Balcilar, Mehmet; Mukherjee, Zinnia; Gupta, Rangan et al.

Book

Effect of temperature on the spread of contagious diseases : evidence from over 2000 years of data

University of Pretoria
Department of Economics Working Paper Series

Effect of Temperature on the Spread of Contagious Diseases: Evidence from over 2000 Years of Data

Mehmet Balcilar
University of New Haven and Eastern Mediterranean University and OSTIM Technical University
Zinnia Mukherjee
Simmons University
Rangan Gupta
University of Pretoria
Sonali Das
University of Pretoria
Working Paper: 2023-22
July 2023

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413

Effect of Temperature on the Spread of Contagious Diseases: Evidence from over 2000 Years of Data

Mehmet Balcilar[1], Zinnia Mukherjee[2]*, Rangan Gupta[3], and Sonali Das[4]

Abstract:

The COVID-19 pandemic led to a surge in interest among scholars and public health professionals in identifying the predictors of health shocks and their transmission in the population. With temperature increase becoming a persistent climate stress, our aim is to evaluate how temperature specifically impacts the incidences of contagious disease. Using annual data from 1 AD to 2021 AD on incidence of contagious disease and temperature anomalies, we apply both parametric and non-parametric modelling techniques, and provide estimates on the contemporaneous, and as well as lagged effects, of temperature anomalies on the spread of contagious diseases. A non-homogeneous Hidden Markov Model is then applied to estimate the time-varying transition probabilities between hidden states where the transition probabilities are governed by covariates. For all empirical specifications, we find consistent evidence that temperature anomalies in fact have statistically significant effect of the incidence of the contagious disease in any given year covered in the sample period. The best fit model further indicates that the contemporaneous effect of a temperature anomaly on the response variable is the strongest, and that given temperature anomaly predictions are becoming very accurate, one can prepare effectively with necessary public health response for at least contagious diseases. These findings further have implications for designing cost effective infectious disease control policies for different regions of the world.

Keywords: Temperature anomaly, contagious disease, General additive model, Nonhomogeneous Hidden Markov Model, climate change, public health

JEL Codes: C1, H1, Q0

[1] Department of Economics and Business Analytics, University of New Haven, 300 Boston Post Road, West Haven, CT 06516, USA; Department of Economics, Eastern Mediterranean University, Northern Cyprus, via Mersin 10, Turkey; Department of Economics, OSTIM Technical University, Ankara, Turkey.

[2] Corresponding author. Department of Economics, Simmons University, Boston, MA 02115, U.S.A. Email: zinnia.mukherjee@simmons.edu.

[3] Department of Economics, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa.

[4] Department of Business Management, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa.

I. Introduction:

The global scale and impact of the COVID-19 pandemic led to renewed interests and efforts among scientists, governments, and academicians across different fields to identify ways to minimize the cumulative damage from such future events to human health and the global economy. It is extremely challenging to compare epidemics and pandemics given that a myriad of contextual factors lies behind each occurrence. For example, each pandemic is characterized by a specific point of origin and shaped by a unique line of historical events leading to the onset of that event. The extent of damage to human lives also varies greatly among the major pandemics over the last 2000-plus years. For example, the World Health Organization estimates that the total number of deaths from COVID-19 is close to 7 million[5], whereas the estimated number of deaths caused by the Black Death plague of Central Asia between 1347 and 1351 has been estimated to be between 75 - 200 million (Cirillo and Taleb, 2020). However, in spite of all the challenges inherent in this type of analysis, the effort is always worthwhile given that any lesson learned and applied could potentially lead to scores of human lives being saved during a future occurrence.

A pandemic is defined as a low probability global event with an epicenter located in some part of the world. It can happen decades and even centuries after the occurrence of an earlier pandemic, which may have a different epicenter located thousands of miles away. Given the rarity and uniqueness of such events, it is extremely challenging to identify a common set of contributory factors that cause these events and influence the transmission rates through space and time. Nonetheless, all pandemics are major disruptive events that result in large shocks affecting one or more economies, and the society's overall wellbeing. The expected damages associated with such events can be multifaceted, sometimes affecting multiple generations of the population, albeit in different ways. Unlike earlier pandemics, the 2003 SARS-COVID pandemic and the 2019 COVID-19 happened at a time when ample scientific evidence was available on global warming (Norris et al., 2016), which has spurred interests among scientists and policymakers alike about the relationship between ambient temperature and the transmission rate of contagious diseases.

The Climate Change 2021 report presented by the Intergovernmental Panel on Climate Change presents some startling changes in the global environment that have happened over the past few decades, which are unprecedent in recent human history. For example, Figure 1 below shows that the global surface temperature increased sharply between 1950 and 2020 when contrasted with the relatively moderate rise in surface temperature between 1850 and 1950. The left panel in the following figure[6] shows the changes in the global surface temperature over the past 2000 years using both reconstructed and observed temperature data. The reconstructed data covers the period from 1 AD until the year 2000, whereas the observed data covers the 1850 to 2020 timeframe. The panel on the right focuses on the 1850-2020 period showing the changes in global surface temperature. The black line indicates observed temperature data. The brown line shows simulated temperature data accounting for both natural and human related factors, while the blue line indicates simulated surface temperature data that only accounts from natural climate change related factors. The gap between the brown and the blue lines represents an approximate measure of the rise in global surface temperature that stem from factors related to anthropogenic activities, particularly those following the first industrial revolution.

[5] Source: WHO COVID Dashboard https://COVID19.who.int/.
[6] Source: IPCC Climate Change 2021: The Physical Science Basis.

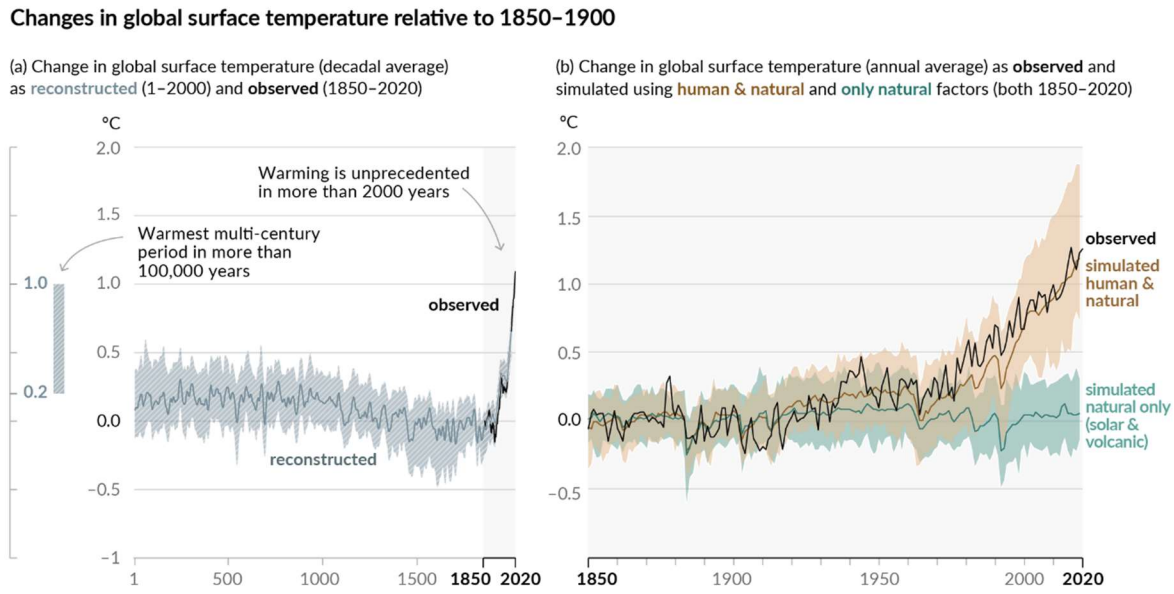


Figure 1: History of global temperature change and causes of recent warming.
Source: IPCC, 2021: Climate Change 2021: The Physical Science Basis (page 6).

Between 1970 and 2020, the global surface temperature increased faster than any other 50-year period over at least the past 2000 years. Hot extreme events, such as heatwaves, have become more frequent and intense since 1950, whereas cold extreme events have become less frequent and less severe. Human induced climate events have led to droughts. While the facts about the changes in the global environment are gravely concerning by themselves, they however do not capture the full magnitude of potential adversity that can stem from such changes in the future, such as hastening the spread of infectious diseases in the future (McDermott, 2022). Geographic boundaries of disease ranges are climate sensitive. They can both shift and expand with changes in temperature through effects of various disease carrying vectors. For example, valley fever is a fungal disease that is endemic to southwestern United States (Gorris et al., 2019), with the regions temperature and precipitation affecting the number of Valley fever cases, and the extent of the spread of the disease across the region. Using climate projections for the 21st century along with a climate niche model derived from contemporary climate and disease incidence data, Corris et al. (2019) predict that throughout this century the endemic region will spread north reaching up to the Canadian border covering western U.S. states and resulting in 50% more cases.

Since contagious diseases often affect the human population through carrier organisms, it is essential to understand the effect of temperature changes on the spread of contagious diseases among wildlife. While the frequency of infectious disease outbreaks among wildlife has increased in recent decades paralleling global climate, the exact mechanisms through which climate change affects the spread of infectious disease largely remains unknown. To address this gap in knowledge, using both laboratory experiments and field prevalence estimates, Cohen et a. (2017) and Cohen et al. (2020) tested the thermal mismatch hypothesis, which posits that cool-adapted host species are more susceptible to pathogen infection during warm temperature periods whereas warm-adapted host species are more susceptible to pathogens during periods of cool temperatures. The datasets used in these studies include a large and highly diverse spectrum of wildlife hosts and parasites that vary in ecologically important traits across a worldwide climatic gradient. Their results confirmed the thermal mismatch hypothesis, which suggests that as climate change shifts hosts away from their optimal temperature ranges, hosts can become more susceptible to infectious diseases though the exact effect will be

dependent on the particular host and the direction of the shift in climate patterns. Another example is that from Morens et al., (2020) who compared the 1918 influenza pandemic with the 2019 COVID pandemic, two disastrous health emergencies caused by different viruses that occurred a century apart from each other. The authors were able to identify similarities in both the clinical, pathological, and epidemiological features of the two pandemics, and in the civic, medical, and public health responses to these events.

In this paper, we take a historical perspective to understand the relationship between contagious disease outbreaks and changes in ambient temperature. Using alternative model specifications, both parametric and nonparametric, we first derive estimates for the causal relationship between temperature anomalies and contagious disease outbreak in any given year, modelled as a binary variable. The time evaluation of the transition probabilities of switching between contiguous disease and non- contiguous disease states or time periods are further studied using a non-homogenous hidden Markov model, under the assumption that the data generated follows a Markov process. Bearing the name of Russian mathematician Andrey Markov, the Hidden Markov Model (HMM) is a stochastic model that is assumed to involve a Markov process in which a sequence of events is characterized by their dependance only on the state that has occurred prior to that event, and not on any preceding states. The Markov process is essentially a stochastic process with a memoryless property, implying that if the current state is given, past states do not play any role in its transition from a current state to a future state. Note, the transition process itself remains unobserved, and is assumed to follow a Markov process. The probability of the process transitioning from a given state in the present to a future state is referred to as a transition probability. A transition matrix provides the set of transition probabilities that describe the likelihood of transition from any present state to all possible future states. The "hidden" in the term refers to the prior states remaining unobserved. HMM models have wide ranging applications in finance (Dias et al., 2015; Mamon and Elliott, 2007), statistics (Genon-Catalot et al., 2000; Scott et al., 2005), cognitive science (Nock and Young, 2002; Dasgupta and Gershman, 2021), mobile communication (Yap and Chong, 2017; Gani et al., 2009), and climatology (Zucchini and Guttorp, 91, Green et al., (2011)). The HMM model can be extended to a nonhomogeneous HMM model (NHMM) if we relax the assumption of homogeneity among the transitions and allow them to depend on additional variables.

The analysis presented in this paper contributes to the strand of academic literature that aims to develop our understanding of how environmental factors affect the outbreak and spread of contagious diseases, both contemporarily and temporally. While the extent of damages associated with every pandemic is characterized by a unique set of contributory factors, comparing different pandemics and identifying similar features can offer valuable lessons. We demonstrate that temperature changes have always played a fundamental role is the spread of contagious diseases. Our results identify a common factor in pandemics covering the past two millenniums. These findings are particularly relevant in developing effective public health strategies to manage future outbreaks and the spread of contagious diseases. These strategies must be designed and implemented while considering the pace of changes in the global environment, driven by global economic expansion, geographic shifts in economic activities, and population growth.

The paper is organized as follows: Section II provides our data sources and presents a description of the dataset. In Section III, we present the methodology used in the analysis, which is followed by a discussion of the results in Section IV. In Section V, we include some reflections and concluding remarks.

II. Data Sources and Description:

The data used in this paper was obtained from the data on contagious diseases presented in Table 1 in Cirillo and Taleb (2020), which runs till 2019. We then include the years 2020 and 2021 as periods associated with the COVID-19 pandemic. Table 1 in this paper follows Cirillo and Taleb (2020) and lists the contagious disease

events included in our analysis.[7,8,9] The table provides information about the primary regions of the world that were affected and estimated death tolls.

[Insert Table 1]

The temperature anomaly data from 1 AD until 2019 were acquired from Hawkins[10] (2020), and then updated from the National Oceanic and Atmospheric Administration (NOAA) until 2021 AD. Table 2 describes the data characteristics for the different variables used in the analysis. The complete dataset including observations on temperature anomalies and contagious disease breakouts contains 2021 observations. It has been divided into two subsamples comprising the "Non-Disease" and "Disease" periods. A temperature anomaly occurs either when the observed temperature is higher than a reference value such as the long-run average value of temperature or (a positive anomaly), or when it is lower than the reference value (a negative anomaly).[11] A high temperature anomaly occurs when the standard deviation between the observed temperature and the reference value is greater than 0.25 points whereas a low temperature anomaly occurs when the difference is less than 0.25 points.

[Insert Table 2]

III. Methodology

Let, $t = 1,2,\dots,2021$ denote the year of observation, d_t denote a binary variable taking value of 1 if a contiguous disease occurred in year t and zero, otherwise; h_t denotes temperature anomaly; and τ_t denote a linear time trend, i.e., $\tau_t = t$.

We start with the linear probability model:

$$d_t = \beta_0 + \beta_1 h_t + \beta_2 h_{t-1} + \beta_3 h_{t-2} + \beta_4 \tau_t + \varepsilon_t \tag{1}$$

where ε_t is an identically and independently distributed error term with zero mean and constant variance σ^2, $\varepsilon_t \sim iid(0, \sigma^2)$. Defining $\boldsymbol{x}_t = (1, h_t, h_{t-1}, h_{t-2}, \tau_t)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$, we can write Eq. (1) as

$$d_t = \boldsymbol{\beta}'\boldsymbol{x}_t + \varepsilon_t, \qquad t = 1,2,\dots,T \tag{1'}$$

Defining $\pi_t = \pi(\boldsymbol{x}_t) = P(d_t = 1|\boldsymbol{x}_t)$, linear probability model implies that $\pi_t = E(d_t = 1|\boldsymbol{x}_t) = \boldsymbol{\beta}'\boldsymbol{x}_t$ while $E(d_t = 0|\boldsymbol{x}_t) = 1 - \pi_t = 1 - \boldsymbol{\beta}'\boldsymbol{x}_t$.

We use the linear probability model as one of the benchmark models. The second benchmark model we use is the logistic probability model defined as:

$$d_t = \frac{\exp\{g(\boldsymbol{x}_t)\}}{1 + \exp\{g(\boldsymbol{x}_t)\}} + \varepsilon_t \tag{2}$$

[7] Source: Cirillo and Taleb (2020) and https://en.wikipedia.org/wiki/List_of_epidemics_and_pandemics#cite_note-38 (retrieved July 24, 2023)
[8] Source: https://www.worldhistory.org/article/1532/plagues-of-the-near-east-562-1486-ce/ (retrieved July 24, 2023)
[9] Source: https://listfist.com/list-of-epidemics-compared-to-coronavirus-COVID-19 (retrieved July 24, 2023)
[10] Source: https://web.archive.org/web/20200202220240/https://www.climate-lab-book.ac.uk/2020/2019-years/.
[11] Source: https://www.ncei.noaa.gov/access/monitoring/global-temperature-anomalies#:~:text=The%20term%20temperature%20anomaly%20means,cooler%20than%20the%20reference%20value.

where the logistic link function $g(\boldsymbol{x}_t)$ is defined as

$$g(\boldsymbol{x}_t) = \log\{\pi(\boldsymbol{x}_t)/[1-\pi(\boldsymbol{x}_t)]\} = \boldsymbol{\beta}'\boldsymbol{x}_t \tag{3}$$

Thus, we can write $d_t = \pi(\boldsymbol{x}_t) + \varepsilon_t$, where $\pi(\boldsymbol{x}_t) = \exp\{g(\boldsymbol{x}_t)\}/[1+\exp\{g(\boldsymbol{x}_t)\}]$. Here, ε_t is distributed with mean zero and variance equal to $\pi(\boldsymbol{x}_t)[1-\pi(\boldsymbol{x}_t)]$.

A generalized additive model (GAM) replaces the logistic link function in Eq. (3) with

$$g(\boldsymbol{x}_t) = \beta_0 + s_1(h_t) + s_2(h_{t-1}) + s_3(h_{t-2}) + s_4(\tau_t) \tag{4}$$

where $s_i(\cdot)$, $i = 1,2,\dots,4$, are univariate smooth functions of their arguments. For the GAM model in Eq. (4), we specify the smooth terms $s_i(\cdot)$, as nonparametric functions, which are estimated using thin-plate regression splines (Wood, 2003). We also specify a first order serially correlated GAM specification where ε_t follows a first order autoregressive process [AR (1)], i.e., $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ with $v_t \sim iid(0, \sigma_v^2)$.

The time evaluation of the probabilities of switching between contiguous disease and non- contiguous disease states (periods) can be studies using a Hidden Markov Model (HMM), which is a statistical model that defines a probability distribution over possible sequences of observations in which each observation is a member of a discrete set of outcomes. It is often used to model time-varying processes. The model is based on the assumption that the underlying process that generates the data is a Markov process, and that the hidden states of the process are unobserved. In our case, the binary variable d_t, which indicates the presence or absence of a contiguous disease in year t, is a two-state process, with d_t taking values 0 or one. Let these finite state states be $\Lambda = \{1,2\}$. The HMM model expresses Markov evolution on the measurable space Λ in terms of a regular Markov chain using the latent variable $S_t \in \{1,2\}$, where $S_t = 1$ denoting the non-disease stated and $S_t = 2$ denoting the disease state. In general, S_t may have M states, $S_t \in \Lambda = \{1,2,\dots,M\}$, with the evolution of the state-space expressed with transition probability matrix $\boldsymbol{P} = [p_{ij}]$, $i,j = 1,2,\dots,M$, and stationary probability distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_M)'$. The transition probability of switching from state i in year $t-1$ to state j in year t is defined with the following properties:

$$\begin{aligned} &p_{ij} = P(S_t = j | S_{t-1} = i) \in (0,1), \quad \forall i,j \\ &\textstyle\sum_{j=1}^{M} p_{ij} = 1, \quad \forall i = 1,2,\dots,M \end{aligned} \tag{5}$$

In our case with $M = 2$, we have two free transition probabilities $p_{12} = (S_t = 2|S_{t-1} = 1)$ and $p_{21} = (S_t = 1|S_{t-1} = 2)$ with $p_{11} = 1 - p_{12}$ and $p_{22} = 1 - p_{21}$. The stationary probabilities $= (\pi_1, \pi_2, \dots, \pi_M)'$ are defined with the following properties:

$$\begin{aligned} &\pi_i = P(S_t = i) \in (0,1), \quad \forall i \\ &\textstyle\sum_{i=1}^{M} \pi_i = 1, \end{aligned} \tag{6}$$

which implies one free state probability π_2 since $\pi_1 = 1 - \pi_2$ with $M = 2$.

If the transition probabilities p_{ij} are independent of time, then the HMM is time invariant or homogenous. However, the homogenous HMM is quite restrictive for many real-world cases where the transition probabilities change over time, likely due to the effects of some underlying factors. We can relax this restrictive assumption by allowing the transition probabilities to be time-varying which leads to a non-homogenous HMM model (NHMM). The time-varying transition probabilities model is an extension of the standard HMM. In the standard Markov model, the transition probabilities between states are constant over time. In the time-varying transition probabilities model, the transition probabilities can vary over time. The time-varying transition probabilities model is a more accurate representation of reality than the standard HMM. It can be used to model

processes that change over time, such as the spread of disease, the growth of a population, or the price of a stock. An attractive approach making transition probabilities time-varying is to allow them to depend on some other covariate. The NHMM model with a time-varying transition probabilities and covariates $\boldsymbol{z}_t = (z_{1t}, z_{1t}, \dots, z_{Kt})'$ can be represented as $p_{ij}(\boldsymbol{z}_t) = P(S_t = j | S_{t-1} = i, \boldsymbol{z}_t)$. The transition probabilities between hidden states are allowed to vary over time and are governed by covariates $\boldsymbol{z}_t$. In this model, the probability of transitioning from one hidden state to another at any given time t depends on both the value of the covariate at that time and the values of the transition probabilities at previous times.

Given that the observed state variable d_t is binary, we use a NHMM with a logistic link function. The covariates are specified to include the temperature anomaly series h_t and a linear time trend in addition to a constant vector, i.e., $\boldsymbol{z}_t = (1, h_t, \tau_t)'$. The logistic HMM model specifies the transition probabilities $\{p_{ij}, i, j \in \Lambda\}$ and stationary distribution components $\{\pi_i, i \in \Lambda\}$ with the following logistic models:

$$p_{ij}(\boldsymbol{z}_t) = \frac{\exp\{\boldsymbol{\alpha}_{ij}'\boldsymbol{z}_t\}}{1 + \exp\{\boldsymbol{\alpha}_{ij}'\boldsymbol{z}_t\}} = \frac{\exp\{\alpha_{0,i} + \alpha_{1,ij}h_t + \alpha_{2,ij}\tau_t\}}{1 + \exp\{\alpha_{0,i} + \alpha_{1,ij}h_t + \alpha_{2,ij}\tau_t\}}, \qquad i, j \in \Lambda \tag{7}$$

$$\pi_i(\boldsymbol{z}_t) = \frac{\exp\{\boldsymbol{\gamma}_i'\boldsymbol{z}_t\}}{1 + \exp\{\boldsymbol{\gamma}_i'\boldsymbol{z}_t\}} = \frac{\exp\{\gamma_0 + \gamma_{1,i}h_t + \gamma_{2,i}\tau_t\}}{1 + \exp\{\gamma_0 + \gamma_{1,i}h_t + \gamma_{2,i}\tau_t\}}, \qquad i \in \Lambda \tag{8}$$

where $\boldsymbol{\alpha}_{ij} = (\alpha_{0,i}, \alpha_{1,ij}, \alpha_{2,ij})'$ and $\boldsymbol{\gamma}_i = (\gamma_0, \gamma_{1,i}, \gamma_{2,i})'$ are parameter to be estimated. In reality, only two sets of transition probabilities and one set of stationary state probability as estimated for a two-state model, since probabilities sum to one.

There are a number of ways to characterize the statistical properties of a logistic hidden Markov model for binary time series. One common approach is to consider the model's ability to correctly predict the next time step in the series, given the previous time steps. Another approach is to consider the model's ability to accurately estimate the underlying probabilities of the time series. Using the later approach, we estimate the parameters of the NHMM model using maximum likelihood (ML) estimation where the maximization is performed using the expectation maximization (EM) algorithm. Once the NHMM model is estimated, there are a number of methods for decoding the states and obtaining the relevant probabilities. For our purposes, smoothed probabilities are appropriate as they us the full sample information for inference in each time point locally.

IV. Results and discussions:

Figures 2-9 showcase various attributes of the data. In panel Figure 2 (a), the dummy variable indicating a contagious disease in a given year is plotted, while Figure 2 (b) shows both the temperature anomalies and years with contagious diseases (shaded bars) between 1 AD and 2021 AD. We use density plots to provide a visualization of the data. In Figure 3, we plot the conditional distributions of temperature anomaly using kernel density estimates and box plots. The distribution of the temperature anomaly is conditional on the contagious disease periods with high temperature anomaly levels (above 0.25) and low temperature anomaly levels (below 0.25) respectively. Using a Gaussian kernel, the probability density at each data point is estimated, which are then smoothed to generate continuous curves. In panel Figure 3 (a), the kernel density estimates are displayed for the conditional probability distribution function of the temperature anomaly series, while in Figure 3 (b), the boxplots with overlayed observations conditional on the contagious disease status are presented.

[Insert Figure 2]

[Insert Figure 3]

The autocorrelation (ACF) and the partial autocorrelation function (PACF) of the contagious disease variable and the temperature anomaly behavior are presented in Figure 4 along with the cross correlation and partial cross correlation functions. Together, they provide insights about the time series characteristics of the data. The ACF provides the correlation of between the current and the lagged values of a variable whereas as the PACF is used to measure the correlation between the current observation of the variable and an observation from a previous time period, after controlling for the observations at the intermediate lags. In Figure 4, the gradual declining ACFs and the PACFs together help to define the autoregressive process of the two variables. The cross correlation and partial cross correlation plots show the relationship between the two time series used in the model. To provide a visual representation of the assessments of the three models, i.e., the logistic generalized additive model (GAM – Logistic), a logistic model, and the benchmark linear model are presented in figure 5. We plot the predicted probabilities of the occurrence of a contagious disease against the estimated residuals ($d_t - \widehat{\pi_t}$), temperature anomalies, and time, in panels (a), (b), and (c) respectively.

[Insert Figure 4]

To provide a visual representation of the assessments of the three models, i.e., the logistic generalized additive model (GAM – Logistic), a logistic model, and the benchmark linear model are presented in Figure 5. We plot the predicted probabilities of the occurrence of a contagious disease against the estimated residuals ($d_t - \widehat{\pi_t}$), temperature anomalies, and time, in panels (a), (b), and (c) respectively of Figure 5. The receiver operating curves (ROC) in panel (d) plot the model sensitivity (true positive rate) against the false positive rate. The true positive rate represents the proportion of observations that are predicted to be positive when the observations are positive whereas the false positive rate indicates the proportion of observations that are predicted to be positive when they are, in fact, negative. The area under the curve indicates the quality of a model in predicting the observations. The GAM-Logistic model with the highest area under the curve indicates the best fit among three models.

[Insert Figure 5]

In Figure 6, we plot the Quantile-Quantile (QQ) plot the histograms of the residuals of the logistic generalized additive model (GAM-Logistic). The points in the QQ plot falls on a straight line indicating the residuals of the model approximately follow the normal distribution. The histogram of the residuals indicates the residuals are centered around zero.

[Insert Figure 6]

Figures 7, 8, and 9 provide further visualization of various features of the logistic generalized additive model. The smoothed transition and state probabilities estimated using the non-homogenous hidden Markov Model are plotted in figure 10.

[Insert Figures 7, 8, 9, and 10]

The estimation results from the alternative parametric and nonparametric model specifications are presented in Tables 3-6. The tables indicate when the null hypothesis of zero effect of temperature anomaly on disease spread can be rejected at the 1% (***), 5% (**), and 1% (*) levels. The R-square, log likelihood function, Akaike Information Criterion (AIC), and the Schwartz Bayesian Information Criterion (BIC) values provide measures for the quality of the respective models and help us to compare them. The models with the better fits have lower AIC and BIC values.

In Table 3, the results of the benchmark Linear Probability Model (LPM) are presented with the first column providing the estimates of the unrestricted model. Columns 2 through 7 represent the restricted versions of the model with estimates of the core model under different zero restrictions on the parameters of the contemporaneous and different lagged terms along with the trend variable. The first column represents the estimated coefficients of the unrestricted model. None of the coefficients that show the relationship between temperature anomalies and the dependent variable are statistically significant. Column 2 presents the coefficient of the contemporaneous effect (β_1) is restricted to zero. In column (5), the results of a restricted version of the model with both the coefficients of h_t and h_{t-2} set to zero are presented. A comparison of the alternative versions indicates the restricted models in columns 5 and 6 are closely comparable. However, column 6 with the coefficients of h_{t-1} and h_{t-2} set to zero gives the best results qualitatively in terms of the information requirement, as confirmed by the AIC and BIC scores. The coefficient for the contemporaneous effect is statistically significant at 1%.

[Insert Table 3]

While relatively straightforward to specify and estimate, linear probability models are often not a suitable choice because the predicted probability values can end up being below zero or greater than 1. To counter the standard limitations of the LPM, a logistic model was estimated. The results are presented in Table 4. Qualitatively, the results from the logistic model are in line with our findings from the benchmark model for both the unrestricted and restricted versions. The best results are for the model that includes a contemporaneous effect of temperature anomalies and a trend term.

[Insert Table 4]

The results of the nonparametric logistic general additive model (GAM) are presented in Table 5. A GAM is a powerful analytical tool because of its ability to fit many types of non-linear data. However, because of this flexibility, it can be easy to overfit the data. The goal of the model is to strike a balance between two objectives. First, the model must capture the relationship exhibited in the data as closely as possible. This is indicated by the "Likelihood" function, which indicates how well a model captures patterns in the data it is fitted to. Second, we want to avoid overfitting the data, which is captured by the "wiggliness" in the fit. In the model, the smoothing functions, *s(.)*, are represented by penalized regression splines to avoid complex overfitting of the model. A smooth or a spline is essentially a function that can take a wide variety of shapes. The smoothing functions are estimated with thin plate splines, which do not depend on the prior knowledge of the functional form of the data. Thin plate regression splines can be computationally more costly relative to other smoothing options such as cubic splines. However, they have the advantage of not requiring knots placements that are a feature of conventional regression spine modelling (Crawley, 2013).

We estimate equation (4) with various restrictions imposed on the smooth functions. In column (2), the results presented are conditioned on the smooth function for h_t set to zero. Similarly, the results in column (5) are derived based on the assumption that the two-period lagged effect and the contemporaneous effect of temperature anomaly of d_t are assumed to be zero. The estimates presented in column (6) indicate the contemporaneous effect of temperature anomaly on the incidence of a contagious disease in a given year is statistically significant at the 1% level. This restricted model also provides the lowest AIC and BIC values indicating a better fit than the alternative versions of the nonparametric model we estimated.

[Insert Table 5]

In order to account for the possibility that the error term in equation (4) might be correlated over time, we also run a version of the GAM specificizing a first order autoregressive process for the error structure. The results are presented in table 6, which also includes estimates of ρ, the autocorrelation parameter. The estimates from

both specifications of the nonparametric model are similar with column 6 (in both tables) indicating the best fit to the data compared to the alternative restricted versions of the core model. In fact, the results from the parametric and nonparametric models are qualitatively consistent. In all specifications, the restricted version of the model that includes the contemporaneous effect of temperature anomaly and the linear trend term provide the best fit compared to the complete unrestricted specification and the alter zero-restriction variations we imposed.

[Insert Table 6]

Table 7 includes the estimates of the parameters of the transition probability expression in equation (7) and the parameters in (8) used to derive a set of stationary state probability for the logistic HMM model. The transition probabilities are calculated at the zero values of the covariates. The sum of the estimated probabilities of a particular state (disease or non-disease) in any time period evolving into either the same state or the alternative state in the following period adds up to 1. Regardless of the initial state, the estimated transition probabilities imply that the probability of transitioning from a given state in a year to the same year in the following year is significantly higher than the probability of transitioning to the other state (0.999 versus 0.0010).

[Insert Table 7]

V. Conclusion

In this paper, using annual data from 1 AD to 2021 AD, we investigated the role of temperature anomalies in the spread of contagious diseases. While the pace and extent of transmission of any contagious disease depend on many contextual factors such as the availability of healthcare related services, governmental efficacy in management of the spread, nature of the diseases, local and regional socioeconomic and environmental conditions at the epicenter, etc., this research provides evidence that temperature anomalies have played an influential role in the spread of transmissible diseases over the last two thousand years, thereby identifying a common cause among different disease spreads over time. The key finding is robust to model specification issues as it is confirmed by the results from the parametric and nonparametric specifications of the core model. These results are of particular significance within the context of developing effective climate change adaptation strategies, particular those that involve public health related initiatives. Changes in weather and climate patterns involve considerable uncertainty. However, our results indicate that given that temperature anomalies have a significant bearing on the bearing of contagious diseases, region specific climate forecasting results can be combined with demographic information to develop location specific, cost-effective disease control policy responses and transmission-based precautionary measures. This is particularly important given that regions across the world vary greatly in available resources that can be dedicated to mitigates damages associated with the transmission of infectious diseases. Future avenues of research could potentially focus on this line of interdisciplinary work.

References:

Cirillo, P., & Taleb, N.N. (2020). Tail risk of contagious diseases. *Nature Physics,* 16, 606–613. https://doi.org/10.1038/s41567-020-0921-x

Cohen, J.M., Sauer, E.L., Santiago, O., Spencer, S., & Rohr, J.R. (2020). Divergent impacts of warming weather on wildlife disease risk across climates. *Science*, 370 (6519), DOI:10.1126/science.abb1702

Cohen, J.M., Venesky, M.D., Sauer, E.L., Civitello, D.J., Taegan, A.M., Roznik, E.A., & Rohr, J.R. (2017). The thermal mismatch hypothesis explains host susceptibility to an emerging infectious disease. *Ecology Letters,* 20 (2), 184 – 193.

Crawley, M.J. (2013). *The R book*. Chichester, West Sussex, United Kingdom: Wiley.

Dasgupta, I. & Gershman, S. (2021). Memory as a Computational Resource. *Trends in Cognitive Sciences*, 25(3), 240 – 251.

Gani, M.O., Sarwar, H. & Chowdhury, M.R. (2009). Prediction of State of Wireless Network Using Markov and Hidden Markov Model. *Journal of Networks*, 4(10), 976-984.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli*, 6(6), 1051 – 1079.

Gorris, M. E., Treseder, K. K., Zender, C. S., & Randerson, J. T. (2019). Expansion of coccidioidomycosis endemic regions in the United States in response to climate change. *GeoHealth*, 3(10), 308 – 327. https://doi.org/10.1029/2019GH000209

Greene, A.M., Robertson, A.W., Smyth, P. & Triglia, S. (2011). Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Quarterly Journal of Royal Meteorological Society,* 137(655), 347 – 359.

IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2391 pp. https://doi:10.1017/9781009157896.

McDermott, A. (2022). Climate change hastens disease spread across the globe. *Proceedings of the National Academy of Sciences*, https://doi.org/10.1073/pnas.220048111

Morens, D.M., Taubenberger, J.K., & Fauci, A.S. (2021). A Centenary Tale of Two Pandemics: The 1918 Influenza Pandemic and COVID-19, Part I. *American Journal Public Health*. 111(6):1086-1094. https://doi:10.2105/AJPH.2021.306310. PMID: 33950739; PMCID: PMC8101587.

Norris, J.R., Allen, R.J., Evan, A.T., Zelinka, M.D., O'Dell, C.W. & Klein, S.A. (2016). Evidence for climate change in the satellite cloud record, *Nature*, 536, 72 – 75. https://doi.org/10.1038/nature18273

Scott, S.L, James, G.M., & Sugar, C.A. (2005). Hidden Markov Models for Longitudinal Comparisons, *Journal of the American Statistical Association*, 100(470), 359-369.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 65(1), 95 – 114.

Mamon, R. S., & Elliott, R.J. (2014). *Hidden Markov Models in Finance.* Springer New York, NY. https://doi.org/10.1007/978-1-4899-7442-6

Yap, K. L. & Chong. Y.W. (2017). Optimized access point selection with mobility prediction using hidden Markov Model for wireless network, *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, Milan, Italy, 38-42, https://doi:10.1109/ICUFN.2017.7993744.

Zucchini, W. & Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resources Research*. 27, 1917 – 1923.

Tables and Figures:

Table 1: Contagious disease events and dates included in the sample.

Event	Start Year	End Year	Location	Estimated Deaths
Plague of Athens	-429	-426	Greece, Libya, Egypt, Ethiopia	75,000 – 100,000
Antonine Plague	165	180	Roman Empire	5 – 10 million
Plague of Cyprian	250	266	Europe	310,000
Plague of Justinian	541	542	Europe, West Asia	15 – 100 million
Plague of Amida	562	562	Mesopotamia (modern day Turkey)	30,000
Roman Plague of 590	590	590	Rome, Byzantine Empire	Unknown
Plague of Sheroe	627	628	Bilad al-Sham	25000+
Plague of the British Isles	664	689	British Isles	Unknown
Plague of Basra	688	689	Basra (southeast Turkey)	200,000
Japanese smallpox epidemic	735	737	Japan	2 million
Black Death	1331	1353	Eurasia and North Africa	75 – 200 million
Sweating sickness	1485	1551	Britain	10,000+
Smallpox Epidemic in Mexico	1520	1520	Mexico	5 – 8 million
Cocoliztli Epidemic of 1545-1548	1545	1548	Mexico	5 – 15 million
1563 London plague	1562	1564	London, England	20,100
Malta plague epidemic	1592	1593	Malta	3,000
Plague in Spain	1596	1602	Spain	600,000 – 700,000
New England epidemic	1616	1620	New England	Unknown
Italian plague of 1629-1631	1629	1631	Italy	1 million
Great Plague of Sevilla	1647	1652	Spain	500,000
Plague in Kingdom of Naples	1656	1658	Italy	1,250,000
Plague in the Netherlands	1663	1664	Amsterdam, Netherlands	24,148
Great Plague of London	1665	1666	England	100,000
Plague in France	1668	1668	France	40,000
Malta plague epidemic	1675	1676	Malta	11,300
Great Plague of Vienna	1679	1679	Vienna, Austria	76,000
Great Northern War plague outbreak	1700	1721	Denmark, Sweden, Lithuania	164,000
Great Smallpox Epidemic in Iceland	1707	1709	Iceland	18,000+
Great Plague of Marseille	1720	1722	France	100,000
Great Plague of 1738	1738	1738	Balkans	50,000
Russian plague of 1770 - 1772	1770	1772	Russia	50,000
Ottoman Plague Epidemic	1812	1819	Ottoman Empire	300,000+
Caragea's plague	1813	1813	Romania	60,000
Malta plague epidemic	1813	1814	Malta	4,500
First cholera pandemic	1816	1826	Asia, Europe	100,000+
Second cholera pandemic	1829	1851	Asia, Europe, North America	100,000+
Typhus epidemic in Canada	1847	1848	Canada	20,000+
Third cholera pandemic	1852	1860	Worldwide	1 million+

Cholera epidemic of Copenhagen	1853	1853	Copenhagen, Denmark	4,737
Third plague pandemic	1855	1960	Worldwide (India, China)	12 – 15 million
Smallpox in British Columbia	1862	1863	Pacific Northwest, Canada, US	20,000+
Fourth cholera pandemic	1863	1875	Middle East	600,000
Fiji Measles outbreak	1875	1875	Fiji	40,000
Yellow Fever	1880	1900	Mississippi, New Orleans, US	17,000+
Fifth cholera pandemic	1881	1896	Asia, Africa, Europe, South America	298,600
Smallpox in Montreal	1885	1885	Montreal, Canada	3,164
Russian flu	1889	1890	Russia, worldwide	1 million
Sixth cholera pandemic	1899	1923	Europe, Asia, Africa	800,000
China plague	1910	1912	China	40,000
Encephalitis lethargica pandemic	1915	1926	Worldwide	500,000
American polio epidemic	1916	1916	United States	7,130
Spanish flu	1918	1920	Worldwide	17-100 million
HIV/AIDS pandemic	1981	2023	Worldwide	42 million
Poliomyelitis in USA	1946	1946	United States	9,000
Asian flu	1957	1958	Worldwide	1-4 million
Hong Kong flu	1968	1969	Worldwide	1-4 million
London flu	1972	1973	United States	1,027
Smallpox epidemic of India	1974	1974	India	15,000
Zimbabwean cholera outbreak	2008	2009	Zimbabwe	4,293
Swine flu	2009	2009	Worldwide	151,700 – 575,400
Haiti cholera outbreak	2010	2020	Haiti	10,075
Measles in D.R. Congo	2010	2014	Democratic Republic of Congo (DRC)	4,500
Ebola in West Africa	2013	2016	Worldwide (Guinea, Liberia, Sierra Leone)	11,323+
Indian swine flu outbreak	2015	2015	India	2,035
Yemen cholera outbreak	2016	2020	Yemen	3,981
2018-2019 Kivu Ebola epidemic	2018	2020	DRC and Uganda	2,280
Measles in D.R. Congo	2019	2020	DRC	7,018
Dengue fever	2019	2020	Asia-Pacific, Latin America	3,930
COVID-19 Pandemic	2019	To date	Worldwide	7 – 29.3 million

Table 2. Descriptive statistics

	(1) Temperature Anomaly: Full Sample	(2) Temperature Anomaly: Non-Disease Periods	(3) Temperature Anomaly: Disease Periods	(4) Low Temperature Anomaly: Disease Periods	(5) High Temperature Anomaly: Disease Periods	(6) Contagious Disease
Observations	2021	1662	359	342	17	2021
Mean	-0.2565	-0.2439	-0.3148	-0.3630	0.6542	0.1776
S.D.	0.1626	0.1315	0.2544	0.1315	0.1789	0.3823
Min	-0.7128	-0.6688	-0.7128	-0.7128	0.4428	0.0000
Max	1.0071	0.5680	1.0071	0.0774	1.0071	1.0000
Skewness	1.8552	0.6627	2.8153	0.6760	0.4781	1.6856
Kurtosis	10.3064	4.2776	9.2488	0.4030	-1.2498	0.8417
JB	10128.7200***	1394.1920***	1776.7710***	28.8250***	1.5200	1018.6720***
$Q(1)$	1608.9177***	1218.9419***	291.7578***	208.1254***	9.8364***	1574.3151***
$Q(4)$	5750.1233***	4171.6226***	981.4622***	629.5901***	17.7115***	4927.1886***
ARCH(1)	1844.5608***	1297.6202***	335.4321***	136.2012***	4.5872**	1574.9812***
ARCH(4)	1876.8272***	1346.6727***	337.1974***	157.6724***	4.6465	1585.6803***

Note: The table reports descriptive statistics for the temperature anomaly (h_t) and contiguous disease variables (d_t), with annual data covering the period from 1 AD to April 2021 (2021observations). In addition to the full sample (column 1), the descriptive statistics for the temperature anomaly are reported for four additional sub-samples: periods of non-contiguous disease ($d_t = 0$; column 2), periods of contiguous disease ($d_t = 1$; column 3), periods of low temperature anomaly and contiguous disease ($d_t = 1$ and $h_t \leq 0.25$; column 4), and), periods of high temperature anomaly and contiguous disease ($d_t = 1$ and $h_t > 0.25$; column 5). The table reports mean, standard deviation (S.D.), minimum, maximum, skewness, and kurtosis, as well as the Jarque-Bera normality test (JB), first [Q(1)] and fifth [Q(5)] order Ljung-Box portmanteau test for serial correlation, and first [ARCH(1)] and fifth [ARCH(5)] order autoregressive conditional heteroskedasticity tests. **, and *** denote rejection at 10%, 5%, and 1% level, respectively.

Table 3. Linear probability model estimates

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-0.102*** (0.019)	-0.101*** (0.019)	-0.100*** (0.018)	-0.099*** (0.018)	-0.098*** (0.018)	-0.098*** (0.018)	0.077*** (0.016)
h_t	-0.080 (0.121)		-0.108 (0.115)			-0.185*** (0.050)	-0.393*** (0.052)
h_{t-1}	-0.034 (0.138)	-0.084 (0.115)	-0.087 (0.117)		-0.186*** (0.051)		
h_{t-2}	-0.089 (0.123)	-0.115 (0.117)		-0.192*** (0.052)			
τ_t	0.00022*** (0.00001)	0.00022*** (0.00001)	0.00023*** (0.00001)	0.00022*** (0.00001)	0.00023*** (0.00001)	0.00023*** (0.00001)	
R-squared	0.139	0.139	0.139	0.139	0.139	0.139	0.028
Log L	-772.311	-772.532	-772.573	-773.186	-773.021	-772.851	-895.195
AIC	1556.622	1555.064	1555.146	1554.372	1554.042	**1553.703**	1796.389
BIC	1590.284	1583.116	1583.198	1584.421	1576.483	**1576.144**	1813.220

Note: The table reports the estimates for the linear probability model in Eq. (1) with various zero restrictions on the parameters. The variable h_t denotes the temperature anomaly in year t, $t = 1,2,\ldots,2021$, and τ_t denotes a linear time trend for year t. The table also reports McFadden's pseudo-R squared (R-squared), logarithm of the likelihood (Log L), Akaike information criterion (AIC), and Schwarz's Bayesian information criterion (BIC). The standard errors of the estimates are given in brackets. Boldface denotes the minimum AIC and BIC values. *** denotes rejection of the null hypothesis of zero effect at the 1% level.

Table 4. Logistic model estimates

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-4.254*** (0.217)	-4.247*** (0.217)	-4.247*** (0.217)	-4.237*** (0.216)	-4.234*** (0.216)	-4.237*** (0.216)	-2.518*** (0.149)
h_t	-0.609 (0.902)		-0.767 (0.859)			-1.237*** (0.334)	-3.501*** (0.451)
h_{t-1}	-0.191 (1.045)	-0.592 (0.860)	-0.518 (0.875)		-1.240*** (0.339)		
h_{t-2}	-0.527 (0.919)	-0.717 (0.875)		-1.271*** (0.345)			
τ_t	0.002*** (0.0001)	0.002*** (0.0001)	0.002*** (0.0001)	0.002*** (0.00001)	0.002*** (0.0001)	0.002*** (0.000)	
R-squared	0.164	0.163	0.164	0.163	0.163	0.163	0.036
Log L	-790.317	-790.545	-790.481	-790.782	-790.881	-790.657	-911.140
AIC	1590.634	1589.090	1588.963	1587.564	1587.763	**1587.315**	1826.280
BIC	1618.686	1611.532	1611.404	1604.395	1604.594	**1604.146**	1837.501

Note: The table reports the estimates for the logistic probability model in Eq. (2) with various zero restrictions on the parameters. The variable h_t denotes the temperature anomaly in year t, $t = 1,2,\ldots,2021$, and τ_t denotes a linear time trend for year t. The table also reports McFadden's pseudo-R squared (R-squared), logarithm of the likelihood (Log L), Akaike information criterion (AIC), Schwarz's Bayesian information criterion (BIC). The standard errors of the estimates are given in brackets. *** denotes rejection of the null hypothesis of zero effect at the 1% level.

Table 5. Logistic generalized additive model estimates

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-4.739*** (0.750)	-4.767*** (0.755)	-4.753*** (0.754)	-4.829*** (0.764)	-4.784*** (0.760)	-4.791*** (0.760)	-1.822*** (0.075)
$s_1(h_t)$	12.337** (3.933)		13.788** (3.964)			73.866*** (5.246)	251.475*** (6.633)
$s_2(h_{t-1})$	9.124 (4.186)	29.620*** (5.257)	10.414* (4.145)		70.326*** (5.429)		
$s_3(h_{t-2})$	0.499 (1.243)	3.048 (2.053)		64.645*** (4.986)			
$s_4(\tau_t)$	207.781*** (12.176)	208.388*** (12.188)	208.685*** (12.182)	209.577*** (12.214)	210.482*** (12.196)	209.530*** (12.199)	
R-squared	0.393	0.390	0.393	0.384	0.389	0.388	0.167
Log L	-544.991	-549.450	-545.648	-556.882	-551.469	-550.085	-789.674
AIC	1141.834	1145.625	1138.606	1151.717	1141.915	**1138.542**	1595.723
BIC	1287.286	1276.693	1271.321	1258.182	1251.255	**1246.180**	1641.660
UBRE	595.411	598.140	595.064	603.269	598.528	596.767	803.486

Note: The table reports the estimates for the logistic probability model in Eq. (4) with various restricted variants. The variable h_t denotes the temperature anomaly in year t, $t = 1,2,\ldots,2021$, and τ_t denotes a linear time trend for year t. The smooth terms $s_i(\cdot)$are represented using penalized regression splines with smoothing parameters selected by unbiased risk estimator (UBRE) criterion. The table reports estimate of the intercept with its standard error in brackets. For the smooth terms $s_i(\cdot)$, the table reports the approximate significance χ^2 statistics with effective degrees of freedom in brackets. The table also reports McFadden's pseudo-R squared (R-squared), logarithm of the likelihood (Log L), Akaike information criterion (AIC), Schwarz's Bayesian information criterion (BIC), and unbiased risk estimator (UBRE) score. *** denotes rejection of the null hypothesis of zero effect at the 1% level.

Table 6. Logistic generalized additive model estimates with serial correlation

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-4.682*** (0.697)	-4.683*** (0.697)	-4.695*** (0.701)	-4.729*** (0.705)	-4.713*** (0.705)	-4.719*** (0.705)	-1.823*** (0.074)
$s_1(h_t)$	11.569** (3.546)		13.269*** (3.595)			75.069*** (4.931)	250.085*** (6.239)
$s_2(h_{t-1})$	9.393** (3.832)	67.878*** (5.178)	9.887** (3.713)		69.784*** (5.051)		
$s_3(h_{t-2})$	0.862 (1.000)	2.127 (1.000)		67.738*** (4.676)			
$s_4(\tau_t)$	192.115*** (12.023)	192.891*** (12.029)	193.256*** (12.030)	195.289*** (12.056)	195.721*** (12.043)	195.017*** (12.045)	
ρ	0.891	0.858	0.858	0.854	0.858	0.885	0.878
R-squared	0.393	0.389	0.393	0.383	0.388	0.388	0.167
Log L	-599.653	-602.125	-600.074	-607.753	-603.168	-601.435	-802.978
AIC	1217.306	1218.251	1214.148	1225.505	1216.337	**1212.870**	1611.955
BIC	1267.799	1257.523	1253.421	1253.557	1244.389	**1240.922**	1628.786

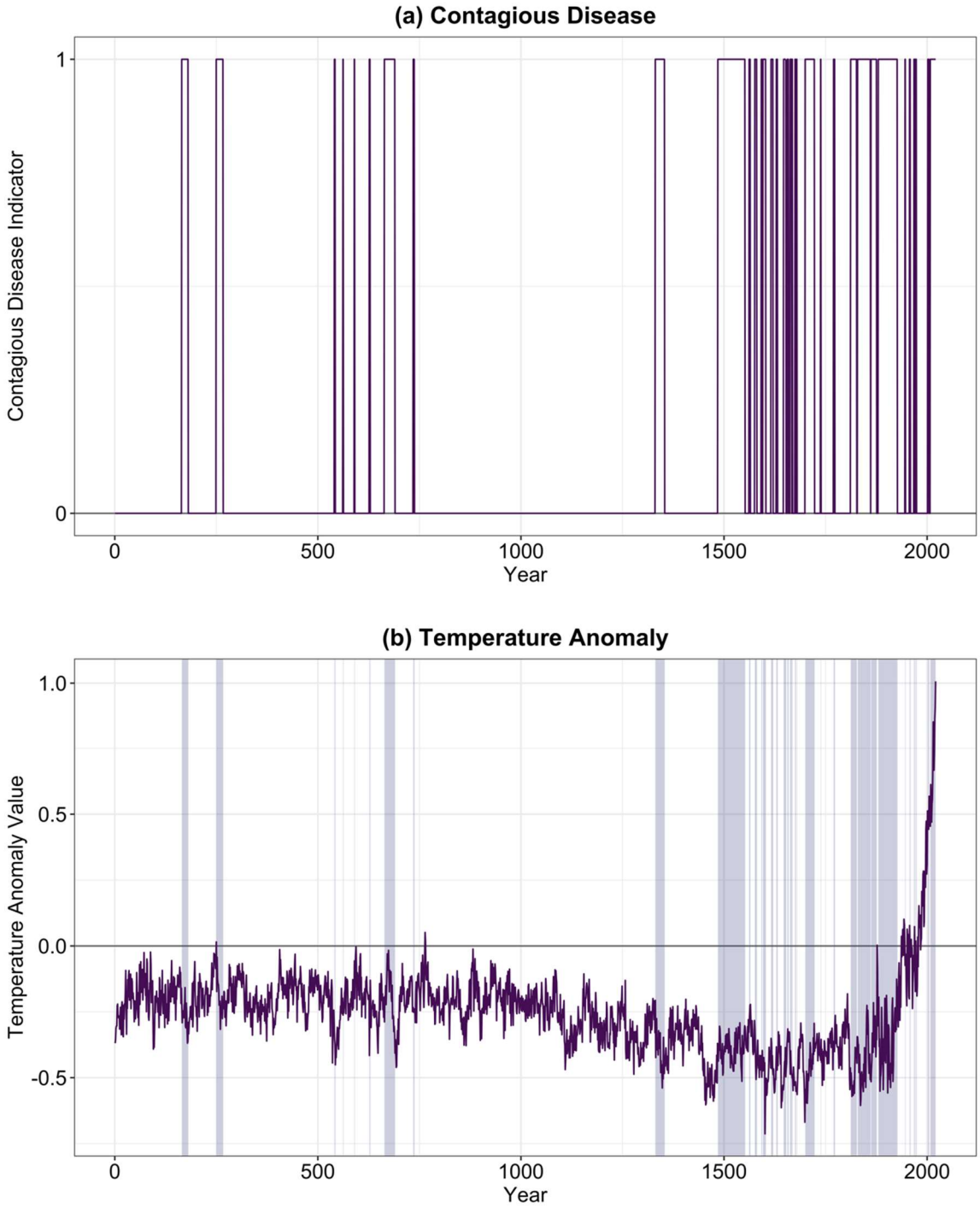
Note: The table reports the estimates for the logistic probability model with AR(1) error structure in Eq. (4) with various restricted variants. The variable h_t denotes the temperature anomaly in year t, $t = 1,2,\ldots,2021$, and τ_t denotes a linear time trend for year t. The smooth terms $s_i(\cdot)$are represented using penalized regression splines with smoothing parameters selected by generalized cross-validation (GCV). The table reports estimate of the intercept with its standard error in brackets. For the smooth terms $s_i(\cdot)$, the table reports the approximate significance χ^2 statistics with effective degrees of freedom in brackets. The table also reports McFadden's pseudo-R squared (R-squared), logarithm of the likelihood (Log L), Akaike information criterion (AIC), and Schwarz's Bayesian information criterion (BIC). Parameters are estimated using a generalization of the penalized quasi likelihood algorithm. ** and *** denote rejection of the null hypothesis of zero effect at the 5% and 1% levels, respectively.

Table 7. Estimates of the hidden Markov model

Parameter	Estimate	Probabilities at zero values of the covariates	
$\alpha_{0,2}$	-6.9516*** (0.7727)		
$\alpha_{1,12}$	-0.8445 (0.9618)	p_{11}	0.9990
$\alpha_{2,12}$	0.0023*** (0.0004)	p_{12}	0.0010
$\alpha_{0,2}$	2.3465*** (0.0016)	p_{21}	0.0010
$\alpha_{1,22}$	-0.0184 (0.1223)	p_{22}	0.9990
$\alpha_{2,22}$	0.00002*** (0.000002)		
γ_0	-683.0342 (23.2476)		
$\gamma_{1,2}$	17.7412 (18.0686)		
$\gamma_{2,2}$	0.4651*** (0.0155)		
Log L	-262.0879		
AIC	542.1758		
BIC	592.6779		

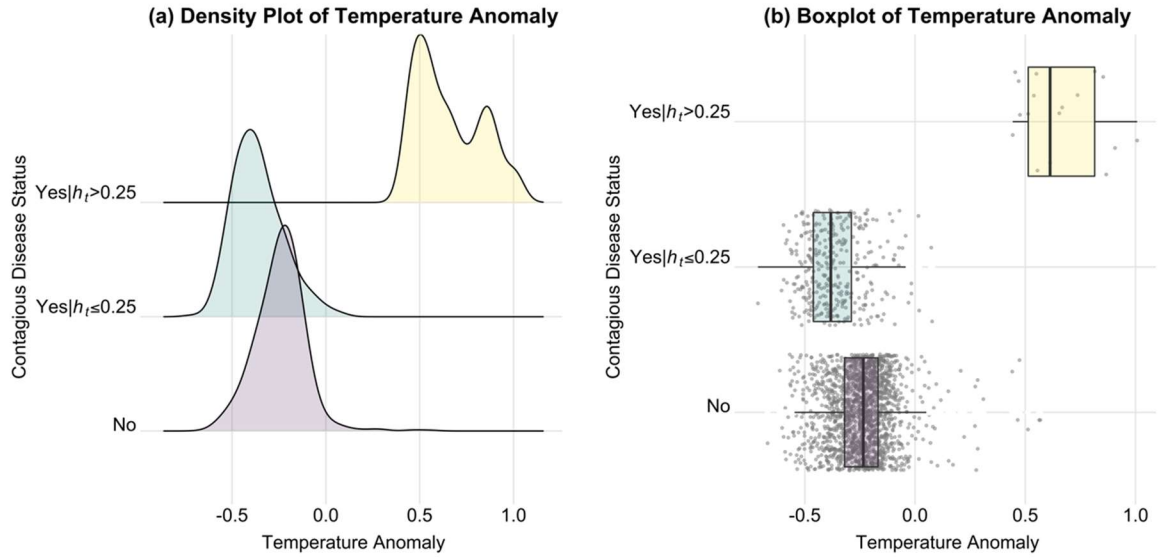
Note: The table reports the estimates for the non-homogenous hidden Markov model defied in Eqs. (5)-(8) The variable h_t denotes the temperature anomaly in year t, $t = 1,2,\ldots,2021$, and τ_t denotes a linear time trend for year t. The table also reports the logarithm of the likelihood (Log L), Akaike information criterion (AIC), and Schwarz's Bayesian information criterion (BIC). score. *** denotes rejection of the null hypothesis of zero effect at the 1% level.

Figure 2. Contagious disease and temperature anomaly series



Note: The figure plots the dummy indicator for the contiguous disease and the temperature anomaly over the years from 1 AD to 2021. Shaded regions in Figure (b) indicate presence of contiguous disease

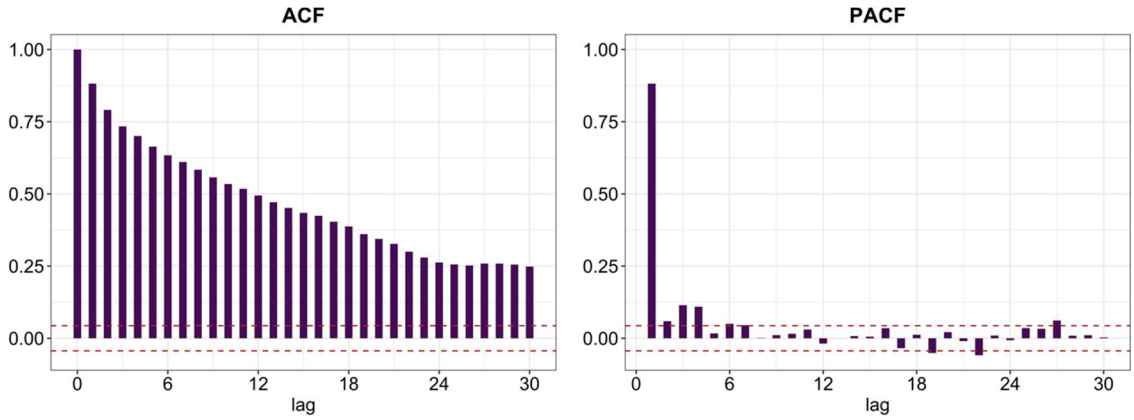
Figure 3. Conditional distribution and temperature anomaly series



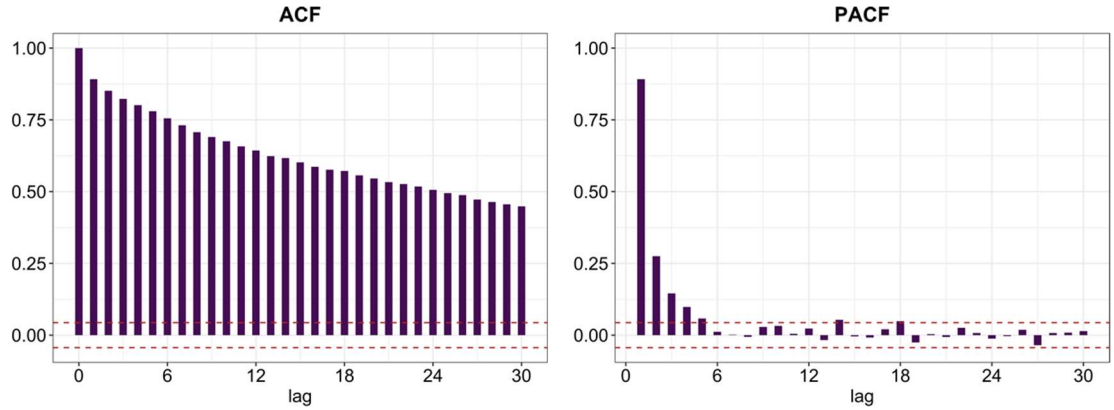
Note: The figure displays the density and boxplots of the temperature anomaly conditional on the status of the contagious disease with high and low temperature anomaly levels. High and low anomaly levels are defined values above 0.25 and below 0.25, respectively, which is the value naturally splits continues disease occurrences into these classes. Panel (a) displays kernel density estimates with a gaussian kernel. Panel (b) displays boxplots with overlayed observations conditional on the contagious disease status.

Figure 4. Autocorrelations and cross correlations of contagious disease and temperature anomaly

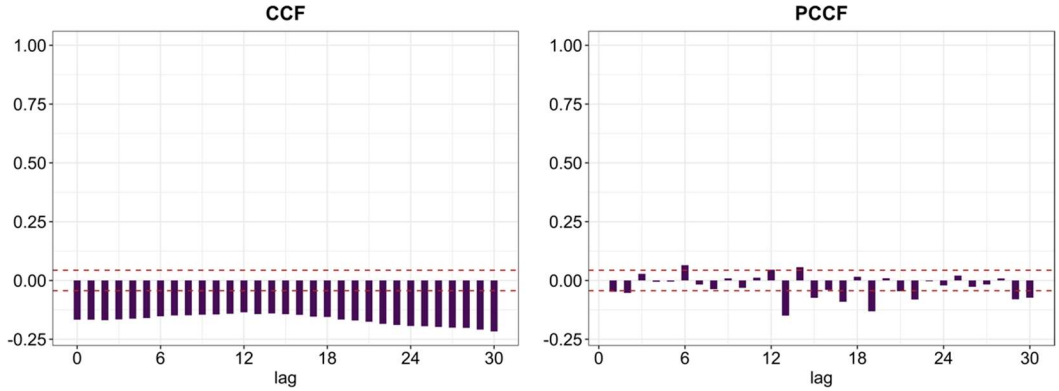
(a) Autocorrelation and partial autocorrelation of contagious disease



(b) Autocorrelation and partial autocorrelation of temperature anomaly

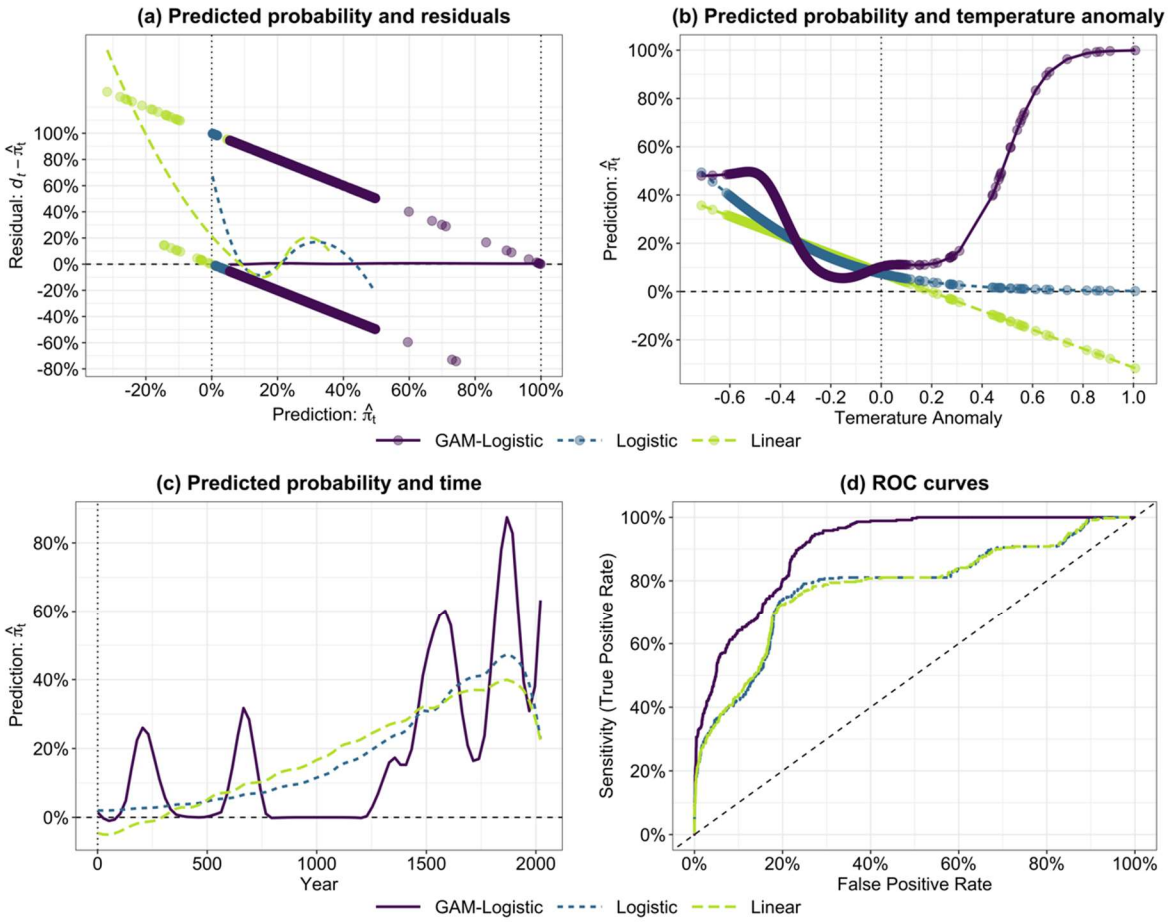


(c) Cross correlation and partial cross correlation of contagious disease and temperature anomaly



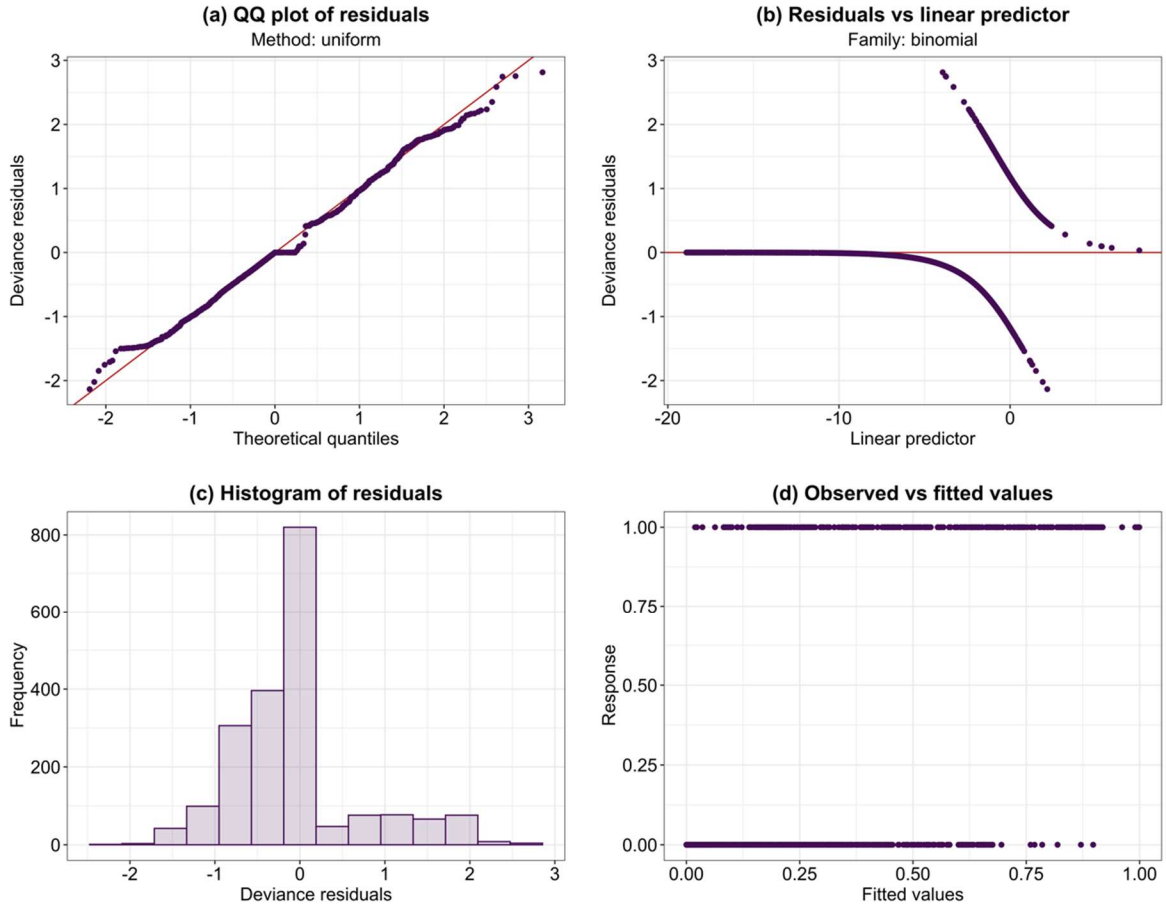
Note: The figure displays the autocorrelation function (ACF), partial autocorrelation function (PACF), cross correlation function (CCF), and partial cross correlation function (PCCF) of contiguous disease and temperature anomaly series. All four measures (ACF, PACF, CCF, and PCCF) when a binary contiguous disease series is involved are obtained using Cohen's κ statistic (see Weiss, 2018, p. 130), a measure of signed serial dependence for discrete-valued time series.

Figure 5. Linear, logistic, and logistic generalized additive model fit assessment



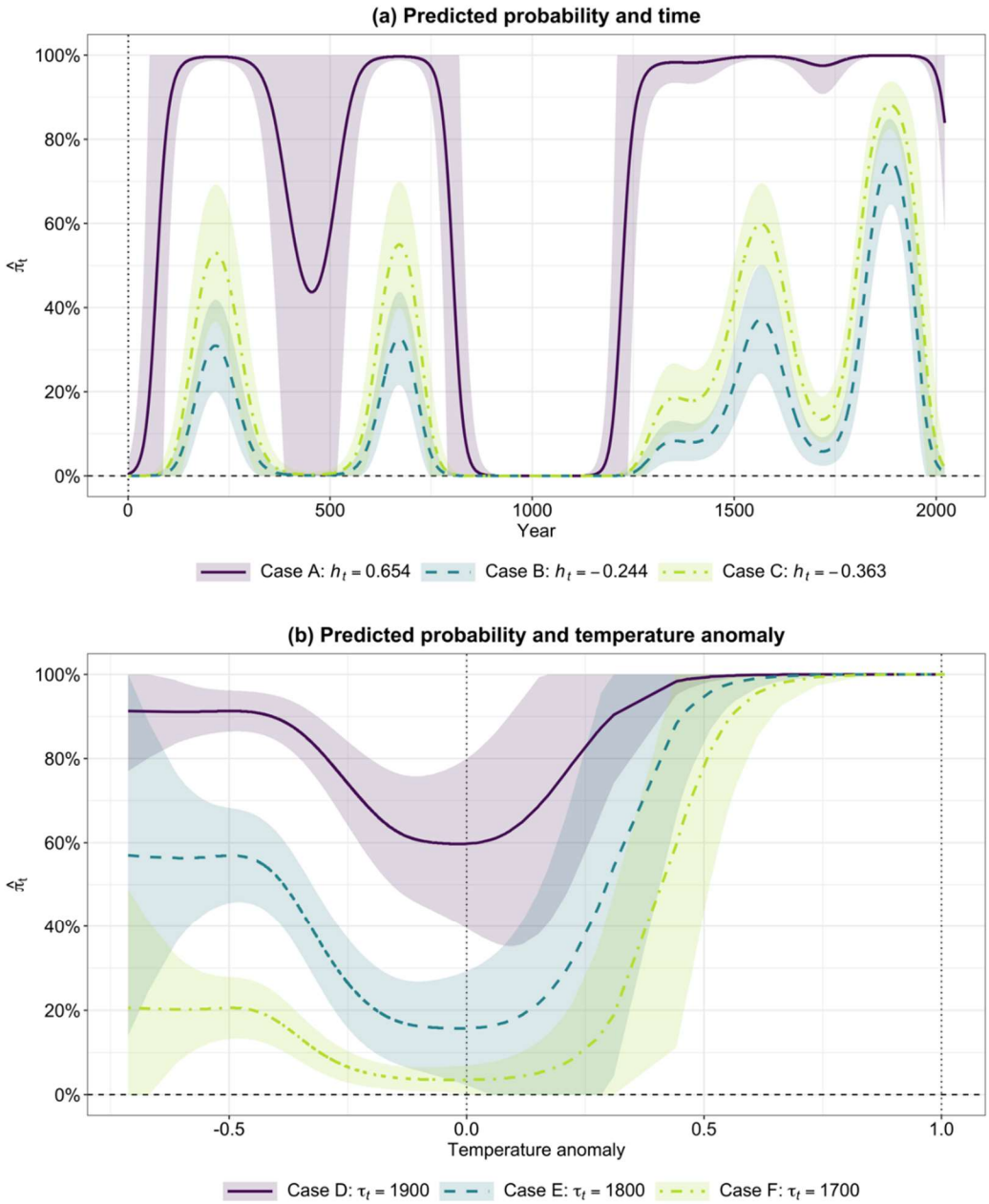
Note: The figure presents model assessment for the best, among all models considered, logistic generalized additive model (GAM-Logistic), a logistic model that has the best AIC among the logistic models, and a benchmark linear model. Panel (a) plots predictions ($\hat{\pi}_t$) against the residuals ($d_t - \hat{\pi}_t$) with a local polynomial regression (LOESS) fits using a second-degree polynomial. Panel (b) plots the predicted probability of the occurrence of a contiguous disease by temperature anomaly with the trend variable set equal to zero. Panel (c) plots the predicted probability of the occurrence of the occurrence of a contiguous disease by the time trend with the temperature anomaly set equal to zero. Panel (d) plots the receiver operating curves.

Figure 6. Diagnostics for the logistic generalized additive model



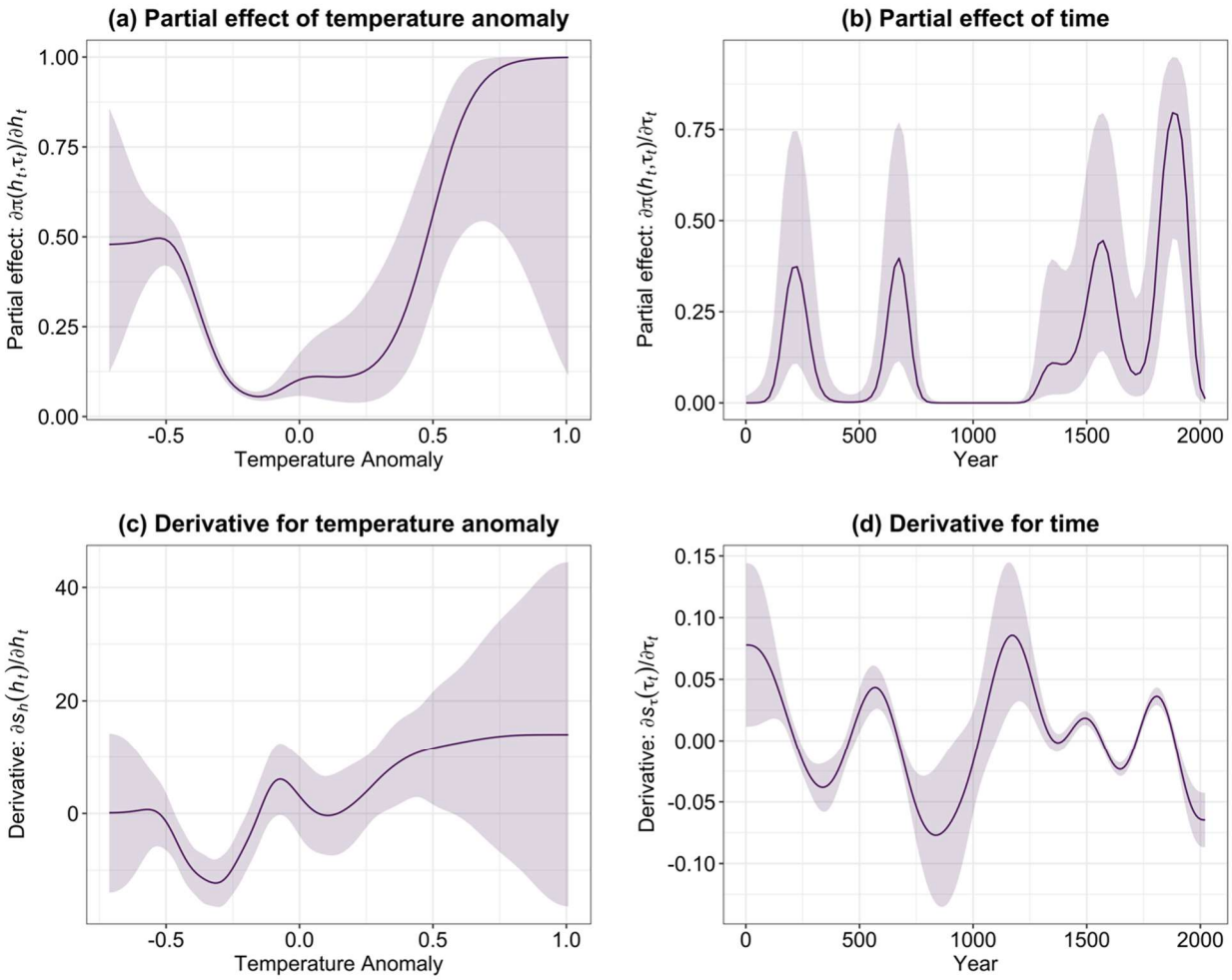
Note: The figure presents model diagnostics for the selected logistic GAM model. Quantile-quantile (QQ) plot of the model residuals are obtained by generating reference quantiles that associate each data point with a quantile of the uniform distribution. The residuals vs. linear predictor plot is based on fitted model prediction of a binomial link function of expected value for each data point.

Figure 7. Conditional predictions from the logistic generalized additive model



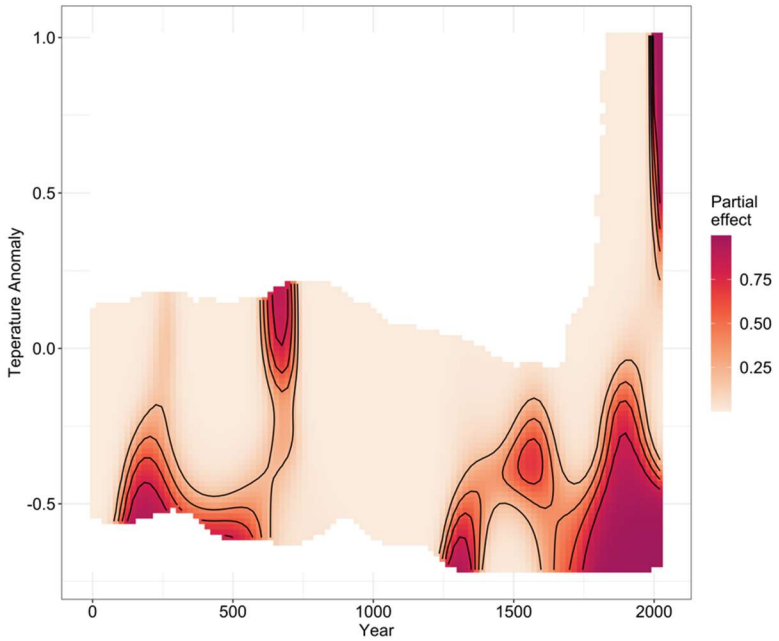
Note: The figure displays conditional predictions for the probability of contagious disease from the logistic generalized additive model. In Panel (a), predictions for the time trend τ_t conditional on three specific values of temperature anomaly are displayed: high temperature anomaly (Case A: $h_t = 0.654$) corresponding to the mean temperature anomaly in high temperature contagious disease periods ($h_t > 0.25$ and $d_t = 1$), medium temperature anomaly (Case B: $h_t = -0.244$) corresponding to the mean temperature anomaly in no contagious disease periods ($d_t = 0$), and low temperature anomaly (Case C: $h_t = -0.363$) corresponding to the mean temperature anomaly in low temperature contagious disease periods ($h_t \leq 0.25$ and $d_t = 1$). In Panel (b), predictions for the temperature anomaly (h_t) conditional on three specific values of time are displayed. The time periods that the predictions are conditioned on are: $\tau_t = 1900$ (Case D), $\tau_t = 1800$ (Case E), and $\tau_t = 1700$ (Case F).

Figure 8. Partial effects and partial derivatives in the logistic generalized additive model



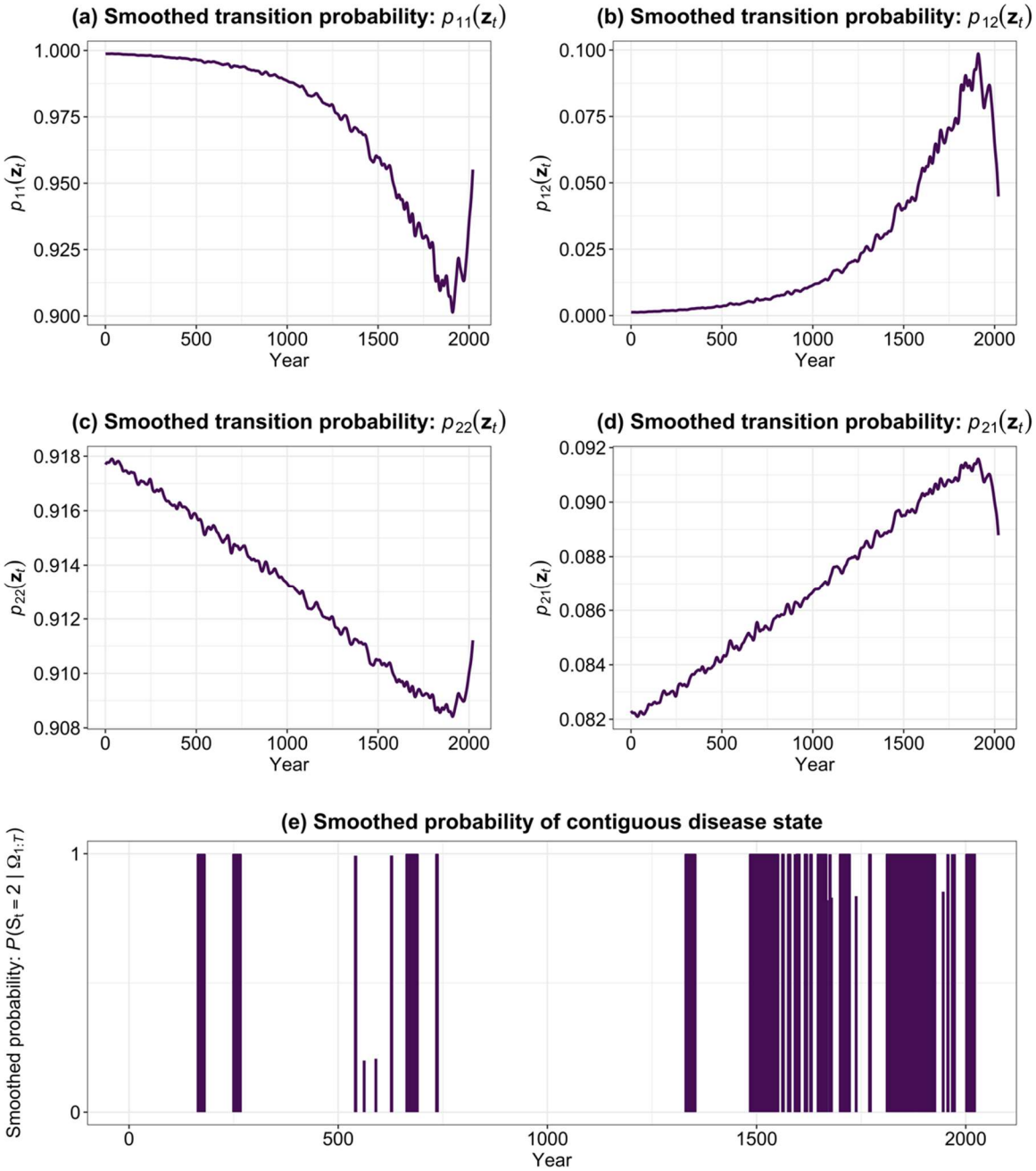
Note: The figure depicts the partial effects and partial derivatives of the temperature anomaly h_t and time trend τ_t in the logistic GAM model, which is specified as the $g(h_t, \tau_t) = c + s_h(h_t) + s_\tau(\tau_t)$, where the function $g(\cdot)$ is a logistic link function defined as $g(\cdot) = \log\{\pi(\cdot)/[1 - \pi(\cdot)]\}$, where $\pi(h_t, \tau_t) = P(d_t = 1|h_t, \tau_t) = \exp\{g(h_t, \tau_t)\}/(1 + \exp\{g(h_t, \tau_t)\})$.

Figure 9. The joint partial effects of temperature anomaly and trend in the logistic generalized additive model



Note: The figure presents the full joint effects of temperature anomaly and trend variables with over-imposed contour lines. The partial effect estimates are obtained from a tensor product smoother with a logistic link function defines as $g(h_t, \tau_t) = c + s(h_t, \tau_t)$, where the function $g(\cdot)$ is a logistic link function defined as $g(\cdot) = \log\{\pi(\cdot)/[1 - \pi(\cdot)]\}$ with $\pi(h_t, \tau_t) = P(d_t = 1|h_t, \tau_t)$. The tensor product smooth $s(\cdot)$ is constructed using row Kronecker products.

Figure 10. Smoothed transition and state probability estimated from a non-homogenous hidden Markov model.



Note: The figure depicts the time-varying transition and state probabilities from a two-state, $S_t \in \{1,2\}$ with $S_t = 1$ denoting the non-contagious disease state and $S_t = 2$ contagious disease state, non-homogenous hidden Markov model. The transition probability estimates are given in Panels (a)-(d) are specified as $p_{ij}(\boldsymbol{z}_t) = P(S_t = j | S_t = i, \boldsymbol{z}_t) = \exp\{\boldsymbol{\alpha}'_{ij}\boldsymbol{z}_t\}/(1 + \exp\{\boldsymbol{\alpha}'_{ij}\boldsymbol{z}_t\})$, $i, j \in \{1,2\}$, where $\boldsymbol{z}_t = (1, h_t, \tau_t)'$ and $\boldsymbol{\alpha}_{ij} = (\alpha_{0i,}, \alpha_{1,ij}, \alpha_{2,ij})'$. The estimates are obtained using maximum likelihood based on the expectation maximization (EM) algorithm.