One Health 17 (2023) 100636

Contents lists available at ScienceDirect

One Health

journal homepage: www.elsevier.com/locate/onehlt

Joint extremes in precipitation and infectious disease in the USA: A bivariate POT study☆

Zhiyan Cai [a,b], Yuqing Zhang [c], Tenglong Li [d], Ying Chen [d], Chengxiu Ling [d,*]

[a] Department of Bioinformatics, Xi'an Jiaotong-Liverpool University, SIP 215123, China
[b] Institute of Health Informatics, University College London, WC1H 9BT, UK
[c] Zhejiang Lab, 311121, China
[d] Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, SIP 215123, China

ARTICLE INFO

Keywords:
Extremal dependence
Multivariate peaks-over-threshold
ARIMA
Extreme precipitation
Infectious diseases

ABSTRACT

Mounting heavy precipitation events (HPEs) caused by the climate change have drawn wide attention. Increased incidences of infectious diseases are known as the common following health impact, while little has been studied about the extremal relationship in between. Therefore, this study aims to investigate the joint extremes of precipitation and infectious disease mortality rate in the USA, using publicly accessible data from the National Centers for Environmental Information and the Centers for Disease Control and Prevention. The study reveals the positive association between heavy precipitations and infectious diseases with slight national and regional differences using multivariate Peaks-Over-Threshold modelling. The strength of extremal dependence is measured by the extreme parameter α from a logistic dependence model in multivariate extreme value theory. The Midwestern USA shows an excessive impact of HPEs on infectious disease mortality ($\alpha = 0.7524$), while the other regions show similar extremal dependence strength with the national one (α values all approximate 0.77). The study also discovered spatial disparities in the extremal dependences for five sub-categories of infectious diseases in each census region, among which mycoses show the strongest extremal dependence with precipitation in almost all regions. These spatial differences of extremal dependence may be attributed to geographic, social-economic factors and the self-inherited characteristics of certain diseases. The findings are expected to assist in developing strategies counteracting extreme risks resulting from weather events and health issues as well. The cutting-edge multivariate Peaks-Over-Threshold (POT) approach employed herein also shows promise for a wide range of extreme risk assessment topics.

1. Introduction

Climate change has been causing mounting threats from extreme weather events and environmental changes [1]. Heavy precipitation events (HPEs) are expected to become more frequent and intense due to global warming, leading to a rise in the water-carrying capacity of the atmosphere and convective storm events. This trend has resulted in increased occurrences of flooding and rainfall events, particularly in the Eastern USA [1,9,10]. These HPEs are more likely to increase with the climbing intensity of tropical cyclones [5].

According to Aune, Davis, and Smith [1], heavy precipitation events (HPEs) are often the primary cause of river and flash flooding, while coastal flooding events can also arise partly from tropical cyclone-induced precipitation or storms. These HPEs and resulting flooding events frequently lead to significant financial losses, displacement, psychological impacts, and increased mortality rates [1,22]. For instance, unpredictable heavy rainfall and the subsequent flooding event in Zhengzhou, China, in July 2021 resulted in economic losses of approximately RMB 53.2 billion [14,16,21].

Incidences of waterborne and vector-borne diseases are known to increase following HPEs in both developing and developed countries [1]. This is likely due to contamination of water resources, displacement and crowding resulting from flooding, and abundant breeding sites for vectors [5,7,11]. Municipal water systems, particularly the outdated

☆ This project is supported by the research Development Fund at XJTLU (RDF1912017), the Post-graduate Research Scholarship (PGRS2112022) and Jiangsu Qinglan Talent in 2022.

* Corresponding author.
E-mail address: Chengxiu.Ling@xjtlu.edu.cn (C. Ling).

https://doi.org/10.1016/j.onehlt.2023.100636
Received 21 April 2023; Received in revised form 27 September 2023; Accepted 29 September 2023
Available online 4 October 2023

ones, are overloaded during HPEs, increasing the risk of contamination of the water supply for drinking and recreation through combined sewer outflows [1,5,11]. The displacement of affected populations often leads to crowded conditions, overburdening wastewater handling infrastructure. Bacterial and viral contaminants can spread to human communities and impact humans and animals in the aspects of facilitating the transmission of waterborne diseases such as diarrhea and cholera through dispersed surface water after flooding [1]. Vector-borne infectious diseases also tend to surge following HPEs, with mosquito-borne illnesses being one of the most common examples, as optimal breeding sites are left for their growth [5,7].

Previous studies have investigated the existence of significant associations between heavy precipitation and infectious disease morbidity in specific regions. Smith et al. [20] conducted a case-crossover study using logistic regression analysis to identify a positive association between extreme precipitation events ($\geq$ 99th percentile) and influenza emergency room visits in Massachusetts, USA with the odds ratio being 1.23 (95% CI: 1.16, 1.30). Phung et al. [18] found significant increases in hospitalization due to infectious intestinal diseases following HPEs ($\geq$ 95th percentile) in Vietnam. Chen et al. [5] conducted Poisson regression analysis to categorize daily accumulated precipitation into four levels and found that the extreme torrential level ($>$ 350 mm/d) posed the highest risks for waterborne infections in Taiwan. Singh et al. [19] focused on Pacific Islands and found a positive relationship between extreme rainfalls and diarrhea infections in Fiji. Curriero et al. [9] reported a positive association in the USA between waterborne disease outbreaks and extreme precipitations ($\geq$ 90th percentile) using Chi-squared and Fisher's Exact Tests.

Although previous research has generally shown a positive association between HPEs and infectious diseases, definitions for extreme precipitation vary depending on different thresholds or quantiles, which may be subjective. Furthermore, studies have typically focused on outbreaks and morbidity with few investigating the mortality of overall infectious diseases in recent years. Besides, basic mean association between advanced statistical modelling on the topic of extreme weather events and infectious diseases is rare. Extreme conditions on mortality are also seldom discussed at either country or regional levels [17]. Therefore, this paper aims to fill this research gap by focusing on the extreme associations between precipitation and the corresponding mortality of infectious diseases at both national and regional levels.

The study will explore category-specific extremal relationships in each region to identify differences among sub-categories. The USA is the study object due to its high-quality database on both HPEs and infectious diseases. Considering the upward trends of heavy precipitation due to climate change in the USA in Figs. 2 and A.1, this study is in high demand for exploring potential linked causes and for providing suggestions for the environmental and public agency in public strategy at a regional level.

The objective of our research is to study the joint extremes of precipitation and the corresponding mortality of infectious diseases. To better evaluate tail risks that are typically difficult to model due to the scarcity of extreme values, we will use extreme value theory (EVT), an ideal tool for such analyses [3,23]. In this paper, we will apply bivariate extreme value theory, specifically the bivariate Peaks-Over-Threshold (POT) method, to analyze the extremal dependence between monthly precipitation and mortality of infectious diseases at national and regional levels. Firstly, we will adjust seasonality or trends from these non-stationary time series data using autoregressive integrated moving average (ARIMA) models since the POT method is suitable for independent and identically distributed variables [15]. The resulting residuals will be fitted to the bivariate generalized Pareto (GP) distribution, and their extremal dependences will be quantified to understand short-term dynamics. The potential relationship explored could inform improvements in risk mitigation measures in the healthcare industry to handle increasing disease infections resulting from extreme weather events in the future. The results may also be useful for related fields, such as the insurance industry in designing catastrophic bonds, as well as policymakers, and possibly engineers and architects to alter structures that help mitigate risks [14,16].

2. Materials and methods

2.1. Database

This research utilized monthly regional precipitation and mortality rate data on infectious diseases in the USA from 1999 to 2019. The joint extreme analysis was conducted for the whole country as well as separately for four census regions in the USA, namely the Northeast, Midwest, South, and West (excluding Hawaii) according to the divisions defined by the USA Census Bureau. The monthly precipitation data from 1999 to 2019 is derived from the Climate At a Glance report released by National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information. National precipitation data was directly obtained. With regard to regional data, statewide data is first downloaded and then assigned to corresponding census regions. The monthly precipitations for each region were calculated as their arithmetic mean.

The mortality rate of infectious diseases is another research object herein, of which the whole category and sub-categories are considered separately. According to the International Classification of Diseases, Tenth Revision (ICD-10), the whole category refers to certain infectious diseases and parasitic diseases (ICD-10: A00-B99), which contains 22 sub-categories. The death numbers of both whole and sub-categories are collected correspondingly from Underlying Cause of Death on Wide-ranging Online Data for Epidemiologic Research (WONDER) database available on the Centers for Disease Control and Prevention (CDC) official website (*https://wonder.cdc.gov/*). Because monthly population data is not available, assuming the overall population does not change greatly annually, annual population is adopted to calculate the mortality rate. The population data by country and by census region are requested from Bridged-Race Population Estimates on the same WONDER CDC platform. Considering the reliability of results, only the sub-categories with $<$20% missing values are passed to subsequent analysis, which is listed as Intestinal infectious diseases (ICD-10: A00-A09), Other bacterial diseases (ICD-10: A30-A49), Viral hepatitis (ICD-10: B15-B19), Mycoses (ICD-10: B35-B49), and Sequelae of infectious and parasitic diseases (ICD-10: B90-B94).

2.2. ARIMA and GARCH models

The auto-regressive integrated moving average (ARIMA) models are widely used for time series analysis, which captures the autocorrelations in the data, and therefore is able to explain the underlying trend and seasonality [12]. If an ARIMA model could precisely capture these two components in a time series data, the residuals are supposed to be white noises (namely independent and identically distributed). Since the monthly precipitation and mortality rate data within a relatively wide year range are adopted herein, a yearly seasonality pattern, as well as upward or downward trends, could be possibly observed, resulting in the non-stationary time series data and thus against the assumption of POT methods. To solve this problem, the ARIMA model will be applied. It is an ideal choice here to generate white noises for subsequent POT analysis. If the residuals are not white noises (i.e., existing conditional variances), The premise of applying the bivariate POT method is that the samples should be independent and identically distributed. Thus, the ARIMA model is adopted herein to remove the trend, seasonality as well as non-stationarity from two time series. If heteroscedasticity remains in the residuals, the generalized autoregressive conditional heteroscedasticity (GARCH) model is then utilized to cope with such a situation. GARCH model could be applied to deal with such irregularity.

Hyndman and Athanasopoulos [12] detailly described an ARIMA(p,d,q) model with three components: Autoregressive (AR), Integrated (I),

and Moving average (MA) with order p, d and q, accordingly. It can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where c is a constant value; y'_t is the differenced data; p, d, and q are the order of the AR model, the order of differencing, the order of the MA model, respectively. The non-seasonal ARIMA model focuses on modelling non-seasonal data. With regard to seasonal data, the ARIMA model with an addition of seasonal terms is applied. A seasonal ARIMA model is presented as

$$\text{ARIMA}(p,d,q) \times (P,D,Q)_S, \tag{1}$$

in which p, d, and q in the nonseasonal part have identical meanings to that in the non-seasonal ARIMA model; P, D, and Q denote the counterparts in the seasonal part of the model with the time span of the seasonality S. The white noises of the residuals of (seasonal) ARIMA models might be checked according to the Ljung-Box test. If further the Autoregressive Conditional Heteroskedasticity (ARCH) effect is confirmed via Portmanteau Q and Lagrange Multiplier (LM) tests, then the addition of the GARCH component is required. A GARCH(p,q) model specifies the conditional variance of the error term ϵ_t as $\epsilon_t = \sigma_t z_t$, with standard variation σ_t following an ARMA(p,q). Finally, the white noise residuals of z_t's will pass to the extreme analysis.

The procedures to fit the ARIMA model to two time series variables mostly follow the Box-Jenkins method [4]. Extra steps of examining whether the data is white noise or not are performed by applying the Ljung-Box test before stationary test and after differencing. The automated function auto.arima() in the forecast package in R is also used for the model fitting by proving more potential choices [24]. The model with the lowest Akaike information criterion (AIC) values would be chosen for obtaining the residuals. If the returned residuals are not white noises which means failing the Ljung-Box test with p- value <0.05, ARCH effect would be tested by Arch.test() function in aTSA package and a GARCH component would be added to the final model [25]. The detailed procedures of model fitting are shown in Fig. A.*1*.

2.3. Bivariate POT approach

After fitting ARIMA/ARIMA-GARCH models, residuals (i.e., white noises) of mortality and precipitation are extracted to apply with the bivariate POT method. Beirlant et al. [2] introduced systematically the multivariate extreme value theory (MEVT). MEVT is widely applied in handling many problems when we are interested in how extremes in one variable relate to those in another. The relationship between those extreme values of variables can be described by the approach of "tail dependence", also known as "extremal dependence" or "asymptotic dependence". Dependence may occur if the processes are studied at neighboring spatial locations during their temporal evolution or share common meteorological conditions. It is very convenient if we handle first the margins and then transform them to the common scale, e.g., unit Fréchet. Then the dependence structure will be studied independently. Below, we introduce first the univariate extreme value theory and then the bivariate extreme value theory with a focus on the dependence measure.

Univariate extreme value theory. EVT aims to study rare events with extremely large influences. Its wide applications range from finance [26], environmental study [8] and pharmacometrics [27]. There are two common approaches to addressing extreme data and its distribution behavior. One is the block maxima (BM) method, which characterizes the maximum value in each block using the Generalized Extreme Value (GEV) distribution. Another one is the Peaks-Over-Threshold (POT) method, which analyzes the excesses over an appropriate threshold by generalized Pareto (GP) distribution. As the POT model can make full use of the extreme value data in comparison with the BM model, in this paper, we mainly employ the POT-based methods for analysing the extremal dependence of temperature and mortality of relevant diseases.

Suppose that $X_1, X_2, \ldots, X_n, \ldots$ is a random sample from parent $X \sim F(x)$, i.e., X_i's are independently and identically distributed with common distribution function (df) $F(x)$. Given a high threshold u, GP distribution is adopted to fit the threshold excess $Y^{[u]} = X - u | X > u$, namely [8]

$$\mathbb{P}\{X - u > y | X > u\} \approx \overline{GP}(y;\sigma,\xi) = \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi}, \quad y > 0.$$

Hence, the tail of the potential distribution F for $x > u$ is approximated as below (recall $\overline{F} = 1 - F$).

$$\overline{F}(x) = \zeta \mathbb{P}\{X > x | X > u\} = \zeta \left(1 + \xi \frac{x-u}{\sigma}\right)^{-1/\xi}, \quad x > u, \tag{2}$$

in which $\zeta = \mathbb{P}\{X > u\}$. Here $\xi \in \mathbb{R}$ and $\sigma > 0$ are the shape and scale parameters of GP distribution $G_{\xi,\sigma}(y) = 1 - [1 + \xi y/\sigma]^{-1/\xi}$, $y > 0$. In practice, the exceedance probability $\overline{F}(x)$ gives insight into the potential risk. Its estimate can be obtained through the extrapolation approach via Eq.(2): to get the approximated tail probability of the GP model using the maximum likelihood estimation of ξ, σ based on excesses $\left(x_{(i)} - u\right)'s$ with $x_{(1)} \geq \cdots \geq x_{(n_u)}$ exceeding the threshold u and the estimate of $\zeta = \overline{F}(u)$ as n_u/n.

With regard to the POT method, threshold selection is of the key role which represents a trade-off between variance and bias. Both the mean residual life plot and parameter stability plot could assist in the selection of a proper threshold, namely, we first get a range of thresholds where the empirical mean excess function is likely to vary linearly in threshold u

$$e_n(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{(i)} - u\right)$$

and its derived stable estimates of both scale and shape parameters.

Bivariate threshold excess model. As mentioned before, we will focus on the dependence for a bivariate vector $(X, Y) \sim F(x, y)$ with margins of F_X and F_Y. These marginal distributions can be transformed to be a common unit Fréchet distribution as follows. For proper threshold u_x and u_y, the margin has an approximation of Eq.(2) with parameter sets of $(\zeta_x, \sigma_x, \xi_x)$ and $\left(\zeta_y, \sigma_y, \xi_y\right)$. Denote by

$$\widetilde{X} = -\left(log\left\{1 - \zeta_x \left[1 + \frac{\xi_x (X - u_x)}{\sigma_x}\right]^{-1/\xi_x}\right\}\right)^{-1}$$

and

$$\widetilde{Y} = -\left(log\left\{1 - \zeta_y \left[1 + \frac{\xi_y (Y - u_y)}{\sigma_y}\right]^{-1/\xi_y}\right\}\right)^{-1}.$$

Consequently, the transformed vector $\left(\widetilde{X}, \widetilde{Y}\right) \sim \widetilde{F}(\widetilde{x}, \widetilde{y})$ has the marginal distribution function which is approximately standard Fréchet for $x > u_x$ and $y > u_y$.

Suppose that F is in the max-domain of attraction of a bivariate extreme value distribution, i.e., the limiting normalized componentwise maxima of observation from F follows a bivariate extreme value distribution. This is equivalent to

$$F(x,y) = \left\{\widetilde{F}^n(\widetilde{x}, \widetilde{y})\right\}^{1/n} \approx G(\widetilde{x}, \widetilde{y}) = exp\{-\ell(1/\widetilde{x}, 1/\widetilde{y})\}, \quad x > u_x, y > u_y,$$

where $\widetilde{x}$ and $\widetilde{y}$ are defined in terms of x and y by the same transformation between X and $\widetilde{X}$. Here ℓ is the so-called stable tail dependence function ([2], chap. 8). In contrast with univariate extreme value theory, there are infinite parametric forms for ℓ such that G is a multivariate extreme

value distribution. Another equivalent form of measuring the dependence is to utilize spectral measure (d.f.) H on $[0,1]$ such that

$$\ell(v_1, v_2) = 2\int_0^1 max(\omega v_1, (1-\omega)v_2)dH(\omega),$$

where H satisfies the mean restraint of $1/2$, namely, $\int_0^1 \omega dH(\omega) = 1/2$. The sample form (i.e., the spectral measure plot) based on this restriction can assist in the selection of threshold [3]. There are several classical choices for the parametric family of G such as logistic, negative logistic, and bilogistic. Among these models, the logistic model is the most popular one given its simplicity. It is commonly used as the dependence model in the survival analysis literature and other applications [28,29]

$$\ell(v_1, v_2) = \left(v_1^{1/\alpha} + v_2^{1/\alpha}\right)^{\alpha}, \quad \alpha \in (0,1], \tag{3}$$

where the smaller value of α indicates a stronger dependence between the two margins. The limiting case of $\alpha \to 0$ corresponds to the variables being totally dependent with

$$\ell(v_1, v_2) = max(v_1, v_2);$$

when $\alpha = 1$, the variables are independent: $\ell(v_1, v_2) = v_1 + v_2$. According to [13], the $1-\alpha$ value can be interpreted as the probability that the maximum values occur simultaneously in a sequence of observations for the same size larger enough, which also equals Kendall's rank correlation coefficient τ. Many alternative approaches to describe the dependence structure of G have been developed in the literature. Quite popular is *Pickands' dependence function*

$$A(\omega) = \ell(1-\omega, \omega) : [0,1] \mapsto [0,1],$$

which equals $\left(\omega^{1/\alpha} + (1-\omega)^{1/\alpha}\right)^{\alpha}$ corresponding to bivariate logistic model. We have

$$max(\omega, 1-\omega) \le A(\omega) \le 1, \quad \omega \in (0,1).$$

Note the lower and upper bounds of the Pickands dependence functions above correspond to the total dependence and independence, respectively. Therefore, we can show graphically the dependence by examining the closeness of the Pickands' dependence function. In addition, a tractable quantity to summarize the main properties of the dependence structure is the upper tail dependence coefficient (UTDC) $\chi \in [0,1]$, giving a rough but representative picture of the full dependence structure [32]. The UTDC is a limiting measure of the tendency for one variable to be large conditional on the other variable being large, i.e.,

$$\chi = \lim_{u \to 1} \mathbb{P}\{F_Y(Y) > u \,|\, F_X(X) > u\}.$$

In the case that $\chi = 0$, the variables are said to be asymptotically independent, and $\chi = 1$ corresponds to the total dependence. It can be given through the Pickands' dependence function as $\chi = 2 - 2A(1/2)$, which equals $2 - 2^{\alpha}, \alpha \in [0,1]$ for bivariate logistic dependence model [8].

In the application of bivariate POT method, the selection of a pair of suitable thresholds (u_x, u_y) is of vital importance in this method. There is no universal threshold selection method for bivariate POT. The spectral measure plot of is one of the potential choices developed by [3], which was developed [29]. The evd package in R could achieve this [30]. Spectral measure plot helps decide k_0, which is the rank of upper order statistics of the radius $r_i = \widetilde{x}_i + \widetilde{y}_i$. In other words, there should be at least k_0 observations which exceed at least one marginal threshold for analysis. The pair of thresholds with their values corresponding to the $(n-k_1)^{th}$ data ranked in ascending order can then be selected, where $k_1 = \lfloor (k_0+1)/2 \rfloor$. With the properly selected thresholds, excess residuals are fitted by bivariate GP distribution adopting the logistic model as its dependence model using POT package [31]. The strength of the extremal dependence is measured by the value of α and finally visualized via Pickands' dependence function plots.

3. Results

3.1. Descriptive analysis

Figs. 1 and A.2 displayed the trends and seasonality of monthly precipitation and mortality rate of infectious diseases at the national and regional scales, respectively. The red smoothing lines indicate the underlying trends. In regard to monthly precipitation, a clear upward trend was observed at the national level and in the Midwestern and Southern USA. Although an overall downward trend is presented in the monthly mortality rate, at the regional level, mortality rates increase in almost all regions, excluding the Northeastern USA. Seasonal patterns are also visible from these time series plots as regularly repeated variations. The yearly seasonality in monthly precipitation is only seen in the Midwestern and Western USA, where precipitation surges in the mid of the year in the Midwestern USA whereas relatively low in a similar period in the Western USA. The yearly pattern for the mortality rate of the whole infectious diseases category is almost identical at both national and regional levels, with the valley occurring in the middle of the year and peaks reaching in the winter.

In Fig. A.3, we examine the seasonality and trends of monthly mortality rate for sub-categories of infectious diseases at the national and regional levels for the Northeastern and Southern USA (the other regions are deferred in the Appendix). All sub-categories generally show similar patterns all over the country and in the Northeastern USA. Aside from intestinal infectious diseases and other bacteria disease categories, the mortality rate of other categories shows a decreasing trend. The annual seasonality exhibits only in the mortality rate of intestinal infectious diseases and other bacteria diseases. Regarding the sub-categories in the Southern USA, with similar overall trends, the trend changes for all sub-categories are more prominent compared to those at the national and Northeastern levels. The data ranges for each sub-category are wider in this region as well.

Table 1 presents the descriptive statistics of monthly precipitation and mortality rate of the whole category at both national and regional levels. The distributions of all monthly precipitation and mortality rates are right-skewed, with greater mean than median values. The range of monthly precipitation is wider in the Northeast and narrower in the South. The monthly mortality rate of the general infectious diseases category in the South is much higher than the national level, while the mortality rate in the West is lower than the national level. Table A.1 shows measurements for the mortality rate of sub-categories. Among the five sub-categories, A30-A49 (other bacterial diseases) has a relatively higher mortality rate in all regions. The mortality rates of other bacteria diseases and viral hepatitis vary among all regions with higher values in the Southern USA but are similar for the remaining diseases.

3.2. Seasonality and trends of precipitation and mortality

The premise of applying the bivariate POT method is that the samples should be independent and identically distributed. Thus, the ARIMA model is adopted herein to remove the trend, seasonality as well as non-stationarity from two time series. If heteroscedasticity remains in the residuals, the GARCH model is then utilized to cope with such a situation.

The stationary tests for monthly precipitation and mortality rate at the national level show p-values >0.05, indicating the presence of trends in both data, as well as the seasonality in mortality data revealed from Fig. 1, A.2 and A.3. For regional data, monthly precipitation in the Northeastern USA passes the Ljung-Box test with p-value of 0.08, implying that the raw data is already white noise. However, non-stationarity presents in other regional data with non-significant

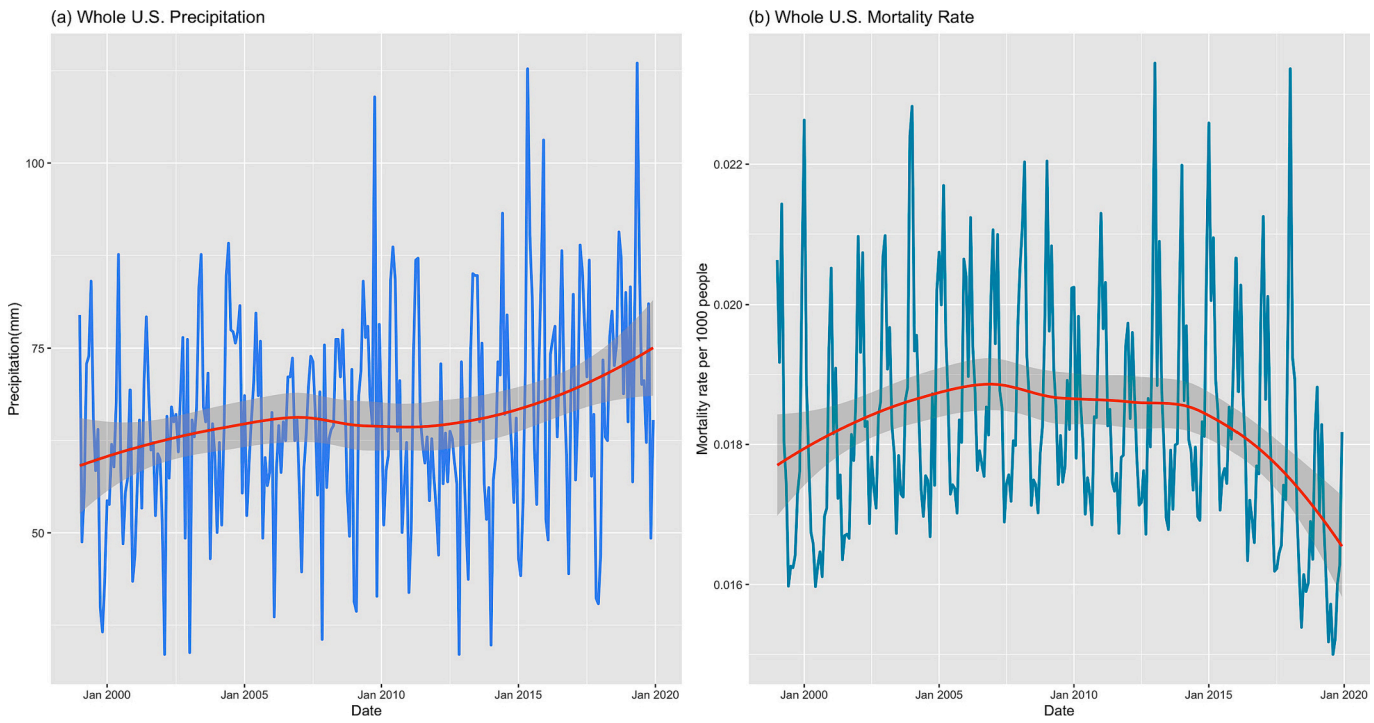

Fig. 1. National monthly precipitation (a) and mortality rate (b). Sources: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/ for precipitation and https://wonder.cdc.gov/ucd-icd10.html for mortality.

Table 1
Descriptive statistics of monthly mortality of the whole category per 1000 people and precipitation in mm (Size = 252).

Variable	Region[1]	Min	Median	Mean	Max	IQR	SD
Precipitation	National	33.53	65.02	65.50	113.54	17.34	14.04
	R1	27.21	95.79	101.23	300.31	49.64	38.65
	R2	15.73	66.70	70.00	164.27	46.98	31.56
	R3	28.57	102.38	102.38	175.59	41.33	29.64
	R4	26.48	44.71	48.42	93.51	16.49	13.14
Mortality	National	0.0150	0.0180	0.0183	0.0234	0.0021	0.0016
	R1	0.0156	0.0219	0.0222	0.0304	0.0304	0.0025
	R2	0.0149	0.0198	0.0199	0.0291	0.0030	0.0023
	R3	0.0342	0.0432	0.0434	0.0606	0.0606	0.0042
	R4	0.0042	0.0134	0.0136	0.0184	0.0022	0.0016

[1] R1: Northeast; R2: Midwest; R3: South; R4: West.

p-values. Differences for removing such non-stationarities in both regional and national data are required and their orders are shown in Table 2, presenting the final ARIMA model parameters for the whole infectious diseases case. The orders of AR and MA components are determined based on the ACF and PACF plots of stationary data after differencing, coupled with results provided by the automated algorithm.

Table 2
Selected seasonal ARIMA models for monthly precipitation in mm and mortality rate of whole categories per 1000 people at both national and regional levels.

Variable	Region	Non-seasonal			Seasonal		
		p	*d*	*q*	*P*	*D*	*Q*
Precipitation	National	3	1	1	0	0	2
	R1	–	–	–	–	–	–
	R2	0	0	0	2	1	0
	R3	2	0	2	2	0	0
	R4	0	0	2	0	1	1
Mortality rate	National	2	1	2	2	1	0
	R1	1	0	1	0	1	1
	R2	2	0	0	1	1	3
	R3	2	1	2	2	1	0
	R4	3	1	1	2	1	1

All residuals of these models pass the Ljung-Box test indicating they are white noises.

Table A.2 provides selected parameters of ARIMA models for the mortality rate of sub-categories of infectious diseases obtained following the same procedures as previous models. Apart from the mortality rate of sequelae of infectious and parasitic diseases in the Northeast, the time series for other diseases all contain seasonal components. It can be inferred from the orders of differencing in these models that different diseases share similar trend and seasonality patterns in all regions, except for some exceptions in the Northeastern and Southern regions for intestinal infectious diseases, mycoses, and sequelae of infectious and parasitic diseases. The residuals of these models are all white noises with *p*-values of Ljung-Box tests >0.05.

3.3. *Extreme association of excess precipitation and mortality*

Spectral measure plots were employed to assist with threshold selections for bivariate data for whole category case and resulted corresponding k_0 are presented in Table 3. Pairs of thresholds are selected via calculating and mapping their indexes to the data in ascending order, following the instruction in Methods section. These thresholds are listed

Table 3
Thresholds and results for bivariate POT analysis of whole infectious diseases category.

Region	k_0	Thresholds		Proportion above threshold		α
		Precipitation	Mortality	Marginal	Joint	
National	145	7.6524	0.0002	0.2900	0.0754	0.7736
R1	92	129.9351	0.0006	0.1825	0.0397	0.7744
R2	118	15.1134	0.0008	0.2341	0.0595	0.7524
R3	152	13.5832	0.0005	0.3016	0.0754	0.7727
R4	125	5.8657	0.0004	0.2500	0.0397	0.7732

in Table 3 alongside marginal and joint proportions above them. The values in the rightmost column α, the maximum likelihood estimate of the parameter involved in the logistic model in Eq.(3), imply the strength of extremal dependence. The extremal dependence between monthly precipitation and mortality rate of the whole infectious diseases category at the national level could be considered weak, with the α being 0.7736 which is closer to 1. The α values for the Northeast, South, and West are similar to the national value, whereas the α value is slightly smaller in the Midwest. This indicates that although still showing a weak pattern, the extremal dependence is relatively stronger in the Midwest than in other regions and at the national level.

Pickands' dependence function plots are helpful to illustrate the strength of dependence (Figs. 2 and 3). The colored lines in these plots represent calculated extremal dependence. The horizontal line at the top of the triangle indicates independence, and the other two grey lines forming the triangle indicate full dependence. Fig. 2 demonstrates that the Midwest (represented by the green dashed line) is more inclined towards the two sides of the triangle, while other lines almost overlap with each other, indicating the same finding.

Table A.3 presents the results of bivariate POT analysis for the sub-categories of infectious diseases at both national and regional levels. The strengths of extremal dependence for different sub-categories in each region are compared and visualized in Fig. 3. In general, the differences in the strength of the extremal dependence for different sub-categories are minor, with lines being compact in each subfigure in Fig. 3. However, in the Northeastern and Western regions, all sub-categories have the same or stronger extremal dependence with monthly precipitation than the whole category, with mycoses having the strongest in the Northeast. A similar situation occurs in the West as well. In the Midwest, intestinal infectious diseases show relatively strong extremal dependence, while no prominent sub-categories were observed with stronger dependence. Among all sub-categories in all regions, the viral hepatitis category (shown by the blue dotted line) has weaker dependence.

4. Discussion

In this study, a combined bivariate Peaks-Over-Threshold and ARIMA methodology was employed to identify a strong positive association between HPEs and infectious diseases. The extremal dependence identified herein differs from mean associations in past studies. The analysis is scientifically significant because it suggests extreme cascading effects which need contingency response strategies. These findings provide new insights for comprehensive risk management in the healthcare industry, public authorities, and environmental agencies.

The study's novelty lies in developing and illustrating bivariate extreme value modelling methods that identify extreme weather environments and their post-influences on public health issues like infectious diseases. Further research on the influence of extreme environments on public health can be examined using this modelling framework, which employs objective threshold selection based on spectral measure plots from extreme value theory. The strength of extremal dependence is quantitatively analyzed using tail dependence index and graphical toolbox, such as Pickands' dependence plots. However, caution is needed when drawing conclusions from this study due to limited data size. Likewise, infectious diseases in animals with higher induced mortality in rainy period could also be investigated in terms of their extremal dependences following a similar framework owing to their natural transmissions from animals to humans, especially zoonoses [6].

In addition, cross-region and cross-disease comparisons revealed that the strongest dependence was present in the Midwestern USA among the four census regions, with mycoses showing the strongest extremal

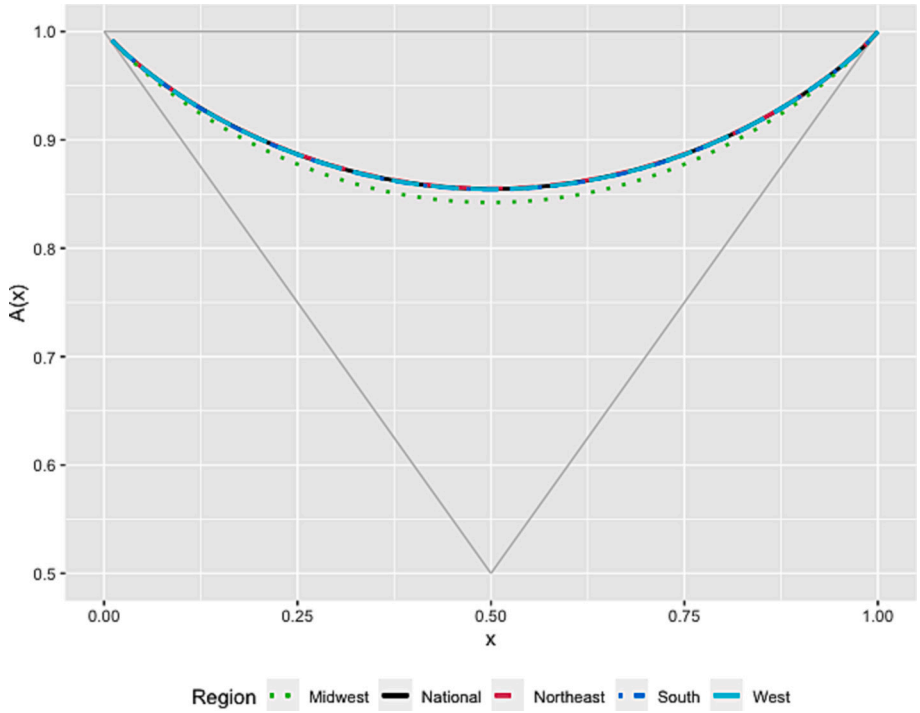

Fig. 2. The plot of Pickands' dependence function for whole infectious diseases category nationally and regionally.

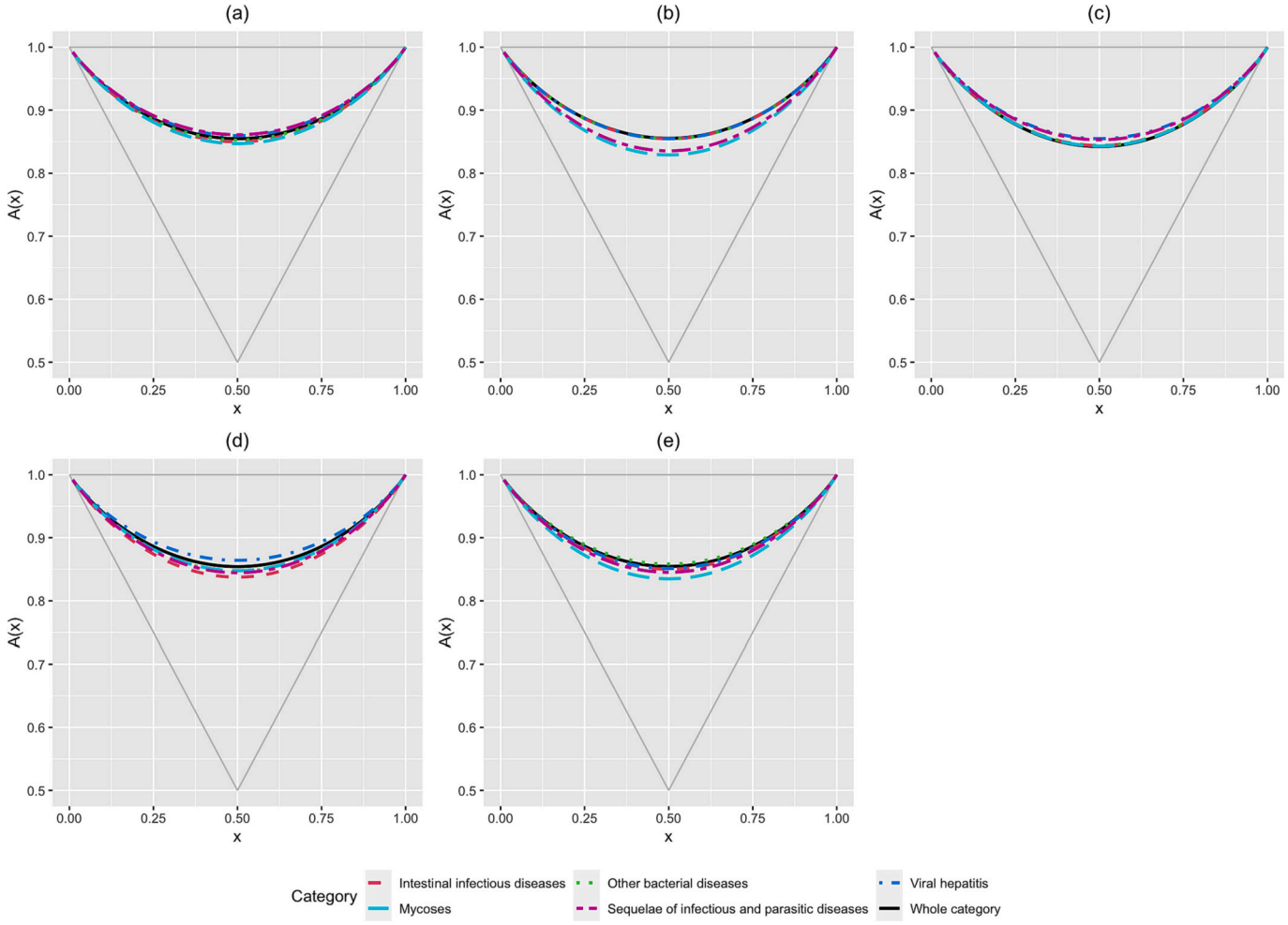

Fig. 3. The plots of Pickands' dependence function for sub-categories of infectious diseases at national and regional levels: (a) National, (b) Northeast, (c) Midwest, (d) South, (e) West.

dependence among the five sub-categories. The findings are in accordance with the geographic, socio-economic factors, and self-inherited disease characteristics. For example, the Western USA's coastal moist subtropical mid-latitude climate experiences frequent extreme precipitation, which can boost infectious disease transmission. Mycoses tend to grow in humid environments with moderate temperature and moisture, making them more prevalent after heavy precipitation events. Additionally, social-economic factors, such as access to quality treatments and income, may vary across Northeastern, Western, Midwestern, and Southern regions, leading to slight differences in extremal dependence between HPEs and infectious diseases in terms of region-specific and disease-specific relationships.

5. Conclusion

Extreme weather events can cause negative consequences, including health issues. This study utilizes multivariate Peaks-Over-Threshold models to examine joint extremes of monthly precipitation and infectious disease in the USA from 1999 to 2019. The extremal dependence found at regional/national levels for various diseases can inform the development of prediction system for health management after disasters caused by extreme events. Moreover, the multivariate extreme value analysis methodology can be useful in studying the extremal dependence structure between multiple extreme weather events and their various health-related consequences.

CRediT authorship contribution statement

Zhiyan Cai: Investigation, Methodology, Data curation, Software, Visualization, Writing – original draft. **Yuqing Zhang:** Writing – review & editing. **Tenglong Li:** Writing – review & editing. **Ying Chen:** Formal analysis, Writing – review & editing. **Chengxiu Ling:** Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Data availability

The data is available via the given link in the paper

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.onehlt.2023.100636.

References

[1] K.T. Aune, M.F. Davis, G.S. Smith, Extreme precipitation events and infectious disease risk: a scoping review and framework for infectious respiratory viruses, Int. J. Environ. Res. Public Health 19 (1) (2022) 165.

[2] J. Beirlant, Y. Goegebeur, J. Segers, J.L. Teugels, Statistics of Extremes: Theory and Applications, John Wiley & Sons, 2006.
[3] J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, Statistics of Extremes: Theory and Applications, Wiley, 2005.
[4] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day Series in Time Series Analysis and Digital Processing, Holden-Day, 1970.
[5] M. Chen, C. Lin, Y. Wu, P. Wu, S. Lung, H. Su, Effects of extreme precipitation to the distribution of infectious diseases in Taiwan, 1994-2008, PLoS One 7 (6) (2012), e34651.
[6] Richard Chepkwony, Carolina Castagna, Ignas Heitkönig, Severine van Bommel, Frank van Langevelde, Associations between monthly rainfall and mortality in cattle due to East Coast fever, Anaplasmosis and Babesiosis, Parasitology 147 (2020) 1743–1751.
[7] G. Chowell, K. Mizumoto, J.M. Banda, S. Poccia, C. Perrings, Assessing the potential impact of vector-borne disease transmission following heavy rainfall events: a mathematical framework, Philos. Trans. R. Soc. B 374 (1775) (2019) 20180272.
[8] S. Coles, An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics, Springer, 2001.
[9] F.C. Curriero, J.A. Patz, J.B. Rose, S. Lele, The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994, Am. J. Public Health 91 (8) (2001) 1194–1199.
[10] A.T. DeGaetano, H. Tran, Recent changes in average recurrence interval precipitation extremes in the mid-Atlantic United States, J. Appl. Meteorol. Climatol. 61 (2) (2022) 143–157.
[11] B.R. Guzman Herrador, B. Freiesleben de Blasio, E. MacDonald, G. Nichols, B. Sudre, L. Vold, J.C. Semenza, K. Nygard, Analytical studies assessing the association between extreme precipitation or temperature and drinking water-related waterborne infections: a review, Environ. Health: Glob. Access Sci. Source 14 (1) (2015) 1–12.
[12] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, 3nd ed, OTexts, Australia, 2021.
[13] A.W. Ledford, J.A. Tawn, Concomitant tail behaviour for extremes, Adv. Appl. Probab. 30 (1) (1998) 197–215.
[14] H. Li, H. Liu, Q. Tang, Z. Yuan, Pricing extreme mortality risk in the wake of the COVID-19 pandemic, Insur.: Math. Econ. 108 (2023) 84–106.
[15] H. Li, Q. Tang, Joint extremes in temperature and mortality: a bivariate POT approach, N. Am. Actuar. J. 26 (1) (2022) 43–63.
[16] J. Li, Z. Cai, Y. Liu, C. Ling, Extremal analysis of flooding risk and its catastrophe bond pricing, Mathematics 11 (1) (2023) 114.
[17] R. Peng, Y. Wang, J. Zhai, J. Zhang, Y. Lu, H. Yi, H. Yan, Y. Peng, T. Sharav, Z. Chen, Driving effect of multiplex factors on human brucellosis in high incidence region, implication for brucellosis based on one health concept, One Health 15 (2022) 100449.
[18] D. Phung, C. Chu, S. Rutherford, H. Lien, T. Nguyen, M.A. Luong, C.M. Do, C. Huang, Heavy rainfall and risk of infectious intestinal diseases in the Most Populous City in Vietnam, Sci. Total Environ. 580 (2017) 805–812.
[19] R.B. Singh, S. Hales, N. de Wet, R. Raj, M. Hearnden, P. Weinstein, The influence of climate variation and change on diarrheal disease in the Pacific Islands, Environ. Health Perspect. 109 (2) (2001) 155–159.
[20] G.S. Smith, K.P. Messier, J.L. Crooks, T.J. Wade, C.J. Lin, E.D. Hilborn, Extreme precipitation and emergency room visits for influenza in Massachusetts: a case-crossover analysis, Environ. Health 16 (1) (2017) 108.
[21] Q. Xu, L. Han, K. Xu, Causal analysis and prevention measures for extreme heavy rainstorms in Zhengzhou to protect human health, Behav. Sci. 12 (6) (2022) 176.
[22] S. Zhao, Y. Deng, R.X. Black, A dynamical and statistical characterization of US extreme precipitation events and their associated large-scale meteorological patterns, J. Clim. 30 (4) (2017) 1307–1326.
[23] L. Wang, X. Li, L. Fang, D. Wang, W. Cao, B. Kan, Association between the incidence of typhoid and paratyphoid fever and meteorological variables in Guizhou, China, Chin. Med. J. 125 (3) (2012) 455–460.
[24] R.J. Hyndman, Y. Khandakar, Automatic time series forecasting: the forecast package for R, J. Stat. Softw. 26 (3) (2008) 1–22.
[25] D. Qiu. Alternative Time Series Analysis, 2015. URL, https://cran.r-project.org/web/packages/aTSA/aTSA.pdf.
[26] P. Embrechts, C. Klüppelberg, T. Mikosch. Modelling Extremal Events: for Insurance and Finance 33, Springer Science & Business Media, 2013.
[27] H. Southworth, J.E. Heffernan, Multivariate extreme value modelling of laboratory safety data from clinical studies, Pharm. Stat. 11 (5) (2012) 367–372.
[28] P. Hougaard, A class of multivanate failure time distributions, Biometrika 73 (3) (1986) 671–678.
[29] L. Zheng, K. Ismail, T. Sayed, T. Fatema, Bivariate extreme value modeling for road safety estimation, Accid. Anal. Prev. 120 (2018) 83–91.
[30] A. Stephenson, evd: Extreme value distributions. R news 2 2, 2002, pp. 31–32.
[31] M. Ribatet, C. Dutang. POT: Generalized Pareto Distribution and Peaks Over Threshold, r package version 1.1-10, 2022. URL, https://CRAN.R-project.org/package=POT.
[32] S. Coles, J. Heffernan, J. Tawn, Dependence measures for extreme value analyses, Extremes 2 (1999) 339–365.