scientific reports

Check for updates

OPEN

# Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany

Naleen Chaminda Ganegoda[1], Karunia Putra Wijaya[2], Miracle Amadi[3], K. K. W. Hasitha Erandi[4] & Dipo Aldila[5]✉

**COVID-19 pandemic continues to obstruct social lives and the world economy other than questioning the healthcare capacity of many countries. Weather components recently came to notice as the northern hemisphere was hit by escalated incidence in winter. This study investigated the association between COVID-19 cases and two components, average temperature and relative humidity, in the 16 states of Germany. Three main approaches were carried out in this study, namely temporal correlation, spatial auto-correlation, and clustering-integrated panel regression. It is claimed that the daily COVID-19 cases correlate negatively with the average temperature and positively with the average relative humidity. To extract the spatial auto-correlation, both global Moran's $\mathscr{I}$ and global Geary's $\mathscr{C}$ were used whereby no significant difference in the results was observed. It is evident that randomness overwhelms the spatial pattern in all the states for most of the observations, except in recent observations where either local clusters or dispersion occurred. This is further supported by Moran's scatter plot, where states' dynamics to and fro cold and hot spots are identified, rendering a traveling-related early warning system. A random-effects model was used in the sense of case-weather regression including incidence clustering. Our task is to perceive which ranges of the incidence that are well predicted by the existing weather components rather than seeing which ranges of the weather components predicting the incidence. The proposed clustering-integrated model associated with optimal barriers articulates the data well whereby weather components outperform lag incidence cases in the prediction. Practical implications based on marginal effects follow posterior to model diagnostics.**

Viral diseases emerge with complex transmission dynamics, and they are hard to eradicate challenging capacity of testing, diagnosis, and cure[1,2]. Such complexity is generated by various factors such as genetic changes of the virus, environmental influences, and host behavior[3,4]. COVID-19 caused by the coronavirus SARS-CoV-2 has also shown its revolutionary dynamics via all those routes, leaving the world at a standstill in many aspects. The transmission of coronavirus occurs and escalates in diverse means. Most notable drivers include direct contact with infectious individuals[5], fomite transmission via contaminated surfaces[6,7], transmission via virus-carrying aerosols[8,9], congested living and mobility leading to superspreading events[10–13], and lack of compliance to health guidelines[14–17]. Though both direct and indirect transmission are recognized, the influence of outdoor aerosol transmission is not properly understood[18,19]. Meanwhile, within-household is much higher compared to cross-household transmission leaving home quarantine also at risk[20]. Thus, planning healthcare and interventions has also become challenging. It is further problematic due to the presence of asymptomatic cases[21].

Transmission and morbidity of COVID-19 can be worsened when co-infections with other respiratory viruses are present. Several clinical studies from different countries have observed the co-infection of COVID-19 with other viral infections[22–24]. The most common respiratory viruses are influenza virus, respiratory syncytial virus, parainfluenza viruses, metapneumovirus, rhinovirus, adenoviruses, bocaviruses, and coronaviruses[25]. These

[1]Department of Mathematics, University of Sri Jayewardenepura, Nugegoda 10250, Sri Lanka. [2]Mathematical Institute, University of Koblenz, 56070 Koblenz, Germany. [3]Department of Mathematics and Physics, Lappeenranta University of Technology, 53851 Lappeenranta, Finland. [4]Department of Mathematics, University of Colombo, Colombo 00300, Sri Lanka. [5]Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia. ✉email: aldiladipo@sci.ui.ac.id

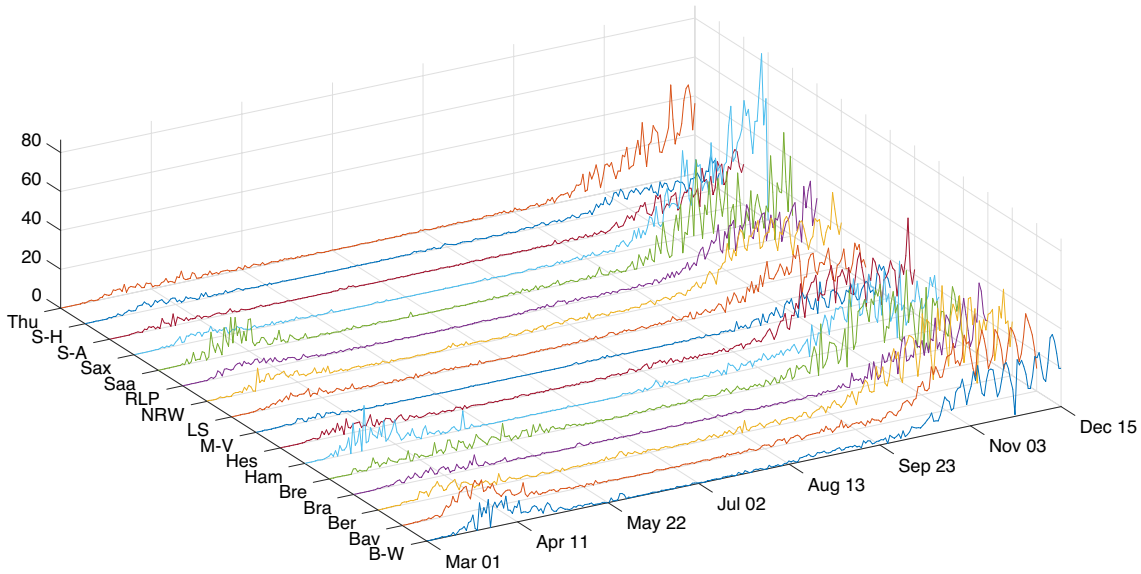
viral infections share common symptoms such as sneezing, cough, sore throats, and fever while following similar ways of transmission[26,27]. Influenza viruses that cause seasonal flu would easily co-exist with COVID-19 in the winter season[28]. This is motivated by the fact that most respiratory pathogens are seasonal[29,30]. Thus, given that many COVID-19 infected cases are undetected[31], sneezing and cough due to another infection may allow passing respiratory droplets carrying SARS-CoV-2 too. Although the information is still limited, one cannot set aside the possible risk of excessive COVID-19 spread due to co-infection[32,33]. In this regard, timely detection is important to curtail issues of missed diagnoses[34].

The influence of weather components such as temperature and relative humidity on the transmission of SARS-CoV-2 is investigated recently. Related studies have been motivated by the fact that temperature and relative humidity also regulated the survival of coronaviruses of SARS[35–38] and MERS[39,40]. Respiratory droplets play a key role in transmission, subsequently more structured with aerosols and fomites[41,42]. Due to other confounding factors related to specific geographical areas, mixed findings can be expected with different levels of temperature and relative humidity[43–46]. Using panel regressions, a study of 20 countries having the most number of confirmed cases[47] suggested that high temperature and relative humidity reduce transmission, while low temperatures are contributory for activation and infectivity of the virus. A low temperature range (− 6.28 °C to + 14.51 °C) has been identified as favorable to COVID-19 growth in[48] via a statistical estimation. This study also found that a 1 °C rise in temperature can reduce the number of cases by 13–17 per day. On the contrary, a study covering many cities in China[49] using a generalized additive model found no evidence supporting the decrease in the number of cases in warmer weather. Moreover, an SEIR model calibrated for 202 locations in 8 countries[50] showed no significant changes in the number of COVID-19 confirmed cases with a broad range of meteorological conditions. Another study in New South Wales, Australia[51], revealed a weak correlation between COVID-19 cases and temperature, but a negative correlation between cases and relative humidity. Studies using data for the earlier infections in Jakarta with average temperature (26.1–28.6 °C)[52] and Bangladesh with average temperature (23.6–31.1 °C) and minimum temperature (17.3–29.3 °C)[45] indicated significantly positive correlation. In addition, COVID-19 cases in China showed negative correlations with both temperature and relative humidity as investigated in[53] while those in 190 countries revealed non-linear correlations with both daily temperature and relative humidity as in[54]. In Iran, also according to[55], there was no clear evidence to relate the number of confirmed cases with warm or cold weather in different provinces, leaving population size to be a determinant factor. A related study for India was carried out using minimum temperature, maximum temperature, average temperature, and specific humidity (ratio of the mass of water vapor to the total mass of the air parcel) as the weather components[56]. The results showed a high positive correlation between COVID-19 cases and temperature measures and a low positive correlation between COVID-19 cases and specific humidity. In Germany, the confirmed cases hit 17 million by the first week of January 2021. The second wave escalation began in autumn and continued in winter. Daily cases exceed 20,000 in many days at the latter stage, where it was over 15,000 for other days in the last two months of 2020. The long-standing plateau of total deaths has also altered since November to a sharp increase and reached 35,000 at the beginning of 2021.

Motivated by the increase of morbidity during autumn and winter, this study employed panel COVID-19 incidence data from Germany and scrutinized their relationship with weather data. In some studies, weather components like temperature were collected in categories such as average, maximum, and minimum level[52,56–58], while others used daily average extracted on a defined regular interval[50,59]. Furthermore, in some other studies, either absolute humidity[59,60] or specific humidity[56] was employed instead of relative humidity. Ours utilized the average of daily average temperature and relative humidity from January 31, 2020 to December 15, 2020, from three representative weather stations in Germany. Besides data availability and similarity with other studies[61,62], the rationale behind the choice of the weather components lies in their readability throughout academia and the fact that no prior and posterior transformation are needed to obtain marginal effects. Extensive investigation on Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ statistics then followed so as to cover spatial auto-correlation and related practical implications. The difference with previous studies is that the temporal progression of the statistics is presented. Subsequently, this study brought forward a random-effects model with a clustering strategy. Our holistic idea lies in which ranges of the incidence are well predicted by the weather components. This is somewhat contrasting to determining the ranges of the weather components that can predict the incidence. Our clustering is based on the method of stratifying incidence data into an arbitrary number of clusters, separated by barriers. The temperature and relative humidity data were also grouped corresponding to the clustered incidence data. This not only improves fitting by providing more explanatory variables but also screens incidence clusters where the weather components fail to predict. Relevant implications using marginal effects for sample cases then followed posterior to model diagnostics.

## Data and methods

**COVID-19 and weather situation in Germany.** According to the official 2018 census, the German states considerably vary in population, with North Rhine-Westphalia and Bremen having the highest and lowest population size of about 17,932,651 and 682,986, respectively, out of the total population size of 83,019,213. The states also have varied economic capacities in business, industries, tourism, and education, which affect their population size. For instance, the largely populated states like Bavaria and Baden-Württemberg have booming economy and offer plenty of employment opportunities due to the situation of renowned business centers and industries, whereas low-populated states e.g. Bremen are laid behind (see[64,65]). Apparently, the number of cases and fatalities relatively depends on the population size. For instance, based on the report from Robert Koch Institute (RKI) on December 16, 2020, the largest populated state shared the highest 7-day incidence cases, and the smallest populated state shared the lowest. Given that the cases are population-driven, the dataset used for this study includes the daily confirmed COVID-19 cases for all the states from the official website of RKI[66], which was later normal-




| State | B-W | Bav | Ber | Bra | Bre | Ham | Hes | M-V |
|---|---|---|---|---|---|---|---|---|
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Max | 38.05 | 40.95 | 53.77 | 40.53 | 44.22 | 39.16 | 49.75 | 18.20 |
| Mean | 5.96 | 6.89 | 7.55 | 3.80 | 5.93 | 5.57 | 6.04 | 1.79 |
| StDev | 8.43 | 9.77 | 11.68 | 6.91 | 9.37 | 8.01 | 9.44 | 3.34 |
| Population | 11,069,533 | 13,076,721 | 3,644,826 | 2,511,917 | 682,986 | 1,841,179 | 6,265,809 | 1,609,675 |
| State | LS | NRW | RLP | Saa | Sax | S-A | S-H | Thu |
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Max | 26.47 | 39.22 | 36.77 | 54.01 | 86.37 | 29.34 | 18.26 | 48.25 |
| Mean | 3.73 | 6.19 | 4.85 | 5.67 | 7.34 | 2.97 | 2.19 | 94.34 |
| StDev | 5.46 | 8.79 | 7.72 | 9.73 | 15.06 | 5.65 | 3.12 | 8.45 |
| Population | 7,982,448 | 17,932,651 | 4,084,844 | 990,509 | 4,077,937 | 2,208,321 | 2,896,712 | 2,143,145 |

**Figure 1.** Daily COVID-19 cases per 100,000 inhabitants from all 16 states in Germany from March 01 until December 15, 2020: B-W (Baden-Württemberg), Bav (Bavaria), Ber (Berlin), Bra (Brandenburg), Bre (Bremen), Ham (Hamburg), Hes (Hesse), M-V (Mecklenburg-Vorpommern), LS (Lower Saxony), NRW (North Rhine-Westphalia), RLP (Rhineland-Palatinate), Saa (Saarland), Sax (Saxony), S-A (Saxony-Anhalt), S-H (Schleswig-Holstein), Thu (Thuringia). Population data come from the 2018 census by the Federal Statistical Office of Germany[63].

ized per 100,000 inhabitants using the 2018 population census, see Fig. 1. This dataset spans the time window from March 01, 2020 to December 15, 2020. The normalization was intentional toward making the number of cases comparable across the states so as to allow for appropriate comparison with weather components that do not depend on the population (see similar treatments in[59,67,68]). Here, the daily cases were defined as the difference of the confirmed cases since the earliest time of the report. As for the accompanying weather components, temperature and relative humidity data were retrieved from climate environment open data[69]. Time series of average temperature and relative humidity were obtained using the records of three weather stations Berlin-Marzahn (Berlin), München-Stadt (Bavaria) and Stuttgart-Schnarrenberg (Baden-Württemberg). This choice was motivated by data availability and the fact that the weather pattern throughout Germany is more or less the same, except in the alps where a negligible percentage of humans live. Average temperature ranges from − 0.766 to 27.13, and average relative humidity ranges from 39.38 to 93.53%. It seems the two weather components have a negative correlation showing equivalence between low temperature and high relative humidity or vice versa. Moreover, looking at the plot of cases by month in Fig. 1 in comparison with the weather components in Fig. 2, it can be seen that cases are generally higher in colder season and considerably reduce during the hot season.

In addition to the reported incidence, the spatial movement of the largest outbreak over the 16 states is also worth investigating. As depicted in Fig. 3, several stages in the timeline can be identified according to the dominance shown by different states. In the first three weeks in March, the largest incidence mainly altered between Hamburg and Baden-Württemberg. Bavaria and Saarland replaced them in the next three weeks. Bavaria hold a local election on March 15, and in the next day, a state of emergency was declared for 14 days with mobility restrictions[70]. Moreover, it is the first state to declare curfew that was imposed on March 20[71]. Saarland neighboring with badly affected French region Grand Est also incurred the same situation at midnight on the same day[72]. Lack of protective clothing and closure of medical practices were also reported from Bavaria[73]. Thus, Bavaria owed the largest incidence from time to time, even after the first few weeks. Outbreaks in initial reception facilities also contributed to the increase of cases in Bavaria. The largest incidence in May and in the first two weeks of June was dominated by Bremen. It was followed by Berlin and North Rhine-Westphalia until the end

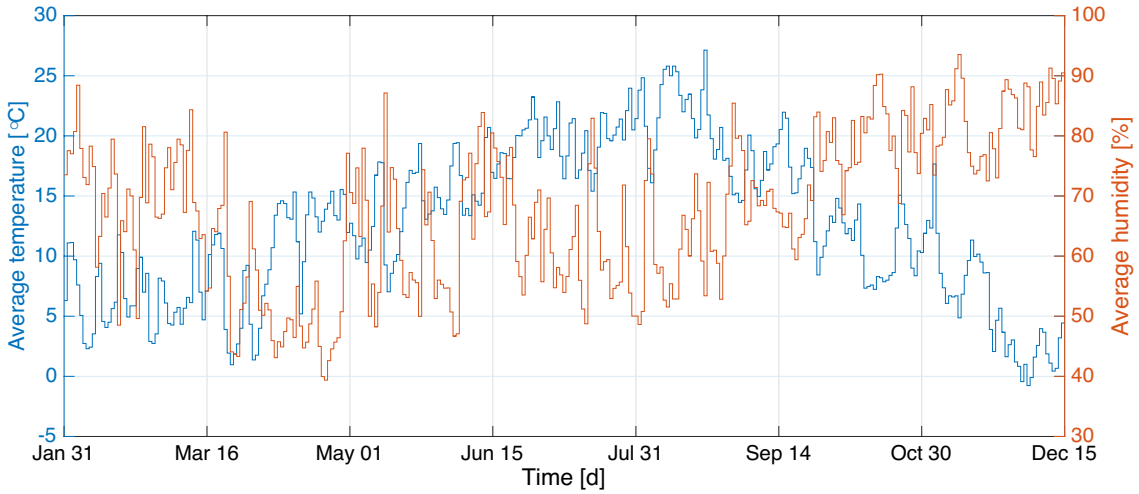



**Figure 2.** Average from the daily average temperature and relative humidity from the three weather stations in Germany: Berlin-Marzahn (Berlin), München-Stadt (Bavaria), Stuttgart-Schnarrenberg (Baden-Württemberg). Time window spans from January 31 until December 15, 2020. The tuples (Min, Max, StDev) are given by (– 0.766 °C, 27.13 °C, 6.45 °C) for the temperature and (39.38%, 93.53%, 12.71%) for the relative humidity, respectively.

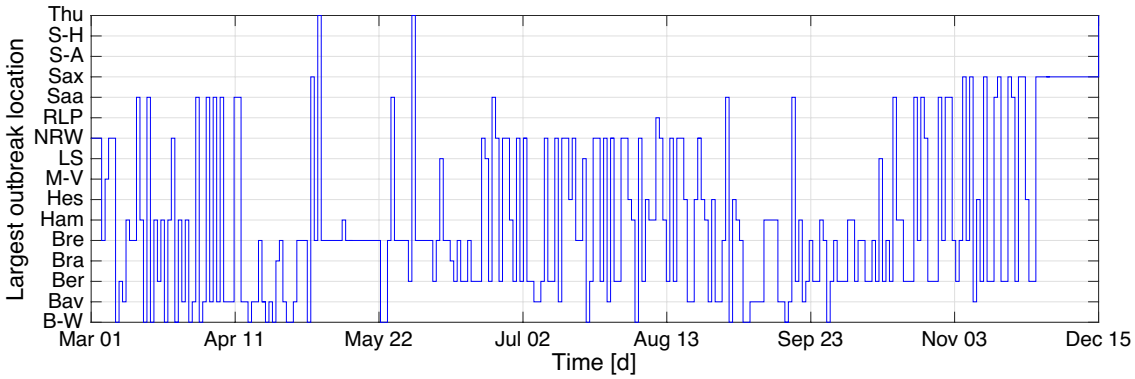



**Figure 3.** Spatial concurrence of the largest outbreak.

of August. A sudden increase of cases was reported in North Rhine-Westphalia due to proactive case tracing, in particular at a meat factory in Coesfeld[74]. Later another cluster occurred on June 17 in a slaughterhouse in Gütersloh, North Rhine-Westphalia, leaving superspreading the main cause of spread[75]. Hamburg and Bremen also came to notice in September and October. The latter stage of October was dominated by Saarland and Berlin. In November, the largest incidence altered between Saxony and Berlin, while Saxony kept the dominance for the first two weeks of December. Saxony had shown early signs of vulnerability, prohibiting residents from leaving their dwellings similar to Bavaria and Saarland. Berlin prevailed as the most responsible state in the latter two-third of the timeline. A large-scale protest was held on August 1 in Berlin against preventive measures. This hints lack of compliance to wearing face masks and keeping physical distance that supports increasing incidence[76].

**Correlation studies.** Referred studies in "Introduction" illustrate how meteorological factors correlated with the transmission of COVID-19. Highly transmissible disease like COVID-19 requires pathogens to remain active outside of the host body and relative humidity and temperature affect the virus's survival in the environment[44,77]. Another study engineering a SARS-CoV-2 isolate came across the fact that the virus can survive at least 28 days at ambient temperature 20 °C and 50% relative humidity on non-porous surfaces and is sensible to the variation of the weather components[78]. Therefore, it is considered noteworthy to examine the interrelationship between COVID-19 cases and meteorological factors. Many statistical methods have been used in earlier studies. According to the recent review in[61], applicable methods other than descriptive analysis are Pearson correlation coefficient, linear, and non-linear regression, LOESS, two-way ANOVA, etc. Wavelet coherency analysis was used in[50]. This study used the Spearman-rank correlation so as to evaluate both the linear and monotonic relationship between two tested covariates. Additionally, auto-correlation between reported COVID-19 cases was also done by piling the spatiotemporal data into one time series, considering that normalized data vary in relatively small numbers. Lags up to 7 days from presently were selected. Therefore, every

covariate augments 16 times 283 observations where the lag-0 time series consists of time window from March 8, 2020 to December 15, 2020. Both the Pearson and Spearman-rank correlation coefficients were computed.

**Spatial pattern.** Of special interest in this study is the degree of interconnection between all states in raising or decreasing the number of cases. The global Moran's $\mathscr{I}$[79] in comparison with the global Geary's $\mathscr{C}$[80,81] and its local decomposition known as Moran's scatter plot were used. The global measures serve to indicate the overall correlation between daily COVID-19 cases per 100,000 inhabitants in every state with the weighted average of the cases in neighboring states, which refers to the *spatial lag* of the state[82]. The spatial pattern is commonly seen to lie between three extreme cases: *locally clustered*, *random*, and *locally dispersed*. Locally clustered refers to the situation where neighboring states are similar in the level of daily new cases, under which spatial dependency rules out the spatial pattern. Locally dispersed refers to the inverse spatial dependency where neighboring states are dissimilar. Something in between is then referred to as random. Representation of these spatial patterns can be understood with the aid of a chessboard. If the spatial profile of daily cases in all states resembles the chessboard, then the spatial pattern is completely locally dispersed. If all the black cells would have gathered in one spot, then the spatial pattern is completely locally clustered. The random spatial pattern is then recognized from the way the black and white cells locate randomly on the board. This is extreme binary stratification that could never occur in the realism of epidemics, from which the corresponding global measure rarely reaches its bounds.

Let us suppose that time is fixed and the daily cases from all states are reported as $C = (c_1, \ldots, c_S)^\top$ with mean $\bar{c}$. The other main ingredient in spatial auto-correlation is the spatial weight matrix $W = (w_{ij})$, which measures the degree of contiguity among all the states. This study used the binary adjacency matrix, where $w_{ij}$ is 1 in case $i$ and $j$ share a common border or 0 in case otherwise (including diagonal entries). This definition is commonly used in the literature (referred to as "queen case") in contrast to distance-based proximity measure where central locations play a significant role as well as a definition of being a "center" is required to define the distances. Let us write $Z = (z_1, \ldots, z_S)^\top := C - \bar{c}$ and define $|W| := \sum_{i,j} w_{ij}$. The global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ statistic are given by

$$\mathscr{I} := \frac{S}{|W|} \cdot \frac{Z^\top W Z}{Z^\top Z} \quad \text{and} \quad \mathscr{C} := \frac{S-1}{2|W|} \cdot \frac{\sum_{i,j} w_{ij}(c_i - c_j)^2}{Z^\top Z}$$

respectively. According to the formulas, the global Moran's $\mathscr{I}$ represents the standardized spatial autocovariance by the variance of the data, while the global Geary's $\mathscr{C}$ replaces the autocovariance by the sum of the squared differences in all data values. Both formulas then differ in sensitivity controlled by the autocovariance. In terms of stability against uncertainty in the data, Wijaya et al. in[68] describe how Geary's $\mathscr{C}$ tends to vary less significantly than Moran's $\mathscr{I}$ when data are perturbed using noise of any kind. The current study presented Geary's $\mathscr{C}$ only for the sake of comparison. A measurement $0 < \mathscr{I} \to 1$ (similarly $1 > \mathscr{C} \to 0$) indicates the direction toward locally structured spatial pattern; $\mathscr{I} = 0$ (or $\mathscr{C} = 1$) random spatial pattern; and $0 > \mathscr{I} \to -1$ (or $1 < \mathscr{C} \to 2$) locally dispersed spatial pattern. Statistical inference is usually done under a total randomization assumption to have a decision outcome based on the values of the statistics[83]. The p-value is generated after normalization using the expected values $\mathbb{E}(\mathscr{I}) = -1/(S-1)$, $\mathbb{E}(\mathscr{C}) = 1$ and variances $\mathbb{V}(\mathscr{I})$, $\mathbb{V}(\mathscr{C})$ reported in the original studies[79,80]. The null hypothesis is that there is no spatial auto-correlation of the daily cases on the observed $S$ states, meaning that $\mathscr{I} \simeq \mathbb{E}(\mathscr{I})$ and $\mathscr{C} \simeq \mathbb{E}(\mathscr{C})$. Therefore, a p-value smaller than a predefined significance level $\alpha$ rejects the null hypothesis whereby either a locally structured or a locally dispersed spatial pattern occurs.

In contrast to the global measures, Moran's scatter plot measures the extent to which a state is considered a "hot spot" or "cold spot" or something in between[83]. It reports the coordinates $(Z/\sigma_C, WZ/\sigma_C)$ for all states, with $\sigma_C = \sqrt{Z^\top Z/S}$ denoting the standard deviation of $C$. As a row-standardized weight matrix is utilized, i.e., $|W| = S$, the pooled estimator of the regressing linear line for these coordinates passing through the origin is given by $(0, \mathscr{I})$. In the present context, a hot spot is defined as a state with a large number of daily cases surrounded by those with large numbers of cases (*high-high*). In the 2-dimensional Euclidean space, the coordinates of hot spots locate in the upper-right quadrant Q1. A cold spot, on the contrary, defines a state with a small number of cases surrounded by those with small numbers of cases (*low-low*). The coordinates of cold spots gather in the lower-left quadrant Q3. Other than these, local dispersion may occur falling into the following categories: a state with a small number of cases surrounded by those with large numbers (*low-high*) in the upper-left quadrant Q2, and a state with a large number of cases surrounded by those with small numbers (*high-low*) in the lower-right quadrant Q4. From the practical point of view, being a hot spot or cold spot may only rely on the health care capacity to ameliorate the disease burdens without imposing further restrictions to travel around neighboring states, except for those who travel across the border between scattered hot spots and cold spots. A state in a high-low or low-high spatial pattern, however, requires more restriction in traveling to neighboring states as the disease may diffuse (in case of high-low) or be absorbed (in case of low-high).

**Simple case–weather relation.** Let $i$ and $j$ denote the state and time index where $i \in \{1, \ldots, S = 16\}$ and $j \in \{1, \ldots, N\}$. Our approach to modeling daily COVID-19 cases in all states in Germany was based on directly relating collected entities. These include presently (lag-0) reported cases $C := (c_{ij})$, cases reported on the past seven days (lag-1, $\cdots$, lag-7) from presently $C_{-1} := (c_{i,j-1}), \ldots, C_{-7} := (c_{i,j-7})$, average air temperature $T := \mathbb{1}_S \otimes (t_j)$, and lag average relative humidity $H := \mathbb{1}_S \otimes (h_{j-25})$ corresponding to the cross-correlation result in Fig. 4. The notations $\mathbb{1}_S$ and $\otimes$ denote the column vector of size $S$ whose entries are 1 and the Kronecker product between two matrices, respectively. The final size of our observations is the entire time window length minus the maximal autoregressive lag, which is $N := 290 - 7 = 283$ (i.e. from March 8 until December 15, 2020). Let us denote $\beta_0$ as the intercept, $\beta_{\text{ind}} := (\beta_1, \ldots, \beta_{S-1})$ as the individual-specific effects (cut down by

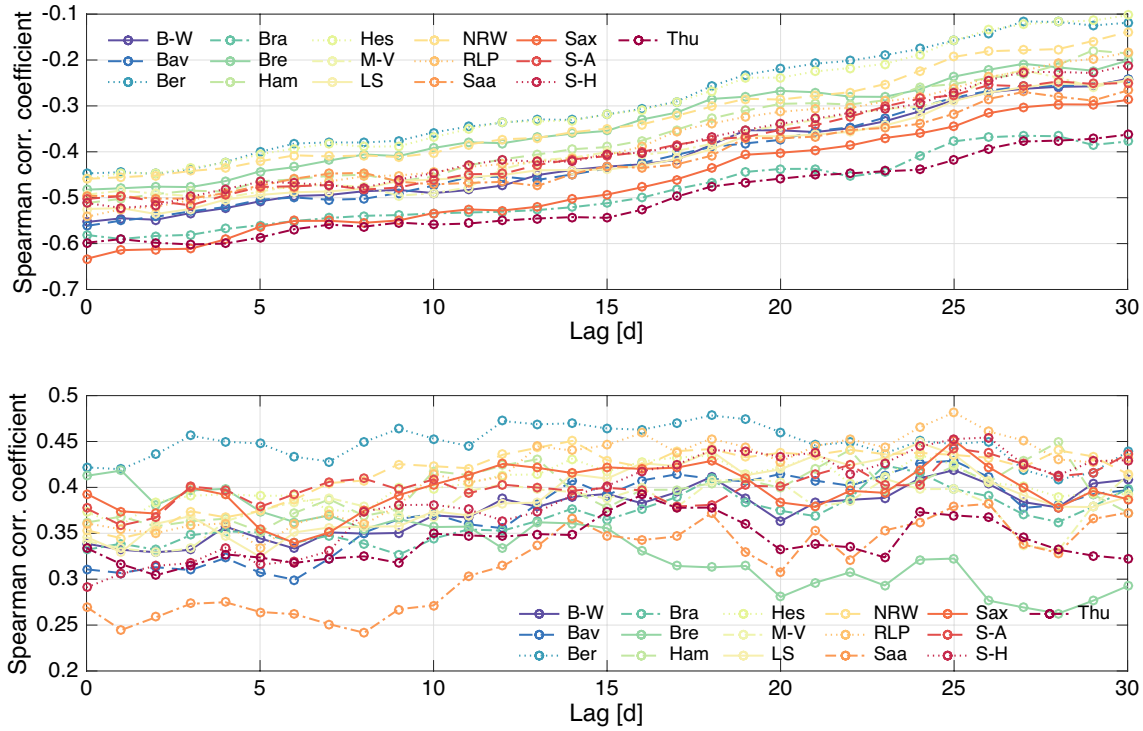



**Figure 4.** Spearman-rank correlation coefficients between daily cases from all states in Germany with the average temperature (above) and average humidity (below) on a moving window of 290 observations. Averaging throughout the states obtains the minimum of − 0.5223 (temperature) and maximum of 0.4194 (humidity) corresponding to the lags 0 and 25, respectively.

one term to avoid linear dependence with the intercept), $\beta_{-i}$ (for $i = 1, \ldots, 7$) as the marginal effects of the lag incidence cases, $\beta_T$ as the marginal effect of the temperature, $\beta_H$ as the marginal effect of the relative humidity, and $\varepsilon = (\varepsilon_{ij})$ as the idiosyncratic error. The direct relationship among these covariates intends to not only skip additional transformations but also return direct marginal effects represented by the coefficients of the corresponding explanatory variables. This reads as

$$C = \beta_0 \mathbb{1}_{S\times N} + \sigma^{(0)} \mathbb{1}_N^\top \otimes [\beta_{\text{ind}}\, 0]^\top + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} C_{-i} + \sigma^{(8)} \beta_T T + \sigma^{(9)} \beta_H H + \varepsilon, \tag{1}$$

which folds

$$\begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{S1} & \cdots & c_{SN} \end{pmatrix} = \begin{pmatrix} \beta_0 & \cdots & \beta_0 \\ \vdots & \ddots & \vdots \\ \beta_0 & \cdots & \beta_0 \end{pmatrix} + \sigma^{(0)} \begin{pmatrix} \beta_1 & \cdots & \beta_1 \\ \vdots & \ddots & \vdots \\ \beta_{S-1} & \cdots & \beta_{S-1} \\ 0 & \cdots & 0 \end{pmatrix} + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} \begin{pmatrix} c_{1,1-i} & \cdots & c_{1,N-i} \\ \vdots & \ddots & \vdots \\ c_{S,1-i} & \cdots & c_{S,N-i} \end{pmatrix}$$
$$+ \sigma^{(8)} \beta_T \begin{pmatrix} t_1 & \cdots & t_N \\ \vdots & \ddots & \vdots \\ t_1 & \cdots & t_N \end{pmatrix} + \sigma^{(9)} \beta_H \begin{pmatrix} h_{-24} & \cdots & h_{N-25} \\ \vdots & \ddots & \vdots \\ h_{-24} & \cdots & h_{N-25} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1N} \\ \vdots & \ddots & \vdots \\ \varepsilon_{S1} & \cdots & \varepsilon_{SN} \end{pmatrix}.$$

The indicator parameters $\sigma^{(i)}$ take binary values and will serve to drop certain variables in the model specification (by value 0), whenever necessary. This model represents, perhaps, the simplest panel regression model in the following sense. The marginal effects of the lag incidence cases and those of the weather components could have been raised to matrices like in vector autoregression with exogenous variables (VAR-X) models[84]. Besides appending too many parameters (entries of the endogeneous matrices), which may lead to overfitting, VAR-X models also require all the explanatory variables to be covariance stationary (see[85] for details), which is rarely the case for disease and weather data in the subtropics. As the only random spatial pattern was observed from the incidence data for almost all observations, no essential state-crossing marginal effects were expected. State-dependent marginal effects for the weather components were also not considered due to data aggregation and limitation, also to the intention to have unified marginal effects that work on the national level. Moreover, all lags smaller than the optimal values for the weather components were not considered for complexity reduction. For the reason of having straight-forward marginal effects, prior transformations were not applied to any of the variables. Despite its simplicity, the model (1) treats omitted variable bias by including individual-specific effects. These are the simplest terms assuming that the omitted variables only have constant effects on the daily

COVID-19 cases in all the states. After all, the present study draws forth an outlook for compiling temperature and relative humidity data from all eligible stations as well as data of other confounding factors (e.g. other weather components, human mobility, employment opportunities, mapping of manufactures or public gatherings, etc) that not only add more explanatory variables but also clear up the heteroscedasticity issue.

**Model including incidence clustering.** Previous studies based their investigation on asking which ranges of weather components correctly predict incidence cases. This study asks a slightly different question: which ranges of incidence cases are correctly predicted by the existing values of the weather components. The values that fail to predict certain incidence cases due to insignificance would deem dropping. In[68], this clustering strategy was designed to eliminate the weather dependency on the zero incidence cases, handling the zero-inflation problem appropriately. In the context of COVID-19, some extreme cases might have never been related to weather, for example superspreading events[10–13] and indoor aerosol transmission[8,9]. The basic aim of the clustering is then to correctly place the role of weather where it should have never predicted such events. The use of a transient function to replace this functionality was inapplicable to us, for which bias may arise from the functional choice and its related extension strategy for prediction.

The clustering idea departs from stratifying the incidence data into $M$ clusters $(\Omega_k)_{k=1}^{M}$ separated by barriers $\theta := (\theta_k)_{k=1}^{M-1}$. In the closed forms, the clusters are given by $\Omega_k = \{c : \max\{0, \theta_{k-1}\} \leq c < \min\{\theta_k, \max_{i,j} c_{ij}\}\}$. Let us define the function $\delta_k(C;\theta) := (\mathbb{1}_{\Omega_k} c_{ij})$, where $\mathbb{1}_{\Omega_k}$ denotes the characteristic function, taking value 1 in case $c_{ij}$ belongs to $\Omega_k$ or 0 in case otherwise. Let us denote $P \circ Q = (p_{ij} q_{ij})$ as the Hadamard product between two matrices and define $T^k = T^k(\theta) := \delta_k(C;\theta) \circ T$, $H^k = H^k(\theta) := \delta_k(C;\theta) \circ H$. The latter return the original entries of the matrices $T$, $H$ in case their pairing incidence cases belong to the corresponding cluster or 0 in case otherwise. Under this decomposition it always holds $\sum_k T^k = T$ and $\sum_k H^k = H$. Including clustering, a new model revises model (1) in the following fashion

$$C = \beta_0 \mathbb{1}_{S\times N} + \sigma^{(0)} \mathbb{1}_N^\top \otimes \beta_{\text{ind}}^\top + \sum_{i=1}^{7} \sigma^{(i)} \beta_{-i} C_{-i} + \sum_{i=1}^{3} \sigma^{(7+i)} \beta_T^i T^{(i)} + \sum_{i=1}^{3} \sigma^{(10+i)} \beta_H^i H^{(i)} + \varepsilon. \tag{2}$$

Here, the incidence data were classified into three clusters ($M = 3$) on the basis of practicality to call for lower, middle, and upper cluster. In principle, the specification is not bound to such a small number as fitting would be better with more explanatory variables. However, questions regarding complexity and practical interpretations might arise when using a large number of clusters. On the present choice, when for instance $T^{(2)}$ has to be dropped due to insignificance, this simply means that the average temperature fails to predict incidence cases in the range defined by the middle cluster $\Omega_2$. This model then allows the lone cases to be "unexplained by temperature".

The fact that $T^k$ and $H^k$ change with the lower and upper barrier $\theta = (\theta_\text{l}, \theta_\text{u})$, so does the pooled estimator $\hat{\beta} = \hat{\beta}(\theta)$ where $\beta = (\beta_0, \beta_{\text{ind}}, \beta_{-1}, \ldots, \beta_{-7}, \beta_T^1, \ldots, \beta_H^3)$. Our aim is to find the optimal barriers such that the squared error between data $C = (c_{ij})$ and the model approximate $C[\hat{\beta}](\theta)$ achieves its minimum. Mathematically, the preceding statement translates to the following problem

$$\min_{\theta} \qquad \sum_{i,j} (c_{ij}[\hat{\beta}](\theta) - c_{ij})^2 \tag{3a}$$

$$\text{subject to} \quad \min_{i,j} c_{ij} \leq \theta_\text{l} \leq \theta_\text{u} \leq \max_{i,j} c_{ij}. \tag{3b}$$

The pooled estimator $\hat{\beta}$ follows from the straightforward formula in terms of matrix inverse and multiplication involving explanatory and response variable.

## Results

**Case–weather cross-correlation and case-specific auto-correlation.** Figure 4 represents the correlation coefficients on a moving window of 290 observations with time lags from 0 to 30 days for each state. Notice that the reported daily COVID-19 cases correlated negatively with the average temperature and positively with the average relative humidity. The magnitude of the correlation coefficient with average temperature shows decreasing trends with lag for all the states. With no lag introduced, the correlations are negative and significant for all the states (p-values from $6.27 \times 10^{-34}$ to $1.17 \times 10^{-15}$). Averaging the correlation coefficients throughout the states, the minimum of − 0.5223 was obtained. This negative correlation is comparable up to certain ranges of minimum, maximum and average temperature to the studies in Brazil (with both average ranging from 20.9 to 27 °C and maximum temperature from 23.1 to 34.2 °C in[57] and with average temperature ranging from 16.8 to 27.4 °C in[86]) as well as the data in 127 countries (with average temperature from − 17.8 to 42.9 °C in[87]). In New York[88], the correlation was positive and insignificant for average and minimum temperature but positive and insignificant for the maximum temperature. In Oslo, Norway[89], the correlation was negative and insignificant for all maximum, minimum, and average temperature with 14 days time lag, but positive and significant correlation was obtained for normal temperature with 0, 5, 6, and 14 days lag. The temperature in Oslo ranged from − 7.5 to 21.9 °C during the study period. COVID-19 cases in Russian Federation exhibited positive significant correlation with minimum (− 17.78 °C to 8.89 °C), maximum (0.56 °C to 27.2 °C) and average temperature (− 2.78 °C to 16.1 °C)[46].

As far as relative humidity is concerned, it can be observed from Fig. 2 that its average varies from 39.38 to 93.53%. The best lag was found 25 days with the correlation coefficient value of 0.4194 from averaging throughout

| $\rho$ | lag-0 | lag-1 | lag-2 | lag-3 | lag-4 | lag-5 | lag-6 | lag-7 |
|---|---|---|---|---|---|---|---|---|
| lag-0 | 1 | | | | | | | |
| lag-1 | 0.87, 0.83 | 1 | | | | | | |
| lag-2 | 0.83, 0.81 | 0.87, 0.83 | 1 | | | | | |
| lag-3 | 0.80, 0.79 | 0.83, 0.81 | 0.87, 0.83 | 1 | | | | |
| lag-4 | 0.79, 0.79 | 0.81, 0.79 | 0.83, 0.79 | 0.87, 0.83 | 1 | | | |
| lag-5 | 0.82, 0.80 | 0.80, 0.79 | 0.81, 0.79 | 0.83, 0.80 | 0.87, 0.83 | 1 | | |
| lag-6 | 0.87, 0.82 | 0.83, 0.80 | 0.80, 0.79 | 0.80, 0.79 | 0.83, 0.80 | 0.87, 0.83 | 1 | |
| lag-7 | 0.89, 0.83 | 0.87, 0.82 | 0.83, 0.79 | 0.79, 0.78 | 0.80, 0.79 | 0.83, 0.80 | 0.87, 0.83 | 1 |

**Table 1.** Pearson and Spearman-rank correlation coefficients from the incidence data, rounded to two digits after comma.

the states. With this lag, the correlations are positive and significant for all states (p-values from $2.98 \times 10^{-18}$ to $1.92 \times 10^{-8}$). For the relative humidity, different results preceded ours. A previous study in New York[88] concluded that average relative humidity was insignificantly negatively correlated with the daily new cases. It was found that average humidity was significantly negatively correlated and relative humidity was insignificantly negatively correlated with the number of the ICU daily patients, according to data from Milan (14–100% for relative humidity, 1–23 g $m^{-3}$ for average humidity), Florence (10% to 100% for relative humidity, 1 to 23 g $m^{-3}$ for average humidity) and Trento (16–100% for relative humidity, 1 to 25 g $m^{-3}$ for average humidity) in Italy[90]. Data from Brazil ranging from 69.5 to 90.8% with no lag[50,57] showed that the correlation was positive but not significant with minimum and maximum average humidity. Data from 127 countries[87] led to the conclusion that the relative humidity was correlated negatively and insignificantly with daily new cases.

Table 1 shows the case-specific auto-correlations. Generally, Pearson is higher than Spearman-rank correlation coefficient. In addition, both Pearson and Spearman-rank correlation coefficient are significant with minimum 0.78 (p-values $\simeq 0$). From the column of lag-0, the auto-correlation generally swings from a large value at lag-1, then minima at either lag-3 or lag-4, to another large value at lag-7. The same behavior can be observed from the columns lag-1 until lag-3 where decrement rules out the first 4 lags and minima were found at either lag 3 or 4 days from the time series. This finding will set a basis for those in the panel regression models, as seen shortly.

**Spatial auto-correlation.** Meanwhile previous studies much focused on aggregated data and variation of distances in the spatial weight matrix, this study computed the global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ for all time to see how the spatial pattern changes seasonally since the earliest infection. The corresponding computation results together with the 95% confidence interval $[\mathscr{I} - 1.93\sqrt{\mathbb{V}(\mathscr{I})}, I + 1.93\sqrt{\mathbb{V}(\mathscr{I})}]$ (respectively for $\mathscr{C}$) are presented in Fig. 5. Although the spatial pattern of the daily cases in all the states changes around with time, it is evident that randomness overwhelms the pattern for most of the time. The progression of p-values (especially below $\alpha$) indicates that, generally, no significant difference between Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ was observed except on the duration from November until mid of December where Geary's $\mathscr{C}$ shows more locally clustered spatial pattern.

The Moran's scatter plot for all the states in Germany was determined for all observations, see Fig. 6. For the sake of serial presentation, indexing the coordinates based on the quadrants is more favorable than plotting them. Overall, the results suggest that all the states show randomness with time in to which spatial pattern they belong. If one solely focuses on the recent observations (November 1 to December 15, 2020), then the following states have the tendency to occupy the following quadrants: Baden-Württemberg, Bavaria, Hesse, Thuringia (Q1); Brandenburg, Rhineland-Palatinate, Saxony-Anhalt (Q2); Hamburg, Mecklenburg-Vorpommern, Lower Saxony, Schleswig-Holstein (Q3); Berlin, Bremen, North Rhine-Westphalia, Saxony (Q4).

**Panel regression models.** Variable choices for model specification were investigated. The criteria are based on not only fit and complexity (information-type criterion) but also insignificance, negative marginal effects, and multicollinearity driven by certain variables. For the fit and complexity, a minimal value of Bayesian Information Criterion $\mathrm{BIC} = -2\log(L) + \log(N) \cdot k$[91] was sought. The first term of this criterion expresses maximization over the likelihood function $L$ generated from our model and the second term includes the observation size $N$ as well as the number of parameters $k$. Unlike Akaike Information Criterion (AIC)[92] that would have replaced $\log(N)$ by 2, BIC penalizes the number of parameters much more, especially for large observation sizes. Our study aims to drop certain variables toward cutting down BIC and amending insignificance as well as multicollinearity. The standard $t$-test was used for the significance test. Checking for multicollinearity follows from computing the Inverse Variance Inflation Factor (1/VIF) values for all explanatory variables except the constant. A 1/VIF measures one minus the coefficient of determination derived from an OLS-regression whereby the variable under test serves as the response while the others as the explanatory variables. In this sense, 1/VIF of a value smaller than the rule of thumb 0.1 shows multicollinearity driven by the tested variable[93]. In addition, the p-value of the $F$-statistic is monitored, which measures if the overall variables are simultaneously significant; of which smaller than $\alpha = 0.05$ indicates that they are. Not only can the model be designated to be better than just a constant, but multicollinearity can also be diagnosed. Johnston in[94] hinted the existence of multicollinearity as some p-values from $t$-tests are large while that from $F$-test is radically small, which agrees to the analytical inves-

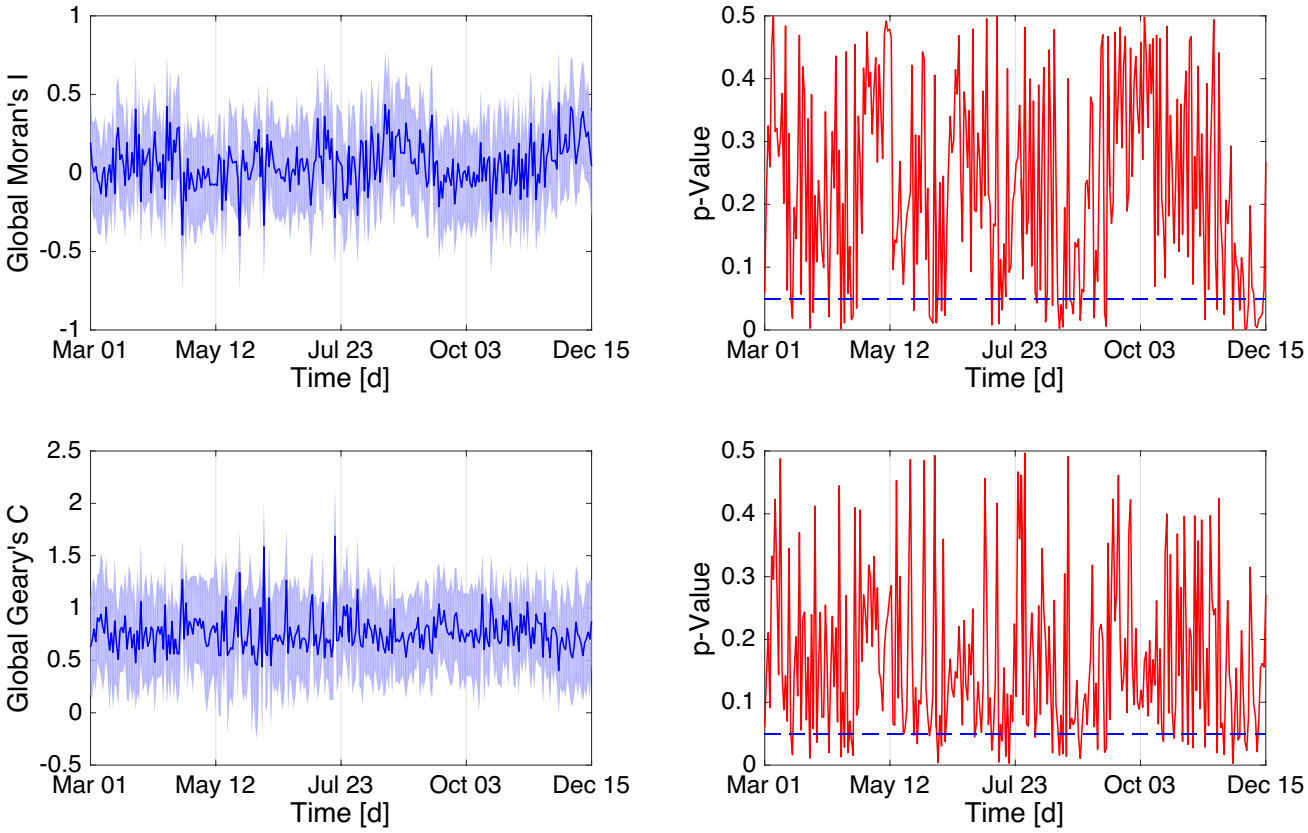
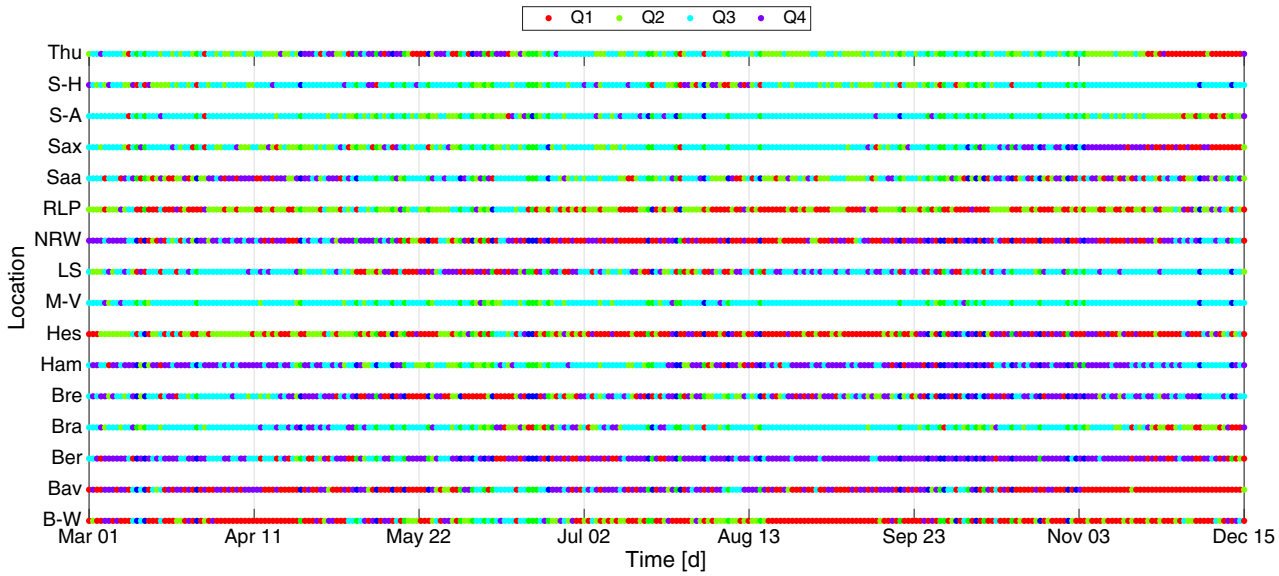



**Figure 5.** Global Moran's $\mathscr{I}$ and Geary's $\mathscr{C}$ computed on a daily basis together with the corresponding 95% confidence interval and p-Value (right) for significance. The blue dashed line represents the significance level $\alpha = 0.05$.




| State | B-W | Bav | Ber | Bra | Bre | Ham | Hes | M-V | LS | NRW | RLP | Saa | Sax | S-A | S-H | Thu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q1 | 62.22% | 88.88% | 24.44% | 22.22% | 0% | 0% | 55.55% | 0% | 0% | 28.88% | 28.88% | 17.77% | 42.22% | 11.11% | 0% | 46.66% |
| Q2 | 24.44% | 4.44% | 4.44% | 35.55% | 6.66% | 0% | 11.11% | 0% | 2.22% | 4.44% | 51.11% | 22.22% | 2.22% | 57.77% | 0% | 35.55% |
| Q3 | 6.66% | 0% | 8.88% | 35.55% | 40% | 53.33% | 11.11% | 100% | 91.11% | 24.44% | 8.88% | 26.66% | 2.22% | 28.88% | 100% | 13.33% |
| Q4 | 6.66% | 6.66% | 62.22% | 6.66% | 53.33% | 46.66% | 22.22% | 0% | 6.66% | 42.22% | 11.11% | 33.33% | 53.33% | 2.22% | 0% | 4.44% |

**Figure 6.** Classification into four quadrants (Q1, Q2, Q3, Q4) equivalent to Moran's scatter plot and the concurrence percentages from November 1 to December 15, 2020.

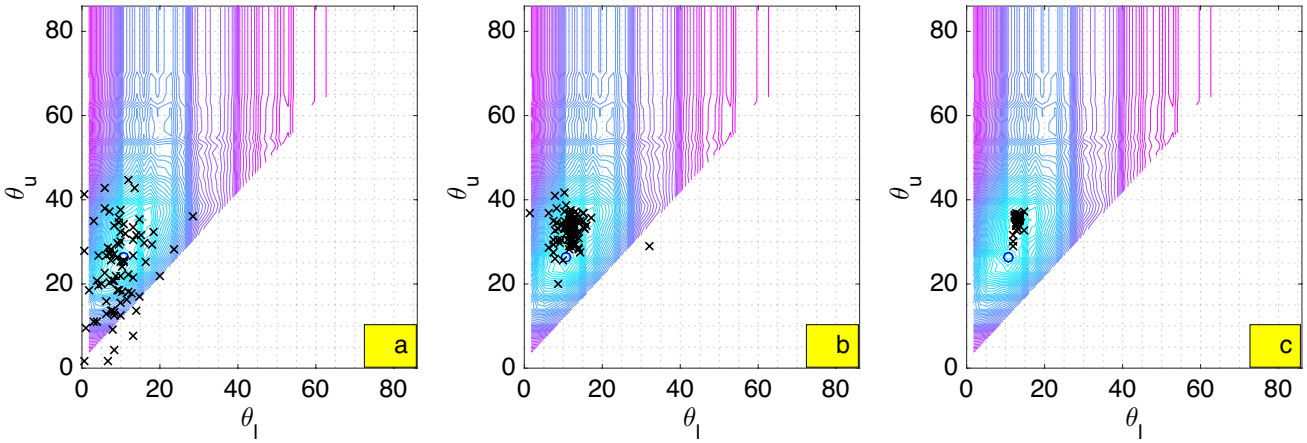



**Figure 7.** Computation of optimal barriers $(\theta_l, \theta_u) \approx (13.3645, 36.0597)$ for the clustering. Blue circle encodes the optimal barriers found by the brute-force computations on the $50 \times 50$ grid. The figures show the evolution of the locations of 100 players (black $\times$) converging to an optimal solution that does not overlap with the grid: (**a**) 5th iteration, (**b**) 10th iteration, (**c**) 20th iteration.

| **Model** (1) | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\sigma^{(i)} = 0$ for $i$ | | 0 | 0, 3 | 0, 4 | 0, 3, 4 | | |
| BIC | 24,049.74 | 23,933.03 | 23,924.93 | 23,953.11 | 23,946.74 | | |
| Issue | S0, S3 | S3, N4 | N4 | S3, N3 | | | |
| **Model** (2) | | | | | | | |
| $\sigma^{(i)} = 0$ for $i$ | | 0, 12 | 0, 12, 10 | 0, 12, 3 | 0, 12, 10, 3 | 0, 12, 10, 4 | 0, 12, 10, 3, 4 |
| BIC | 21,006.56 | 21,578 | 21,569.72 | 21,570.42 | 21,562.14 | 21,587.59 | 21,580.1 |
| Issue | S0, S3, S10, M12 | S3, S10 | S3 | S10 | N4 | N3, S3 | |

**Table 2.** Model specification under variable dropping. BIC values as well as corresponding issues leading to model exclusion are reported: S$i$, N$i$, M$i$ stand for insignificance, negative marginal effect, and multicollinearity driven by the corresponding variable ordered by $\sigma^{(i)}$, respectively.

tigation in[95,96]. Besides these aspects, if certain marginal effects would be consistent with our auto-correlation study were also checked. From Table 1, it is seen how cases in the past 7 days positively predict present cases with the least auto-correlations found from cases from the past 3 and 4 days. This led to dropping negative marginal effects corresponding to lag incidence cases that may occur due to a certain model specification.

To deal with the model including incidence clustering (2), the computation of optimal lower and upper barrier $(\theta_l, \theta_u)$ as in (3) is necessary. The characteristic functions embedded in the objective function make the optimization problem non-smooth. The brute-force computations of the objective function in the upper-left triangle of the $50 \times 50$ grid in the domain $[\min_{i,j} c_{ij}, \max_{i,j} c_{ij}]^2$ and a PSO algorithm[68] were put in comparison. From Fig. 7, PSO outperforms the brute-force computations in locating the optimal barriers that minimize the objective function, also in terms of computation time.

According to Table 2, the BIC value for the simple model (1) is relatively large, exacerbated by large degrees of freedom. The model including incidence clustering (2) gives the least BIC value due to a minimal likelihood function. Additionally, the insignificance of the entire individual-specific effects for both models was spotted. The rationale behind this can be connected to the fact that the entire profile of global and local spatial auto-correlation as well as the largest outbreak ("COVID-19 and weather situation in Germany" and "Spatial pattern") show randomness for almost all observations. Therefore, no state was worth constant recruitment (weighting) for its neighborhood to show a consistent spatial pattern throughout the observations.

Post-estimation diagnostics for all the models including those investigated during model specification were performed. Additional to the models including lag incidence cases and weather components, this study considered the models where either of these entities is present. The fitting results are presented in Table 3. For straightforward marginal effects and computation of optimal barries, the pooled estimator was considered subject to its inefficiency. The test was conducted via the comparison between fixed-effects and random-effects estimator and that between random-effects and pooled estimator. To the former, the two estimators were compared using Durbin–Wu–Hausman test[97,98], where the fixed-effects estimator is assumed to be consistent, and the random-effects estimator is efficient and assumed to follow a normal distribution. The null hypothesis suggests that the random-effects estimator is a consistent estimator regardless of the size of the data. According to Table 3, the p-value corresponding to the statistic greater than $\alpha = 0.05$ indicates that the random-effects estimator is equally

| | Val | StDev | *t* p-Val | 1/VIF | *F* p-Val | $R^2$ | Adj $R^2$ | D-W-H | Wo | B-P LM |
|---|---|---|---|---|---|---|---|---|---|---|
| Model (1) | | | | | | | | | | |
| $\beta_0$ | −.8742 | .3787 | .021 | | 0 | .8558 | .8556 | .5355 | 0 | 1 |
| $\beta_{-1}$ | 0.1827 | 0.0142 | 0 | 0.1603 | | | | | | |
| $\beta_{-2}$ | 0.0984 | 0.0128 | 0 | 0.2011 | | | | | | |
| $\beta_{-5}$ | 0.0514 | 0.0135 | 0 | 0.2033 | | | | | | |
| $\beta_{-6}$ | 0.2736 | 0.0149 | 0 | 0.1716 | | | | | | |
| $\beta_{-7}$ | 0.4145 | 0.0155 | 0 | 0.1645 | | | | | | |
| $\beta_T$ | −0.0295 | 0.0099 | 0.003 | 0.6390 | | | | | | |
| $\beta_H$ | 0.0246 | 0.0049 | 0 | 0.7054 | | | | | | |
| $\beta_0$ | 0.1949 | 0.0593 | 0.001 | | 0 | 0.8544 | 0.8543 | 0.5556 | 0 | 1 |
| $\beta_{-1}$ | 0.1918 | 0.0142 | 0 | 0.1619 | | | | | | |
| $\beta_{-2}$ | 0.1054 | 0.0128 | 0 | 0.2026 | | | | | | |
| $\beta_{-5}$ | 0.0604 | 0.0135 | 0 | 0.2056 | | | | | | |
| $\beta_{-6}$ | 0.2791 | 0.0150 | 0 | 0.1723 | | | | | | |
| $\beta_{-7}$ | 0.4166 | 0.0155 | 0 | 0.1646 | | | | | | |
| $\beta_0$ | −2.0767 | 0.7908 | 0.009 | | 0 | 0.3694 | 0.3691 | 1 | 0.0094 | 0 |
| $\beta_T$ | −0.5681 | 0.0185 | 0 | 0.7997 | | | | | | |
| $\beta_H$ | 0.2256 | 0.0097 | 0 | 0.7997 | | | | | | |
| Model (2) | | | | | | | | | | |
| $\beta_0$ | 5.9089 | 0.2162 | 0 | | 0 | 0.9148 | 0.9146 | 0.7646 | 0 | 1 |
| $\beta_{-1}$ | 0.1378 | 0.0109 | 0 | 0.1590 | | | | | | |
| $\beta_{-2}$ | 0.0716 | 0.0098 | 0 | 0.1998 | | | | | | |
| $\beta_{-5}$ | 0.0337 | 0.0104 | 0 | 0.2031 | | | | | | |
| $\beta_{-6}$ | 0.1636 | 0.0117 | 0 | 0.1667 | | | | | | |
| $\beta_{-7}$ | 0.2866 | 0.0123 | 0 | 0.1543 | | | | | | |
| $\beta_T^l$ | −0.1261 | 0.0076 | 0 | 0.4755 | | | | | | |
| $\beta_T^m$ | 0.3158 | 0.0224 | 0 | 0.4380 | | | | | | |
| $\beta_H^l$ | −0.0528 | 0.0026 | 0 | 0.3687 | | | | | | |
| $\beta_H^u$ | 0.2033 | 0.0047 | 0 | 0.6981 | | | | | | |
| $\beta_0$ | 1.8381 | 0.3594 | 0 | | 0 | 0.8682 (within)<br>0.9558 (between)<br>0.8692 (overall) | | 0 | 0.0097 | 0 |
| $\beta_T^l$ | −0.2088 | 0.0092 | 0 | 0.4927 | | | | | | |
| $\beta_T^2$ | −0.1010 | 0.0292 | 0.001 | 0.3878 | | | | | | |
| $\beta_T^3$ | −0.6897 | 0.1037 | 0 | 0.4472 | | | | | | |
| $\beta_H^l$ | 0.0524 | 0.0046 | 0 | 0.1785 | | | | | | |
| $\beta_H^2$ | 0.2627 | 0.0051 | 0 | 0.1243 | | | | | | |
| $\beta_H^3$ | 0.5608 | 0.0085 | 0 | 0.3258 | | | | | | |

**Table 3.** Fitting results and diagnostics for the models (1) and (2). The abbreviations stand for the following: Val (value), StDev (standard deviation), *t* p-Val (p-value of the *t*-test for the variable significance), 1/VIF (Inverse Variance Inflation Factor for multicollinearity), *F* p-Val (p-value of the *F*-test for the overall variable significance), $R^2$ (coefficient of determination), Adj $R^2$ (adjusted coefficient of determination), D–W–H (p-value of Durbin–Wu–Hausman test for random-effects vs. fixed-effects estimator), Wo (p-value of Wooldridge test for the serial correlation), B–P LM (p-value of Breusch–Pagan test for random effect vs pooled estimator).

consistent as the fixed-effects estimator. The two estimators for all presented models confirm equivalence except for model (2) where only weather components are present. For this case, the fixed-effects estimator was kept to handle consistency and panel effect. To the latter, Breusch–Pagan Lagrange Multiplier test was done under no panel effect as the null hypothesis[99], i.e., the model under the random-effects estimator returns zero variance in the state-dependent errors. Apparently, no panel effect was observed for all models except for those that include only weather components, in which case either random-effects or fixed-effects estimator is preferable. The inefficiency of the presented pooled, random-effects, and fixed-effects estimator is confirmed as serial correlation in all the state-dependent errors occurred. Wooldridge test[100] showed this. Therefore, a caveat remains for all models that their standard deviations of the coefficients are smaller and $R^2$'s are larger than they should be. After all, the pooled estimator is always consistent, even for a relatively small data size. As final practical remarks from the models, all the lag incidence cases give the waving effects in terms of lag where the cases 5 days and

7 days from presently predict the present cases the least and the most, respectively. Keeping the lag incidence cases, the weather components from model (1) give a consistent prediction with that from the cross-correlation study. Together with clustering, the marginal effects of weather were corrected for model (2). It was observed that temperature fails to predict cases in the upper cluster while relative humidity fails to cases in the middle cluster. Temperature seems to give a larger positive marginal effect for the middle cluster while relative humidity a negative smaller marginal effect for the lower cluster.

As far as predictive performance is concerned, several findings can be highlighted. As the larger models exhibit no more issues with insignificance and multicollinearity, neither do the smaller models. For the model variant (1), the smaller models gain $R^2 \approx 0.8544$, BIC $\approx$ 23,972.15 (only lag incidence cases) and $R^2 \approx 0.3694$, BIC $\approx$ 30585.35 (only weather components), respectively. Meanwhile the model including only weather components shows the poorest performance; its BIC value is also radically larger than that of the model including only lag incidence cases. For the model (1), the impact of weather is rather small, as the decrease of temperature from a reference value e.g. $T \approx 20\,^{\circ}\mathrm{C}$ to $T \approx 10\,^{\circ}\mathrm{C}$ (i.e. by 50%) is associated to the increase of COVID-19 cases for all states from e.g. $C \approx 20$ by $(|\beta_T|10/20) \cdot 100\% \approx 1.475\%$. When the lag incidence cases were dropped, the increase changes to $(0.5681 \cdot 10/20) \cdot 100\% \approx 28\%$. Moreover, the increase of relative humidity from 60 to 80% (by 33%) is associated to the increase of the cases from $C \approx 20$ by 2.46% (with lag incidence cases) and 22.56% (without lag incidence cases). The overall impression indicates the superiority of the model with only lag incidence cases when one designates fit to significantly matter than the number of parameters. For the model including incidence clustering (2), a different profile was obtained when only using non-dropped weather components: $R^2 \approx 0.7948$, BIC $\approx$ 25517.61. Here, a significant improvement under incidence clustering becomes evident. Surprisingly, the model including the entire weather components even outperforms that including only lag incidence cases by fit and complexity: $R^2 \approx 0.8692$, BIC $\approx$ 23494.94. All marginal effects corresponding to the temperature matrices are negative, and those corresponding to the relative humidity matrices are positive. It was observed that the temperature returns the smallest marginal effect on the COVID-19 cases in the middle cluster and relative humidity in the lower cluster. Besides the significance of the marginal effects, even no multicollinearity was observed. Apart from this, when the predictive ability is evaluated by $R^2$ and BIC amending multicollinearity and inconsistent predictors, it is still argued that combining lag incidence cases and weather components serve as the best models as presented in Table 3. The corresponding graphical fitting can be seen in Fig. 8.

## Discussion

In this study, lags from the cross-correlation between average temperature and relative humidity were extracted to synthesize suitable variables in the regression models. Additionally, case-specific auto-correlation supports the model specification where lag-3 and lag-4 incidence would rather be insignificant predictors for the present incidence. Spatial auto-correlation using global Moran's $\mathscr{I}$ and global Geary's $\mathscr{C}$ was investigated in the framework of analyzing the spatial effect in COVID-19 transmission. The global measures indicate random spatial patterns most of the time, except there were either local clusters or dispersion in recent observations from November 1 to December 15, 2020. Moran's scatter plot was then used to disclose the local behavior of the spatial pattern. The result shows that the distribution of the hot spots and cold spots generally changed with time. The random spatial pattern justifies the model specification where the individual- or state-specific effects that would have served to endow specific states with constant weighting factors, were dropped.

In the simple random-effects model, the average temperature and lag relative humidity were shown to affect the incidence significantly, however, the resulting coefficient of determination is comparably much smaller than whenever only lag incidence cases were used; panel effect also raises in the former case. For the reason of placing the correct role of weather in predicting certain ranges of incidence, the weather components were grouped with the aid of a clustering strategy. The new clustering-integrated model accompanied by optimal barriers shows good agreement with the data whereby weather components outperform lag incidence cases in the prediction. On this matter, the fixed-effects estimator was the only presumably consistent estimator that also tackles the panel effect. For all models, it was observed that every explanatory variable competes against the others to be a significant predictor. Therefore, model choice together with its consequences (marginal effects), depend entirely on the decision-maker. Marginal effects can be guidance when a model is chosen a priori. When $R^2$ and BIC matter a lot, our recommendation is to opt for the clustering-integrated model with lag incidence cases and lag weather components. There it was found that temperature and relative humidity have negative, relatively small marginal effects on the cases in the lower cluster (below 13 cases per 100,000 inhabitants); the temperature has a large positive marginal effect on the cases in the middle cluster (between 13 and 36 cases per 100,000 inhabitants) and no marginal effect on the upper cluster (above 36 cases per 100,000 inhabitants); relative humidity has a large positive marginal effect on the upper cluster but none on the middle cluster. The clustering-integrated model with only weather components is recommended when weather receives more privilege than lag incidence cases. Our result is consistent with the cross-correlation study that temperature has negative marginal effects while relative humidity has positive marginal effects on the incidence in all clusters. The middle cluster receives the smallest marginal effect from temperature and the lower cluster from relative humidity. This hints physical consequences that temperature can only predict incidence cases during hot (summer) and cold season (winter), where cases clearly distinguish against each other from the data, not during transitional seasons (spring and fall). Relative humidity, on the other hand, is less likely to predict sinking cases during the hot season.

## Conclusion

This study focused on the interrelationship between two weather components overlapping in many previous studies (average temperature and relative humidity) and COVID-19 incidence in Germany. Cross-correlation, case-specific auto-correlation, and spatial auto-correlation analysis were done to determine suitable variables

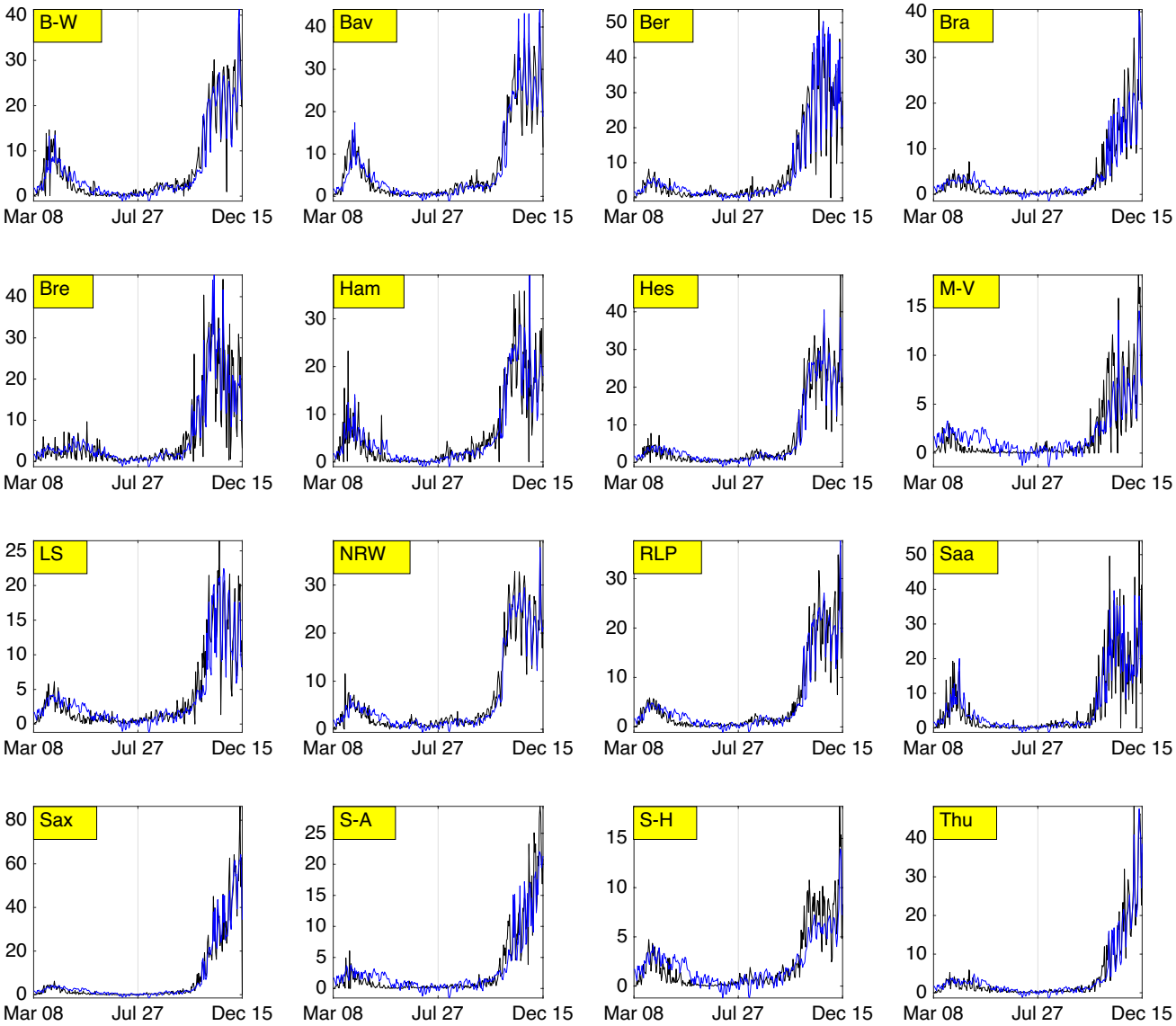



**Figure 8.** Fitting result (in blue) for the model including incidence clustering.

and to explain the negligible panel effect in the panel random-effects models. In addition, the findings from the spatial auto-correlation provide the placement of the 16 states in the four quadrants from Moran's scatter plot and appropriate policy regarding traveling restrictions. The increasing demand for confounding factors to explain various incidence levels has been neutralized by the aid of incidence clustering. This strategy supports the idea of considering only certain hypothetical factors predicting COVID-19 incidence and general regression modeling wherein explanatory variables are limited. This localization of incidence that is correctly predicted by the two weather components has profound implications for public health authorities. The modeling does not only determine the extent of the prediction via marginal effects but also paves the way for precautionary actions amidst upcoming weather.

## Data availability

All the data sources have been included in "COVID-19 and weather situation in Germany".

Received: 18 January 2021; Accepted: 16 May 2021
Published online: 28 May 2021

## References

1. Belser, J. A., Eckert, A. M., Tumpey, T. M. & Maines, T. R. Complexities in ferret influenza virus pathogenesis and transmission models. *Microbiol. Mol. Biol. Rev.* **80**, 733–744 (2016).
2. Storch, G. A. Diagnostic virology. *Clin. Infect. Dis.* **31**, 739–751 (2000).

3. Steinmeyer, S. H., Wilke, C. O. & Pepin, K. M. Methods of modelling viral disease dynamics across the within- and between-host scales: The impact of virus dose on host population immunity. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 1931–1941 (2010).
4. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487 (2008).
5. World Health Organization. Coronavirus disease (COVID-19): How is it transmitted? https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted (2020). Accessed 19 December 2020.
6. Azuma, K. *et al.* Environmental factors involved in SARS-CoV-2 transmission: Effect and role of indoor environmental quality in the strategy for COVID-19 infection control. *Environ. Health Prev. Med.* **25**, 1–16 (2020).
7. Wijaya, K. P. *et al.* An epidemic model integrating direct and fomite transmission as well as household structure applied to COVID-19. *J. Math. Ind.* **11**, 1–26 (2021).
8. Karia, R., Gupta, I., Khandait, H., Yadav, A. & Yadav, A. COVID-19 and its modes of transmission. *SN Compr. Clin. Med.*, 1–4 (2020).
9. Morawska, L. & Milton, D. It is time to address airborne transmission of Coronavirus Disease 2019 (COVID-19). *Clin. Infect. Dis.* **71**, 2311–2313 (2020).
10. Bouffanais, R. & Lim, S. Cities - try to predict superspreading hotspots for COVID-19. *Nature* **583**, 352–355 (2020).
11. Wong, F. & Collins, J. J. Evidence that coronavirus superspreading is fat-tailed. *Proc. Natl. Acad. Sci.* **117**, 29416–29418 (2020).
12. Kain, P. M., Childs, M. L., Becker, A. D. & Mordecai, E. A. Chopping the tail: How preventing superspreading can help to maintain COVID-19 control. *Epidemics* **34**, 100430 (2020).
13. Wang, L. *et al.* Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat. Commun.* **11**, 5006 (2020).
14. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
15. Ebrahim, S. H. & Memish, Z. A. COVID-19—The role of mass gatherings. *Travel Med. Infect. Dis.* **34**, 101617 (2020).
16. World Health Organization. WHO mass gathering COVID-19 risk assessment tool—Generic events. https://www.who.int/publications/i/item/10665-333185 (2020). Accessed 25 October 2020.
17. Assche, J. V., Politi, E., Dessel, P. V. & Phalet, K. To punish or to assist? Divergent reactions to ingroup and outgroup members disobeying social distancing. *Br. J. Soc. Psychol.* **59**, 594–606 (2020).
18. Belosi, F., Conte, M., Gianelle, V., Santachiara, G. & Contini, D. On the concentration of SARS-CoV-2 in outdoor air and the interaction with pre-existing atmospheric particles. *Environ. Res.* **193**, 110603 (2021).
19. Tung, N. T. *et al.* Particulate matter and SARS-CoV-2: A possible model of COVID-19 transmission. *Sci. Total Environ.* **750**, 141532 (2021).
20. Lei, H., Xu, X., Xiao, S., Wu, X. & Shu, Y. Household transmission of COVID-19—A systematic review and meta-analysis. *J. Infect.* **81**, 979–997 (2020).
21. Ooi, E. E. & Low, J. G. Asymptomatic SARS-CoV-2 infection. *Lancet Infect. Dis.* **20**, 996–998 (2020).
22. Lin, D. *et al.* Co-infections of SARS-CoV-2 with multiple common respiratory pathogens in infected patients. *Sci. China Life Sci.* **63**, 1–4 (2020).
23. Richardson, S. *et al.* Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* **323**, 2052–2059 (2020).
24. Kim, D., Quinn, J., Pinsky, B., Shah, N. H. & Brown, I. Rates of co-infection between SARS-CoV-2 and other respiratory pathogens. *JAMA* **323**, 2085–2086 (2020).
25. Boncristiani, H. F., Criado, M. F. & Arruda, E. Respiratory viruses. *Encycl. Microbiol.* **2009**, 500–518 (2009).
26. Dasaraju, P. V. & Liu, C. Infections of the respiratory system. in: *Medical Microbiology* 4th edn (ed Baron, S.) (University of Texas Medical Branch at Galveston, 1996).
27. Azekawa, S., Namkoong, H., Mitamura, K., Kawaoka, Y. & Saito, F. Co-infection with SARS-CoV-2 and influenza A virus. *IDCases* **20**, e00775 (2020).
28. Mossad, S. B. COVID-19 and flu: Dual threat, dual opportunity. *Clevel. Clin. J. Med.* **87**, 651–655 (2020).
29. Dowell, S. F. & Ho, M. S. Seasonality of infectious diseases and severe acute respiratory syndrome—What we don't know can hurt us. *Lancet Infect. Dis.* **4**, 704–708 (2004).
30. Shi, P. *et al.* Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Sci. Total Environ.* **728**, 138890 (2020).
31. Kronbichler, A. *et al.* Asymptomatic patients as a source of COVID-19 infections: A systematic review and meta-analysis. *Int. J. Infect. Dis.* **98**, 180–186 (2020).
32. Ozaras, R. *et al.* Influenza and COVID-19 coinfection: report of six cases and review of the literature. *J. Med. Virol.* **92**, 2657–2665 (2020).
33. Singh, B., Kaur, P., Reid, R. J., Shamoon, F. & Bikkina, M. COVID-19 and influenza co-infection: Report of three cases. *Cureus J. Med. Sci.* **12**, e9852 (2020).
34. Pormohammad, A. *et al.* Comparison of influenza type A and B with COVID-19: A global systematic review and meta-analysis on clinical, laboratory and radiographic findings. *Rev. Med. Virol.*, e2179 (2020).
35. Cai, Q. C. *et al.* Influence of meteorological factors and air pollution on the outbreak of severe acute respiratory syndrome. *Public Health* **121**, 258–265 (2007).
36. Chan, K. H. *et al.* The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* **2011**, 734690 (2011).
37. Casanova, L. M., Jeon, S., Rutala, W. A., Weber, D. J. & Sobsey, M. D. Effects of air temperature and relative humidity on coronavirus survival on surfaces. *Appl. Environ. Microbiol.* **76**, 2712–2717 (2010).
38. Sun, Z., Thilakavathy, K., Kumar, S. S., He, G. & Liu, S. V. Potential factors influencing repeated SARS outbreaks in China. *Int. J. Environ. Res. Public Health* **17**, 1633 (2020).
39. Gardner, E. G. *et al.* A case-crossover analysis of the impact of weather on primary cases of Middle East respiratory syndrome. *BMC Infect. Dis.* **19**, 1–10 (2019).
40. Altamimi, A. & Ahmed, A. E. Climate factors and incidence of middle east respiratory syndrome coronavirus. *J. Infect. Public Health* **13**, 704–708 (2020).
41. Cai, J. *et al.* Indirect virus transmission in cluster of COVID-19 cases. *Emerg. Infect. Dis.* **26**, 1343–1345 (2020).
42. Yeo, C., Kaushal, S. & Yeo, D. Enteric involvement of coronaviruses: Is faecal-oral transmission of SARS-CoV-2 possible?. *Lancet Gastroenterol. Hepatol.* **5**, 335–337 (2020).
43. Chin, A. W. H. *et al.* Stability of SARS-CoV-2 in different environmental conditions. *Lancet Microbe* **1**, e10 (2020).
44. Ahlawat, A., Wiedensohler, A. & Mishra, S. K. An overview on the role of relative humidity in airborne transmission of SARS-CoV-2 in indoor environments. *Aerosol Air Qual. Res.* **20**, 1856–1861 (2020).
45. Islam, A. R. T. *et al.* Effect of meteorological factors on COVID-19 cases in Bangladesh. *Environ. Dev. Sustain.*, 1–24 (2020).
46. Lasisi, T. T. & Eluwole, K. K. Is the weather-induced COVID-19 spread hypothesis a myth or reality? Evidence from the Russian federation. *Environ. Sci. Pollut. Res.*, 1–5 (2020).
47. Sarkodie, S. A. & Owusu, P. A. Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. *Environ. Res.* **191**, 110101 (2020).
48. Sil, A. & Kumar, V. N. Does weather affect the growth rate of COVID-19, a study to comprehend transmission dynamics on human health. *J. Saf. Sci. Resil.* **1**, 3–11 (2020).

49. Xie, J. & Zhu, Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci. Total Environ.* **724**, 138201 (2020).
50. Pan, J. *et al.* Warmer weather unlikely to reduce the COVID-19 transmission: an ecological study in 202 locations in 8 countries. *Sci. Total Environ.* **753**, 142272 (2020).
51. Ward, M. P., Xiao, S. & Zhang, Z. The role of climate during the COVID-19 epidemic in New South Wales, Australia. *Transbound. Emerg. Dis.* **67**, 2313–2317 (2020).
52. Tosepu, R. *et al.* Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Sci. Total Environ.* **725**, 138436 (2020).
53. Qi, H. *et al.* COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci. Total Environ.* **728**, 138778 (2020).
54. Guo, C. *et al.* Meteorological factors and COVID-19 incidence in 190 countries: An observational study. *Sci. Total Environ.* **757**, 143783 (2020).
55. Jahangiri, M., Jahangiri, M. & Najafgholipour, M. The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran. *Sci. Total Environ.* **728**, 138872 (2020).
56. Sharma, P., Singh, A. K., Agrawal, B. & Sharma, A. Correlation between weather and COVID-19 pandemic in India: An empirical investigation. *J. Public Affairs* **20**, e2222 (2020).
57. Rosario, D. K. A., Mutz, Y. S., Bernardes, P. C. & Conte-Junior, C. A. Relationship between COVID-19 and weather: Case study in a tropical country. *Int. J. Hyg. Environ. Health* **229**, 113587 (2020).
58. Mofijur, M. *et al.* Relationship between weather variables and new daily COVID-19 cases in Dhaka, Bangladesh. *Sustainability* **12**, 8319 (2020).
59. Bukhari, Q., Massaro, J., D'Agostino, R. & Khan, S. Effects of weather on coronavirus pandemic. *Int. J. Environ. Res. Public Health* **17**, 5399 (2020).
60. Rashed, E. A., Kodera, S., Gomez-Tames, J. & Hirata, A. Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: Multi-prefecture study in Japan. *Int. J. Environ. Res. Public Health* **17**, 5354 (2020).
61. Mecenas, P., Baston, R., Vallinoto, A. & Normando, D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. *PLoS One* **15**, e0238339 (2020).
62. Malki, Z. *et al.* Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos Solitons Fractals* **138**, 110137 (2020).
63. Federal Statistical Office. Current population. https://www.destatis.de/EN/Home/_node.html (2020). Accessed 05 January 2021.
64. Statistische Ämter des Bundes und der Länder. Bruttoinlandsprodukt (VGR) Ergebnisse der Volkswirtschaftlichen Gesamtrechnungen der Länder. https://www.statistikportal.de/en/node/649 (2020). Accessed 04 January 2021.
65. statista. Arbeitslosenquote in Deutschland nach Bundesländern. https://de.statista.com/statistik/daten/studie/36651/umfrage/arbeitslosenquote-in-deutschland-nach-bundeslaendern/ (2020). Accessed 04 January 2021.
66. Robert Koch Institute. Coronavirus disease 2019 (COVID-19): Daily situation report of the Robert Koch Institute. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html (2020). Accessed 31 December 2020.
67. Adams, A., Chen, X., Li, W. & Zhang, C. The disguised pandemic: The importance of data normalization in COVID-19 web mapping. *Public Health* **183**, 36–37 (2020).
68. Wijaya, K. *et al.* Learning from panel data of dengue incidence and meteorological factors in Jakarta, Indonesia. *Stoch. Environ. Res. Risk Assess.*, 1–20 (2020).
69. CDC-OpenData. Index of /climate_environment/CDC/. https://opendata.dwd.de/climate_environment/CDC/ (2020). Accessed 31 December 2020.
70. Xinhuanews. Germany's Bavaria declares emergency situation effective on Tuesday. http://www.xinhuanet.com/english/2020-03/16/c_138884534.htm (2020). Accessed 09 March 2021.
71. J. Mladek (Nordkurier). Bavaria imposes curfew! https://www.nordkurier.de/politik-und-wirtschaft/bayern-verhaengt-ausgangssperre-2038792303.html (2020). Accessed 04 January 2021.
72. Richard Connor. German states move closer to near-total lockdowns. https://www.dw.com/en/german-states-move-closer-to-near-total-lockdowns/a-52863482 (2020). Accessed 09 March 2021.
73. WELT. First major German city introduces mandatory masking. https://www.welt.de/politik/deutschland/article206911189/Coronavirus-Erste-deutsche-Grossstadt-fuehrt-Maskenpflicht-ein.html (2020). Accessed 04 January 2021.
74. L. Riekhoff and A. Sommer (streiflichter). Coronavirus in the Coesfeld district: 59 new infections with the coronavirus. https://www.streiflichter.com/lokales/coesfeld/coronavirus-kreis-coesfeld-aktuelle-fallzahlen-region-13643612.html (2020). Accessed 04 January 2021.
75. Deutsche Welle (DW). Coronavirus: Over 600 people test positive at German slaughterhouse. https://www.dw.com/en/coronavirus-over-600-people-test-positive-at-german-slaughterhouse/a-53846038 (2020). Accessed 04 January 2021.
76. BBC. Coronavirus: Thousands protest in Germany against restrictions. https://www.bbc.com/news/world-europe-53622797 (2020). Accessed 04 January 2021.
77. Das, P. & Choudhuri, T. Decoding the global outbreak of COVID-19: The nature is behind the scene. *Virus Dis.* **31**, 1–7 (2020).
78. Riddell, S., Goldie, S., Hill, A., Eagles, D. & Drew, T. W. The effect of temperature on persistence of SARS-CoV-2 on common surfaces. *Virol. J.* **17**, 1–7 (2020).
79. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
80. Geary, R. C. The contiguity ratio and statistical mapping. *Inc. Stat.* **5**, 115–146 (1954).
81. Cliff, A. & Ord, J. *Spatial Autocorrelation. Monographs in Spatial and Environmental Systems Analysis* (Pion, 1973).
82. Anselin, L. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Spat. Anal. Perspect. GIS* **4**, 111–116 (1996).
83. Sokal, R. R., Oden, N. L. & Thomson, B. A. Local spatial autocorrelation in a biological model. *Geogr. Anal.* **30**, 331–354 (1998).
84. Ocampo, S. & Rodríguez, N. An introductory review of a structural VAR-X estimation and applications. *Revista Colombiana de Estadística* **35**, 479–508 (2012).
85. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis* (Springer, 2005).
86. Prata, D. N., Rodrigues, W. & Bermejo, P. H. Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil. *Sci. Total Environ.* **729**, 138862 (2020).
87. Yuan, J. *et al.* Non-linear correlation between daily new cases of COVID-19 and meteorological factors in 127 countries. *Environ. Res.* **193**, 110521 (2020).
88. Bashir, M. F. *et al.* Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci. Total Environ.* **728**, 138835 (2020).
89. Menebo, M. M. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. *Sci. Total Environ.* **737**, 139659 (2020).
90. Lolli, S., Chen, Y. C., Wang, S. H. & Vivone, G. Impact of meteorological conditions and air pollution on COVID-19 pandemic transmission in Italy. *Sci. Rep.* **10**, 1–15 (2020).
91. Raftery, A. E. Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).

92. Akaike, H. Information theory and an extension of the maximum likelihood principle. in *Selected Papers of Hirotugu Akaike*. Springer Series in Statistics (Perspectives in Statistics) (eds. Parzen, E. *et al.*) (Springer, 1998).
93. Mansfield, E. R. & Helms, B. P. Detecting multicollinearity. *Am. Stat.* **36**, 158–160 (1982).
94. Johnston, J. *Econometric Methods* 2nd edn (McGraw Hill Higher Education, 1972).
95. Farrar, D. E. & Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **49**, 92–107 (1967).
96. Willis, C. E. & Perlack, R. D. Multicollinearity: Effects, symptoms, and remedies. *J. Northeast. Agric. Econ. Council* **7**, 55–61 (1978).
97. Hausman, J. A. Specification tests in econometrics. *Econometrica* **46**, 1251–1271 (1978).
98. Davidson, R. & MacKinnon, J. G. *Estimation and Inference in Econometrics. OUP Catalogue* (Oxford University Press, 1993).
99. Baltagi, B. H. & Li, Q. A lagrange multiplier test for the error components model with incomplete panels. *Econom. Rev.* **9**, 103–107 (1990).
100. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data* (MIT Press, 2002).

## Acknowledgements

This research has been supported by the Ministry of Research, Technology, and Higher Education/National Research and Innovation Agency, Indonesia through the PUPT research grant scheme 2021.

## Author contributions

N.C.G. and K.P.W. drafted the work and performed the computations; D.A, M.A., K.K.W.H.E, and N.C.G. interpreted data and conducted preliminary analysis; D.A. and K.P.W. managed funding acquisition. All authors reviewed earlier drafts and approved its final version.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to D.A.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.