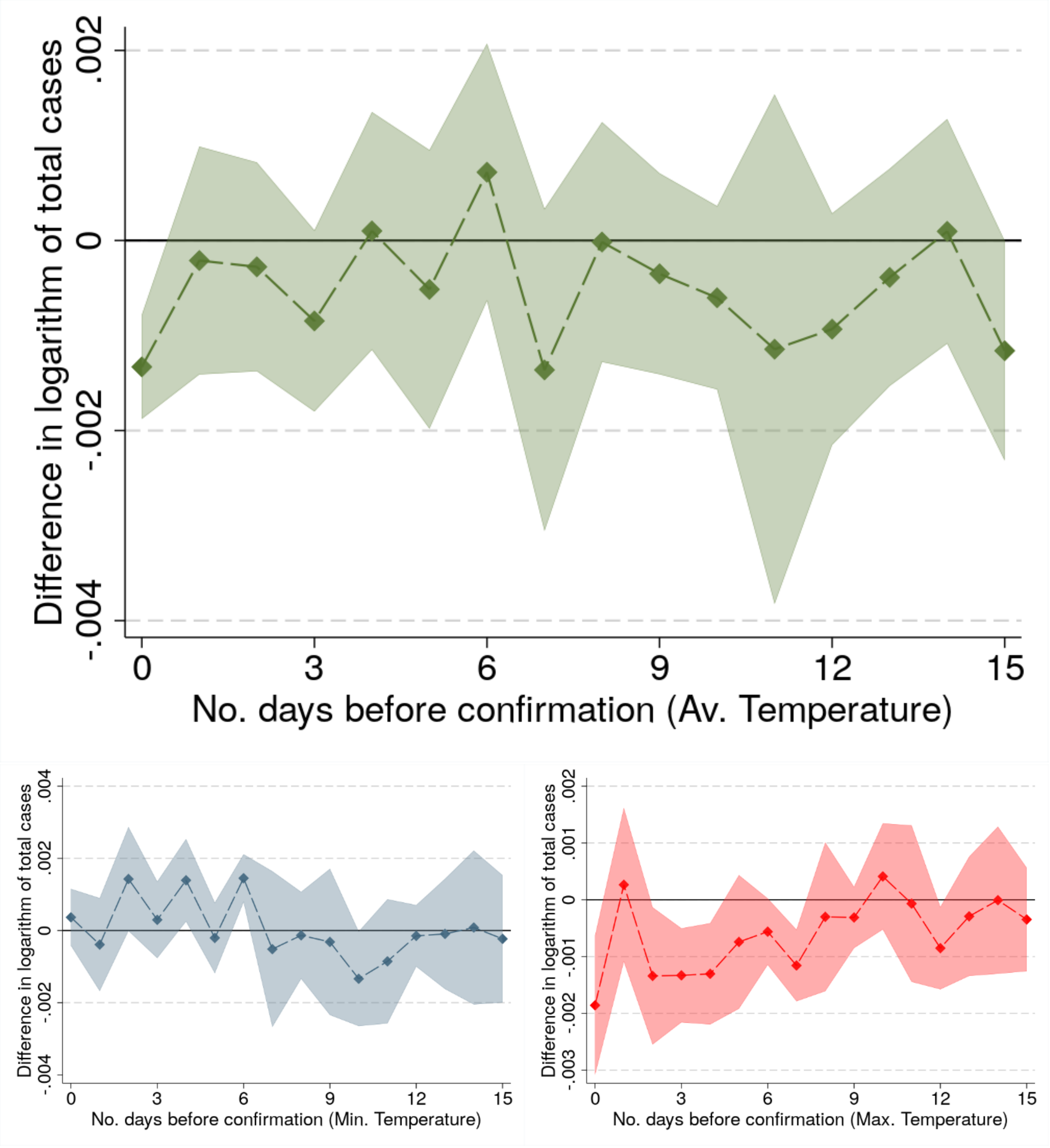
Variable	All observations	China	Outside China
Av. Temperature (°C)	-0.0082***	-0.0090***	-0.0065*
	(0.0026)	(0.0029)	(0.0035)
Observations	76,696	24,932	51,764

The corresponding individual effects are reported in **Appendix Figure A2**. The figure shows that all lags seem to have a similar, negative correlation.

There is a strong and statistically negative impact of temperature on confirmed cases for the temperatures on the day. This could be due to two reasons. As explained in the core of the text, testing is very likely to be influenced by the weather, explaining why we would find effects of temperatures on the day of reporting. Another possibility is that the severity of the COVID-19 infections could increase with cold weather, especially for people with preconditions that are known to be affected by the weather (e.g. diabetes, cardiovascular illnesses), or people that also suffer from other respiratory infections (since these are known to correlate with cold weather). We would then capture some effect of temperature very close to the date of case confirmation for severe cases (which are the ones we record). Coronavirus test results can arrive on the same day as when the test is performed.

Appendix Figure A2 below also provides the individual effects (a_x) for minimum and maximum temperatures separately using another econometric specification reported later, in **Appendix Table A3, column 2**. Effects are clearer and seem mostly determined by temperatures a few days before the cases are confirmed to be COVID-19.

Appendix Figure A2: Values of the individual coefficients (a_x) for average temperature (from Table A2, column 1), and maximum and minimum temperatures (from Table A3, column 2)



Alternative choice of weather variables. Appendix Table A3, column 1 implies that the correlation between temperature and COVID-19 cases is similar at different temperature ranges. Column 2 suggests that the correlation is driven by maximum temperatures. Columns 3 to 5 suggest that relative humidity and precipitation do not strongly correlate with COVID-19 cases. None of the effects below should be interpreted as being causal, as explained in the core of the text.

Appendix Table A3: Alternative choice of weather variables. The estimation is for the full sample. The dependent variable is $\ln(C_{i,t}) - \ln(C_{i,t-1})$. Standard errors are in brackets and clustered at country level. *, **, and *** are for statistical significance at 10%, 5% and 1% respectively. The results displayed for all the weather variables are for the lagged variables combined ($\sum_{x=0}^{X} a_x$). The model includes country-by-day fixed effects (e.g. UK, 6th April, 2020) and area-by-week fixed effects (London, 5th–11th April, 2020).

Column	(1)	(2)	(3)	(4)	(5)
Av. Temperature (°C)	-0.0093***		-0.0099***	-0.0095***	-0.0049
	(0.0034)		(0.0031)	(0.0029)	(0.0051)
x below 0°C	0.0005				
	(0.0043)				
x above 25°C	-0.0025				
	(0.0035)				
Max. Temperature (°C)		-0.0098***			
		(0.0027)			
Min. Temperature (°C)		0.0008			
		(0.0052)			
Relative humidity (%)			-0.0006	-0.0009	-0.0002
			(0.0008)	(0.0008)	(0.0005)
x Av. Temperature (°C)					-0.0001
					(0.0001)
Precipitations (mm)				0.0016	0.0015
				(0.0016)	(0.0015)
Observations	76,696	76,789	76,696	76,696	76,696

Choice of fixed effects. The specifications below illustrate that there are high risks that confounding variables could contaminate the association between the weather and confirmed COVID-19 cases. In columns (1) to (4), the fixed effects do not control for changes in local climates, as well as for differences in sub-regional testing practices. Columns (5) and (6) use area-by-week fixed effects that are able to control for local climates, and results suggest a negative association between the weather and confirmed COVID-19 cases.

These results are only correlations. As explained in the core of the text, these estimates suffer from serious shortcomings.

Appendix Table A4: Robustness checks on the choice of the fixed effects. The estimation is for the full sample. The dependent variable is $\ln(C_{i,t}) - \ln(C_{i,t-1})$. Standard errors are in brackets and clustered at country level. *, **, and *** are for statistical significance at 10%, 5% and 1% respectively. The results displayed for average temperature are for the lagged temperatures combined ($\sum_{x=0}^{X} a_x$).

Column	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Av. temperature	-0.0002	0.0002	-0.0052**	-0.0011	-0.0081**	-0.0082***
(°C)	(0.00070)	(0.00026)	(0.00210)	(0.00145)	(0.00395)	(0.00256)
Fixed effects:						
Day	N	Y	Y	Y	Y	Y
Country	N	Y	Y	Y	Y	Y
Area	N	N	Y	Y	Y	Y
Day by country	N	N	N	Y	N	Y
Area by week	N	N	N	N	Y	Y

Model dynamics. We change the number of lags below. The effect is stable over specific periods (0-5 lags and 12-18 lags) but lose significance between the 6th and the 10th lag. Either results are unstable because of multicollinearity, or there is no to little effect on these lags and therefore the standard deviation increases when they are included. This could be the case if it takes 11-15 days for most people to develop symptoms, get tested and receive their confirmation. In Appendix Table A6, we reduce the number of parameters to be estimated by assuming that the effect of temperature on the outcome variable is the same for ranges of lags over 4 day time windows (1st to 4th lags, 5th to 8th and so on). This allows us to check that our main results are not driven by multicollinearity issues. The results of Appendix Tables A5 and A6 are similar, suggesting that multicollinearity is not driving our results.

Appendix Table A5: Alternative number of day lags. All the specifications are based on Table A2, column 1, but using different number of daily lags. The coefficients are for the lagged temperatures combined ($\sum_{x=0}^{X} a_x$). Standard errors are in brackets and clustered at country level. *, **, and *** are for statistical significance at 10%, 5% and 1% respectively.

Specification	**Coefficient**	**Standard error**
20 daily lags	-0.0046*	(0.0026)
19 daily lags	-0.0058	(0.0035)
18 daily lags	-0.0062**	(0.0029)
17 daily lags	-0.0074***	(0.0022)
16 daily lags	-0.0056**	(0.0027)
15 daily lags	-0.0082***	(0.0026)
14 daily lags	-0.0061**	(0.0027)
13 daily lags	-0.0053**	(0.0022)
12 daily lags	-0.0050**	(0.0020)
11 daily lags	-0.0038*	(0.0021)
10 daily lags	-0.002	(0.0015)
9 daily lags	-0.0012	(0.0015)
8 daily lags	-0.0013	(0.0014)
7 daily lags	-0.0021*	(0.0011)
6 daily lags	-0.0008	(0.0014)
5 daily lags	-0.0019**	(0.0009)
4 daily lags	-0.0022***	(0.0006)
3 daily lags	-0.0022**	(0.0010)
2 daily lags	-0.0016**	(0.0007)
1 daily lags	-0.0014**	(0.0007)
No lag	-0.0012***	(0.0003)

Appendix Table A6: Estimating jointly the effect of groups of lags over 4-day time windows. All the specifications are based on Table A2, column 1, but assuming that groups of lags have the same effect on the outcome variable to reduce multicollinearity. We also vary the number of lags in the model. The coefficients are for the lagged temperatures combined ($\sum_{x=0}^{X} a_x$). Standard errors are in brackets and clustered at country level. *, **, and *** are for statistical significance at 10%, 5% and 1% respectively.

Specification	Coefficient	Standard error
20 daily lags	-0.0041	(0.0035)
16 daily lags	-0.0064**	(0.0030)
12 daily lags	-0.0051**	(0.0023)
8 daily lags	-0.0016	(0.0015)
4 daily lags	-0.0023***	(0.0009)
No lag	-0.0012***	(0.0004)

4. Information on Projections

The estimates above could mislead researchers and policymakers. If misinterpreted as being causal, they could lead them to make unreliable forecasts. A forecast with our main estimate would suggest that warmer weather in summer would attenuate the spread of COVID-19, while colder weather in autumn and winter could lead to flare ups.

In the process of developing this manuscript, we produced such projections. Given the major concerns outlined in this article, however, we have chosen not to provide the results of such a projection since they are misleading and should not be used to inform for policy-decisions. We are, however, happy to share our projection approach and the associated results for research purposes upon request. Below we present our projection methodology for transparency.

We followed the following steps to produce the projections. First, we used the 10-year average (2010-2019) of daily temperature and relative humidity from ERA5 for March to December and aggregated this data to the country-level using 2020 population weighting (24). Second, we constructed a daily anomaly dataset of the weather conditions for March 1st –December 31st with respect to the monthly mean values for March, as most confirmed cased considered in this study have been observed in March this year. This gave us a very rough estimate of the difference in the expected weather conditions by country with respect to March until the end of 2020. Third, we used the baseline model estimates to project changes in the daily growth rate of COVID-19 confirmed cases. Fourth, we inserted these estimates of the weather impacts on the daily growth rate of infection into a simple susceptible-infectious-recovered (SIR) compartment model (25, 26, 27), using the parameters provided in Walker et al. (2020) for COVID-19.

REFERENCES

[16] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." The Lancet infectious diseases (2020).

[17] Xu, Bo, et al. "Epidemiological data from the COVID-19 outbreak, real-time case information." *Scientific Data* 7.1 (2020): 1-6.

[18] Kraemer, Moritz UG, et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China." *Science* (2020).

[19] Correia, S. (2018). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects.

[20] Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., ... & Leone, M. (2015). Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, *386*(9991), 369-375.

[21] Gasparrini, A., Guo, Y., Sera, F., Vicedo-Cabrera, A. M., Huber, V., Tong, S., ... & Ortega, N. V. (2017). Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health*, *1*(9), e360-e367.

[22] Barnett, A. G., Hajat, S., Gasparrini, A., & Rocklöv, J. (2012). Cold and heat waves in the United States. *Environmental research*, *112*, 218-224.

[23] Deschenes, Olivier, and Enrico Moretti. "Extreme weather events, mortality, and migration." *The Review of Economics and Statistics* 91.4 (2009): 659-681.

[24] Center for International Earth Science Information Network - CIESIN - Columbia University. 2016. Gridded Population of the World, Version 4 (GPWv4.11): Population Count. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4X63JVC . Accessed [1st April, 2020].

[25] Schneider, T. (2020). Flatten the Curve. Code available at: https://github.com/tinu-schneider/Flatten_the_Curve [Accessed 30th March 2020].

[26] Höhle, M. (2020). Flatten the COVID-19 curve. Blog post available at: https://staff.math.su.se/hoehle/blog/2020/03/16/flatteningthecurve.html [Accessed 30th March 2020].

[27] Kermack, W. O., and A. G. McKendrick. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society, Series A* 115: 700–721.