BMC Public Health

RESEARCH

Open Access

Measuring the impact of climate and air quality on COVID-19 transmission dynamics: a tale of two cities in pandemic preparedness

Gayathry R. Menon[1], Sayooj Aby Jose[2*], Asaph Young Chun[2*], Anuwat Jirawattanapanit[3] and Hsuan-Wei Lee[1]

Abstract

This comparative study investigates the non-linear influence of nine meteorological and air quality factors on COVID-19 transmission in Daegu and Seoul, South Korea, from January to July 2020. Employing Generalized Additive Models (GAMs), percentile-based risk zone identification, and K-means clustering, our analysis demonstrated that environmental factors are highly region-specific, challenging universal public health approaches. In Daegu, the risk profile was strongly shaped by climatic factors (GAM Explained Deviance: 63.0%), exhibiting a U-shaped temperature effect with elevated risk at both extremes. A primary high-risk zone appeared in cool, moderately dry conditions, averaging 278.8 cases per day, consistent with the Cold environmental regime. Conversely, Seoul's risk profile indicated significant air pollutant influence (GAM Explained Deviance: 48.3%), showing monotonic positive relationships for temperature, NO_2, and O_3. Crucially, the data revealed a strong positive correlation between temperature and O_3 (r = +0.794), indicating warmer periods feature elevated O_3, increasing respiratory sensitivity. A primary high-risk zone in Seoul occurred in warm, humid conditions (averaging 29.5 cases per day), reinforcing higher counts in Warm environmental regimes directly opposite to Daegu. These findings underscore the need for tailored, data-driven environmental early warning systems (EEWS) specific to regional characteristics, enhancing preparedness for future pandemic threats.

Keywords Atmospheric pollution, COVID-19, Generalized additive models (GAM), Meteorological determinants, Non-linear dynamics, South Korea

Introduction

Infectious diseases continue to pose a significant public health challenge, placing substantial burdens on healthcare systems even in industrialized nations like South Korea [1], with epidemiological patterns evolving over time. The World Health Organization (WHO) consistently projects the regular emergence of novel diseases, emphasizing the constant need for vigilance, adaptive public health strategies, and continuous scientific inquiry into disease transmission determinants [2].

The COVID-19 pandemic exposed deep systemic vulnerabilities, forcing a rapid reassessment of traditional models for pathogen dispersal [3–6]. Beyond immediate

*Correspondence:
Sayooj Aby Jose
sayooj@snu.ac.kr
Asaph Young Chun
swiftasaph_snu@snu.ac.kr
[1]Department of Biostatistics & Health Data Science, Lehigh University, Bethlehem, USA
[2]Institute for Pandemic Sciences AI.celerator (IPSAI), Artificial Intelligence Institute, Seoul National University, Seoul, South Korea
[3]Department of Mathematics, Faculty of Education, Phuket Rajabhat University, Phuket, Thailand

crisis management, the pandemic revealed a sophisticated interplay between atmospheric conditions and viral transmission, establishing environmental epidemiology as a cornerstone of modern public health inquiry.

Daegu, in particular, was the epicenter of South Korea's initial and most severe COVID-19 outbreak, largely due to a single superspreading event, which led to thousands of infections in a short period, with Daegu accounting for a substantial 63.7% of the entire country's caseload by March 24, 2020 [7]. In contrast, Seoul, the nation's political and economic center with nearly a quarter of the population, experienced a different pattern of initial spread [8]. While studies like Peng Shi et al. [9] identified temperature as a significant environmental drivers potentially contributing to COVID-19 transmission globally, a common limitation is the reliance on linear assumptions. Such models often rely on simplistic correlations or limited environmental variables, oversimplifying the complex, non-linear dynamics between environmental factors and viral transmission and failing to capture true intricate interacting effects. Such approaches often overlook the significant role of air pollutants, which are known to impact respiratory health and susceptibility to viral infections [8, 10–12]. Lim et al. (2021) [13] previously suggested contradictory environmental influences, indicating a complex effect that required further investigation. The application of high-parameter Generalized Additive Models (GAMs) is a well-established approach for capturing the multi-factorial and non-linear nature of COVID-19 transmission. Recent studies in diverse geographical contexts have successfully utilized GAMs with 10 to 12 environmental parameters, demonstrating that the primary limitation on parameter count is multicollinearity rather than an absolute cap. For instance, Zhao et al. (2022) [14] modelled transmission across four Chinese urban agglomerations using a 10-parameter GAM, reporting explained deviance (ED) values ranging from 35.2% to 75.4% depending on regional spatial heterogeneity. Similarly, Sahoo et al. (2021) [15] employed a 12-parameter Gaussian GAM with a log-transformation to successfully isolate the independent effects of four air pollutants and eight meteorological factors in India. Furthermore, Hoang et al. (2021) [16] utilized a 10-parameter framework (6 pollutants and 4 meteorological factors) specifically within South Korea to contrast environmental factors associated with increased between the Seoul and Daegu clusters. By carefully screening for multicollinearity, our 9-parameter model aligns seamlessly with these peer-reviewed frameworks. Moreover, our model's performance achieving an explained deviance of 63.0% for Daegu and 48.3% for Seoul falls closely within established benchmarks, confirming the robustness of our regional explanatory methodology. While social and behavioral models have clarified how distancing and network interventions alter epidemic spread [17, 18]. We introduce critical methodological innovations and an expanded scope for a deeper exploration of specific environmental variables and their independent and interactive impacts. To ensure clarity regarding our contributions, Table 1 will directly compare our methodology and findings with Lim et al. [13].

Table 1 Comparison of methodological frameworks: Advancements of the current Generalized Additive Model (GAM) approach over the baseline correlational study (Lim et al., 2021) in assessing environmental factors lead to of COVID-19 transmission

Methodological aspect	**Baseline study: Lim et al. (2021) [13]**	**Present study**
Modeling Technique	Correlation analysis (Spearman, Kendall), t-tests	Generalized Additive Model (GAM) for robust explanatory modeling, supplemented with K-means clustering.
Regional Focus	Seoul (SMR) and Daegu-Gyeongbuk (DGR) regions	Seoul (SMR) and Daegu-Gyeongbuk (DGR) regions
Environmental Factors	Limited meteorological variables (temp, humidity, sunshine, rainfall)	Comprehensive suite of 9 environmental variables, including meteorological factors and key air pollutants.
Temporal Scope	February to July 2020	January to July 2020, covering the complete first wave of the pandemic in Daegu and Seoul.
Relationship Analysis	Primarily linear correlations	Investigation of complex, non-linear relationships using penalized smoothing splines within the GAM.
Study Objective	Descriptive and correlational study	Explanatory modeling to identify, quantify, and visualize environmental risk zones and non-obvious transmission risk factors.
Air Quality Impact	Qualitative assessment, focused mostly on ozone	Quantitative inclusion and assessment of four major air pollutants (CO, NO_2, SO_3, O_3) within the multi-variable GAM.

The main contributions of this study are to quantify how combined environmental conditions modify COVID-19 risk in each region, map high-risk micro-regions for targeted interventions, and demonstrate a transferable, data-driven prototype for environmental early warning systems that can inform public health planning for future respiratory pandemics [2, 19–21].

The conceptual framework of this study illustrated in Fig. 1 delves into how specific environmental elements namely meteorological conditions and air pollution influence the transmission dynamics of respiratory viruses like SARS-CoV-2. Each environmental factor plays a distinct role in shaping these transmission patterns. Lower temperatures, for instance, are known to enhance the stability of coronaviruses, allowing them

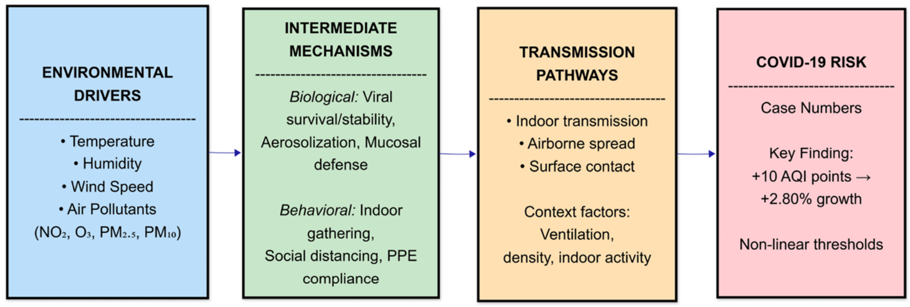

Fig. 1 Conceptual framework illustrating the pathways through which environmental factors influence COVID-19 transmission dynamics in Seoul and Daegu metropolitan regions, South Korea

to persist longer on surfaces and in the air [22]. Moreover, colder weather often encourages people to gather indoors, where ventilation is limited, thereby increasing the risk of transmission [23]. The influence of climate variables on virus viability and transmission dynamics is illustrated in Fig. 1. Relative humidity is another critical variable, influencing how respiratory droplets behave in the air. Low humidity can lead to quicker evaporation of droplets into smaller, more enduring aerosols, increasing airborne transmission potential, while higher humidity may cause larger droplets to settle more rapidly [24, 25]. Furthermore, the impact of wind speed is more complex, as it can help disperse viral particles outdoors. However, it might also contribute to aerosol recirculation in densely built urban areas or poorly ventilated indoor spaces. Alongside these meteorological factors, ambient air pollutants further associated with transmission risks. These pollutants can potentially serve as carriers for virus particles, facilitating their spread through the air, while also compromising respiratory health. In particular, a 10-point increase in the Air Quality Index has been linked to a 2.80% point rise in the daily COVID-19 growth rate in Chinese cities [12]. Ultimately, COVID-19 case counts reflect how weather and air quality alter both viral biology and human behavior. For instance, cold temperatures or heavy pollution drive people into poorly ventilated indoor spaces, significantly increasing the risk of infection. At the same time, poor air quality warnings can prompt protective habits, such as mask wearing or avoiding the outdoors. Finally, government interventions like lockdowns often coincide with these environmental shifts, further altering transmission patterns.

We selected temperature, relative humidity, wind speed, and key air pollutants due to their established effects on respiratory disease transmission and public health interventions. Since environmental factors often exhibit non-linear and threshold-based relationships, a flexible statistical framework is essential to identify specific thresholds where transmission risk intensifies. Climatically, Daegu experiences milder winters and hotter summers, whereas Seoul exhibits larger daily temperature variations, stronger winds, and erratic pressure profiles. Both cities also face sudden, high-magnitude seasonal disruptions, such as monsoon humidity surges and abrupt atmospheric pressure drops [26]. These rapidly shifting patterns demonstrate why simple linear models are inadequate for capturing the true complexity of environmental transmission dynamics. Relying solely on national-level trend lines obscures localized pandemic dynamics, as disease transmission is heavily shaped by region-specific environmental, demographic, and behavioral conditions. South Korea provides a clear justification for a localized experimental framework. Daegu's rapid outbreak, driven by a church-linked superspreading event, generated over 3,900 secondary cases and accounted for 55% of the local cluster [27]. Such anomalous hotspots disproportionately skew national epidemiological models, diluting unique signals from other regions. As a critical counterpoint, we also examine Seoul. Unlike Daegu's cluster-driven experience, Seoul represents the complex challenge of diffuse, sustained transmission within a densely populated megacity. Its high human mobility, varied micro climates, and chronic air quality issues provide a unique testing ground. Contrasting these divergent urban profiles demonstrates the crucial value of localized modeling in accurately capturing specific environmental drivers.

Methods

We employed a systematic analytical framework to investigate the complex, non-linear relationships between a suite of environmental variables and COVID-19 transmission patterns in the Daegu and Seoul regions, as shown in Fig. 2. The workflow begins with exploratory data analysis to understand initial trends. This is followed by feature engineering and selection, where multicollinearity is assessed using the Variance Inflation Factor to ensure model stability. The core of the analysis is the Generalized Additive Model (GAM), a powerful statistical technique chosen for its ability to capture non-linear effects without prior assumptions. This explanatory model allows us to quantify the impact of nine distinct

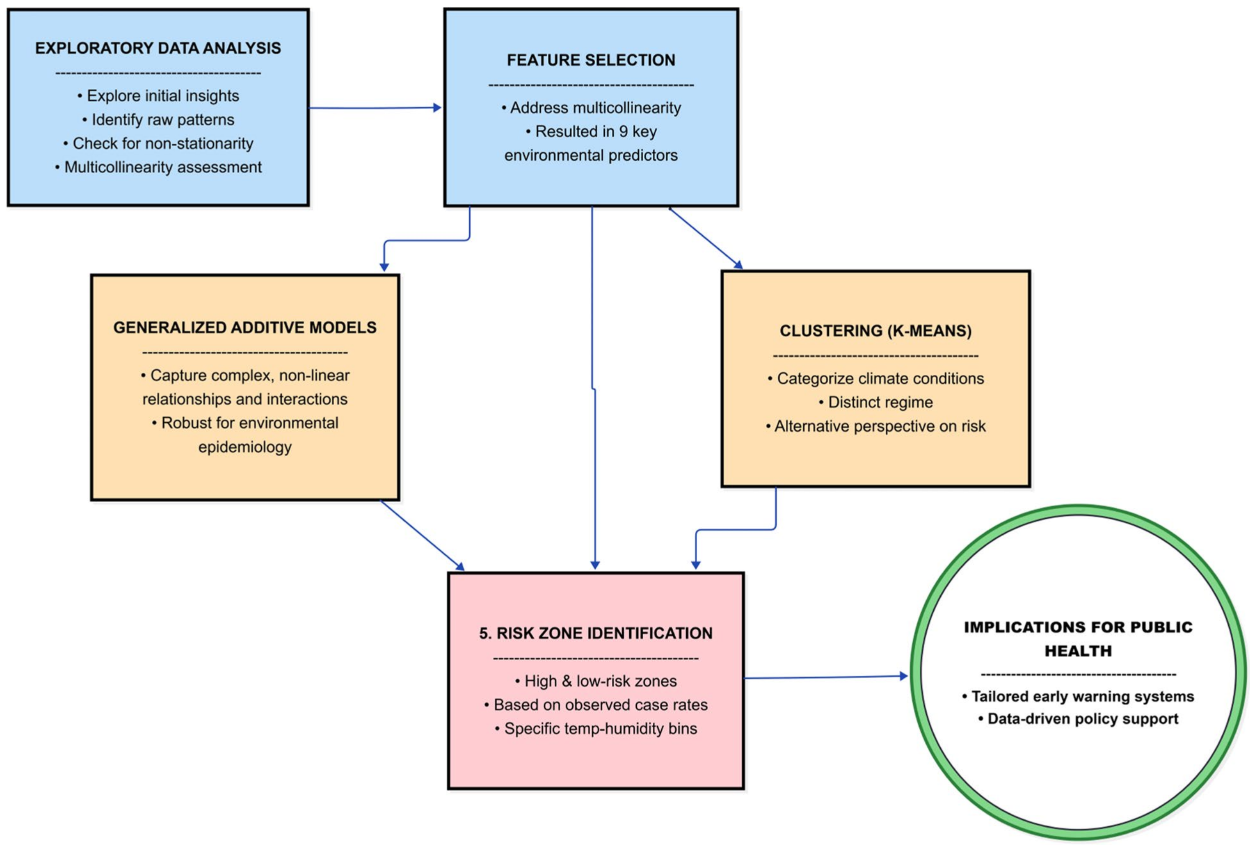

Fig. 2 Operational flowchart illustrating the proposed analytical framework. The diagram details the sequence of data processing steps, exploratory analysis, non-linear modeling (GAM), unsupervised clustering (K-means), and the subsequent derivation of environmental risk zones for Daegu and Seoul

environmental factors. To supplement this, clustering is used to identify and categorize distinct climate regimes within the data. Finally, a threshold effect analysis is conducted by visualizing the GAM's predictions as a contour map. This risk map is used to identify specific hotspots as and provide real-world examples of high-risk and low-risk environmental conditions, providing actionable insights for public health.

Data representation and feature selection

This study utilizes two datasets: one from the metropolitan city of Daegu and another from Seoul, South Korea, regions that experienced a significant and well-documented early outbreak of COVID-19. The daily time series data from January 1, 2020, to July 31, 2020, were taken from official, publicly available sources to encompass the complete first wave of the pandemic in the region. Daily confirmed COVID-19 case counts were obtained from the Korea Centers for Disease Control and Prevention (KCDC) [28]. Corresponding daily meteorological data (average temperature, rainfall, wind speed, relative humidity, pressure) were sourced from the Korea Meteorological Administration (KMA) database [29], and daily air quality data for four key pollutants (CO, NO_2, O_3, and SO_3) were acquired from the Air Korea platform [8]. The relevance of these environmental parameters is underscored by their theoretical link to viral transmission, as they influence pathogen survival and dispersal dynamics. For example, Lim et al. [13] demonstrated correlations between climatic fluctuations and COVID-19 case patterns. The analytical framework focuses on twelve key variables: ambient temperature, precipitation, relative humidity, (CO), (NO_2), (SO_3), ozone (O_3), PM10 and PM2.5, atmospheric pressure, wind velocity, and COVID-19 cases as the dependent variable. To ensure the stability and interpretability of the final explanatory model, a systematic multicollinearity assessment was conducted on the nine environmental predictors. A 7-day moving average was utilized to stabilize daily fluctuations and capture cumulative exposure. While lacking an explicit lag structure, this framework provides a robust, interpretable baseline for identifying fundamental non-linear atmospheric drivers.

Statistical analysis

All statistical analyses were conducted in Python (version 3.9) utilizing standard libraries, pandas for data manipulation, scikit-learn for data preprocessing, pygam for

Generalized Additive Model fitting, matplotlib and seaborn for data visualization, and statsmodels for Variance Inflation Factor (VIF) calculation. The primary analytical tool employed in this study is a Generalized Additive Model (GAM). This model was selected for its ability to effectively model complex, non-linear relationships between environmental predictors and COVID-19 cases without imposing restrictive, predefined assumptions about the shape of these relationships. Before model fitting, the target variable, a 7-day moving average of cases, was transformed using a natural logarithm ($\log(x+1)$) to mitigate positive skewness and stabilize variance, thereby improving residual normality. The predictor variables were standardized using z-score normalization, which aids in model convergence and enhances the comparability of smoothing parameters across different scales. To strictly assess multicollinearity, particularly between meteorological variables and pollutants, we performed a sensitivity check using a stringent Variance Inflation Factor (VIF) threshold of 5.0. In Daegu, both average temperature (VIF = 4.87) and O_3 (VIF = 3.64) remained below this strict limit, confirming their independence. In Seoul, while average temperature showed moderate collinearity (VIF = 6.22) reflecting strong seasonal patterns, O_3 remained well below the threshold (VIF = 3.07). This confirms that the O_3 signal is statistically distinct from temperature in both models, ensuring that the estimated environmental effects are effectively disentangled. The selected predictors included temperature, rainfall, wind speed, relative humidity, pressure, and air pollutants such as CO, NO_2, O_3 and SO_3. While count data is typically modelled using Poisson or Negative Binomial distributions, we employed a Gaussian GAM with a log link function ($\log(x+1)$) for this analysis. This decision was driven by the continuous nature of the moving average data. To properly account for the estimated dispersion parameter and the finite sample size of our daily observations, the statistical significance of the non-linear smooth terms in the GAM was evaluated using approximate F-statistics. We fitted a Gaussian GAM with an identity link to the log-transformed 7-day moving average of cases $\log(\text{ma.case}+1)$, as follows:

$$\begin{aligned}\log(\text{ma.case}+1) &= \text{intercept} + s(\text{temp.avg}) \\ &\quad + s(\text{rainfall}) \\ &\quad + \cdots + s(\text{so3}) + \epsilon\end{aligned} \tag{1}$$

where $\log(\text{ma.case}+1)$ represents the log-transformed 7-day moving average of cases, $s()$ denotes a penalized smoothing spline for each of the standardized predictor variables, and ϵ is the error term assumed to be normally distributed. The number of splines for each smooth term was set to 4 to balance model flexibility with the risk of overfitting, with smoothing parameters determined using generalized cross-validation (GCV) during model fitting. This approach enabled the quantification of the isolated non-linear effect of each environmental factor while controlling for the influence of all others. Model performance was evaluated using explained deviance (analogous to R-squared for GLMs), with model stability and generalization assessed through 3-fold cross validation. Residual diagnostics confirmed model assumptions (Figs. 13 and 14 in Appendix).

Preliminary Poisson and Negative Binomial models exhibited non-convergence due to the continuous nature of the 7-day moving average. Consequently, the Gaussian family was selected as the most stable framework for the log-transformed data. To complement the primary GAM analysis, K-means clustering was applied as an exploratory technique to categorize the data into distinct climate regimes based on the primary climatic drivers, providing additional context to the main model's findings.

Risk zone identification

To translate the non-linear relationships identified by the GAM into actionable insights, a threshold effect analysis was conducted on key environmental predictors. This process involved identifying critical points of change in the partial dependence functions, specifically where the first derivative of the smoothing spline, $s'(\cdot)$, exhibited significant shifts in magnitude. These inflection and turning points were then used to delineate distinct risk zones for each environmental variable. For example, temperature ranges were categorized into optimal, low risk, and high-risk zones based on the sign and slope of the partial dependence function. The 75th and 25th percentiles serve as non-parametric thresholds to normalize city-specific distributions for unbiased comparability. This approach isolates environmental regimes associated with relative transmission extremes, regardless of absolute baseline differences. To ensure the robustness of these identified thresholds, the 95% confidence intervals of the partial dependence curves were examined to confirm that the observed changes in the predictor's effect were statistically distinguishable from zero or from adjacent risk zones.

Results

Initial exploratory data analysis

An initial exploratory data analysis (EDA) was conducted to examine the daily patterns and distributions of Initial Exploratory Data Analysis in Seoul and Daegu from January 1 to July 31, 2020. This step offered useful insights that helped guide the development of the Generalized Additive Models. From the Fig. 3 the two cities followed different epidemic patterns. Daegu saw a sharp peak of about 800 daily cases in late February to early March

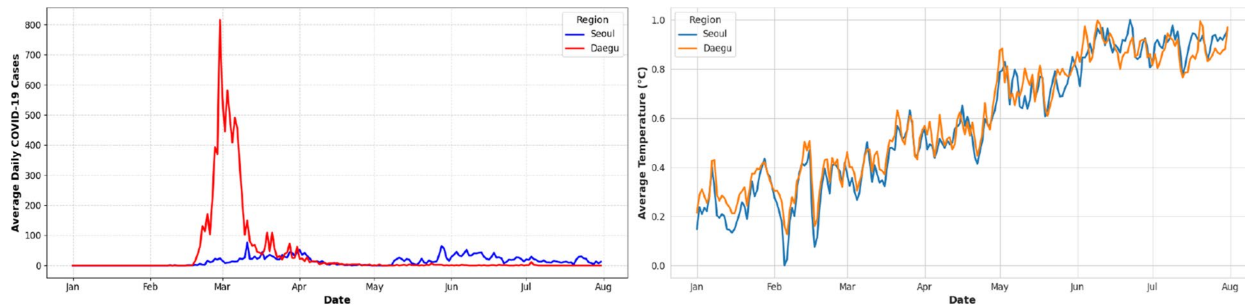

Fig. 3 Left: Initial COVID-19 case counts and Right: Temperature levels over Daegu and Seoul over year 2020

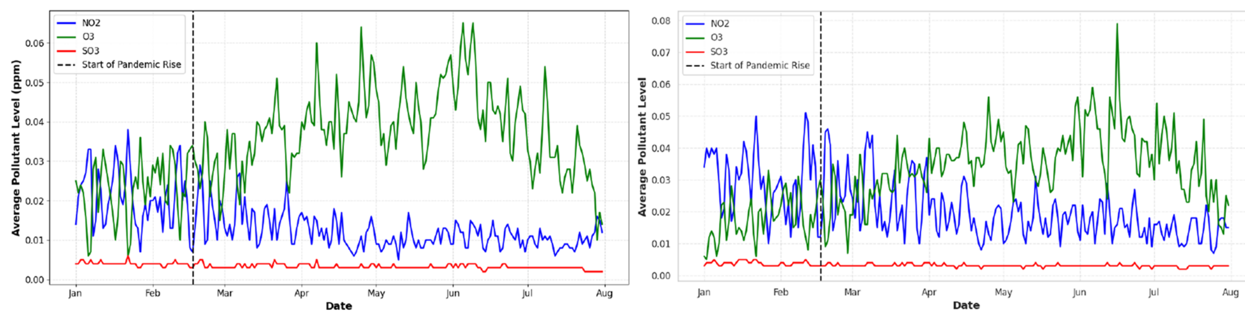

Fig. 4 Initial Air pollution concentration in Daegu (Left) and Seoul (Right) over Daegu over year 2020

2020, followed by a quick decline. Seoul, on the other hand, had lower but more steady case numbers without a similar early spike. Daily average temperatures in Fig. 3 exhibited clear seasonal warming patterns consistent across both cities. Air pollutant concentrations NO_2, O_3, CO, SO_3 in Fig. 4 generally mirrored each other in both cities. Notably, CO concentrations showed a decreasing trend after the onset of the pandemic, potentially linked to reduced human activity, while NO_2 and O_3 levels remained relatively low and stable. Our localized *CO* reduction aligns with the 5–10% global decrease in anthropogenic hotspots reported by Pathak et al. (2023) [30], whereas our findings of stable NO_2 levels in Seoul are supported by Amritha et al. (2025) [31], who identified Seoul as a global exception with a post-lockdown rebound of $< 10\%$ compared to the 10–30% spikes observed elsewhere. In addition to the timing, there were clear differences in how the data were spread out between the cities. Daegu's temperatures were often in the 0 °C to 15 °C range, while Seoul's showed two peaks, with more days at colder and warmer extremes. Relative humidity patterns were similar overall, but Daegu had more days with moderate humidity (50–70% RH). Atmospheric pressure in Daegu was more clustered around 1000 hPa than in Seoul. For precipitation and wind, Seoul had more days with strong winds and no rain, even though total rainfall was similar. There were also notable differences in air quality. Daegu typically had higher CO levels, reaching about 0.35 - 0.4 ppm, while Seoul had higher NO_2. SO_3 levels were mostly low in Seoul but higher in Daegu. O_3 patterns varied as well, with Daegu showing a bimodal and more even spread, while Seoul exhibited a two peak distribution, despite having similar average levels.

To further motivate the use of Generalized Additive Models, bivariate analyses were conducted to explore the raw interactive effects of temperature and relative humidity on COVID-19 case counts in Fig. 5. These contour plots visually highlighted regions of varying case incidence across different temperature–humidity combinations. In Daegu, the contour map revealed complex, non-linear patterns. High-risk zones emerged at moderate temperatures of 5 - 10°C with moderate to high relative humidity of 60–80%, and also around 0°C with 60–70% RH. low-risk zones were observed at higher temperatures of 10 - 15°C with lower relative humidity. For Seoul, a similarly highly non-linear and varied landscape was evident. Prominent high-risk zones were found across warmer temperatures of 15 - 25°C, combined with moderate to high relative humidity of 60–90%, with another notable area around 5 - 10°C with moderate humidity. low-risk zones generally occurred at very low temperatures (<0°C) and at moderate temperatures (10 - 15°C) with lower relative humidity. These visualizations compellingly demonstrate that the combined influence of temperature and relative humidity on COVID-19 cases is neither simple nor consistent, thus affirming the necessity of advanced non-linear modeling techniques like GAMs to accurately characterize these complex, interactive relationships.

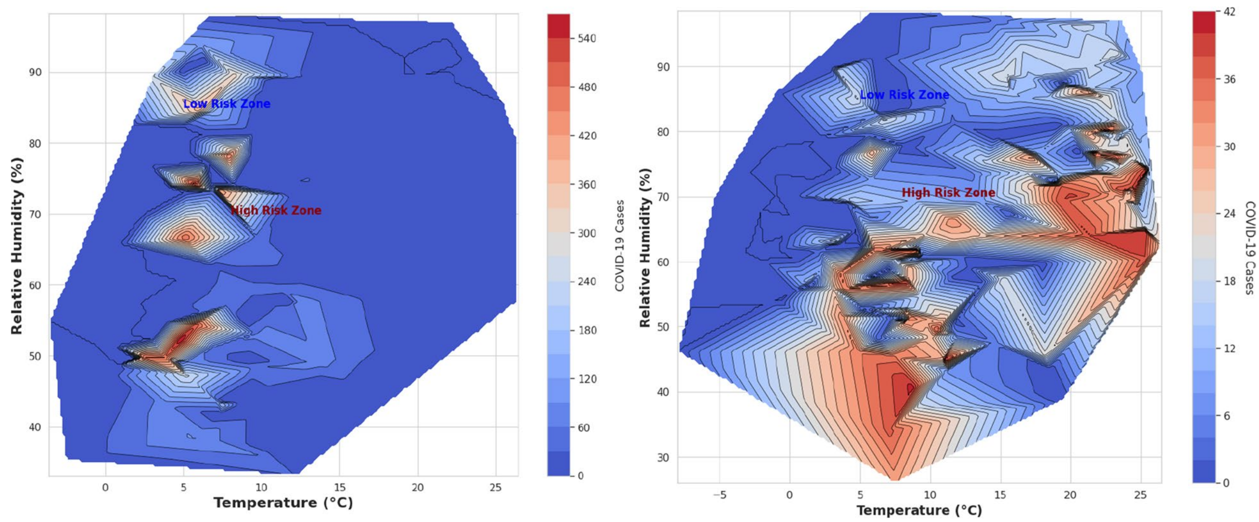

Fig. 5 Bivariate analysis: Raw interactive effect of temperature and humidity on COVID-19 case in Daegu (Left) and Seoul (Right)

Non-linear analysis and Independent climatic effects on COVID-19 case counts

Following the exploratory data analysis, Generalized Additive Models (GAMs) were fitted to characterize the non-linear relationships between environmental factors and log-transformed COVID-19 case counts for both Daegu and Seoul. This section first presents the model performance metrics, followed by a detailed interpretation of the partial dependence plots (PDPs), highlighting the estimated non-linear effects of each environmental predictor for both cities. Prior to non-linear modeling, Spearman's rank correlation coefficients were calculated to assess linear associations between environmental variables. In Seoul, a notable strong positive correlation was observed between average daily temperature and O_3 ($r = 0.794, p < 0.001$). This indicates a significant meteorological influence on secondary pollutant formation in the capital region, a pattern not strongly observed in Daegu.

GAM model performance

The performance of the GAMs for both cities were evaluated, including Explained deviance (ED) on training, validation, and full datasets, along with the optimal smoothing parameters (λ), is summarized in Table 2. The correlation analysis is provided in Appendix A.3 (Fig. 15)

Table 2 Summary of generalized additive model performance for Daegu and Seoul

City	Train ED (CV Avg ± SD)	Val ED (CV Avg ± SD)	Full data ED	Optimal λ
Daegu	0.641 ± 0.038	0.500 ± 0.067	0.630	0.0010
Seoul	0.511 ± 0.052	0.394 ± 0.078	0.483	1.0000

For Daegu, the GAM demonstrated strong performance, explaining 63.0% of the variability in log-transformed COVID-19 cases on the full dataset, with a robust cross-validation average of 0.500 ± 0.067 on unseen data. The optimal smoothing parameter for Daegu was $\lambda = 0.0010$, indicating a preference for more flexible smooths. In Seoul, the model achieved an explained deviance of 48.3% on the full dataset and 0.394 ± 0.078 on validation data. Seoul's model selected a higher optimal $\lambda = 1.0000$, suggesting a preference for smoother, less complex functions compared to Daegu's. While the explained deviance decreased during cross-validation for both cities, this reduction is an expected outcome of generalizing to unseen data segments. The validation scores remain robust, confirming that the models retain substantial predictive power without overfitting to the training dataset. The Partial Dependence Plots (PDPs) illustrate the estimated non-linear effect of each environmental factor on log-transformed COVID-19 cases, isolating the impact of each variable while holding others constant. The shaded areas in the figures represent the 95% confidence intervals.

Figure 6 presents the PDPs for Daegu. A prominent non-linear, U-shaped relationship was observed for temperature, where the estimated effect on cases was lowest at moderate temperatures and sharply increased at both extremes. The PDP reveals a U-shaped relationship on the log scale, with estimated effects increasing at both temperature extremes reaching a minimum marginal effect of −0.32 [95% CI: −1.45, 0.45] at −3.60°C and a maximum of 3.34 [95% CI: 2.80, 3.90] at the highest observed temperature of 26.4°C. However, the risk zone analysis shows that epidemiologically dominant high-case events were concentrated under cool-to-moderate conditions (2.4°C to 11.4°C). The elevated model effect

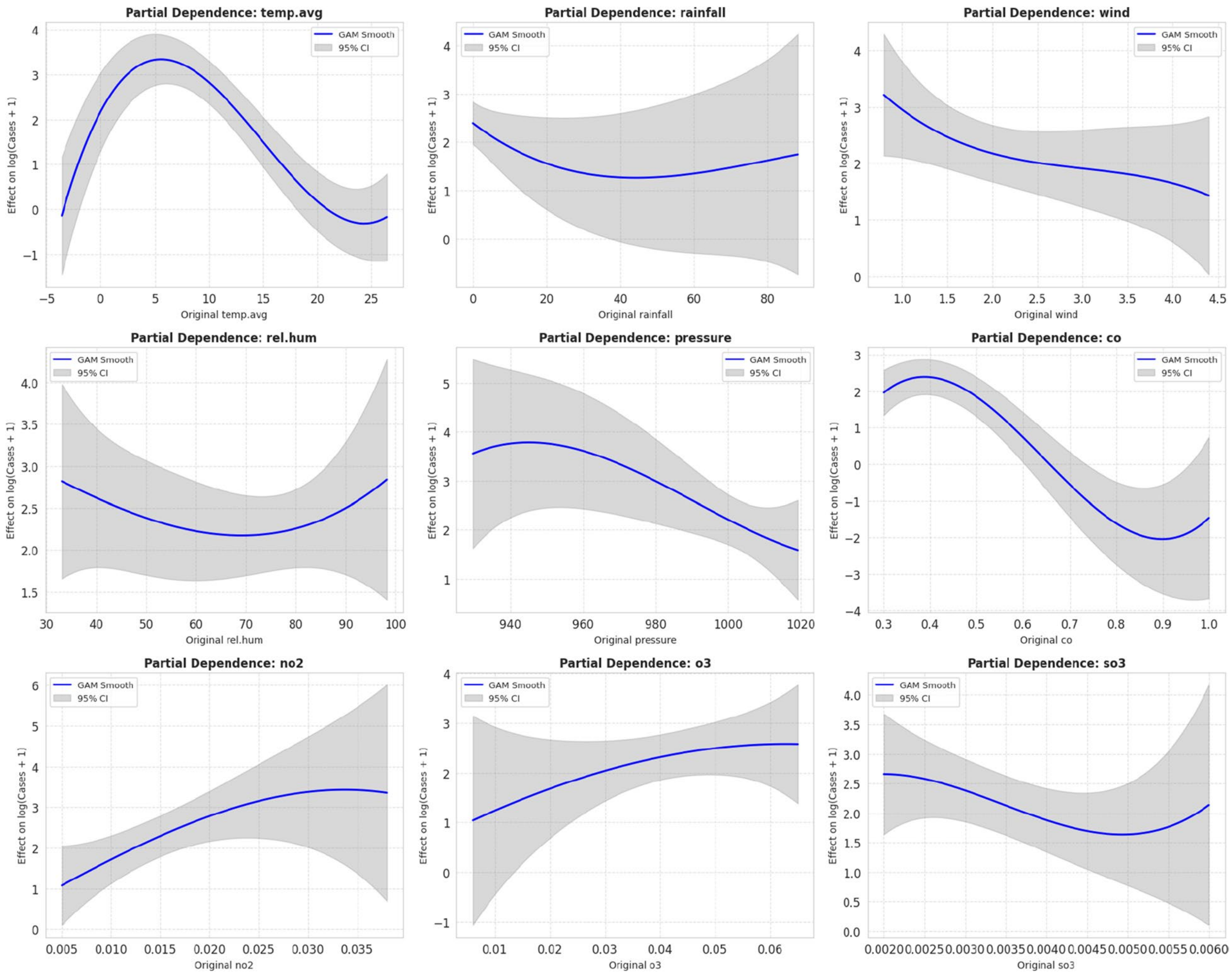

Fig. 6 PDPs of the GAM model for Daegu, illustrating non-linear environmental influences on log-transformed COVID-19 cases. Blue lines show fitted smooth effects with 95% confidence bands, revealing heterogeneous response patterns across all variables

at the warm extreme (26.4°C) should be interpreted cautiously, as observations at that temperature range correspond to low observed case rates in the data, placing those bins firmly in the low-risk zone. Relative Humidity displayed a more subtle pattern with a modest range of effect, varying from a minimum of 2.17 [95% CI: 1.40, 2.64] at lower humidity (33.1%) to a maximum of 2.84 [95% CI: 1.79, 4.27] at high humidity (98.2%). Wind speed and Pressure exhibited generally positive associations in this model. The effect of wind speed increased from 1.43 [95% CI: 0.03, 2.57] at the lowest speed (0.8 m/s) to a maximum of 3.22 [95% CI: 2.14, 4.29] at the highest speed (4.4 m/s). Atmospheric pressure showed an effect ranging from 1.59 [95% CI: 0.57, 2.46] at the lowest pressure (929.6 hPa) to 3.78 [95% CI: 2.46, 5.48] at the highest (1019.2 hPa). Rainfall showed a relatively flat and uncertain effect across its range (0 to 88.2 mm), with wide confidence intervals that crossed zero (Effect range: 1.26 to 2.39; 95% CI lower bound as low as −0.74). Among pollutants, CO demonstrated a pronounced non-linear relationship, with the effect dropping sharply to a minimum of −2.05 [95% CI: −3.72, −0.65] at the highest concentration of 1.00 ppm. NO_2 and O_3 both showed generally positive, non-linear relationships, with cases tending to increase as concentrations rose. The effect of NO_2 increased to a maximum of 3.43 [95% CI: 2.25, 6.01] at the highest observed level (0.04 ppm), and O_3 reached a maximum effect of 2.58 [95% CI: 1.97, 3.77] at its peak concentration (0.07 ppm). Finally, SO_3 revealed a positive effect ranging from 1.64 [95% CI: 0.10, 2.34] at the lowest level to 2.66 [95% CI: 1.93, 4.17] at the highest level (0.01 ppm). These diverse non-linear patterns with their quantified effects underscore the complex interplay between environmental conditions and COVID-19 transmission dynamics in Daegu.

The PDPs for Seoul from Fig. 7, present revealing relationships distinctly different from those in Daegu, characterized largely by monotonic or mildly curved effects. Several variables exhibited strong, clear monotonic trends. Relative Humidity and Pressure both displayed

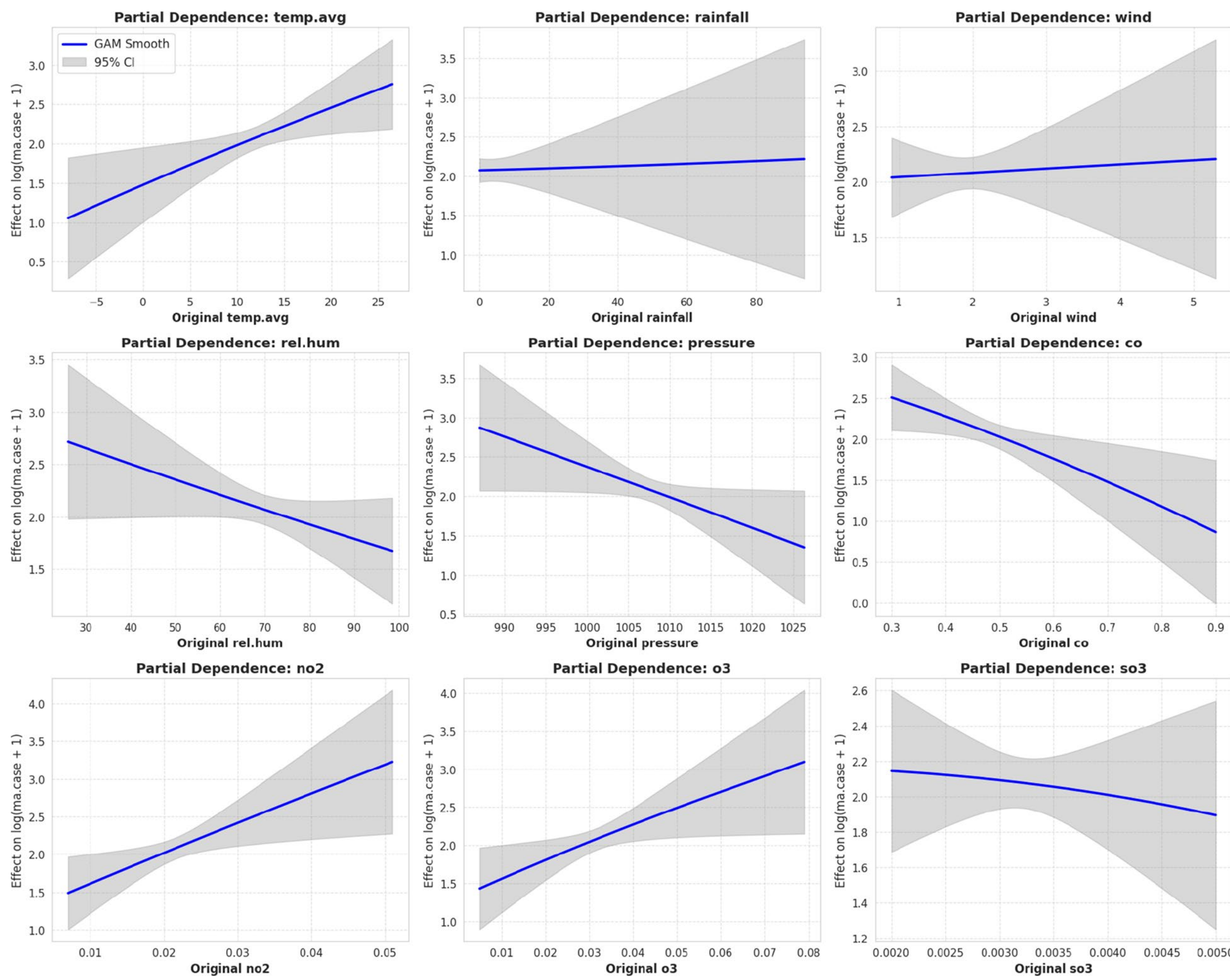

Fig. 7 PDPs of the GAM model for Seoul, illustrating non-linear environmental influences on log-transformed COVID-19 cases. Blue lines show fitted smooth effects with 95% confidence bands, revealing heterogeneous response patterns across all variables

consistent negative relationships. The effect of humidity decreased from a maximum of 2.72 [95% CI: 1.98, 3.45] at the lowest level (26.0%) to a minimum of 1.67 [95% CI: 1.17, 2.15] at the highest level (98.5%). Similarly, atmospheric pressure showed a decreasing effect from 2.87 [95% CI: 2.07, 3.67] at the lowest pressure (987.0 hPa) to 1.35 [95% CI: 0.63, 2.07] at the highest (1026.3 hPa). Conversely, temperature, NO_2, and O_3 all demonstrated clear positive monotonic relationships, where an increase in the variable's value steadily increased the predicted effect on transmission. The effect of temperature rose from 1.06 [95% CI: 0.29, 1.82] at the coldest point (−8.0°C) to 2.76 [95% CI: 2.19, 3.33] at the warmest (26.5°C). NO_2 showed an effect increasing from 1.49 [95% CI: 1.00, 1.97] to 3.23 [95% CI: 2.27, 4.18] across its range (0.01 to 0.05 ppm). O_3 exhibited a similar trend, with its effect rising from 1.43 [95% CI: 0.90, 1.97] to 3.10 [95% CI: 2.15, 4.04] as concentrations increased from 0.01 to 0.08 ppm. Unlike these consistent trends, CO revealed a distinct non-linear relationship, with the effect decreasing from a maximum of 2.51 [95% CI: 2.11, 2.91] at the lowest concentration (0.30 ppm) to a minimum of 0.86 [95% CI: −0.02, 1.74] at the highest (0.90 ppm). SO_3 showed a weaker, mildly curved effect, with values ranging from 1.90 [95% CI: 1.25, 2.22] to 2.14 [95% CI: 1.94, 2.60]. Finally, Rainfall and Wind speed exhibited weak or uncertain effects; rainfall's trend was slightly positive but with very wide confidence intervals (Effect range: 2.07 to 2.22), while the wind speed effect was also modest (Effect range: 2.04 to 2.20). Seoul's PDP results indicate a relatively flat marginal dependence on rainfall and wind speed in the Seoul model, with marginal effects on the log scale of just 0.15 and 0.16, respectively. These effects are substantially smaller than those of direct pollutants like NO_2 (1.74), O_3 (1.67), CO (1.65), and primary meteorological drivers like temperature (1.70). The prevalence of these monotonic and distinct non-linear patterns in Seoul, when compared to Daegu's U-shaped climate curves, highlights the significantly region-specific nature of environmental drivers.

Table 3 Relative contribution of meteorological and air pollutant variables to COVID-19 incidence in the Daegu and Seoul clusters, ranked by GAM derived Estimated Degrees of Freedom (EDoF)

Daegu model			**Seoul model**		
Rank	**Variable**	**EDoF**	**Rank**	**Variable**	**EDoF**
1	temperature	3.2***	1	temperature	2.6***
2	Rainfall	2.0	2	Rainfall	1.3
3	Wind Speed	1.9	3	Wind Speed	1.2
4	Relative Humidity	1.5	4	Carbon Monoxide (CO)	1.1***
5	Atmospheric Pressure	1.2***	5	Relative Humidity	1.0
6	Carbon Monoxide (CO)	1.2**	6	Atmospheric Pressure	1.0
7	Nitrogen Dioxide (NO_2)	1.2	7	Sulfur Trioxide (SO_3)	0.9
8	Ozone (O_3)	1.0	8	Nitrogen Dioxide (NO_2)	0.8
9	Sulfur Trioxide (SO_3)	0.8	9	Ozone (O_3)	0.7

Note: EDoF indicates the complexity and relative contribution of each variable in the GAM. Significance codes derived from F-statistics : *** $p < 0.001$, ** $p < 0.01$

To explicitly quantify the relative contribution of these environmental and meteorological drivers, Table 3 ranks the variables based on their Estimated Degrees of Freedom (EDoF) derived from the GAM summaries. Because higher EDoF values indicate that a variable is exhibiting a more complex non-linear relationship and performing more work within the model, this metric serves as an objective proxy for relative importance. The rankings demonstrate a clear spatial heterogeneity in transmission drivers. In the Daegu model, primary meteorological factors wielded the most significant influence, with temperature ranking highest (EDoF = 3.2), followed by rainfall (2.0) and wind speed (1.9). Conversely, the hierarchy in the Seoul model demonstrates lower overall complexity for precipitation and wind (EDoF = 1.3 and 1.2, respectively), highlighting the dominance of temperature (2.6) alongside anthropogenic pollutants like CO (1.1) in the highly urbanised environment.

Heterogeneity in transmission dynamics: the case for region-specific analysis

To provide a broader contextual understanding of how distinct meteorological conditions influence COVID-19 transmission, and to complement the granular insights from the risk zone analysis, K-means clustering was applied to classify daily temperature and relative humidity into three distinct climatic regimes for each city as Cold & Humid, Mild & Humid and Warm & Humid. This method offers a generalized categorization of prevailing climatic patterns and their overall association with COVID-19 case incidence. The distribution of these clusters in the temperature humidity space and their associated COVID-19 case incidence visualized is presented in Figs. 8 and 9.

In Daegu illustrated in Fig. 8, three distinct climatic regimes were identified, revealing a clear pattern in average daily COVID-19 cases, and the Cold & Humid regime was associated with the highest average case count of 97.87 (95% CI: 45.69 - 150.04), followed by the Mild & Humid regime with an average of 45.66 cases (95% CI: 22.30 - 69.01). The lowest average case count, 0.77 (95% CI: 0.59 - 0.95), was recorded in the Warm & Humid regime. A Kruskal – Wallis test confirmed significant differences in case distributions across these clusters ($H = 13.144, p = 0.0014$), with an effect size (ϵ^2) of 0.053. Dunn's post-hoc test revealed that the differences between the Mild & Humid and Warm & Humid regimes were statistically significant ($p = 0.0011$). While the Cold

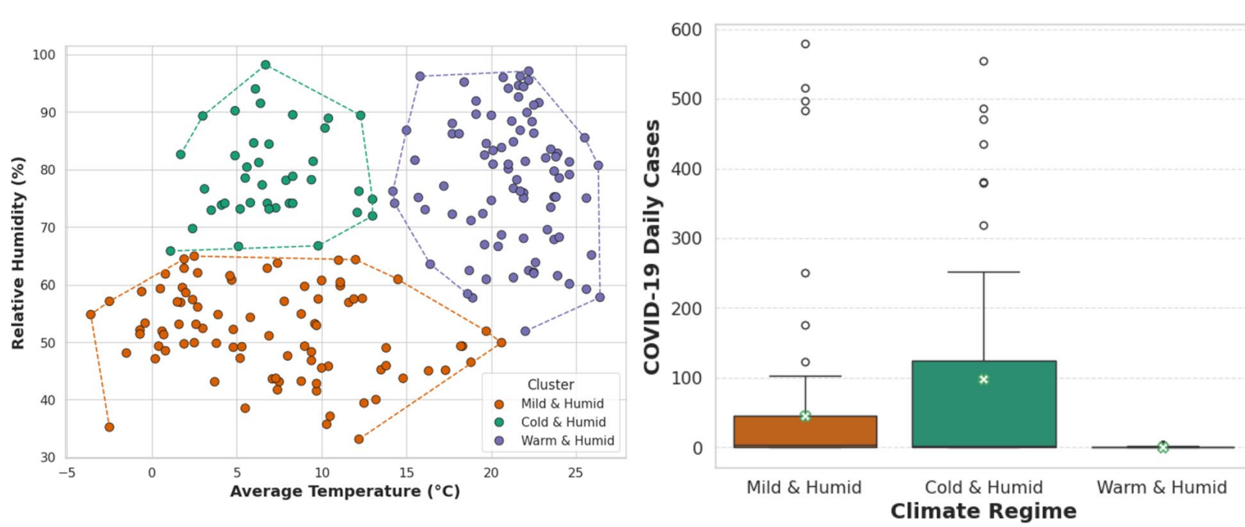

Fig. 8 Climatic regimes and COVID-19 case distributions in Daegu (left) daily temperature and humidity clustered (right) shows distribution of daily COVID-19 cases across three identified regimes

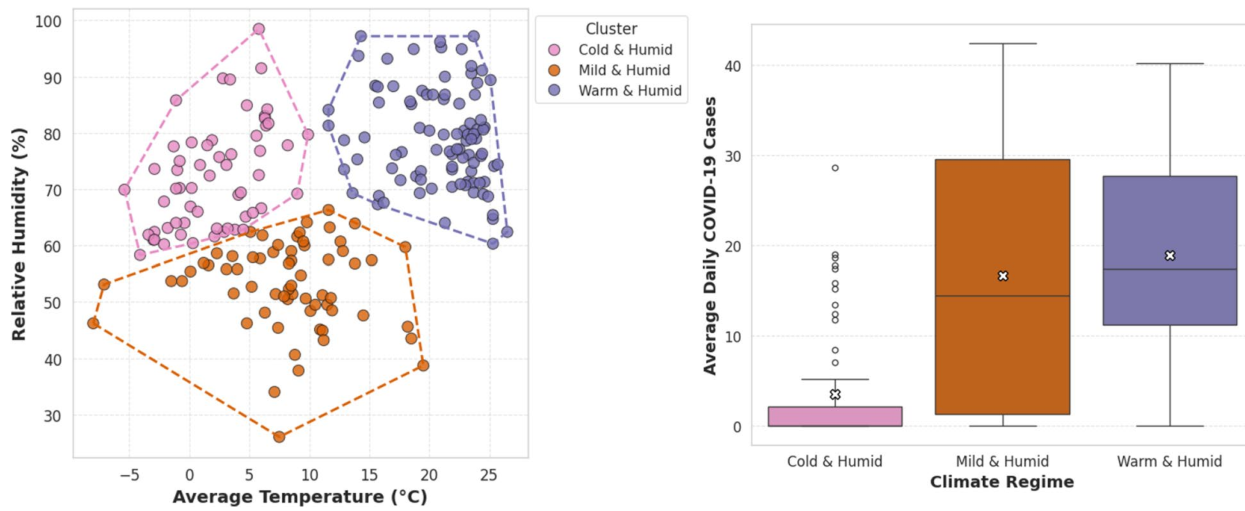

Fig. 9 Climatic regimes and COVID-19 case distributions in Seoul(left) along with daily temperature and humidity clustered (right) shows the distribution of daily COVID-19 cases across these three identified regimes

& Humid regime had the highest mean, its large standard deviation resulted in non-significant post-hoc comparisons with the other two regimes. These findings suggest that in Daegu, colder and milder humid conditions generally correlated with higher COVID-19 case numbers, while warm and humid conditions were associated with significantly lower transmission.

Seoul also exhibited three distinct climatic regimes and their association with COVID-19 cases presented a contrasting profile in Fig. 9. The highest average case count was associated with the Warm & Humid regime (18.92 cases, 95% CI: 16.59–21.25), followed closely by the Mild & Humid regime with an average of 16.64 cases (95% CI: 13.05–20.22). The Cold & Humid regime was associated with the lowest average case count of 3.51 (95% CI: 1.79 - 5.23). A Kruskal – Wallis test confirmed highly significant differences in case distributions across these regimes ($H = 64.763, p < 0.0001$), with a stronger effect size (ϵ^2) of 0.299 compared to Daegu. Dunn's post-hoc test indicated statistically significant differences between the Cold & Humid regime and both Mild & Humid ($p < 0.0001$) and Warm & Humid ($p < 0.0001$) regimes. No significant difference was observed between the Mild & Humid and Warm & Humid regimes ($p = 0.3535$). These results suggest that in Seoul, warmer and milder humid conditions are associated with higher COVID-19 cases, whereas colder and humid conditions are linked to significantly lower transmission, a pattern in contrast to Daegu. The distribution of days across these climatic regimes further highlights regional differences. In Daegu, the Mild & Humid regime accounted for 41.31%, while the Warm & Humid regime comprised 39.91%, and Cold & Humid days constituted 18.78%. Conversely, in Seoul, the Warm & Humid regime dominated (42.72%), followed by Mild & Humid (30.05%) and Cold & Humid (27.23%) days. This differential prevalence of climatic conditions is critical for interpreting the unique environmental risk profiles and contrasting COVID-19 dynamics observed between the two cities.

Risk zone identification

Building upon the non-linear relationships identified by the GAMs, a specific Risk Zone Identification framework was applied to temperature-humidity bins to define statistically reliable high- and low-risk environmental conditions for COVID-19 case incidence in Daegu and Seoul. This approach provides granular, empirically derived thresholds for exceptionally high or low observed case rates within specific temperature–humidity combinations. For each city, environmental data were binned by temperature and relative humidity, and the average observed case rate within each bin was calculated. High - Risk Zones were defined as bins exceeding the 75th percentile of observed rates, while low-risk zones were defined as bins falling below the 25th percentile.

In Daegu the Fig. 10, the analysis identified a distinct high-risk threshold of 26.84 cases. The primary high-risk hotspot was situated in cooler conditions, with the most dominant risk cluster exhibiting an exceptionally high observed case rate of 278.80. This occurred within a temperature range of 2.4°C to 5.4°C combined with relative humidity between 46.12% and 52.63%. Additional significant high-risk zones were identified at slightly warmer temperatures (5.4°C to 8.4°C) across a range of humidity levels, with case rates reaching as high as 181.52 at high humidity (78.67% to 85.18%). Conversely, the most reliable low-risk zones in Daegu, defined by a threshold below 0.60 cases, were consistently identified in both the coldest sub-zero conditions (e.g., −0.6°C to 2.4°C) and,

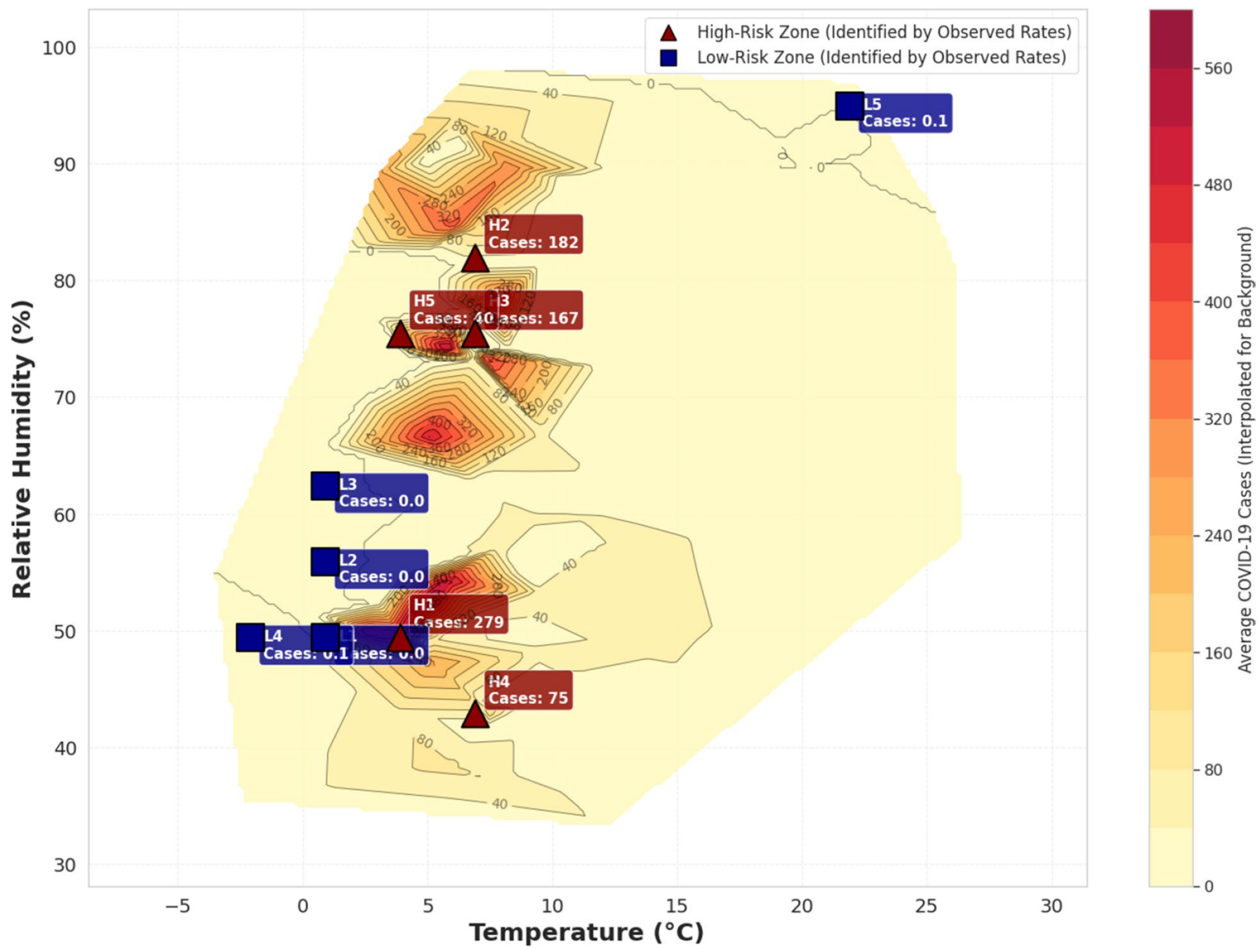

Fig. 10 Risk zones in Daegu based on temperature and humidity bins

notably, in warmer ranges > 20°C, where observed case rates approached zero, irrespective of humidity levels.

Applying the identical risk zone identification framework to Seoul revealed a significantly different environmental fingerprint. The thresholds for Seoul were defined by a High-Risk Rate exceeding 20.38 cases and a Low-Risk Rate below 4.69 cases. The resulting multimodal risk profile in Fig. 11 displayed high and low-risk zones distributed differently than in Daegu. The primary high-risk hotspot was identified in warm conditions, with the most dominant risk clusters, H1 and H2, exhibiting the highest observed case rates of 29.76 and 29.29, respectively. These occurred at warm temperatures ranging from 19.6°C to 26.5°C combined with moderate to high humidity of 69.5% to 76.8%. In addition to this primary warm hotspot, a second set of distinct high-risk zones was found at cooler temperatures in Seoul, with case rates ranging from 26.20 to 27.53 across temperatures between 2.35°C and 12.7°C. In sharp contrast to Daegu's warm low-risk zones, the most reliable low-risk zones in Seoul were almost exclusively identified in cold, sub-zero conditions (e.g., −4.55°C to 2.35°C), a pattern that was consistent across various humidity levels. The profound differences in identified risk zones between Daegu and Seoul underscore the highly localized nature of environmental risk factors for COVID-19 transmission and highlight the critical need for region-specific public health interventions.

Sensitivity analysis

To address potential confounding driven by specific epidemic waves, we conducted sensitivity analyses by excluding the peak outbreak periods February 15 – March 15, 2020 for Daegu & March 5 – March 20, 2020 for Seoul) and refitting the GAMs. The analyses yielded high explained deviance values for both regions 80.2% for Daegu and 76.4% for Seoul, significantly increasing from the primary models' fit. The functional forms of the temperature associations remained consistent with the primary findings in both cities, confirming that the observed environmental associations are robust and not artifacts of specific superspreading events (see Supplementary Figs. 18 and 19).

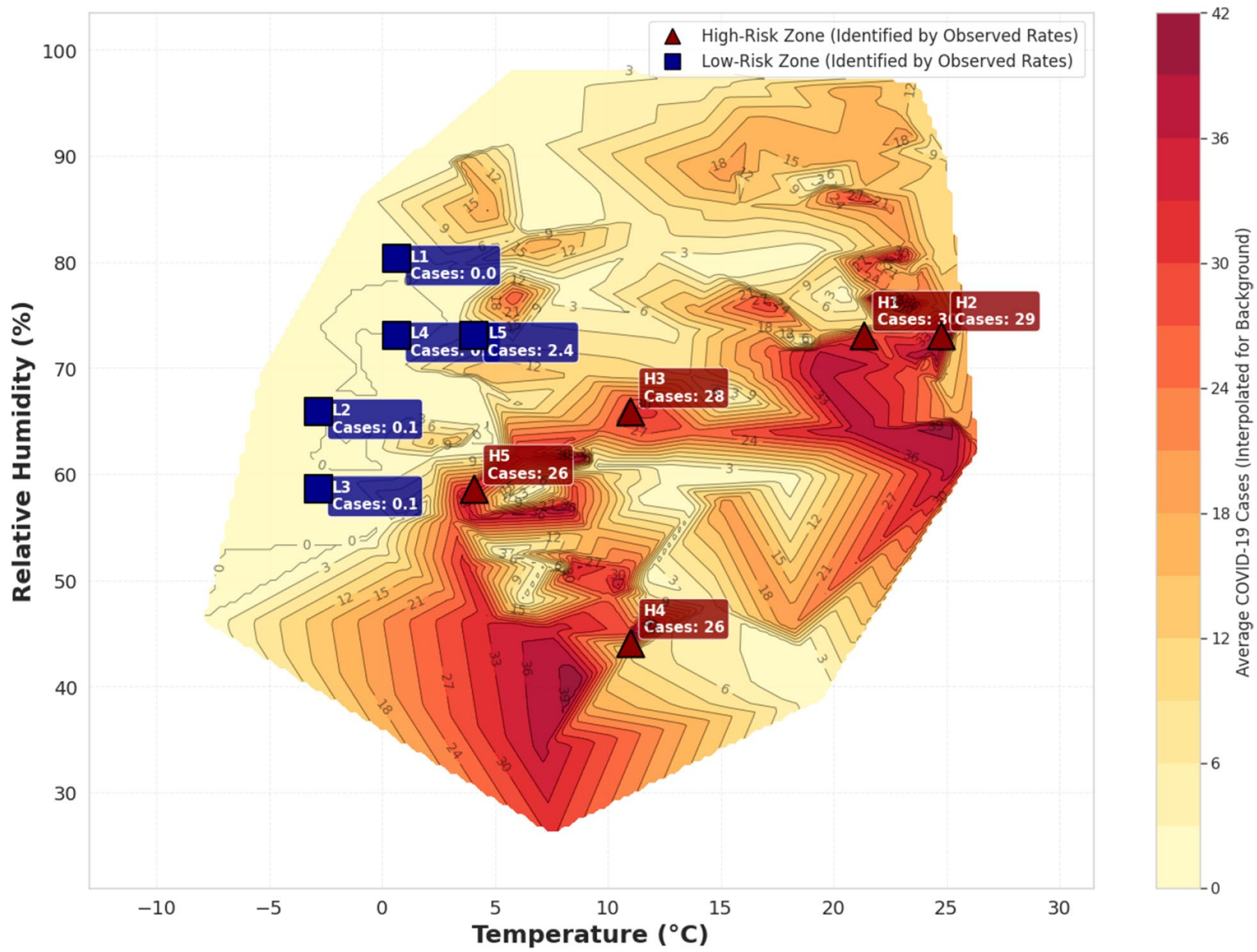

Fig. 11 Risk zones in Seoul based on temperature and humidity bins

Discussion and policy recommendations

This study developed a robust comparative framework to investigate the complex influence of environmental factors on COVID-19 transmission, revealing that the environmental drivers of respiratory pandemics are highly region-specific. Our analysis presents a tale of two cities the inland basin of Daegu exhibited a more climate-sensitive risk profile, whereas Seoul demonstrated stronger associations with air pollution indicators. While person-to-person contact and population mobility remain the primary drivers of viral propagation, our findings demonstrate that environmental factors act as critical secondary modulators. During periods when behavioral interventions such as lockdowns are relatively constant, these environmental conditions become the primary differentiators of regional risk, acting as powerful amplifiers or dampeners of the underlying transmission dynamics.

Daegu: a climate-driven risk profile

In Daegu, the site of South Korea's initial explosive outbreak, our GAM results identified physical climate as the dominant influence. Low temperature emerged as the primary risk factor, demonstrating a strong non-linear U-shaped relationship where the marginal effect dropped at the coldest sub-zero temperatures before rising sharply to high positive values at warmer extremes. This aligns with extensive literature indicating that cooler conditions provide further evidence for the stability and survival of coronaviruses on surfaces and in aerosols, while extreme heat alters transmission dynamics through behavioral changes [22, 24, 32]. To understand the mechanics of this aerosolization, ambient temperature and relative humidity are the primary drivers of respiratory droplet fate [33]. Within specific humidity ranges such as the 50%–90% RH risk zone identified in our model, droplets evaporate into smaller droplet nuclei that remain suspended for extended periods, enhancing long-range transmission risk [25]. Crucially, cooler temperatures preserve viral infectivity significantly longer than higher temperatures [34]. Furthermore, the observed U-shaped relationship likely reflects a behavioral component, extreme temperatures drive occupants indoors and increase close contact clustering, a behavioral pattern documented during quarantine periods [35], further amplifying transmission risk beyond direct viral stability factors [36]. These findings suggest that in the absence of overwhelming urban pollution loads, physical climate parameters act as the primary environmental regulators of transmission intensity, aligning well with previous work by Lim et al. [13] and Shi et al. [9] regarding the general association between temperature and viral spread [37–39].

Seoul: a pollution-driven risk profile

In contrast, the analysis for Seoul revealed a risk profile dominated by air pollutants, particularly NO_2 and O_3, which showed clear positive monotonic relationships with higher case incidence. Crucially, our model uncovered an indirect transmission mechanism, identifying high-risk days during warm, dry weather. Warm, stagnant, and sunny conditions act as a photochemical reactor in the polluted urban basin, accelerating the conversion of primary precursor emissions like NO_x from traffic into high concentrations of secondary pollutants, specifically ground-level ozone [40]. We verified this mechanism by finding a strong positive correlation between average temperature and O_3 levels ($r = +0.794$) in Seoul. Biologically, inhalation of ground-level O_3 induces potent oxidative stress and airway inflammation [41], which likely increases population susceptibility to viral infection [10, 11]. Notably, chronic exposure to these urban pollutants is hypothesized to upregulate ACE2 receptors the primary entry point for SARS-CoV-2 thereby mechanistically increasing viral susceptibility [42]. This specific respiratory vulnerability is strongly corroborated by the pathophysiological mechanisms outlined by Amritha et al., who note that NO_2 exposure explicitly elevates the risk of viral infections. Furthermore, their global assessment highlights that even short-term exposure to NO_2 is positively associated with severe respiratory distress, linking every 10 $\mu g/m^3$ increase in NO_2 to a 10% excess risk in emergency respiratory visits [31]. This observation is supported by research linking small increases in air pollution to significant rises in COVID-19 cases and mortality [12]. While some research has shown a general link between air quality and COVID-19 in Korea [43], our model is the first to demonstrate this substantial influence in a direct comparative analysis.

Methodological complementarity and model strengths

The main strength of our study is how it combines different analytical tools to get a much clearer picture of the data. By pairing Generalized Additive Models (GAMs) with precise risk-zone binning and broad climate clustering, we created a highly complementary approach. While K-means clustering gives us the big picture by identifying broad danger zones such as a Cold & Humid climate, our granular binning zooms in to find the exact sub-zero temperature ranges where peak transmission actually happens. Furthermore, the simultaneous evaluation of variables within our GAM framework clarifies the hierarchy of these drivers, particularly regarding compound meteorological effects. For instance, our Seoul model indicated a relatively flat marginal dependence on rainfall and wind speed. While these parameters do not exhibit a strong independent effect on viral incidence in the highly urbanized multivariate framework, they operate as critical distal drivers. Existing literature demonstrates that meteorological conditions such as wind speed and precipitation have significant effects on the dispersion, dilution, and accumulation of local air pollutants by [15, 31]. Because our model includes both atmospheric conditions and pollutants simultaneously, the explanatory power of the explicit pollutant predictors absorbs the upstream variance attributable to these meteorological drivers. Therefore, the compound effects of wind and rainfall are inherently and accurately captured through the fluctuations they cause in primary pollutant concentrations.

Comparative synthesis and broader public health implications

The notable contrast between the risk profiles of Daegu and Seoul demonstrates that environmental drivers of respiratory pandemics are overwhelmingly shaped by local urban characteristics. This divergence in pollutant profiles can be further contextualized using the NO_2/CO ratio framework introduced by Pathak et al. (2023) [30] to identify emission sources. According to their methodology, high NO_2/CO ratios indicate traffic dominated conditions in cities, whereas smaller ratios suggest contributions from other sources such as domestic or industrial burning. Applying this to our findings, the NO_2 driven risk profile in Seoul points to a heavily traffic dominated transmission environment, whereas Daegu's distinct pollutant mix characterized by higher CO levels indicates a more complex atmospheric baseline driven by alternative regional emissions. The main insight is that the cities differ not only in their profiles, but also in how risks develop. In Daegu, we see a relatively direct pathway where weather conditions drive transmission, while in Seoul, the pathway is more indirect, with weather mainly helping to build up pollution from human activities. This helps explain the unexpected high-risk during warmer weather in the capital it's a sign of trapped pollutants in the air, rather than the heat alone. The broader public health lesson from this tale of two cities is that a one - size - fits all, national level approach to environmental monitoring during a pandemic is inadequate. A national average model underestimates cold-weather risks in Daegu while entirely overlooking warm-weather pollution risks in Seoul. Effective pandemic preparedness requires a paradigm shift toward localized environmental monitoring, utilizing non-linear analytical tools like the one proposed here to customize early warning systems and proactive public health responses.

Policy and recommendations

The profound regional differences in environmental drivers of COVID-19 transmission identified in this study offer critical insights for shaping future public health policy and preparedness strategies against respiratory viral outbreaks. Our findings underscore that a uniform, national-level environmental warning system is likely insufficient and could misallocate resources. Therefore, we propose a multifaceted, region-specific policy framework centered on the implementation of Localized Environmental Early Warning Systems (EEWS) with triggering criteria tailored to specific regional risk profiles. For areas like Daegu, where our analysis identified a climate-driven risk profile, EEWS alerts should be prioritized based on specific meteorological thresholds. Advisories should be issued during forecasts of sustained cold temperatures (−3.6°C to 5.4°C), and heightened when these temperatures coincide with moderate humidity levels (46% to 53%), a combination identified as a primary high-risk zone. Public advisories under these specific conditions should recommend enhanced indoor ventilation strategies to counteract increased indoor crowding and emphasize personal protective measures in cold-weather settings [44].

Conversely, for dense urban centers like Seoul, where air pollution emerges as the dominant driver, EEWS criteria must shift to air quality indices. Alerts should be triggered specifically during meteorological conditions proven to trap pollution. Based on our correlation analysis ($r = +0.794$ for temperature-O_3), alerts should be prioritized during warm, stagnant periods (temperatures >19.6°C with low wind speeds) which potentially contributes to high concentrations of secondary pollutants like O_3 and NO_2. In these scenarios, actionable guidance should focus on reducing outdoor exposure during peak pollution hours, utilizing indoor air purification, and issuing specific warnings for vulnerable populations. Beyond early warning, epidemiological models such as SIR or SEIR variants must be updated to explicitly incorporate local environmental variables as dynamic predictors. This should be coupled with targeted educational campaigns. Citizens must be informed about specific local environmental risks, and clear communication strategies must articulate the distinction between direct climate effects and indirect pollution-driven effects, empowering individuals to make informed decisions regarding their daily activities and protective measures [45].

The necessity of such long-term, localized EEWS frameworks is ultimately validated by the transient nature of the pandemic 'anthropause'. As demonstrated by Pathak et al. (2023) [30] and Amritha et al. (2025) [31], the air quality improvements observed during the COVID-19 lockdowns were strictly temporary. Following the easing of restrictions, both CO and NO_2 emissions rebounded sharply, often returning to or exceeding pre-lockdown levels due to the rapid resumption of traffic and industrial activities. Because temporary mobility restrictions do not provide sustainable environmental safety, our proposed EEWS is an essential, permanent necessity for mitigating future environmentally-driven pandemic threats.

Limitations

Several limitations of this study must be acknowledged. First, as a retrospective observational study focused on a single pandemic wave (January to July 2020), we cannot establish direct biological causation, and this finite temporal scope limited long-term predictive modeling. Second, using daily case counts as the dependent variable introduces potential confounding from testing intensity and the existing infectious pool. While our 7-day moving average partially addresses reporting delays, it does not fully capture the delayed biological response to environmental stressors. Future studies should utilize distributed lag non-linear models (DLNMs) to identify critical exposure windows, and ideally model the effective reproduction number (R_t) to more precisely isolate transmission dynamics. Furthermore, our analysis may be subject to temporal confounding; for instance, Daegu's explosive cold-weather superspreading event may be partially conflated with our observed temperature correlations, which is distinct from Seoul's more prolonged transmission. Methodologically, relying on central outdoor monitoring stations introduces exposure measurement uncertainties by assuming uniform city-wide exposure, potentially missing localized micro-climates. The K-means clustering approach also introduces a degree of inherent subjectivity. Most critically, our models are purely environmental and omit public health interventions (e.g., social distancing, mask mandates) and behavioral changes, which are often the dominant drivers of pandemic trends, as demonstrated by Anderson et al. [7]. To enhance model accuracy, future iterations should integrate our environmental risk framework with mechanistic epidemiological models like SIR or SEIR [46, 47]. Additionally, incorporating indoor environmental data (such as building ventilation and air purifier usage), detailed human mobility metrics, and specific socio-demographic characteristics like age and gender will significantly improve the model's robustness. Finally, while our GAM framework demonstrates strong explanatory power achieving 63.0% and 48.3% explained deviance for Daegu and Seoul, respectively, caution is required if applying this approach to other

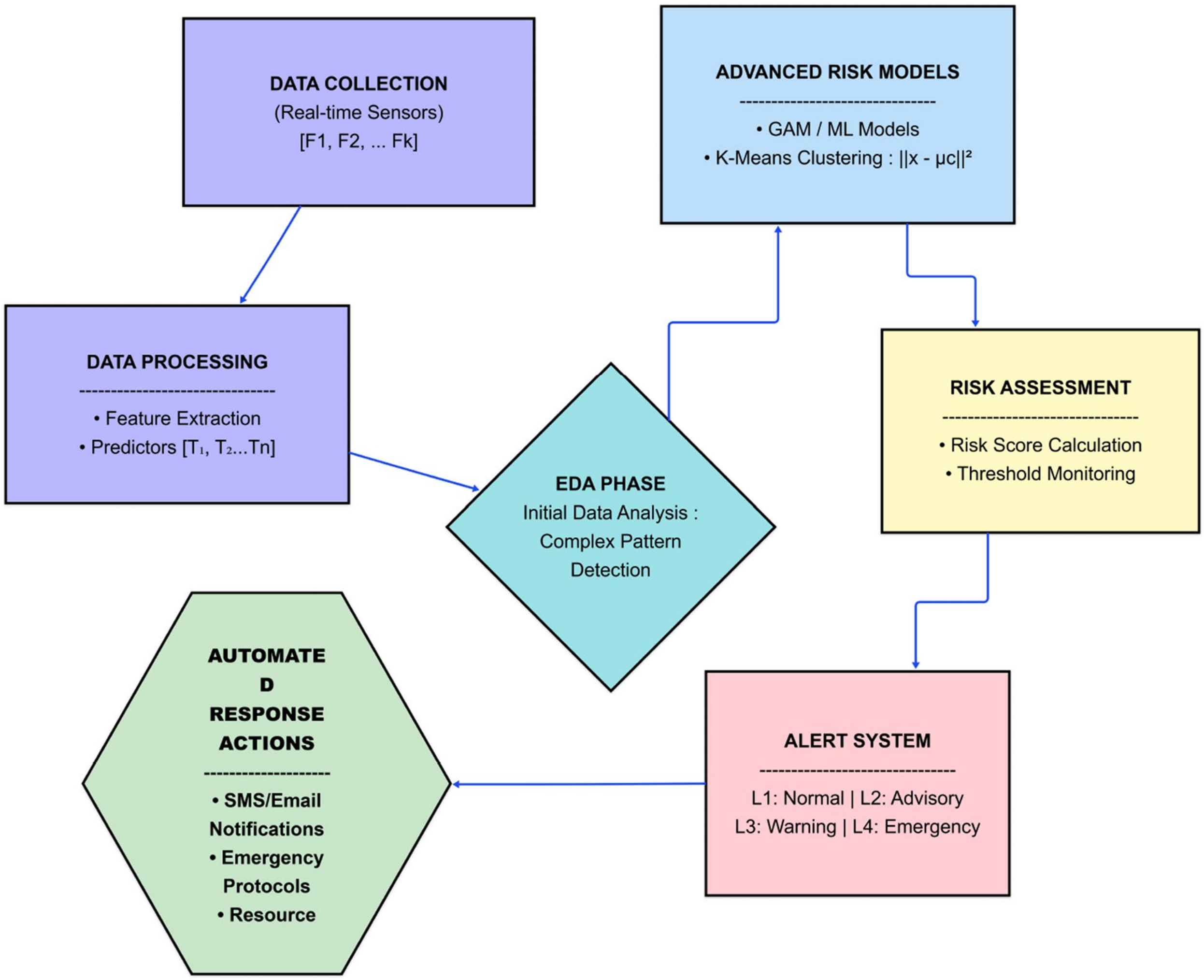

Fig. 12 A robust analytical framework for developing an early warning system for localized disease/disaster outbreaks

global regions. The complex, non-linear relationships between the virus and environmental factors vary significantly based on local climatic, topographic, and socio-economic profiles. Therefore, global application requires that the model's parameters and smoothing functions be strictly tailored to the distinct environmental and emission characteristics of each specific region, rather than applying a universal baseline. Despite these limitations, demonstrating that risk drivers vary significantly by location effectively demonstrates the limitations of a universal model, providing a strong scientific foundation for tailored, localized environmental health surveillance.

Early warning and risk identification system framework

This research study employs a robust comparative analytical framework to systematically identify and model the complex, localized drivers for disease outbreaks or disaster impacts. This framework, visualized in Fig. 12, outlines a transferable, data-driven workflow. Formally, this entire system can be defined as a computational pipeline that transforms a raw dataset D into an automated response A. The process commences with Initial Data Collection (Step 1), where the system ingests a raw dataset D containing k initial factors (features), F : The initial set of factors is defined as:

$$\mathrm{F} = \{f_1, f_2, \ldots, f_k\}.$$

This raw data then undergoes rigorous Data Processing & Feature Selection (Step 2). This stage is critical for refining the dataset. The initial k factors are filtered by a processing function, f_{process}, to yield a final stable set of n predictor variables, T:

$$\mathrm{T} = f_{\text{process}}(\mathrm{F}) = \{t_1, t_2, \ldots, t_n\}$$

Following data refinement, an Initial Exploratory Analysis (Step 3) is conducted to identify complex, non-linear relationships. This step validates the rejection of simple linear models and justifies the necessity for advanced techniques. Formally, this selection step is a conditional choice of the analytical model M based on the output of an exploratory function, f_{eda} :

$$M = \begin{cases} M_{\text{adv}}(\mathrm{T}) & \text{if } f_{\text{eda}}(\mathrm{T}) \text{ is True (Complex Patterns Detected)} \\ M_{\text{base}}(\mathrm{T}) & \text{if } f_{\text{eda}}(\mathrm{T}) \text{ is False (Simple Patterns)} \end{cases}$$

The core of the framework is the Advanced Explanatory Modeling (Step 4), M_{adv}, which employs a dual-pronged approach:

- Predictive Model (M_{gam}) : First, a Generalized Additive Model (GAM) is applied to quantify the impact of each individual driver and generate a 'Real-time Risk Prediction', $\text{Risk}_{\text{pred}}$:

$$\text{Risk}_{\text{pred}} = M_{\text{gam}}(\mathrm{T})$$

- Classification Model (M_{kmeans}) : Concurrently, K-Means Clustering (M_{kmeans}) is applied for 'Risk Zone Classification', partitioning the data into *c* distinct clusters (zones) by minimizing the within-cluster sum of squares (WCSS), *J*:

$$J = \sum_{i=1}^{c} \sum_{\mathrm{t} \in C_i} \|\mathrm{t} - \mu_i\|^2$$

(Where C_i is the i-th cluster and μ_i is its centroid). Finally, the framework synthesizes these two outputs (Step 6) in the "Risk Assessment & Threshold Monitoring" stage. This is performed by a scoring function, f_{score}, which calculates a single, continuous Risk Score, *R*:

$$R = f_{\text{score}}(\text{Risk}_{\text{pred}}, J)$$

This synthesis culminates in the final, actionable output. The continuous Risk Score *R* is mapped to one of four discrete alert levels, *L*, by an alert classification function, f_{alert}:

$$L = f_{\text{alert}}(R) = \begin{cases} \text{Level 1: Normal} & \text{if } R < \theta_1 \\ \text{Level 2: Advisory} & \text{if } \theta_1 \le R < \theta_2 \\ \text{Level 3: Warning} & \text{if } \theta_2 \le R < \theta_3 \\ \text{Level 4: Emergency} & \text{if } R \ge \theta_3 \end{cases}$$

(Where $\theta_1, \theta_2, \theta_3$ are the predefined risk thresholds.) This alert level L then triggers a corresponding Automated Response Action (Step 7), A, via a response function, f_{response}:

$$A = f_{\text{response}}(L)$$

The entire framework, from initial data factors F to the final action A, can thus be expressed as a single composite function:

$$A = (f_{\text{response}} \circ f_{\text{alert}} \circ f_{\text{score}} \circ M_{\text{adv}} \circ f_{\text{process}})(\mathrm{F})$$

The theoretical framework outlines the computational and methodological foundation for a responsive, local alert system. However, turning these analytical results into effective public health practice means translating numerical risk scores into concrete, actionable strategies. Using the specific findings from our local pandemic experiment the climate-driven risk in Daegu and the pollution-driven dynamics in Seoul we now set out targeted policy recommendations. These proposals are meant to connect data driven risk identification with real world operational response, making sure that future intervention strategies are as complex and local as the environmental drivers they aim to reduce.

Conclusion & future directions

This study investigated the complex, non-linear effects of meteorological and air quality factors on COVID-19 transmission in Daegu and Seoul, South Korea, using Generalized Additive Models (GAMs). Our analysis Strongly indicates that environmental drivers of COVID-19 are significantly region-specific, challenging a universal approach to pandemic preparedness. In the Daegu basin, primary meteorological factors wielded the most significant influence on transmission. Specifically, cold temperatures and moderate-to-high humidity created high-risk climatic regimes that drove elevated incidence rates. Conversely, in the highly urbanized environment of Seoul, anthropogenic air pollution particularly NO_2 and CO emerged alongside temperature as the dominant environmental drivers. Here, specific meteorological conditions acted as catalysts, concentrating urban pollutants and increasing population susceptibility. By quantifying these non-linear relationships and ranking variable importance, this research confirms that the interplay between climate, pollution, and viral spread is highly localized. While acknowledging the inherent limitations of an observational study, these findings highlight the necessity of transitioning from broad policies to tailored, data-driven Environmental Early Warning Systems (EEWS). Such systems could proactively leverage local forecasts to trigger targeted public health advisories focusing on cold/humid conditions in areas like Daegu, and high-pollution events in urban centres like Seoul. Future research must expand this framework to diverse urban typologies and integrate non-environmental data streams, such as human mobility patterns, to refine predictive capabilities. Ultimately, this localized methodological approach provides a robust, scalable model for enhancing precision in pandemic response and mitigating the impact of future respiratory disease outbreaks.

Appendix

This appendix provides supplementary information, including detailed model summaries and diagnostic plots, to support the findings presented in the main text.

Daegu GAM summary

Figure 13 shows the full model summary output for the Daegu GAM.

Generalized additive model summaries

This section presents the detailed summary statistics for the Generalized Additive Models (GAMs) fitted for both Daegu and Seoul. These summaries include the significance of smooth terms (approximate significance of smooth terms), estimated degrees of freedom (EDoF), and model convergence information.

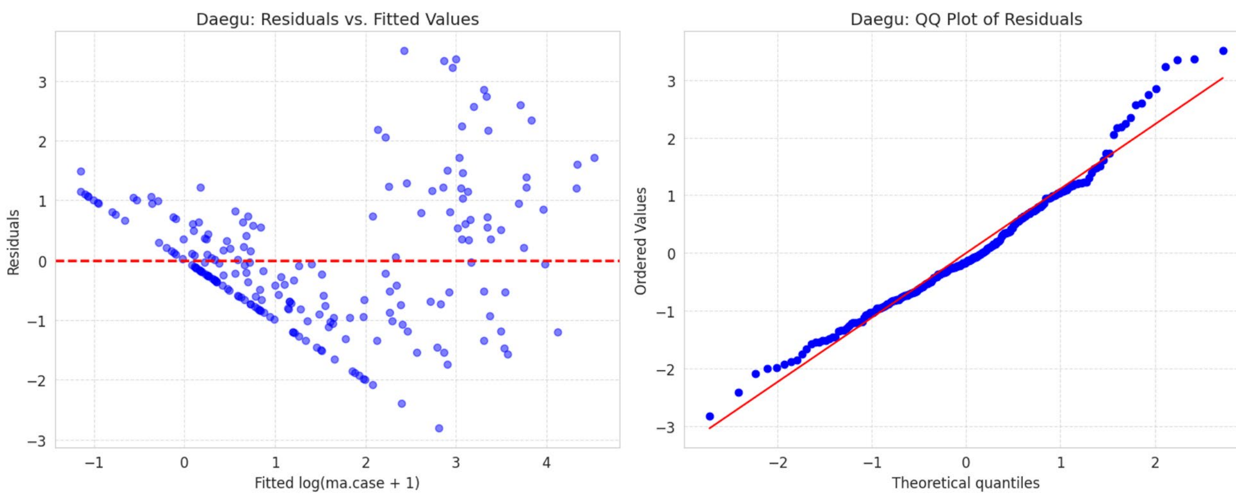

Fig. 13 Diagnostic plots for the Daegu GAM. Left panel: residuals versus fitted values. Right panel: Q-Q plot of residuals. The residuals are dispersed randomly around the horizontal zero line with no systematic pattern, indicating that the Gaussian family assumption is appropriate. In the Q-Q plot, the alignment of residuals along the diagonal reference line confirms the assumption of normally distributed errors

Seoul GAM summary

Figure 14 shows the full model summary output for the Seoul GAM.

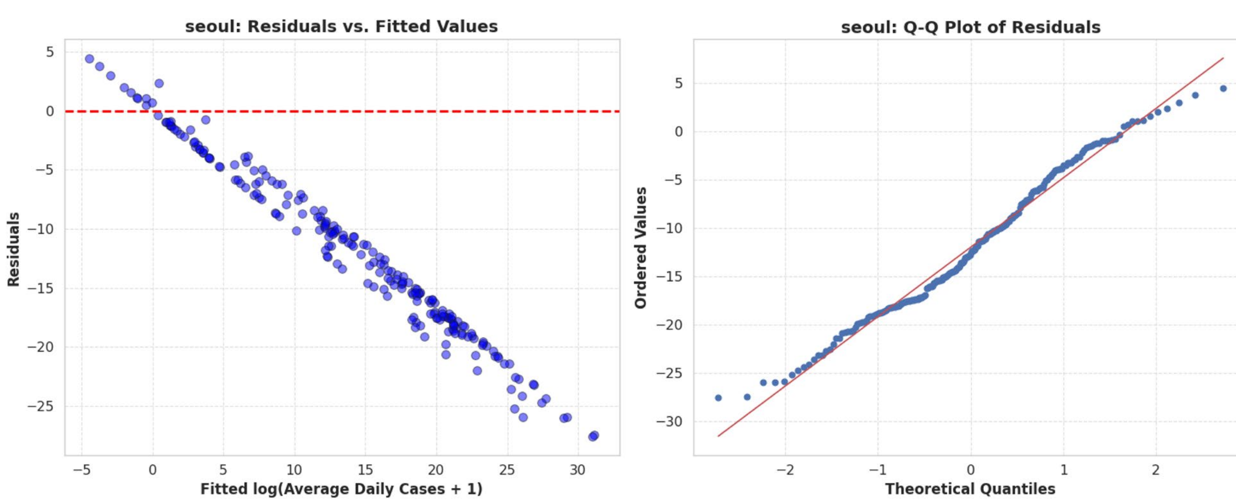

Fig. 14 Diagnostic plots for the Seoul GAM. Left panel: residuals versus fitted values, the downward diagonal trend illustrates the floor effect of zero-case observations in the dataset. Right panel: Q-Q plot of residuals. The close alignment of residuals along the diagonal reference line validates the assumption of normality, confirming that the Gaussian family specification is a reasonable approximation for the Seoul dataset

Simple correlation matrices of Daegu and Seoul

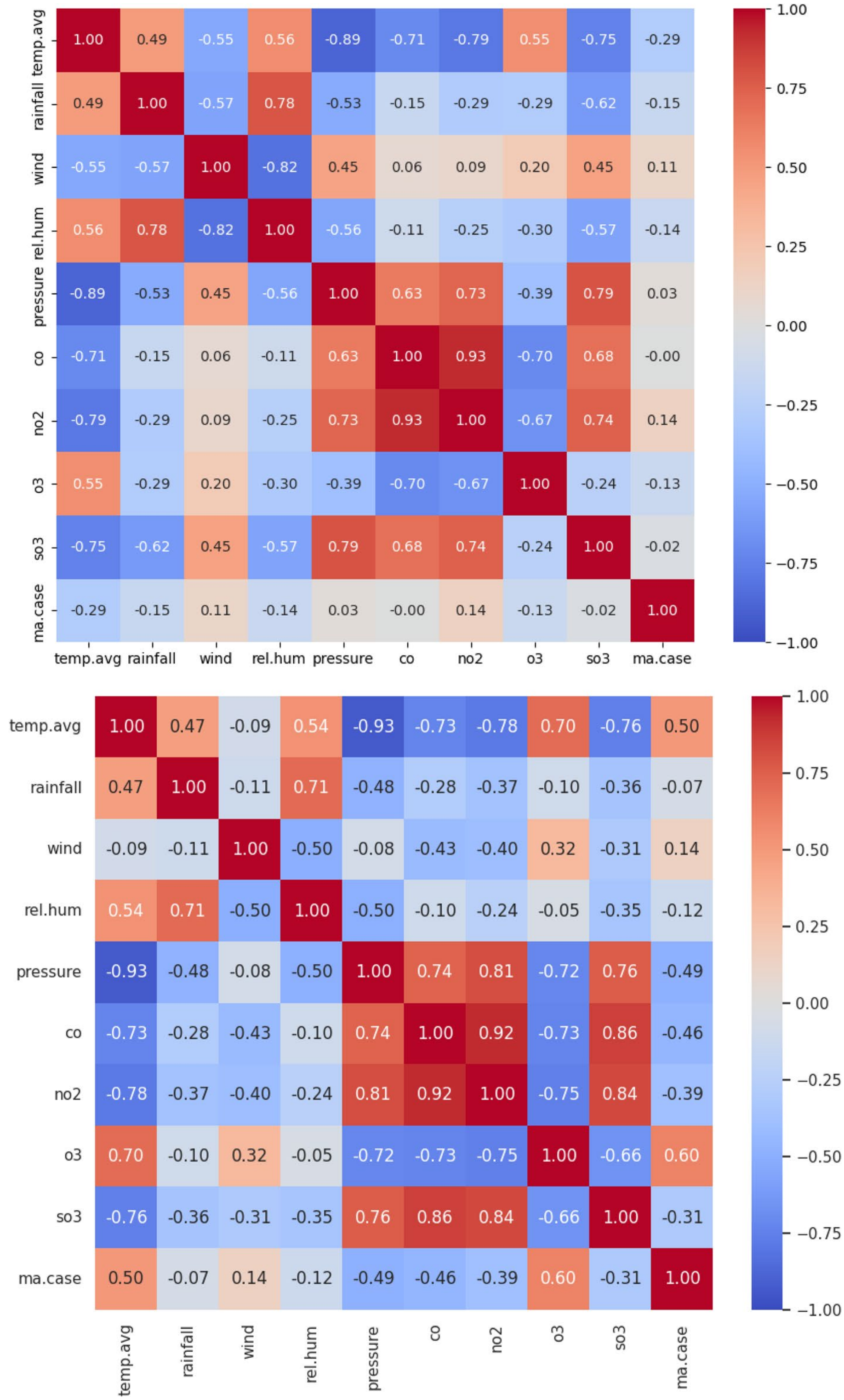

Fig. 15 Spearman correlation heatmaps of daily environmental variables and log-transformed COVID-19 cases for Daegu (first) and Seoul (second) during the study period

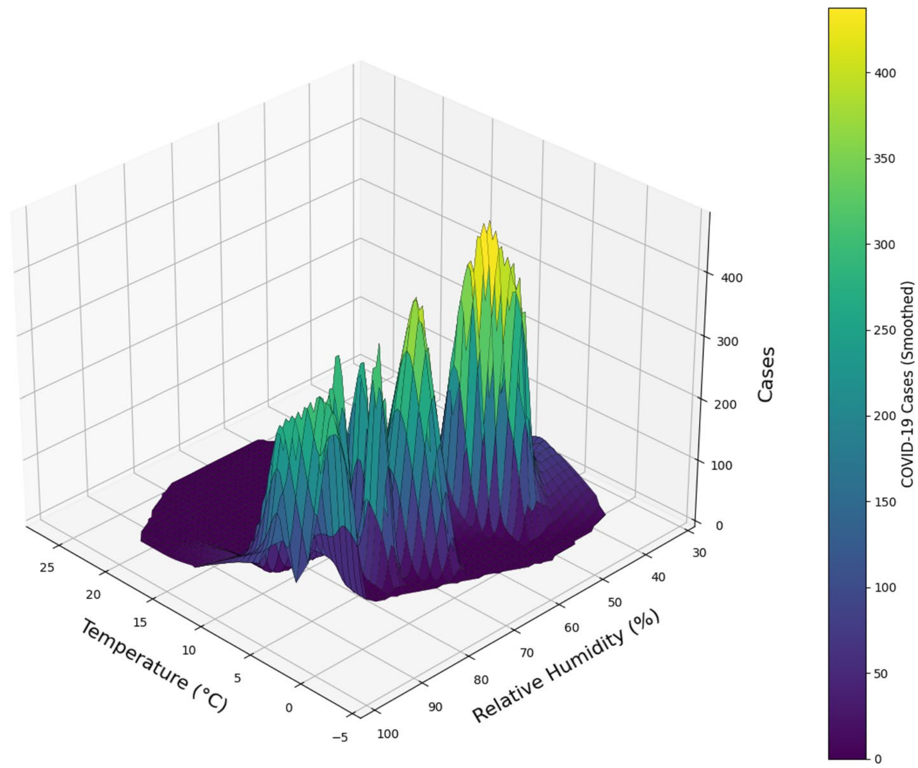

Fig. 16 (A): Daegu bivariate relationship between temperature, humidity and COVID-19 cases

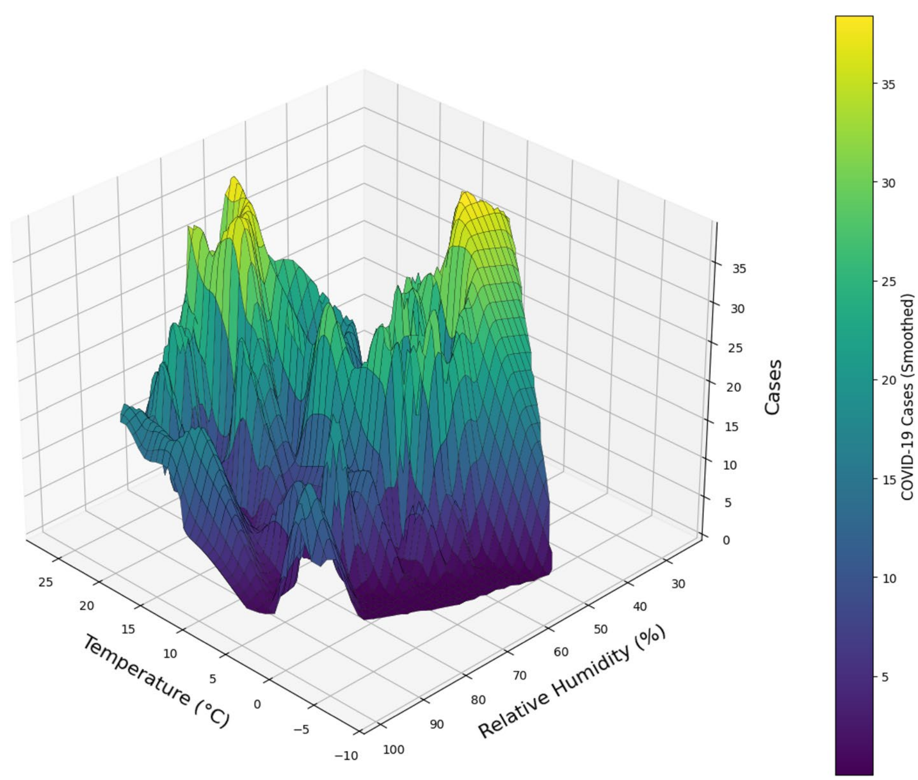

Fig. 17 (B) : Seoul bivariate relationship between temperture, humidity and COVID-19 cases

Identified highest and lowest risk zone combinations of Seoul and Daegu region

Table 4 Identified high-risk and low-risk temperature and humidity zones associated with COVID-19 transmission in Seoul and Daegu regions

City	Temperature (°C)	Humidity (%)	Risk rate	Observations	Risk level
Seoul					
Seoul	(19.6, 23.05]	(69.5, 76.75]	29.76	9	High
Seoul	(23.05, 26.5]	(69.5, 76.75]	29.29	11	High
Seoul	(9.25, 12.7]	(62.25, 69.5]	27.53	3	High
Seoul	(2.35, 5.8]	(55.0, 62.25]	26.20	4	High
Seoul	(9.25, 12.7]	(40.5, 47.75]	26.20	3	High
Seoul	(23.05, 26.5]	(62.25, 69.5]	25.37	6	High
Seoul	(5.8, 9.25]	(55.0, 62.25]	23.25	8	High
Seoul	(19.6, 23.05]	(84.0, 91.25]	21.47	6	High
Seoul	(-1.1, 2.35]	(76.75, 84.0]	0.00	3	Low
Seoul	(-4.55, -1.1]	(62.25, 69.5]	0.05	4	Low
Seoul	(-4.55, -1.1]	(55.0, 62.25]	0.08	5	Low
Seoul	(-1.1, 2.35]	(69.5, 76.75]	0.13	6	Low
Seoul	(2.35, 5.8]	(69.5, 76.75]	2.35	4	Low
Seoul	(-1.1, 2.35]	(62.25, 69.5]	3.35	4	Low
Seoul	(-1.1, 2.35]	(55.0, 62.25]	3.43	7	Low
Seoul	(9.25, 12.7]	(55.0, 62.25]	4.10	4	Low
Daegu					
Daegu	(2.4, 5.4]	(46.12, 52.63]	278.80	7	High
Daegu	(5.4, 8.4]	(78.67, 85.18]	181.52	5	High
Daegu	(5.4, 8.4]	(72.16, 78.67]	167.49	9	High
Daegu	(5.4, 8.4]	(39.61, 46.12]	74.95	4	High
Daegu	(2.4, 5.4]	(72.16, 78.67]	40.00	5	High
Daegu	(11.4, 14.4]	(52.63, 59.14]	39.60	3	High
Daegu	(8.4, 11.4]	(52.63, 59.14]	34.75	4	High
Daegu	(8.4, 11.4]	(59.14, 65.65]	31.68	5	High
Daegu	(8.4, 11.4]	(46.12, 52.63]	31.13	3	High
Daegu	(-0.6, 2.4]	(46.12, 52.63]	0.00	6	Low
Daegu	(-0.6, 2.4]	(52.63, 59.14]	0.00	7	Low
Daegu	(-0.6, 2.4]	(59.14, 65.65]	0.00	5	Low
Daegu	(-3.63,-0.6]	(46.12, 52.63]	0.07	3	Low
Daegu	(20.4,23.4]	(91.69, 98.20]	0.08	8	Low
Daegu	(23.4,26.4]	(72.16, 78.67]	0.32	5	Low
Daegu	(23.4,26.4]	(59.14, 65.65]	0.50	4	Low
Daegu	(23.4,26.4]	(78.67, 85.18]	0.54	7	Low
Daegu	(20.4,23.4]	(85.18, 91.69]	0.60	7	Low
Daegu	(20.4,23.4]	(72.16, 78.67]	0.60	5	Low
Daegu	(14.4,17.4]	(72.16, 78.67]	0.60	3	Low

Sensitivity analysis

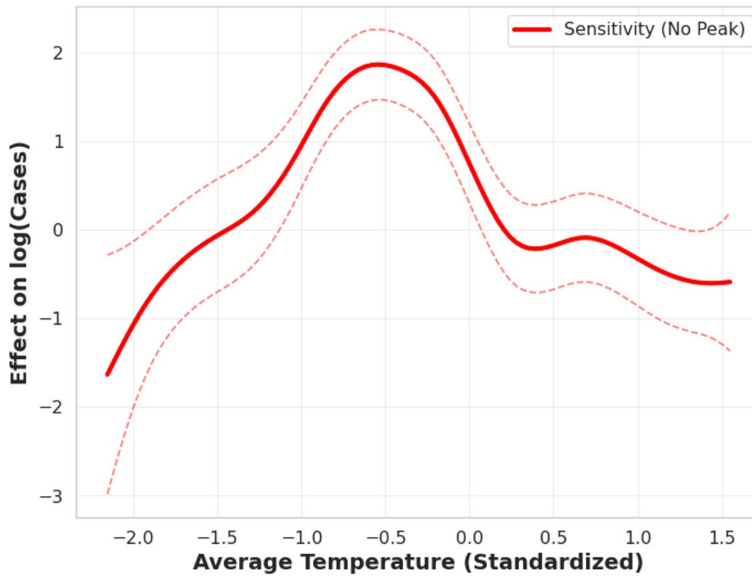

Fig. 18 Sensitivity analysis of the relationship between temperature and COVID-19 cases in Daegu. The red solid line represents the partial dependence of average temperature on the log-transformed 7-day moving average of cases, modelled after excluding the peak outbreak period (February 15 – March 15, 2020). Dashed lines indicate 95% confidence intervals. The consistent non-linear shape confirms that the environmental association is robust to the exclusion of the superspreading event

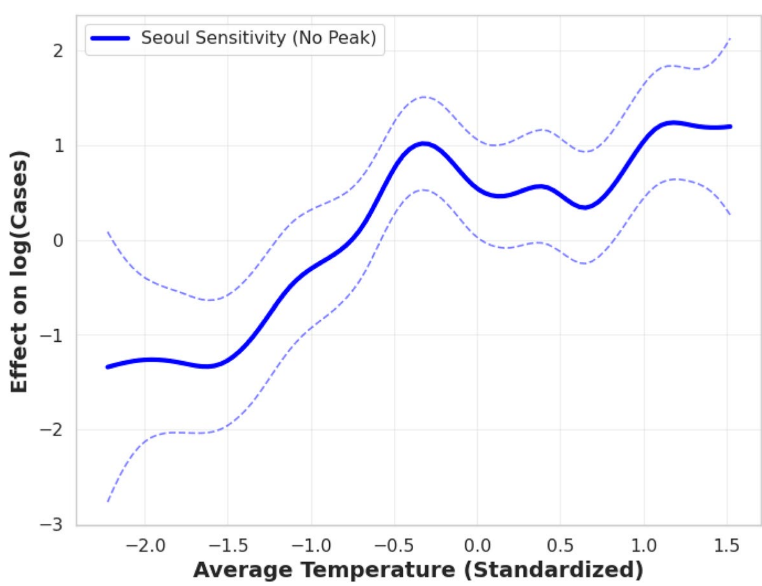

Fig. 19 Sensitivity analysis of the relationship between temperature and COVID-19 cases in Seoul. The blue solid line represents the partial dependence of average temperature modelled after excluding the specific cluster period (March 5 – March 20, 2020). Dashed lines indicate 95% confidence intervals. The persistent trend confirms that the environmental signal remains stable even after removing the primary early infection cluster

Acknowledgements
The research was supported by a grant (RS-2023 00227944) from the National Research Foundation, funded by the Ministry of Science, Technology, and Telecommunication of the government of the Republic of Korea. The research was also supported by a grant of the National Research Foundation of Korea funded by the Korean government (NRF-2022R1A2C2092076). Both grants have been awarded to Asaph Y. Chun, Principal Investigator and Director of the Institute for Pandemic Sciences AI.celerator (IPSAI), in the AI Institute of Seoul National University.

Authors' contributions
G.R.M.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing S.A.J: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing A.Y.C.: Investigation, Methodology, Project administration, Validation, Writing - Review & Editing, Funding acquisition A.J: Methodology, Validation, Investigation, Data Curation, Writing-Review & Editing, H.W.L.: Formal analysis, Investigation, Methodology, Validation, Resources, Supervision, Data Curation, Writing - Review & Editing.

Data availability
Publicly available dataset are utilized for the study for COVID-19 cases, meteorological parameters, and air quality indices were sourced from the KCDC [31], the KMA Open Data Portal [32], and Air Korea [28], respectively. Processed data and analysis code are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 January 2026 / Accepted: 26 May 2026
Published online: 10 June 2026

References

1. Herdicho FF, Jose SA, Jirawattanapanit A, Park T. Fractional derivative models in COVID-19 dynamics: application to symptom severity and hospital resource allocation in South Korea. J Appl Math Comput. 2025;71:3179–209.
2. World Health Organization. Coronavirus disease 2019 (COVID-19): Situation report, 120. 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200519-covid-19-sitrep-120.pdf.
3. World Health Organization. Air pollution. 2018. https://www.who.int/news-room/air-pollution.
4. Hunter RF, Garcia L, Dagless S, Haines A, Penney T, Clifford Astbury C, et al. The emerging syndemic of climate change and non-communicable diseases. Lancet Planet Health. 2024;8(7):e430-1. https://doi.org/10.1016/S2542-5196(24)00112-8.
5. Omede BI, Jose SA, Anuwat J, Park T. Impact of imperfect vaccination and re-infection on the dynamics of Delta and Omicron COVID-19 variants in the USA. Model Earth Syst Environ. 2024;10:7383–420.
6. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan. China. JAMA. 2020;323(11):1061–9. https://doi.org/10.1001/jama.2020.1585.
7. Anderson RM, et al. How will country-based mitigation measures influence the course of the COVID-19 epidemic? Lancet. 2020;395(10228):931–4.
8. Korean Ministry of Environment. AIRKOREA: A Platform for Air Quality Information. 2024. Retrieved from https://www.airkorea.or.kr/eng/AIRKOREA.
9. Shi P, Dong Y, Yan H, Li X, Zhao C, Liu W, et al. The impact of temperature and absolute humidity on the coronavirus disease 2019 (COVID-19) outbreak - evidence from China. medRxiv. 2020. https://www.medrxiv.org/content/10.1101/2020.03.22.20038919v1.
10. Maheswari S, Pethannan R, Sabarimurugan S. Air pollution enhances susceptibility to novel coronavirus (COVID-19) infection - an impact study. Environ Anal Health Toxicol. 2020;35(4):e2020020. https://doi.org/10.5620/eaht.2020020.
11. Pozzer A, Dominici F, Haines A, Witt C, Münzel T, Lelieveld J. Regional and global contributions of air pollution to risk of death from COVID-19. Cardiovasc Res. 2020;116(14):2247–53. https://doi.org/10.1093/cvr/cvaa288.
12. He G, Pan Y, Tanaka T. The causal effect of air pollution on COVID-19 transmission: evidence from China. medRxiv. 2020. https://doi.org/10.1101/2020.10.19.20215236.
13. Lim YK, Kweon OJ, Kim HR, Kim TH, Lee MK. The impact of environmental variables on the spread of COVID-19 in the Republic of Korea. Sci Rep. 2021;11(1):5977. https://doi.org/10.1038/s41598-021-85493-y.
14. Zhao M, Liu Y, Gyilbag A. Assessment of meteorological variables and air pollution affecting COVID-19 cases in urban agglomerations: Evidence from China. Int J Environ Res Public Health. 2022;19(1):531. https://doi.org/10.3390/ijerph19010531.
15. Sahoo MM. Significance between air pollutants, meteorological factors, and COVID-19 infections: Probable evidences in India. Environ Sci Pollut Res. 2021;28:40474–95. https://doi.org/10.1007/s11356-021-12709-z.
16. Hoang T, Nguyen TQ, Tran TTA. Short-term exposure to ambient air pollution in association with COVID-19 of two clusters in South Korea. Trop Med Int Health. 2021;26(4):478–91. https://doi.org/10.1111/tmi.13538.
17. Lee HW, Malik N, Shi F, Mucha PJ. Social clustering in epidemic spread on coevolving networks. Phys Rev E. 2019;6:99.
18. MacIntyre CR, Chen X, Kunasekaran M, Quigley A, Lim S, Stone H, et al. Artificial intelligence in public health: the potential of epidemic early warning systems. J Int Med Res. 2023;51(3):3000605231159335. https://doi.org/10.1177/03000605231159335.
19. Mao K, Zhang K, Du W, Ali W, Feng X, Zhang H. The potential of wastewater based epidemiology as surveillance and early warning of infectious disease outbreaks. Curr Opin Environ Sci Health. 2020. https://doi.org/10.1016/j.coesh.2020.04.006.
20. Villanueva-Miranda I, Xiao G, Xie Y. Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review. Front Public Health. 2025;13:1609615. https://doi.org/10.3389/fpubh.2025.1609615.
21. Feng C. Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada. Spat Stat. 2021;49. https://www.sciencedirect.com/science/article/pii/S2211675321000361.
22. Das M, Ghosh A, Ghosh SK. Does climate variability impact COVID-19 outbreak? An enhanced semantics-driven theory-guided model. SN Comput Sci. 2021;2(6):452. https://doi.org/10.1007/s42979-021-00845-9.
23. James et al. Interactions between climate and COVID-19. Lancet Planet Health. 2022;6. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(22)00174-7/fulltext.
24. Zhang C, Guo Z. The impact of temperature and relative humidity on SARS-CoV-2 airborne transmission in Syrian hamsters. Microbiol Spectr. 2025;13:e00979-25. https://doi.org/10.1128/spectrum.00979-25.
25. Morris DH, Yinda KC, Gamble A, Rossine FW, Huang Q, Bushmaker T, et al. Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. eLife. 2021;10:e65902. https://doi.org/10.7554/eLife.65902.
26. Weather Spark. Climate data computed from Weather Spark. 2025. https://weatherspark.com.
27. Shim E, Tariq A, Choi W, Lee Y, Chowell G. Transmission potential and severity of COVID-19 in South Korea. Int J Infect Dis. 2020;93:339–44. https://doi.org/10.1016/j.ijid.2020.03.031.
28. Korea Centers for Disease Control and Prevention. Korea Centers for Disease Control and Prevention (KCDC) website. https://www.kdca.go.kr/index.es?sid=a3.
29. Korea Meteorological Administration. Korea Meteorological Administration (KMA) database. https://www.kma.go.kr/neng/index.do.
30. Pathak M, Patel VK, Kuttippurath J. Spatial heterogeneity in global atmospheric CO during the COVID-19 lockdown: implications for global and regional air quality policies. Environ Pollut. 2023;335:122269. https://doi.org/10.1016/j.envpol.2023.122269.
31. Amritha S, Patel VK, Kuttippurath J, Gopikrishnan GS, Varikoden H. Global changes in the hazardous atmospheric NO_2 during the COVID-19 lockdown and post-lockdown periods. J Hazard Mater Adv. 2025;20:100911. https://doi.org/10.1016/j.hazadv.2025.100911.

32. Wang J, Tang K, Feng K, Lin X, Lv W, Chen K, et al. Impact of temperature and relative humidity on the transmission of COVID-19: A modeling study in China and the United States. BMJ Open. 2021;11(2):e043863. https://doi.org/10.2139/ssrn.3551767.
33. Rezaei M, Netz RR. Airborne virus transmission via respiratory droplets: Effects of droplet evaporation and sedimentation. J Aerosol Sci. 2021;55. https://www.sciencedirect.com/science/article/pii/S1359029421000558.
34. Pyankov OV, Bodnev SA, Pyankova OG, Agranovski IE. Survival of aerosolized coronavirus in the ambient air. J Aerosol Sci. 2018. https://doi.org/10.1016/j.jaerosci.2017.09.009.
35. Fu Y-C, Lee H-W. Daily contacts under quarantine amid limited spread of COVID-19 in Taiwan. Int J Sociol. 2020;50(5):434–44.
36. Zhang A, Zhen Q, Zheng C, Li J, Zheng Y, Du Y, et al. Assessing the impact of architectural and behavioral interventions for controlling indoor COVID-19 infection risk: an agent-based approach. J Build Eng. 2023. https://doi.org/10.1016/j.jobe.2023.106807.
37. Pica N, Bouvier NM. Environmental factors affecting the transmission of respiratory viruses. Curr Opin Virol. 2012;2(1):90–5. https://doi.org/10.1016/j.coviro.2011.12.003.
38. Liu X, Sun J, Liu X, Fan J, Niu Y, Lang L, et al. Ambient temperature and major infectious diseases in China. In: Ambient temperature and health in China. 2019. pp. 75–93. https://doi.org/10.1007/978-981-13-2583-05.
39. Adekunle IA, Tella SA, Oyesiku KO, Oseni IO. Spatio-temporal analysis of meteorological factors in abating the spread of COVID-19 in Africa. Heliyon. 2020;6(8):e04749. https://doi.org/10.1016/j.heliyon.2020.e04749.
40. Sillman S. The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. Atmos Environ. 1999. https://doi.org/10.1016/S1352-2310(98)00345-8.
41. Enweasor C, Flayer CH, Haczku A. Ozone-induced oxidative stress, neutrophilic airway inflammation, and glucocorticoid resistance in asthma. Front Immunol. 2021;12:631092. https://doi.org/10.3389/fimmu.2021.631092.
42. Paital B, Agrawal PK. Air pollution by NO_{2} and PM2.5 explains COVID-19 infection severity by overexpression of Angiotensin-converting enzyme 2 in respiratory cells: A review. Environ Chem Lett. 2021;19(1):25–42. https://doi.org/10.1007/s10311-020-01091-w.
43. Seo JH, Jeon HW, Sung UJ, Sohn J-R. Impact of the COVID-19 outbreak on air quality in Korea. Atmosphere. 2020;11(10):1137. https://doi.org/10.3390/atmos11101137.
44. Lee HW, Leng CH, Chan TC. Determinants of personal vaccination hesitancy before and after the mid-2021 COVID-19 outbreak in Taiwan. PLoS ONE. 2022;17(7):e0270349.
45. Li Y-T, Chen M-L, Lee H-W. Health communication on social media at the early stage of the pandemic: examining health professionals' COVID-19 related tweets. Soc Sci Med. 2024;347:116748.
46. Thomas R, Jose SA, Raja R, Alzabut J, Cao J, Balas VE, et al. Modeling and analysis of SEIRS epidemic models using homotopy perturbation method: A special outlook to 2019-nCoV in India. Int J Biomath. 2022;15(8):2250059.
47. Lee H, Lee J, Jung H, Lee JY. Power of universal health coverage in the era of COVID-19: A nationwide observational study. Available at The Lancet. 2020. https://doi.org/10.1016/S2666-6065(20)30088-2.

Publisher's Note