# Using Georeferenced Data to Understand the Influence of Weather Conditions on COVID-19

Francois Cohen[[1]](#footnote-2), Anant Jani[[2]](#footnote-3), Sihan Li[[3]](#footnote-4), Yangsiyu Lui and Moritz Schwarzi,[[4]](#footnote-5)

This version: 1st April 2020; first version: 1st April 2020; data update: March 29th, 2020.

## Supplementary material

### Summary statistics of meteorological data

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Std. Dev. | Min | Max |
| Av. temperature (°C) | 9.97 | 8.93 | -30.81 | 38.08 |
| Min. temperature (°C) | 5.41 | 9.46 | -36.25 | 32.62 |
| Max. temperature (°C) | 14.54 | 8.8 | -29.00 | 43.55 |
| Total Precipitation (mm) | 2.48 | 5.63 | 0.00 | 169.39 |
| Relative humidity (%) | 67.09 | 17.19 | 5.74 | 99.97 |
| Temperature bins |  |  |  |  |
| Average temperature: |  |  |  |  |
| <0°C | 0.10 | 0.3 | 0 | 1 |
| 0-10°C | 0.47 | 0.5 | 0 | 1 |
| 10-20°C | 0.27 | 0.44 | 0 | 1 |
| 20-30°C | 0.17 | 0.37 | 0 | 1 |
| >30°C | <0.01 | 0.05 | 0 | 1 |
| Min. temperature: |  |  |  |  |
| <-10 °C | 0.04 | 0.2 | 0 | 1 |
| -10-0°C | 0.25 | 0.43 | 0 | 1 |
| 0-10°C | 0.43 | 0.5 | 0 | 1 |
| 10-20°C | 0.19 | 0.39 | 0 | 1 |
| >20°C | 0.09 | 0.29 | 0 | 1 |
| Max. temperature: |  |  |  |  |
| <10°C | 0.33 | 0.47 | 0 | 1 |
| 10-20°C | 0.40 | 0.49 | 0 | 1 |
| 20-30°C | 0.23 | 0.42 | 0 | 1 |
| 30-40°C | 0.04 | 0.2 | 0 | 1 |
| >40°C | <0.01 | 0.01 | 0 | 1 |

### Statistical model

We use the following econometric model to estimate the impact of the weather on confirmed COVID-19 cases:

(1)

In Eq. (1), is the natural 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 infections in relative terms based on the level of infections 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.

is a matrix of weather-related variables that includes information on the weather at time . We include the lagged values of these weather variables (until ) to capture the effect of the weather of the previous days on infections. is modulable. We run our baseline specifications with average temperature, but also use humidity and precipitation as controls in the robustness checks. We also use daily minimum and maximum temperatures and vary the total number of lags (). Our main models use 21 lags ( and therefore covers 22 days. This should cover most cases for the maximum time reported for the incubation of the disease (14 days) and its detection through testing. The rationale for using 21 lags is that results are stable after the 21st lag. We report alternative models with less lags, and more lags, in **supplementary material 3**.

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. is an area-specific (e.g. regions or cities) fixed effect that is assumed to be different every week and is the error term. The parameters are the vectors of interest to be estimated. With this specification, the effect of temperature on the spread of COVID-19 is identified from sub-regional deviations in confirmed cases within a country and within a day (), accounting for the average differences in the weekly number of cases between areas (controlled by ) and associated changes in expected meteorological conditions. The model is estimated using the estimator developed by Correia (2018).[[5]](#endnote-3) We cluster standard errors at the country level.

The identification strategy primarily relies on the hypothesis that weather variations can be deemed as good as random within a week for a given area after correcting for daily weather anomalies at national level with day by country fixed effects.

We tried another specification before opting for Eq. (1):

(2)

In Eq. (2), is the number of confirmed cases of COVID-19 observed in area i, on day t. All other variables remain unchanged. This specification is a Poisson model and often used to look at count data. The main problem with this specification is that it did not converge when we included the many lags. We therefore ended up choosing a specification that is very close to a count model in principle but offers the advantage of being linear.

To match the COVID-19 data with the meteorological data, we had to handle the following imprecisions in the COVID-19 datasets. For 1.54% 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 days and 14 days, and most of the time lower than 4 days (in 1.21% of the cases, without an exact date, the given period was longer 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 confirmed that both including these observations and not including them has no impact on the results.

In addition, the COVID-19 dataset is not always providing detailed georeferenced information. The information is provided either at national (6% of observations), regional (%), city (%) or postcode level (%). 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.

### Linearized results

**Figure 2** provides a visual representation of our main results and the possible non-linearities between COVID-19 confirmed cases and average temperature. The regressions below use Eq. (1) and average temperatures to directly estimate the linear effect of a change in average temperature on COVID-19 cases.

**Appendix Table A2: Linearized results**

|  |  |  |  |
| --- | --- | --- | --- |
| Column | All areas | In China | Outside China |
| Av. Temperature (°C) | -0.0099\*\*\* | -0.0094\*\* | -0.0077\*\*\* |
|  | (0.0015) | (0.0047) | (0.0018) |
| Observations | 59,268 | 24,849 | 34,419 |

Notes: The dependent variable is . Standard errors are in brackets and clustered at country level (and area level for China). \*, \*\*, and \*\*\* are for statistical significance at 10%, 5% and 1% respectively. The results displayed for the average temperatures are for cumulated effects over 22 days. The model includes country-by-day fixed effects (e.g. UK, April 6th, 2020) and area-by-week fixed effects (London, April 5th–11th, 2020).

We calculate absolute changes in the growth of total cases as follows. On average in all areas, the dependent variable, , increases by 0.0099 with a 1°C decrease in temperature. This is: . Therefore, We apply the same formula to calculate the confidence intervals and the separate effects for China and outside China.

### Robustness checks

**Alternative choice of weather variables.** In Appendix Table A3, column 1 implies that the impact of temperature can be deemed linear. Column 2 suggests that is it driven by maximum temperatures. Table 3 to 5 suggest that relative humidity and precipitation only play a minor role compared to temperature.

**Appendix Table A3: Alternative choice of weather variables**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column | (1) | (2) | (3) | (4) | (5) |
| Av. Temperature (°C) | -0.0123\*\*\* |  | -0.0136\*\*\* | -0.0126\*\*\* | -0.0035 |
|  | (0.0034) |  | (0.0020) | (0.0019) | (0.0055) |
| x below 0°C | 0.0076 |  |  |  |  |
|  | (0.0054) |  |  |  |  |
| x above 30°C | 0.0116 |  |  |  |  |
|  | (0.0151) |  |  |  |  |
| Max. Temperature (°C) |  | -0.0134\*\*\* |  |  |  |
|  |  | (0.0035) |  |  |  |
| Min. Temperature (°C) |  | 0.0051 |  |  |  |
|  |  | (0.0053) |  |  |  |
| Relative humidity (%) |  |  | -0.0040 | -0.0038 | -0.0030 |
|  |  |  | (0.0027) | (0.0025) | (0.0019) |
| x Av. Temperature (°C) |  |  |  |  | -0.0002\* |
|  |  |  |  |  | (0.0001) |
| Precipitations (mm) |  |  |  | 0.0004 | 0.0004 |
|  |  |  |  | (0.0029) | (0.0031) |
| Observations | 59,268 | 59,268 | 59,268 | 59,268 | 59,268 |

Notes: The estimation is for the full sample. The dependent variable is . 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 cumulated effects over 22 days. The model includes country-by-day fixed effects (e.g. UK, April 6th, 2020) and area-by-week fixed effects (London, April 5th–11th, 2020).

**Choice of fixed effects.** The specifications below illustrate the importance of using fixed effects to isolate the impact of temperatures on confirmed COVID-19 cases. Models (1) to (3) strongly underestimate COVID-19 response. Model (4) and (5) provide results that are closer to those of our linearized model.

**Appendix Table A4: Robustness checks on the choice of the fixed effects**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column | (1) | (2) | (3) | (4) | (5) |
| Av. Temperature | 0.00055 | -0.00049 | 0.00008 | -0.00913\*\*\* | -0.00345\* |
| (°C) | (0.00052) | (0.00060) | (0.00012) | (0.00160) | (0.00176) |
| Fixed effects (Y/N): |  |  |  |  |  |
| *Day* | N | Y | Y | Y | Y |
| *Country* | N | N | Y | Y | Y |
| *Area* | N | N | N | Y | Y |
| *Day by country* | N | N | N | N | Y |

Notes: The estimation is for the full sample. The dependent variable is . 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 cumulated effects over 22 days. The model includes country-by-day fixed effects (e.g. UK, April 6th, 2020) and area-by-week fixed effects (London, April 5th–11th, 2020).

**Model dynamics.** We change the number of lags below. The effect of temperature tends to increase with the number of lags included in the model. This might be due to fewer infections leading to fewer infections today. Weather conditions that are favourable for transmission of the virus should have an accumulative effect on increasing the growth rate of coronavirus cases.

We also ran tests where, in addition to the lags, we include leaded values(t+n) in the regressions,. In line with expectations, the introduced leads do not convey a consistent picture, with the linearized coefficients varying greatly in magnitude and sign.

**Appendix Table A5: Alternative number of day lags**

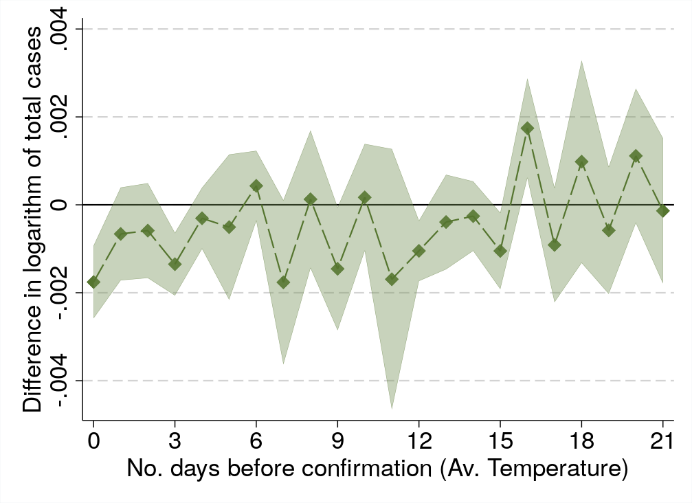
|  |  |  |
| --- | --- | --- |
|  | Average Temperature (°C) | |
| Specification | Coefficient | Standard error |
| 30 daily lags | -0.0078\*\*\* | (0.0028) |
| 29 daily lags | -0.0091\*\*\* | (0.0024) |
| 28 daily lags | -0.0077\*\*\* | (0.0018) |
| 27 daily lags | -0.0085\*\*\* | (0.0017) |
| 26 daily lags | -0.0064\*\*\* | (0.0020) |
| 25 daily lags | -0.0053\*\* | (0.0021) |
| 24 daily lags | -0.0051\*\* | (0.0024) |
| 23 daily lags | -0.0058\*\*\* | (0.0019) |
| 22 daily lags | -0.0092\*\*\* | (0.0023) |
| 21 daily lags | -0.0099\*\*\* | (0.0015) |
| 20 daily lags | -0.0096\*\*\* | (0.0017) |
| 19 daily lags | -0.0113\*\*\* | (0.0026) |
| 18 daily lags | -0.0112\*\*\* | (0.0022) |
| 17 daily lags | -0.0122\*\*\* | (0.0028) |
| 16 daily lags | -0.0108\*\*\* | (0.0024) |
| 15 daily lags | -0.0129\*\*\* | (0.0023) |
| 14 daily lags | -0.0119\*\*\* | (0.0016) |
| 13 daily lags | -0.0104\*\*\* | (0.0016) |
| 12 daily lags | -0.0092\*\*\* | (0.0014) |
| 11 daily lags | -0.0075\*\*\* | (0.0021) |
| 10 daily lags | -0.0050\*\*\* | (0.0014) |
| 9 daily lags | -0.0047\*\*\* | (0.0010) |
| 8 daily lags | -0.0041\*\*\* | (0.0008) |
| 7 daily lags | -0.0045\*\*\* | (0.0011) |
| 6 daily lags | -0.0028\*\*\* | (0.0009) |
| 5 daily lags | -0.0031\*\*\* | (0.0008) |
| 4 daily lags | -0.0037\*\* | (0.0014) |
| 3 daily lags | -0.0037\*\*\* | (0.0014) |
| 2 daily lags | -0.0026\*\* | (0.0010) |
| 1 daily lags | -0.0020\*\*\* | (0.0006) |
| No lag | -0.0015\*\*\* | (0.0005) |

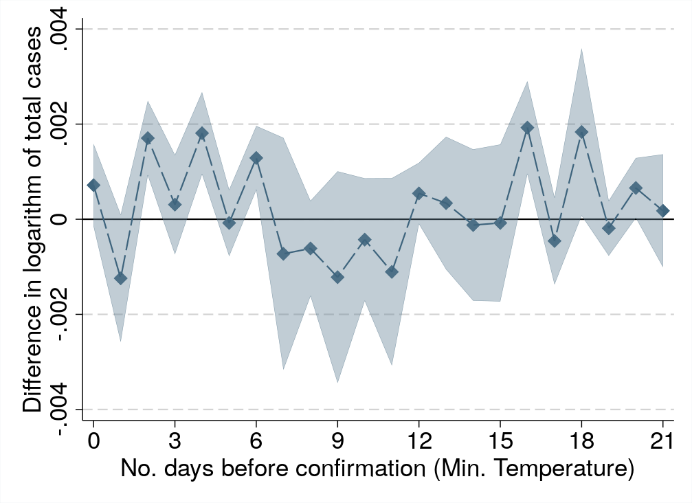
Notes: All the specifications are based on Table A2, column 1, but using different number of daily lags. The coefficients are for the lagged temperatures combined. Standard errors are in brackets and clustered at country level. \*, \*\*, and \*\*\* are for statistical significance at 10%, 5% and 1% respectively.

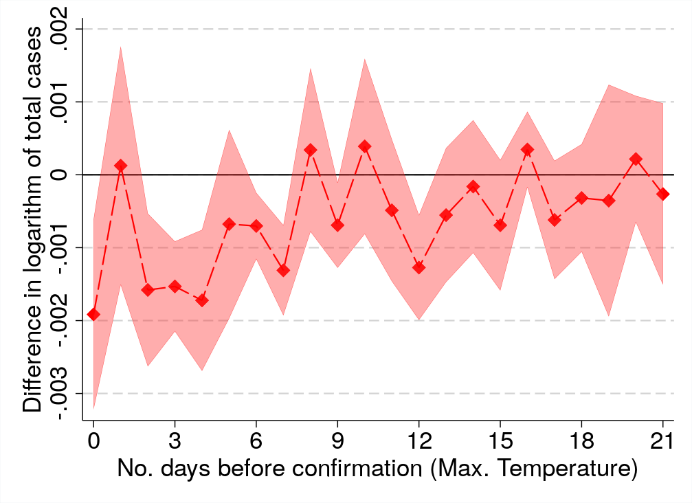
The graphs and figures included in the main text are for the cumulative effect of all the lags included in the model (the sum ). The graph below reports the individual effects of each lag (each single ) for average temperature and the model of **Table A1, column 1**. The pattern seems to represent negative effects until the 12th lags. Surprisingly, there is a negative impact of temperature on confirmed cases for the temperatures on the day. This may be because the severity of the COVID-19 infection might increase with cold weather. This is very likely for people with preconditions that are know 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).

The figure below also provides the individual effects () for minimum and maximum temperatures separately using the model of **Table A1, column 3**. Effects are clearer and seem mostly determined by maximum temperatures a few days before cases are confirmed to be COVID-19.

**Appendix Figure A1: Values of the individual coefficients () for daily average temperatures (from Table 1, column 1), and maximum and minimum temperatures (from Table 1, column 3)**







**Newly infected countries.** Below we check if the spread of the virus correlates with the weather at the start of the contagion, when a country records less than a given number of total cases (as estimated in the dataset). We do not find an association with temperature in the early stages of the disease.

**Appendix Table A6: Restricting the sample to below/above a certain number of national cases**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column | (1) | (2) | (3) | (4) | (5) |
| Only in sample if total cases in country are: | <1000 | <2500 | <5000 | >5000 | All data |
| Av. Temperature (°C) | 0.0103 | 0.0080 | 0.0205 | -0.0156\*\*\* | -0.0099\*\*\* |
|  | (0.0187) | (0.0256) | (0.0238) | (0.0010) | (0.0015) |
| Observations | 11,724 | 15,955 | 19,189 | 39,277 | 59,268 |

Notes: The dependent variable is . 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 the average temperatures are for cumulated effects over 22 days. The model includes country-by-day fixed effects (e.g. UK, April6th, 2020) and area-by-week fixed effects (London, April 5th–11th, 2020).

**Before and after within-country movement restrictions.** Rapidly, countries reacted by reducing movement from/to infected areas, and then ask people to stay at home and leave their house only for essential activities. We test if we observe differences in the response of COVID-19 confirmed cases to weather before and after nationwide restrictions came into force.

We use the Oxford COVID-19 government response tracker[[6]](#footnote-6) to identify the moment when a country decided to enforce a general rule restricting internal movement at national scale. We then estimate separately the response of COVID-19 to weather conditions before and after general restrictions were taken. We run models with average temperature (similar to Table 1, column 1) and with maximum and minimum temperature (similar to Table 1, column 3). Results are similar and do not suggest strong differences in the effect of the weather on infections before and after drastic government response so far. Note that this is not an evaluation of the effectiveness of government measures, but an evaluation of the change of the relative effect of the weather on COVID-19, once government measures have been put in place.

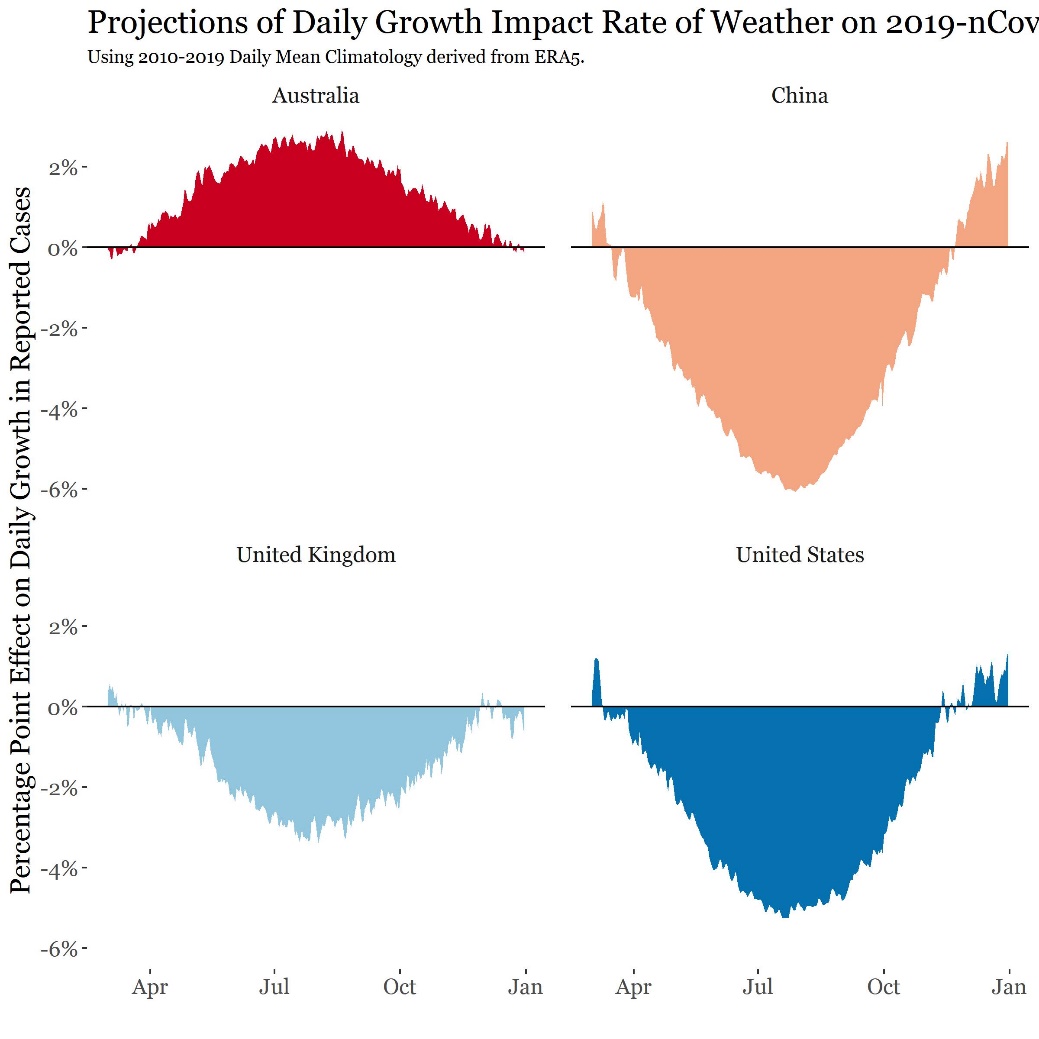
### Projection Details

The projections have been produced in four steps. First, we use the 10-year average (2010-2019) of daily temperature and relative humidity from ERA5 for March to December and aggregate this data to the country-level using 2020 population weighting.[[7]](#endnote-4) We construct 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 gives 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 use the estimates of Table 1, column 1 to predict changes in the average growth rate of COVID-19 confirmed cases. These impacts on the daily growth rate are provided in the Figure A2 below.

**Appendix Figure A2: Changes in the daily growth rate of confirmed cases as a function of the expected weather from 1 March 2020**



Fourth, we insert these estimates of the weather impacts on the daily growth rate of infection into a simple susceptible-infectious-recovered (SIR) compartment model,[[8]](#endnote-5),[[9]](#endnote-6),[[10]](#endnote-7) using the parameters provided in Walker et al. (2020) for COVID-19. This allows us to investigate the evolution of the disease in three social distancing scenarios (no distancing, distancing for the elderly, and distancing for the whole population), as a function of the other parameters in the model, especially the value of the , and the weather.

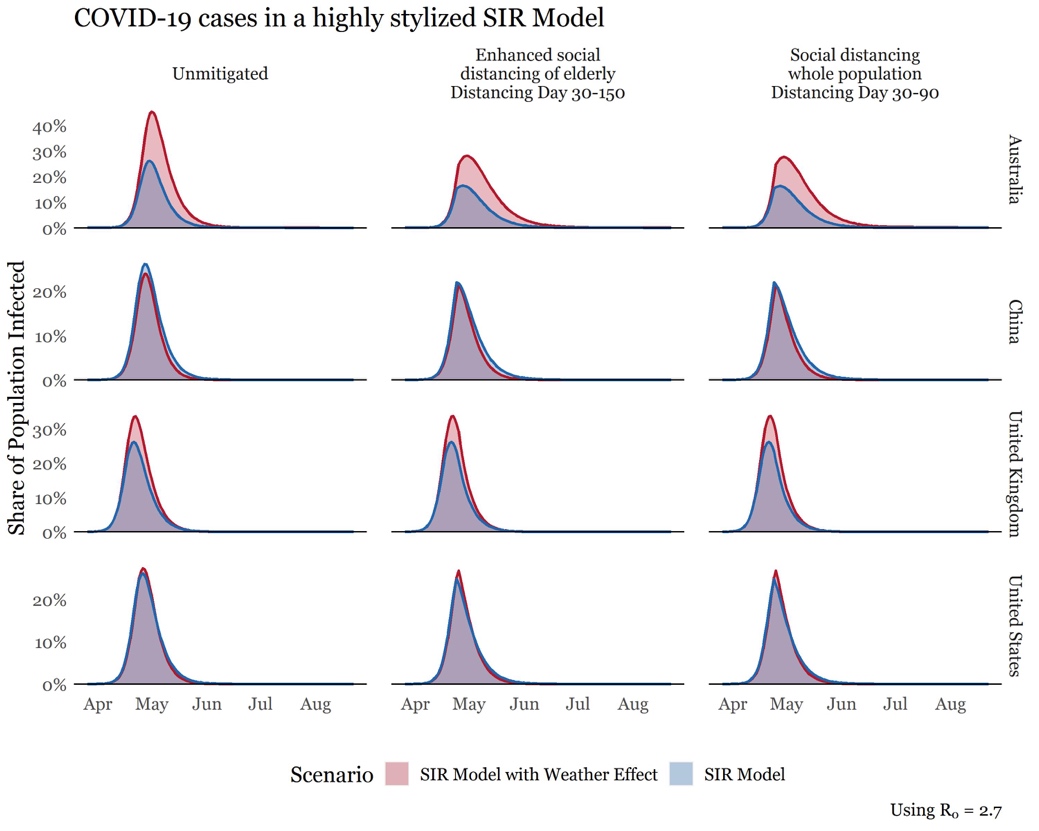
We also rely on Walker et al. (2020) for their parameters of , which are 2.4, 2.7, 3 and 3.3. We provide the results for 2.4 in the core of the text, and the projections for other values hereafter.

The two social distancing scenarios are the ones given by Walker et al. (2020). We use the social contact rates calculated by these authors. Social distancing measures are assumed to come into effect 30 days after the start of the model. To simulate a relaxation of social distancing measures, the scenario that assumes social distancing to the entire population keeps their measures only in place for 60 days and relaxes them again after that. Walker et al. (2020) use scenario-specific contact reduction rates of up to 48%.

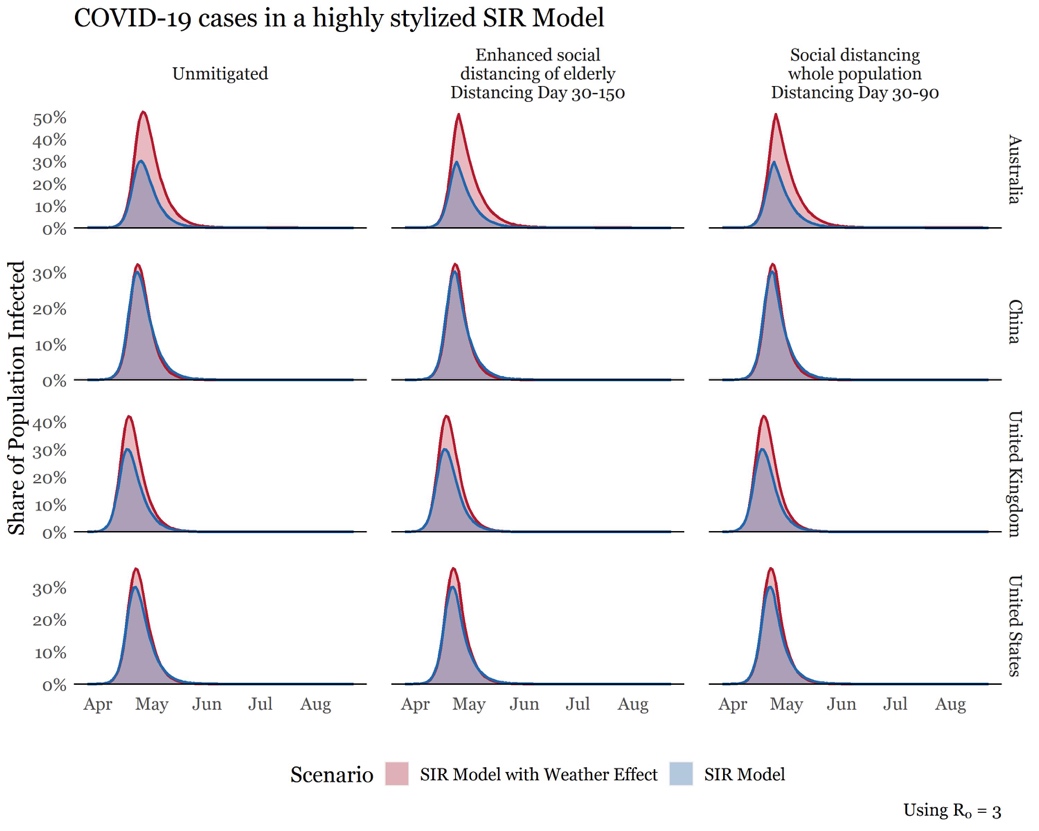
The initial number of cases for each country is the latest recorded date in the Xu et al. (2020) database at the time of writing. Each model is run for 150 days.

The growth projections described above were added to the endogenous daily SIR growth rates after each model was run to produce the weather effect. This constitutes only a first attempt, as more sophisticated insights could be developed by endogenizing the weather effect into the SIR model.

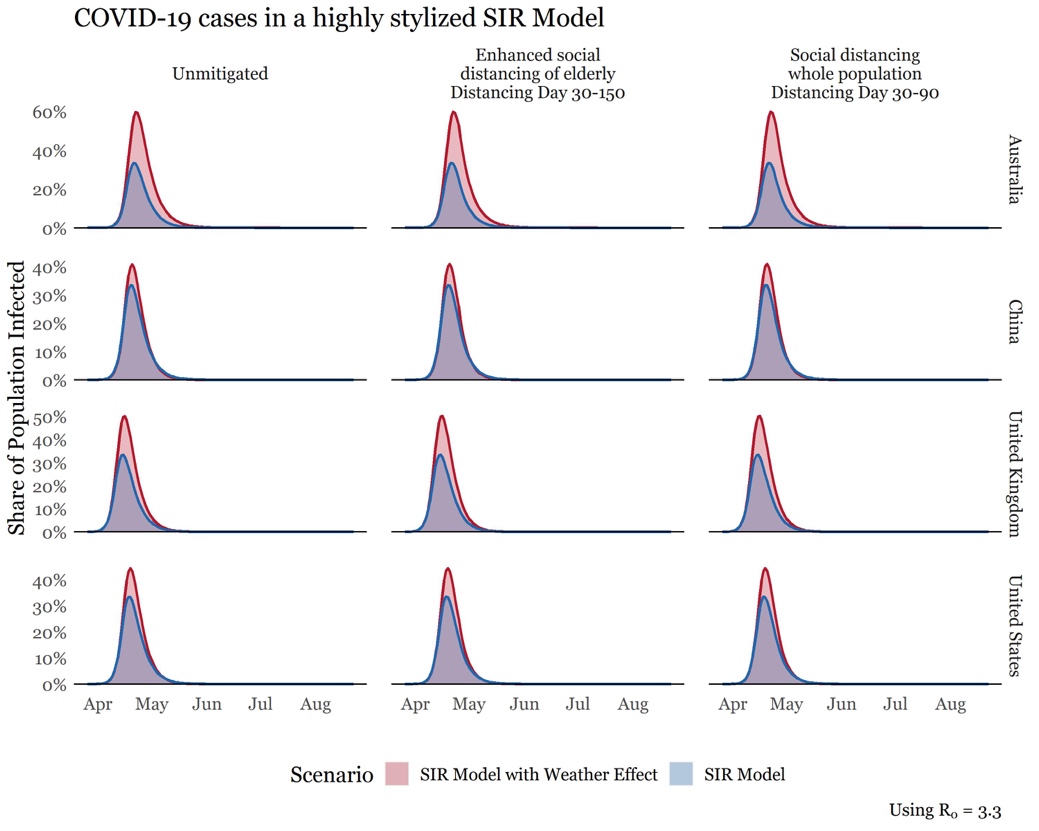
**Appendix Figure A3: Output of SIR model with R0 = 2.7**



**Appendix Figure A4: Output of SIR model with R0 = 3**



**Appendix Figure A5: Output of SIR model with R0 = 3.3**



## ADDITIONAL REFERENCES

1. Smith School of Enterprise and the Environment, University of Oxford; and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford. [↑](#footnote-ref-2)
2. Nuffield Department of Primary Care Health Sciences, University of Oxford. [↑](#footnote-ref-3)
3. Environmental Change Institute, University of Oxford. [↑](#footnote-ref-4)
4. Climate Econometrics, Nuffield College, Oxford [↑](#footnote-ref-5)
5. Correia, S. (2018). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. [↑](#endnote-ref-3)
6. <https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker> [↑](#footnote-ref-6)
7. 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]. [↑](#endnote-ref-4)
8. Schneider, T. (2020). Flatten the Curve. Code available at: <https://github.com/tinu-schneider/Flatten_the_Curve> [Accessed 30th March 2020]. [↑](#endnote-ref-5)
9. 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]. [↑](#endnote-ref-6)
10. 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. [↑](#endnote-ref-7)