GeoHealth

RESEARCH ARTICLE
10.1029/2022GH000587

Web-Based Data to Quantify Meteorological and Geographical Effects on Heat Stroke: Case Study in China

Qinmei Han[1,2,3], Zhao Liu[4], Junwen Jia[5], Bruce T. Anderson[6], Wei Xu[1,2,3], and Peijun Shi[1,2,3,7]

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing, China, [2]Academy of Disaster Reduction and Emergency Management, Ministry of Emergency Management and Ministry of Education, Beijing Normal University, Beijing, China, [3]Faculty of Geographical Science, Beijing Normal University, Beijing, China, [4]School of Linkong Economics and Management, Beijing Institute of Economics and Management, Beijing, China, [5]School of System Science, Beijing Normal University, Beijing, China, [6]Department of Earth and Environment, Boston University, Boston, MA, USA, [7]Academy of Plateau Science and Sustainability, People's Government of Qinghai Province and Beijing Normal University, Xining, China

Abstract Heat stroke is a serious heat-related health outcome that can eventually lead to death. Due to the poor accessibility of heat stroke data, the large-scale relationship between heat stroke and meteorological factors is still unclear. This work aims to clarify the potential relationship between meteorological variables and heat stroke, and quantify the meteorological threshold that affected the severity of heat stroke. We collected daily heat stroke search index (HSSI) and meteorological data for the period 2013–2020 in 333 Chinese cities to analyze the relationship between meteorological variables and HSSI using correlation analysis and Random forest (RF) model. Temperature and relative humidity (RH) accounted for 62% and 9% of the changes of HSSI, respectively. In China, cases of heat stroke may start to occur when temperature exceeds 36°C and RH exceeds 58%. This threshold was 34.5°C and 79% in the north of China, and 36°C and 48% in the south of China. Compared to RH, the threshold of temperature showed a more evident difference affected by altitude and distance from the ocean, which was 35.5°C in inland cities and 36.5°C in coastal cities; 35.5°C in high-altitude cities and 36°C in low-altitude cities. Our findings provide a possible way to analyze the interaction effect of meteorological variables on heat-related illnesses, and emphasizes the effects of geographical environment. The meteorological threshold quantified in this research can also support policymaker to establish a better meteorological warning system for public health.

Plain Language Summary The impact of extreme heat events on population has become of urgent public health concern. In China, the real mortality and morbidity data are not publicly available and data of some diseases are not enough recorded. Thus, it is difficult to build a solid relationship between heat strokes and meteorological variables in large spatial scale. Internet search data, highly correlated with real heat stoke cases, was adopted as a new data set in our research instead of the heat-related morbidity data to investigate the relationship between heat strokes and multiple meteorological variables. According to Random forest model, temperature and relative humidity (RH) were identified as the most important factor, which mainly affected the severity of heat strokes. We quantified the meteorological threshold that affected the severity of heat strokes as well. Heat strokes may start to occur when temperature exceeds 36°C and RH exceeds 58% in China. The temperature of this threshold is higher in the south or low-latitude region or coastal region of China. This work provides new insights for health research and help to timely alert the public in adverse weather conditions.

Key Points:

- Internet search data could help to evaluate the health impact in large scale to address the data shortage
- Increasing temperature and relative humidity (RH) led to sharply increased heat stroke, particularly when maximum temperature exceeds 36°C and RH exceeds 58%
- Meteorological threshold was affected by geographical factors like altitude, latitude and distance from ocean

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P. Shi,
spj@bnu.edu.cn

Citation:

Han, Q., Liu, Z., Jia, J., Anderson, B. T., Xu, W., & Shi, P. (2022). Web-based data to quantify meteorological and geographical effects on heat stroke: Case study in China. *GeoHealth*, *6*, e2022GH000587. https://doi.org/10.1029/2022GH000587

Received 12 JAN 2022
Accepted 28 JUN 2022

Author Contributions:

Conceptualization: Peijun Shi
Formal analysis: Qinmei Han, Zhao Liu
Investigation: Zhao Liu
Methodology: Qinmei Han, Junwen Jia, Bruce T. Anderson, Peijun Shi
Software: Qinmei Han, Junwen Jia
Supervision: Peijun Shi
Validation: Qinmei Han, Zhao Liu, Wei Xu, Peijun Shi
Writing – original draft: Qinmei Han
Writing – review & editing: Qinmei Han, Zhao Liu, Bruce T. Anderson, Wei Xu, Peijun Shi

1. Introduction

Global warming has become a severe problem worldwide. The frequency of extremely hot summers has increased dramatically since the 2003 European heatwave, and events that would occur twice a century up to the early 2000s are now expected to occur twice a decade (Christidis et al., 2015). Extreme heat events are predicted to be more intense, more frequent and last longer over most land areas in the 21st century (Christidis et al., 2015; Gerald & Claudia, 2004).

Exposure to heatwaves is associated with increased mortality. For instance, the European heat wave of 2003 was responsible for an excess of 70,000 deaths in France, Germany, Italy, Spain and other countries (Robine

et al., 2008). In Russia, the death toll of the 2010 summer heatwave totaled 55,000 people (Barriopedro et al., 2011).

There are abundant literature on heat-related morbidity and mortality (Alzeer & Wissler, 2018; Gasparrini et al., 2015, 2017; Kouis et al., 2021; Vicedo-Cabrera et al., 2021; Wu et al., 2020). Since heat-related mortality/morbidity data are not enough, it is difficult to conduct detailed studies of the relationship between heat-related diseases and extreme heat events. Each study tends to use a different set of heat-related diseases and most of them are based on indirectly heat-related mortality/morbidity, such as cardiovascular, cerebrovascular, and respiratory system mortality (Achebak et al., 2018; Basu, 2009; Pan et al., 1995; Salimi et al., 2018; Yang et al., 2018; B. Zhang et al., 2018), the number of emergency department visits, ambulance calls (Dolney & Sheridan, 2006; Schaffer et al., 2012) or hospitalizations (Gronlund et al., 2016). Although these researches used a wide range of health data to examine non-specific health outcomes and clarify a proportion of the total variance, most of the inferred correlations are consistent and practically significant. Heat stroke is a serious heat-related health outcome that can eventually lead to multiple organ tissue injuries, neurologic morbidity, and even death (Alzeer & Wissler, 2018; Dhainaut et al., 2004; Guo et al., 2017). Due to the paucity of heat stroke mortality or morbidity data (Chen et al., 2015; Harlan et al., 2013), it is difficult to perform robust statistical analyses.

Currently, a great deal of attention is being paid to web search query volume data, which could provide a new solution to this problem. Internet search data are widely used for health-related research (Bragazzi et al., 2016; Fazeli Dehkordy et al., 2014; Jung et al., 2019; Lampos et al., 2015). People increasingly use search engines like Google to look for health-related information. Search keywords have become good indicators for understanding activities taking place in the world (e.g., Flu Trends: http://www.google.org/flutrends/). A retrospective observational study carried out in England has found that daily increases in frequency in Google search terms during heatwave events were highly correlated with validated syndromic indicators (Green et al., 2018). Similar to Google, Baidu has become the most popular web search engine service in China in recent years. The Baidu Index, which is similar to Google Trends, allows users to look up the search volume and trends of certain keywords and phrases and can serve as a Baidu keyword research tool. The potential utility of Internet search data to monitor heat-related morbidity was demonstrated in Shanghai (T. Li et al., 2016), where a strong correlation was found between Internet searches for heat stroke and heat stroke deaths and hospitalized cases. Internet search data also have been found to help better predict heat stroke cases in several cities of China (Y. Wang, Song, et al., 2019).

Compared with the wide discussion on temperature and heat-related mortality and morbidity of several kinds of heat-related diseases, studies on the relationship between heat strokes and RH, wind speed and other meteorological variables and their spatial distributions are limited and outdated. Some conclusions from previous studies indicate that heat-related diseases are not only affected by temperature but also by other meteorological factors. For example, the effects of temperature are greater when heat is accompanied by high RH and weak winds (Kunst et al., 1993; Rohat et al., 2019). Based on this understanding of heatwaves and weather, indicators such as apparent temperature and the Heat Index are widely used (Brooke Anderson et al., 2013; Grundstein & Dowd, 2011; Rohat et al., 2019).

Due to the limited availability of large-scale heat stroke morbidity data, no previous studies have quantified the contribution of multiple meteorological variables to heat stroke variations at a large spatial scale. Existing studies mainly focus on one or few cities (T. Li et al., 2016; Y. Wang, Song, et al., 2019). In this article, we used web-based data in place of patient hospitalization caused by heat stroke to analyze the relationships and spatial patterns of the effects of heat with several different meteorological factors in China. Finally, we quantified the relative importance of individual meteorological factors to heat stroke variations and identified the meteorological threshold that affected the severity of heat strokes.

2. Materials and Methods

2.1. Data Sources

This study used two types of data: Baidu Index data and meteorological data. All data were extracted on a daily timescale for the summer periods (May 1 to August 31) from 2013 to 2020, totaling 984 days. The study covered the 333 prefecture-level administrative regions present in China in 2019, which hereinafter are referred to as "cities" (Figure S1 in Supporting Information S1).

The Baidu Index was launched in 2006 by Baidu to provide search records containing different keywords on a daily timescale and reflect the different keyword's "user awareness" and "media attention" (Huang et al., 2016). The Baidu Index provides only normalized data and, as there are no actual scales used in the indices, it is not clear what the exact or absolute numbers are (e.g., if the search index is 120, the real search volume may be 1,000 or 10,000). If the search index is lower than 120, the Baidu Index can be discretized into three ranges (i.e., 0, 0–60 and 60–120). As such, when the search index is lower than 120, it may imprecisely reflect the real search volume. In this study, we only analyzed the search index when it exceeded 120.

Daily Baidu Index data for 333 cities, using the Chinese characters "zhongshu" (which is "中暑" in Chinese) as the keyword, were downloaded from the Baidu website (http://index.baidu.com/). In Chinese, zhongshu has multiple meanings, including heat syncope, heat exhaustion, heat cramps and heat strokes. This keyword has been shown to be highly correlated with heat stroke patients in China (T. Li et al., 2016; Y. Wang, Song, et al., 2019). We have also performed a correlation analysis in this study using real cases of heat strokes from the Chinese Center for Disease Control and Prevention. We collected the daily case numbers of heat strokes in Shanghai, Jinan, Guangzhou and Shenzhen for the study periods of 2015 to 2017, which were used to verify the correlation between HSSI and the number of real cases.

Daily meteorological data for 659 stations, including maximum temperature, RH, evaporation, wind speed, and sunshine duration, were obtained from the China Meteorological Administration. Since there are only 333 cities in this study, it is necessary to find at least one meteorological station for each city. The corresponding principles are as follows: (a) if there is a station in a city jurisdiction, we select this station. (b) if there are more than one station in one city, we calculate the daily mean value of all stations for this city.

2.2. Model Establishment and Validation

First, because the data were not normally distributed, we calculated the Spearman correlation coefficients to quantify the correlations between the HSSI and five daily meteorological variables (maximum temperature, RH, evaporation, wind speed, and sunshine duration) for 333 cities (Figure S1 in Supporting Information S1). Since there are strong correlations between different meteorological factors (Table S1 in Supporting Information S1), we calculated the partial Spearman correlation coefficients between one meteorological variable and HSSI while adjusting the effects of the other controlling variables.

The RF model is an ensemble machine learning method with higher prediction accuracy among currently available algorithms. RF integrates multiple decision trees by averaging the results of each decision tree, which reduces the possibility of over-fitting and increases the robustness of the prediction (Hutengs & Vohland, 2016; Zhao et al., 2021; Zhu et al., 2019). Since the data sets in this research are relatively large and the relationships between meteorological variables are complex, the RF is a good prediction model for our analysis.

Because there are large uncertainties when the search index is lower than 120, we only selected search indices exceeding 120 for further analysis. To ensure enough samples to train the RF model, cities where search data with HSSI >120 exceeded 90% of all the search data in those cities (including HSSI ≤ 120) were selected to establish the RF model. Only 28 cities met this requirement. We also divided the 28 cities into northern and southern cities according to their location with respect to the Qinling Mountains-Huaihe River Line, which is regarded as the geographical divide between northern and southern regions of China (Figure 1). Cities were used as a categorical variable in the model analysis. We randomly chose 95% of the data of each city as training data, and the remaining 5% of the data were treated as testing data. The RF model was established with the scikit-learn Python library, and the optimal hyper-parameters of RF were selected through a grid search optimization method. Root mean square error (RMSE) and determination coefficients (R^2) were used to evaluate the predicted performance (Ali et al., 2020; An et al., 2020; Ng et al., 2020; Zhao et al., 2021; Zhu et al., 2019), calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i^{\mathrm{exp}} - Y_i^{\mathrm{pred}}\right)^2}{N}} \quad (1)$$

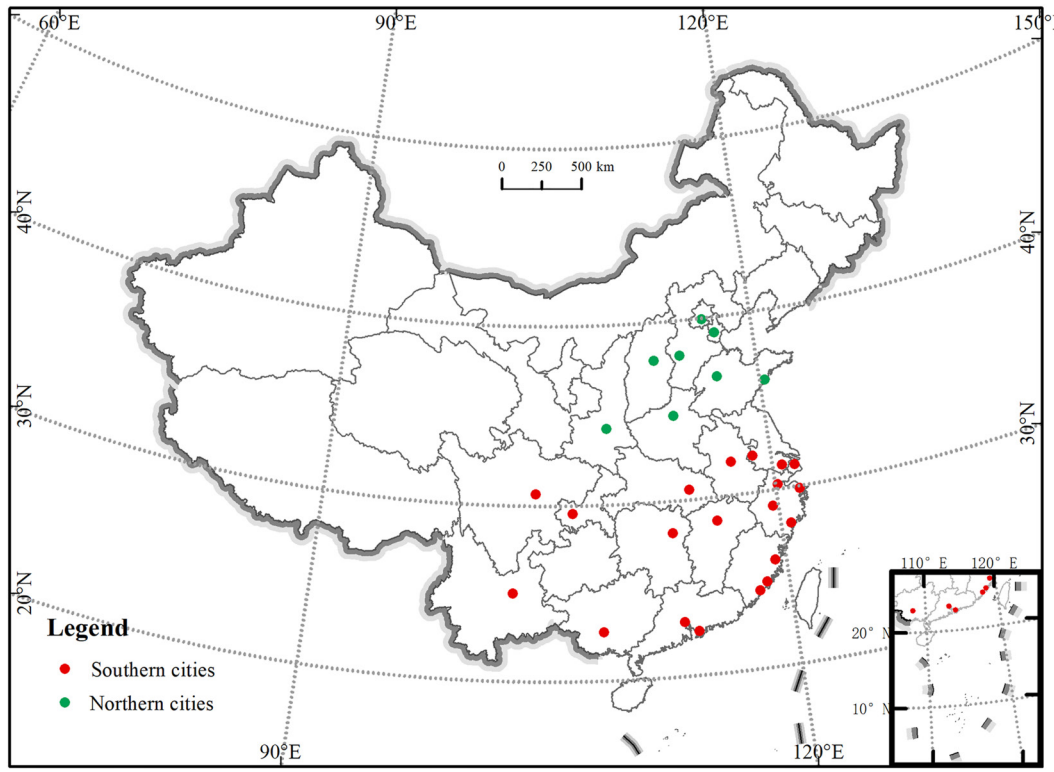


Figure 1. Spatial distribution of northern and southern cities.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i^{\exp} - Y_i^{\text{pred}}\right)^2}{\sum_{i=1}^{N}\left(Y_i^{\exp} - Y_{\text{ave}}^{\exp}\right)^2} \tag{2}$$

where $Y_i^{\exp}$ and Y_i^{pred} are the experimental and predicted values, N is the number of samples, $Y_{\text{ave}}^{\exp}$ is the average of the experimental values.

Partial dependence plots (PDPs) are often used to visualize the relationship between input features and predicted values to explain the machine learning model with the best prediction performance and provide valuable insights of the established model (Zhao et al., 2021; Zhu et al., 2019). In this study, we used PDPs to visualize the marginal effect of one or two meteorological factors on HSSI by varying a single factor and averaging over the values of all other variables.

3. Results

3.1. Correlation Between Heat Stroke Cases and HSSI

We collected real heat stroke cases in Shanghai, Jinan, Guangzhou and Shenzhen for the study period of 2015 to 2017 from National Health Commission of the People's Republic of China. The Pearson correlation coefficients between heat stroke cases and HSSI were computed and shown in Table 1. The coefficients in Jinan and Shanghai exceeded 0.6 which have passed the significance test at 99.9% level. Among these four cities, the larger cases number the higher correlation is.

Table 1
Correlation Coefficients Between Heat Stroke Search Index (HSSI) and Heat Stroke Cases

City	Cases number	Correlation coefficient
Guangzhou	160	0.40***
Jinan	339	0.67***
Shanghai	222	0.66***
Shenzhen	168	0.42***

Note. *** indicates that the coefficients passed the 0.001 significance test.

3.2. Importance Analysis of Meteorological Variables to HSSI

For 333 cities studied in this paper, the correlation coefficients between five meteorological variables and HSSI vary greatly across cities. Most cities that have passed the significance test are located in the east and center of China. Daily maximum temperature, evaporation, wind speed and RH was predominantly positively correlated with HSSI. Sunshine duration was predominantly negatively correlated with HSSI. The higher correlation coefficients were mainly concentrated in the eastern and coastal areas of China (Figure 2).

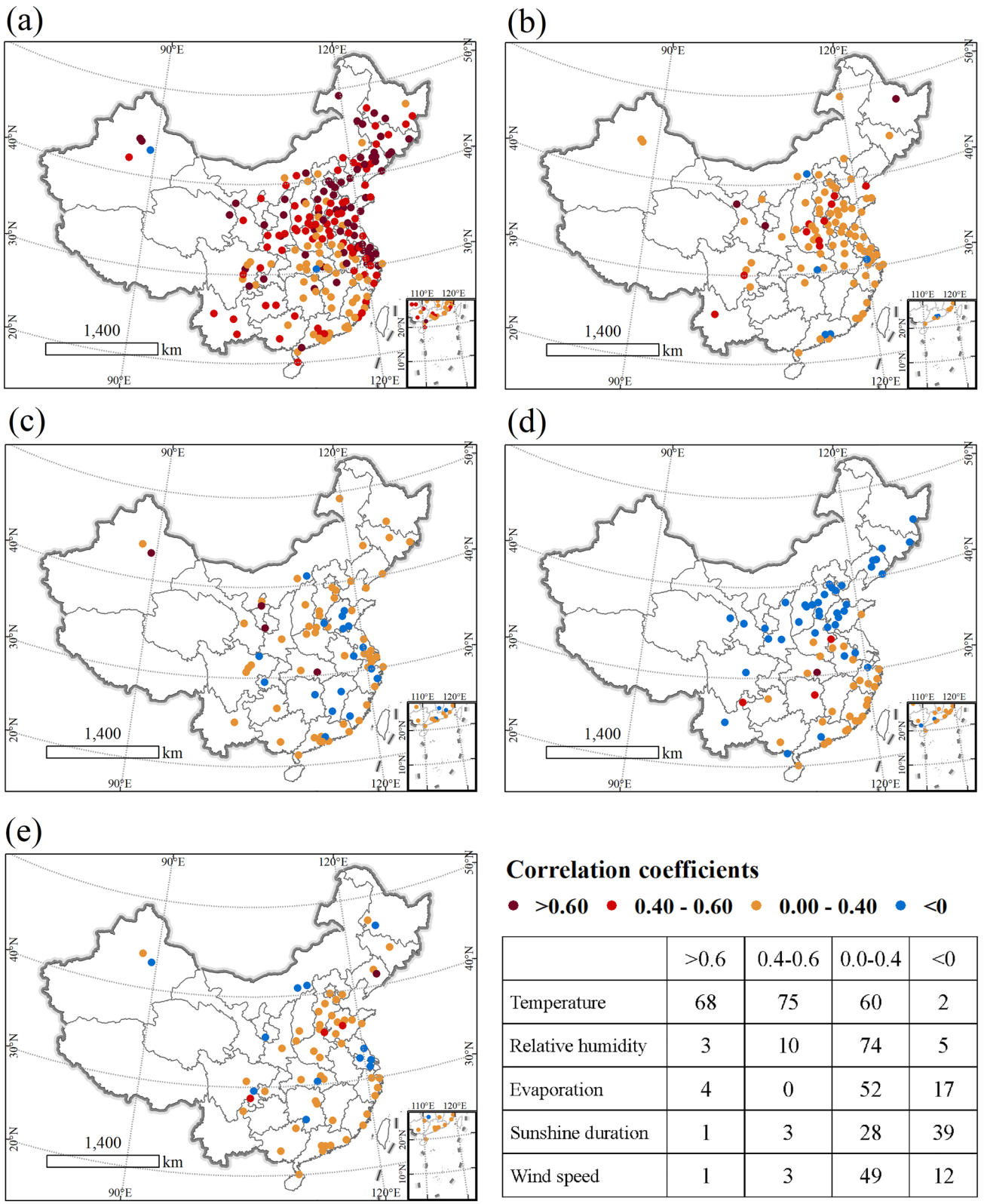
Undoubtedly, temperature is the most important factor. The correlation coefficients for temperature were predominantly positive at 95% significance level in 203 cities distributed in the eastern coastal areas of China (Figure 2a). Cities distributed between 30° and 50°N showed a strong correlation with HSSI. In addition to daily maximum temperature, RH also showed a positive correlation with HSSI in 87 cities, which were mainly distributed in central China (Figure 2b). As seen in Figure 2c, 73 cities had a significant partial correlation between evaporation and HSSI, 77% of which showed a positive correlation. For sunshine duration (Figure 2d), 32 cities mainly located in the south of China were positively correlated with HSSI while 39 cities located in the north of China were negatively correlated with HSSI. For wind speed (Figure 2e), 65 cities passed the 95% confidence test. Only 12 cities showed a negative correlation between wind speed and HSSI, while 53 cities distributed in eastern China were positively correlated with HSSI. To further explore the relationships between heat strokes and meteorological variables, the RF model was applied to analyze the relative importance of each meteorological variable and their interactions in determining heat stroke variations. The R^2 and RMSE of the 28 cities included in the RF analysis are listed in Table S2 in Supporting Information S1. In the training set, the median RMSE and R^2 of 28 cities were 521.41 and 0.94, respectively. In the testing set, the median RMSE and R^2 of 28 cities were 1184.41 and 0.88, respectively. Smaller RMSE and larger R^2 indicate that the model fits better.

The relative importance of temperature, RH, evaporation, wind speed and sunshine duration were assessed in the RF model and are shown in Figure 3. Temperature clearly showed the most significant effect on HSSI, with a median relative importance of 62% among 28 cities. Overall, the importance of RH, wind speed, evaporation and sunshine duration were basically the same, accounting for 9%, 8%, 7% and 9% of the changes of HSSI, respectively.

The importance of daily maximum temperature was approximately the same in the two groups of cities, with median values of 64% and 62% in northern and southern cities, respectively. The importance of RH was evidently higher in northern cities, with a median value of 12%. RH varies greatly in China, gradually decreasing from southeast to northwest (Figure S2 in Supporting Information S1). The RH in northern China is relatively low. This result may indicate people in northern China are more sensitive to RH. It is also observed that wind speed, evaporation and sunshine duration play a slightly more important role in southern China than in the north (Figure 3). Evaporation accounts for 9% and 8% of HSSI changes in southern and northern cities, respectively. Wind speed accounts for 9% and 7% of HSSI changes in southern and northern cities, respectively, and sunshine duration contributes to 11% and 9% HSSI changes in southern and northern cities, respectively.

3.3. Relationship Between Meteorological Variables and HSSI

The partial dependence plots (PDPs) in Figures 4 and 5 show the marginal effect of one or two input variables on the predicted HSSI from the RF model. To isolate the influence of daily maximum temperature on HSSI, only temperature was changed while the constant averages of the other input variables were used in the RF model. In the PDPs with one input variable (Figure 4), the HSSI increased with the increase of temperature, particularly above 20°C. A sharp increase of HSSI appears at temperatures between 30 and 40°C. At temperatures above 40°C, the HSSI remains unchanged, which may be attributable to the fact that the maximum temperature of all samples input in the RF model is 42°C and RF is not efficient at predicting when the input values exceed the range of the training samples. The HSSI started to increase slightly with the increase of RH when this was greater than 60% (Figure 4b). After 80% RH, the HSSI increased rapidly, which indicates that the optimal RH of people may be below 80%. Previous studies have shown that humidity plays an important role in human heat-related discomfort together with temperature (Coffel et al., 2018; Davis et al., 2016; Matthews et al., 2017; Rohat et al., 2019). The skin surface transfers heat to its surroundings through evaporative cooling. This process is more efficient when the temperature and humidity gradients are strong. However, in extremely hot and humid conditions the body may become unable to cool via direct heat exchange. When the body core temperature rises, many adverse health



	>0.6	0.4-0.6	0.0-0.4	<0
Temperature	68	75	60	2
Relative humidity	3	10	74	5
Evaporation	4	0	52	17
Sunshine duration	1	3	28	39
Wind speed	1	3	49	12

Figure 2. Partial Spearman correlation coefficients ($p < 0.05$) between meteorological factors and heat stroke search index (HSSI). (a) Maximum temperature (b) Relative humidity (RH) (c) Evaporation (d) Sunshine duration (e) Wind speed. Points in red are positive and in blue are negative. All points shown on the Figure have passed the significance test. The table in the Figure shows the number of cities in different ranges of correlation coefficients.

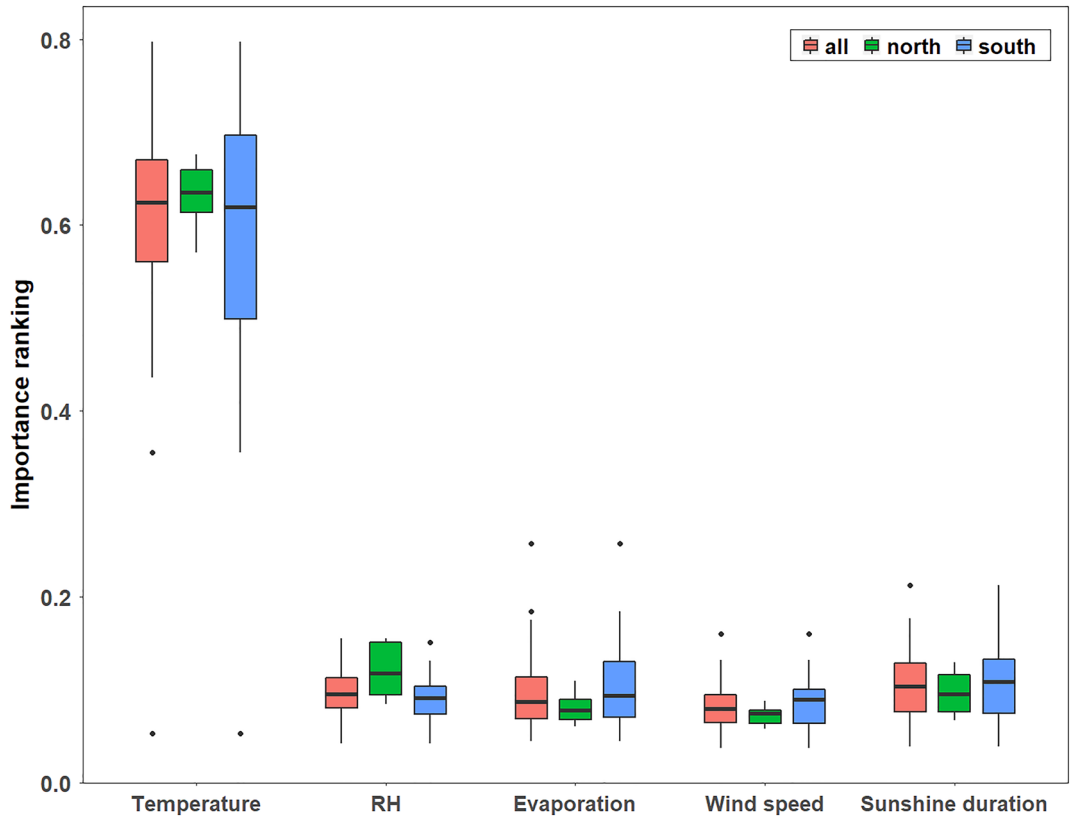


Figure 3. Relative importance ranking of meteorological factors obtained from the Random Forest model. The boxes in red represent all 28 cities, boxes in green and blue represent the cities in the north and south of China, respectively. Black dots are outliers.

outcomes may occur. In our study, the HSSI increased slightly with the increase of evaporation and wind speed (Figures 4c and 4e). On the contrary, the HSSI decreased slightly when sunshine duration increased (Figure 4d). With increased evaporation, sunshine duration and wind speed, the changes in HSSI were not significant (in the range of 180–200), which was consistent with the results of the feature importance analysis.

Our analysis indicates that temperature and RH have significant effects on HSSI changes, which were chosen for further study. The partial dependence plot of two variables is effective for showing the synergy between temperature and RH on HSSI variations (Figure 5). There is a positive interaction between temperature and RH, since increasing temperature and RH lead to increased HSSI when temperature is lower than 40°C, but the dependence of HSSI on the increase of daily maximum temperature is higher. At temperatures between 30 and 40°C, the dependence of HSSI on the increase of RH increased. The HSSI increased markedly at RH greater than 60%, particularly at temperatures greater than 35°C (Figure 5). This method could be used to help different cities formulate risk warning levels of extreme heat events based on multi-meteorological variables and reduce the adverse effects.

After the comparison of real heat stroke cases and HSSI (Table S3 in Supporting Information S1), we selected the 75th and 90th percentile values of HSSI to indicate different levels of heat stroke risk, where the 75th percentile means that heat stroke cases may appear (case number ≥ 2) and the 90th percentile means that a large number of heat stroke cases (case number $\geq$ median of all records) may appear. Then, we computed the meteorological thresholds under different HSSI percentile values of each city in sequence. For northern and southern cities, the mean value of the predicted HSSI for all cities in each group was computed first, and then the thresholds of temperature and RH under different HSSI percentile values were quantified. The thresholds here are the minimum values at which heat strokes may appear or get worse. According to the comprehensive results of 28 cities, cases of heat stroke may start to occur when temperature exceeds 36°C and RH exceeds 58%, and a larger number of heat stroke cases may occur when temperature exceeds 39.5°C and RH exceeds 74% (Table 2). For cities in

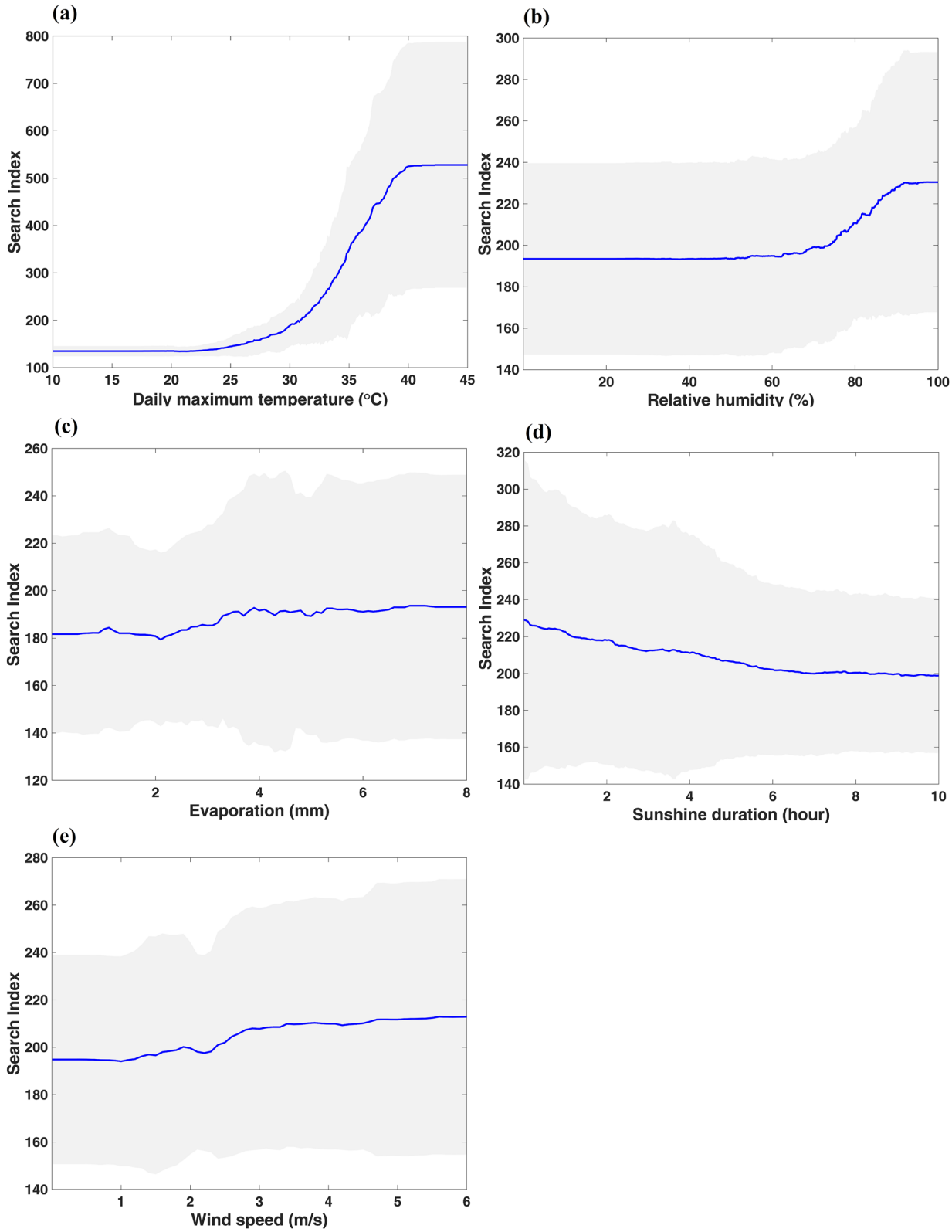


Figure 4. Partial dependence between meteorological variables and heat stroke search index (HSSI). The blue solid lines represent the mean of the variable across 28 cities, and the gray shades represent the standard deviation of the variable across the 28 cities.

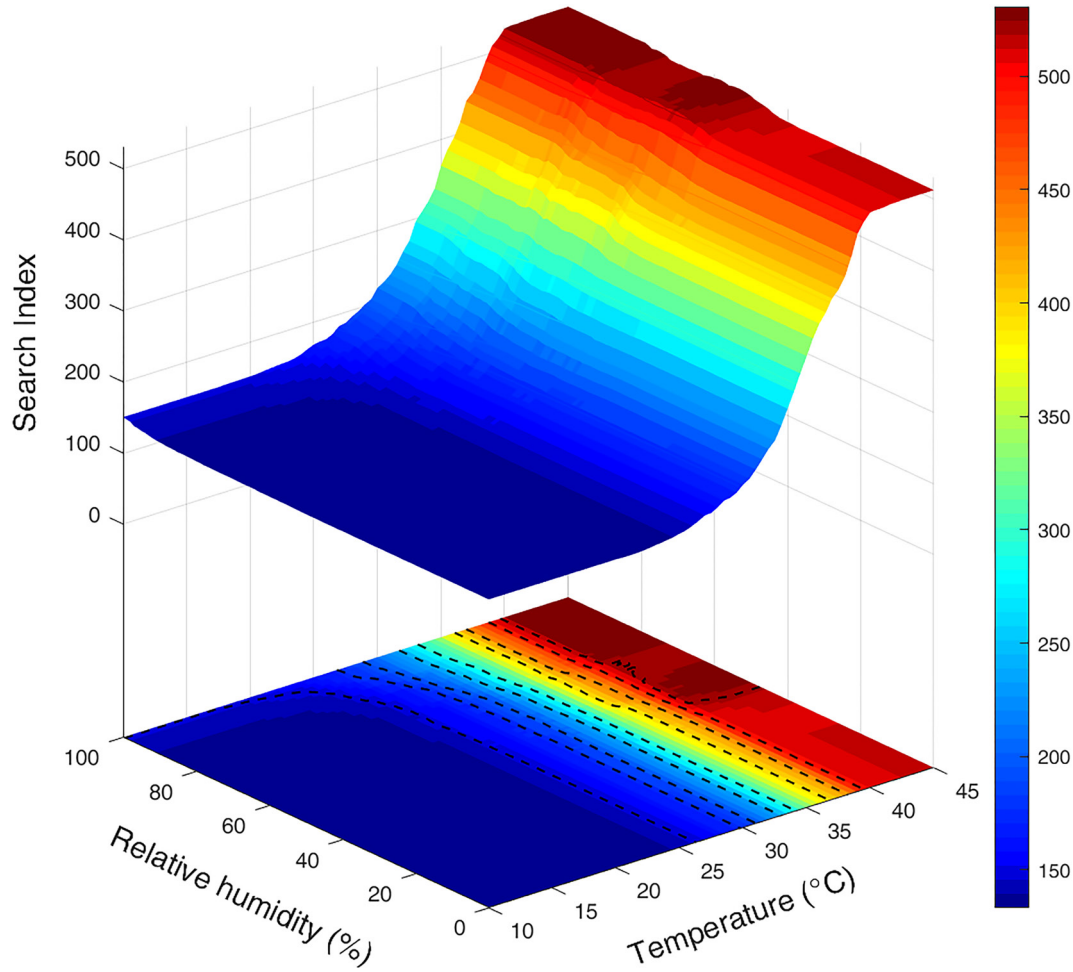


Figure 5. Partial dependence of heat stroke search index (HSSI) on maximum temperature and relative humidity (RH). The color bar refers to values of predicted HSSI.

the north of China, the minimum temperature and RH at which heat stroke cases start to occur are lower than all cities. If temperature exceeds 34.5°C and RH exceeds 79%, there may be some heat stroke cases. Severe heat stroke conditions will occur when temperature is higher than 38°C and RH is higher than 71%. In the south of China, the severity of heat stroke mostly depends on the increase of temperature: 36°C is the minimum temperature at which heat stroke cases start occurring, and 40°C is the minimum temperature at which numerous heat stroke cases start occurring.

Geographical factors such as distance from the ocean and altitude may also affect the threshold of temperature and RH. We computed the mean value of predicted HSSI of coastal and inland cities and quantified the threshold of temperature and RH, respectively. The coastal cities are Fuzhou, Guangzhou, Hangzhou, Jinhua, Ningbo, Qingdao, Shanghai, Shenzhen, Suzhou, Tianjin, Wenzhou, Xiamen, Quanzhou. The others are inland cities. For coastal cities, when temperature exceeds 36.5°C and RH exceeds 1%, cases of heat stroke start to occur, and a larger number of cases of heat stroke may occur when temperature exceeds 40°C and RH exceeds 49%. For inland cities, when temperature exceeds 35.5°C and RH exceeds 68% some cases of heat stroke may occur, and a larger number of cases of heat stroke may occur when temperature exceeds 39°C and RH exceeds 73%. While the temperature threshold in coastal cities is higher than in inland cities, which indicates that people living near the ocean are more adaptable to hot environments. We divided the 28 cities into high-altitude (altitude >200 m) and low-altitude cities according to their mean altitude. High-altitude cities include Chengdu, Chongqing, Jinhua, Taiyuan, Xian and Zhengzhou. We computed the threshold of temperature and RH under different percentile values of HSSI. For high-altitude cities, the minimum temperature and minimum RH at which heat stroke start to

Table 2
Statistics of Meteorological Conditions Affecting the Severity of Heat Stroke

City		Heat stroke cases occurring (75th percentile of HSSI)		Large number of heat stroke cases occurring (90th percentile of HSSI)	
		Temperature (°C)	RH (%)	Temperature (°C)	RH (%)
All cities		36	58	39.5	74
Northern cities	**All**	**34.5**	**79**	**38**	**71**
	Beijing	34.5	75	36.5	71
	Jinan	33	78	37	57
	Qingdao	33.5	69	33.5	79
	Shijiazhuang	33	75	39	73
	Taiyuan	35	68	38	52
	Tianjin	35.5	69	37	62
	Xian	35	55	37.5	71
	Zhengzhou	34.5	73	38	52
Southern cities	**All**	**36**	**48**	**40**	**1**
	Changsha	35.5	72	39	1
	Chengdu	36.5	1	36.5	1
	Chongqing	35.5	1	39	1
	Fuzhou	33.5	77	36	1
	Guangzhou	35.5	76	36	1
	Hangzhou	35.5	75	39.5	59
	Hefei	36	1	39.5	67
	Jinhua	34.5	1	39	50
	Kunming	30.5	1	32	1
	Nanchang	35	65	38.5	60
	Nanjing	36	1	39.5	1
	Nanning	36	1	37.5	71
	Ningbo	35.5	1	38.5	1
	Quanzhou	29.5	74	29.5	74
	Shanghai	36	1	39	57
	Shenzhen	35.5	78	35.5	78
	Suzhou	35.5	1	40	1
	Wenzhou	34.5	82	34.5	83
	Wuhan	36	1	39	65
	Xiamen	36.5	1	38.5	87

occur are 35.5°C and 58%, and minimum temperature and minimum RH of several heat stroke cases appearing are 38°C and 73%. However, the values of low-altitude cities are slightly higher. When temperature exceeds 36°C and RH exceeds 58% the cases of heat stroke may occur, and a larger number of cases of heat stroke may occur when temperature exceeds 39.5°C and RH exceeds 79%. The lower values of meteorological conditions which affect the severity of heat stroke in high-altitude cities also indicate that people living at high-altitudes are more vulnerable to heatwaves.

4. Discussion and Conclusions

This work highlights the potential relationship between meteorological variables and heat stroke using web-based data, providing new insights for countries that lack established public health surveillance systems to monitor heat-related morbidity and help to timely alert the public in adverse weather conditions.

We quantified the partial correlation between meteorological variables and HSSI, and the correlation showed evidently spatial difference especially for temperature, RH and sunshine duration. Daily maximum temperature, evaporation, wind speed and RH are predominantly positively correlated with HSSI. Sunshine duration is predominantly negatively correlated with HSSI. We also analyzed the relative importance of meteorological factors to heat stroke variations. Combined with partial Spearman correlation and machine learning method, our results provide solid evidence of the significant influence of temperature and RH on heat stroke variation, with temperature being the most influential factor. This conclusion is consistent with many existed researches, which suggest that temperature explains the most health outcome variability and select temperature as main influencing factor to study the heat-related health outcomes (Ferrari et al., 2012; Y. Li et al., 2018; Sato et al., 2020; D. Wang, Lau, et al., 2019). The threshold values of temperature and RH were also quantified through the RF model, which were higher in the southern or coastal or low-latitude region of China.

There are still several limitations to our research. Firstly, our results only demonstrate that meteorological variables are significantly correlated with heat stroke and partly contribute to heat stroke variation. However, the mechanism of interaction effects of meteorological factors on the variation of heat stroke is still unclear and needs to be further studied. Humidity plays an important role in heat-related discomfort together with temperature in impact on human health (Coffel et al., 2018; Davis et al., 2016; Matthews et al., 2017; Rohat et al., 2019).The skin surface transfers heat to its surroundings through evaporation of moisture from the skin surface. Even in hot but not humid conditions, human body is still efficient to lose heat through evaporative cooling. However, in extremely hot and humid conditions, body may become unable to cool via direct heat exchange with the surroundings to maintain a stable core temperature. Considering the import influence of humidity, how to choose a proper variable is vital to the research on humidity to health. We selected relative humidity as indicator in our research, which usually refers to the percentage of the vapor pressure in the air and the saturated vapor pressure at the same temperature. However relative humidity can be highly correlated with other atmospheric variables, particularly temperature, making it difficult to identify its accurate contribution to heat stroke in our study. Although relative humidity is the most commonly used moisture variable in epidemiological and environmental health research, in many cases it is inappropriate and it should always be used with caution (Davis et al., 2016). These complexities associated with relative humidity may account for some of the contradictions in the epidemiological literature regarding how humidity influences health outcomes (Gao et al., 2014; K. Zhang et al., 2014). Furthermore, the evaporation data used in this study refer to potential rather than actual water evaporation. As the temperature increases, evaporation increases resulting in atmospheric moisture increases. The effect of evaporation on heat strokes may be indirectly related to increased atmospheric moisture. Studies have shown that wind speed can help reduce skin temperature and reduce the adverse effects of temperature (Alzeer & Wissler, 2018; Lockwood, 1993; Sato et al., 2020). However, the HSSI increased with the increase of wind speed in our analysis, the influence of wind speed need to be further studied. Secondly, our research only quantified the meteorological conditions affecting the severity of heat stroke without considering the demographic, socioeconomic and urban planning factors that may also have critical effects on heat-related health outcomes (Gong et al., 2012; Nayak et al., 2018; Watts et al., 2015).

The search index is highly correlated with the real morbidity of heat stroke in high-income and high internet-usage cities (T. Li et al., 2016). High income means people in these cities may have better access to health care, air conditioning and other house benefits, which can make people less sensitive to heat. The cities included in Random forest model were more developed in China, so that the meteorological threshold quantified by data of these cities may be higher compared with that in relatively undeveloped cities. Exploiting data sources such as the Baidu Index comes at a minimal cost and may therefore be useful in countries with sufficient internet coverage but without an established public health surveillance infrastructure. However, there are notable limitations with these data sources: (a) the data can only describe the phenomenon found but, because demographic information on users is not available, the precise reason for users searching for the terms is unknown (T. Li et al., 2016; Miller & Goodchild, 2014). Moreover, mortality and morbidity data are not publicly available in China. This is why it is difficult to build a solid relationship between the Baidu Index and the total number of patients suffering from heat

stroke. (b) Internet coverage varies greatly among cities. The Internet penetration rate in low and middle-income cities and rural areas is low due to the lower economic and education levels, which results in unreliable data from small and medium-sized cities in China. This explains why the partial correlation coefficients in small cities distributed in the north and center of China are relatively low and insignificant. (c) Many people who are more vulnerable to heat waves, like outdoor workers, older people, young children and homeless people, may not have access to the Internet and therefore may not be captured by Baidu search queries. This may be another limitation of Baidu index data.

As a result of all the mentioned limitations, we have only been able to analyze the correlation and contribution of single meteorological factor to heat strokes. However, in spite of the uncertainties in our analysis, web search data still provide a good alternative way of performing certain kinds of geographic and human-related research. With the rapid development of the Internet, the categories and volumes of big data will continue to grow. When using web-based data, it is vital to choose the appropriate analysis method. Our analysis provides a possible way to analyze the interaction effect of meteorological variables on heat-related illnesses, and emphasizes the effects of geographical environment on heat strokes. The meteorological threshold quantified in this research can support policymakers in timely alerting the public to avoid serious heat-related effects.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The heat stroke search data for the 333 prefecture-level administrative regions in China for the period 2013–2020 can be accessed from the Baidu Index website at https://index.baidu.com/v2/index.html%23/. Daily meteorological data (including maximum temperature, relative humidity, sunshine duration, wind speed and evaporation) can be obtained from the National Meteorological Information Center, China Meteorological Administration at http://data.cma.cn/en/?r%20=%20data/detail%26dataCo-de%20=A.0029.0001 (Climate Daily Data from Surface Meteorological Stations in China V3.0, [Dataset]). Data of real heat stroke cases supporting this research are supplyed by National Health Commission of the People's Republic of China at http://en.nhc.gov.cn/index.html, with restrictions by government policies, and are not accessible to the public. Researchers may gain access or get more information by contacting via chinahealthgov@163.com.

Acknowledgments
This research was supported by the National Key Research & Development program "Global Change Risk of Population and Economic Systems (GCR-PES): Mechanisms and Assessments" of China (Grant No. 2016YFA0602404), and the school-level project of Beijing Institute of Economics and Management (21BSA08). The funding source played a role in data collection and revision of this manuscript. We are grateful to Yue Cai from National Health Commission of the People's Republic of China for providing the real heat stroke cases for our research.

References

Achebak, H., Devolder, D., & Ballester, J. (2018). Heat-related mortality trends under recent climate warming in Spain: A 36-year observational study. *PLoS Medicine*, *15*(7), 1–17. https://doi.org/10.1371/journal.pmed.1002617

Ali, A. M., Darvishzadeh, R., Skidmore, A., Gara, T. W., & Heurich, M. (2020). Machine learning methods' performance in radiative transfer model inversion to retrieve plant traits from Sentinel-2 data of a mixed mountain forest. *International Journal of Digital Earth*, *14*(1), 1–15. https://doi.org/10.1080/17538947.2020.1794064

Alzeer, A. H., & Wissler, E. H. (2018). Theoretical analysis of evaporative cooling of classic heat stroke patients. *International Journal of Biometeorology*, *62*(9), 1567–1574. https://doi.org/10.1007/s00484-018-1551-1

An, G., Xing, M., He, B., Liao, C., Huang, X., Shang, J., & Kang, H. (2020). Using machine learning for estimating rice chlorophyll content from in situ hyperspectral data. *Remote Sensing*, *12*(18), 3104. https://doi.org/10.3390/RS12183104

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., & Garcia-Herrera, R. (2011). The hot summer of 2010: Map of Europe. *Science*, *332*(April), 220–224. https://doi.org/10.1126/science.1201224

Basu, R. (2009). High ambient temperature and mortality: A review of epidemiologic studies from 2001 to 2008. *Environmental Health*, *8*(1), 40. https://doi.org/10.1186/1476-069X-8-40

Bragazzi, N. L., Bacigaluppi, S., Robba, C., Nardone, R., Trinka, E., & Brigo, F. (2016). Infodemiology of status epilepticus: A systematic validation of the google trends-based search queries. *Epilepsy and Behavior*, *55*, 120–123. https://doi.org/10.1016/j.yebeh.2015.12.017

Brooke Anderson, G., Bell, M. L., & Peng, R. D. (2013). Methods to calculate the heat index as an exposure metric in environmental health research. *Environmental Health Perspectives*, *121*(10), 1111–1119. https://doi.org/10.1289/ehp.1206273

Chen, K., Huang, L., Zhou, L., Ma, Z., Bi, J., & Li, T. (2015). Spatial analysis of the effect of the 2010 heat wave on stroke mortality in Nanjing, China. *Scientific Reports*, *5*(October 2014), 10816. https://doi.org/10.1038/srep10816

Christidis, N., Jones, G. S., & Stott, P. A. (2015). Dramatically increasing chance of extremely hot summers since the 2003 European heatwave. *Nature Climate Change*, *5*(January), 3–7. https://doi.org/10.1038/NCLIMATE2468

Coffel, E. D., Horton, R. M., & De Sherbinin, A. (2018). Temperature and humidity based projections of a rapid rise in global heat stress exposure during the 21st century. *Environmental Research Letters*, *13*(1), 014001. https://doi.org/10.1088/1748-9326/aaa00e

Davis, R. E., McGregor, G. R., & Enfield, K. B. (2016). Humidity: A review and primer on atmospheric moisture and human health. *Environmental Research*, *144*, 106–116. https://doi.org/10.1016/j.envres.2015.10.014

Dhainaut, J. F., Claessens, Y. E., Ginsburg, C., & Riou, B. (2004). Unprecedented heat-related deaths during the 2003 heat wave in Paris: Consequences on emergency departments. *Critical Care*, *8*(1), 1–2. https://doi.org/10.1186/cc2404

Dolney, T. J., & Sheridan, S. C. (2006). The relationship between extreme heat and ambulance response calls for the city of Toronto, Ontario, Canada. *Environmental Research*, *101*(1), 94–103. https://doi.org/10.1016/j.envres.2005.08.008

Fazeli Dehkordy, S., Carlos, R. C., Hall, K. S., & Dalton, V. K. (2014). Novel data sources for women's health research: Mapping breast screening online information seeking through google trends. *Academic Radiology*, *21*(9), 1172–1176. https://doi.org/10.1016/j.acra.2014.05.005

Ferrari, U., Exner, T., Wanka, E. R., Bergemann, C., Meyer-Arnek, J., Hildenbrand, B., et al. (2012). Influence of air pressure, humidity, solar radiation, temperature, and wind speed on ambulatory visits due to chronic obstructive pulmonary disease in Bavaria, Germany. *International Journal of Biometeorology*, *56*(1), 137–143. https://doi.org/10.1007/s00484-011-0405-x

Gao, J., Sun, Y., Lu, Y., & Li, L. (2014). Impact of ambient humidity on child health: A systematic review. *PLoS One*, *9*(12), 1–27. https://doi.org/10.1371/journal.pone.0112508

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., et al. (2015). Mortality risk attributable to high and low ambient temperature: A multi country observational study. *The Lancet*, *386*(9991), 369–375. https://doi.org/10.1016/S0140-6736(14)62114-0

Gasparrini, A., Guo, Y., Sera, F., Vicedo-Cabrera, A. M., Huber, V., Tong, S., et al. (2017). Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health*, *1*(9), e360–e367. https://doi.org/10.1016/S2542-5196(17)30156-0

Gerald, A. M., & Claudia, T. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, *305*(5686), 994–997. https://doi.org/10.1126/science.1098704

Gong, P., Liang, S., Carlton, E. J., Jiang, Q., Wu, J., Wang, L., & Remais, J. V. (2012). Urbanisation and health in China. *The Lancet*, *379*(9818), 843–852. https://doi.org/10.1016/S0140-6736(11)61878-3

Green, H. K., Edeghere, O., Elliot, A. J., Cox, I. J., Morbey, R., Pebody, R., et al. (2018). Google search patterns monitoring the daily health impact of heatwaves in England: How do the findings compare to established syndromic surveillance systems from 2013 to 2017? *Environmental Research*, *166*(April), 707–712. https://doi.org/10.1016/j.envres.2018.04.002

Gronlund, C. J., Zanobetti, A., Wellenius, G. A., Schwartz, J. D., & O'Neill, M. S. (2016). Vulnerability to renal, heat and respiratory hospitalizations during extreme heat among U.S. elderly. *Climatic Change*, *136*(3–4), 1–15. https://doi.org/10.1007/s10584-016-1638-9

Grundstein, A., & Dowd, J. (2011). Trends in extreme apparent temperatures over the United States, 1949-2010. *Journal of Applied Meteorology and Climatology*, *50*(8), 1650–1653. https://doi.org/10.1175/JAMC-D-11-063.1

Guo, Y., Gasparrini, A., Armstrong, B. G., Tawatsupa, B., Tobias, A., Lavigne, E., et al. (2017). Heat wave and mortality: A multicountry, multicommunity study. *Environmental Health Perspectives*, *125*(8), 087006. https://doi.org/10.1289/EHP1026

Harlan, S. L., Declet-Barreto, J. H., Stefanov, W. L., & Petitti, D. B. (2013). Neighborhood effects on heat deaths: Social and environmental predictors of vulnerability in Maricopa county, Arizona. *Environmental Health Perspectives*, *121*(2), 197–204. https://doi.org/10.1289/ehp.1104625

Huang, X., Zhang, L., & Ding, Y. (2016). The Baidu Index: Uses in predicting tourism flows –A case study of the Forbidden City. *Tourism Management*, *58*, 1–6. https://doi.org/10.1016/j.tourman.2016.03.015

Hutengs, C., & Vohland, M. (2016). Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, *178*, 127–141. https://doi.org/10.1016/j.rse.2016.03.006

Jung, J., Uejio, C. K., Duclos, C., & Jordan, M. (2019). Using web data to improve surveillance for heat sensitive health outcomes. *Environmental Health: A Global Access Science Source*, *18*(1), 1–13. https://doi.org/10.1186/s12940-019-0499-x

Kouis, P., Psistaki, K., Giallouros, G., Michanikou, A., Kakkoura, M. G., Stylianou, K. S., et al. (2021). Heat-related mortality under climate change and the impact of adaptation through air conditioning: A case study from Thessaloniki, Greece. *Environmental Research*, *199*, 111285. https://doi.org/10.1016/j.envres.2021.111285

Kunst, A. E., Looman, C. W., & Mackenbach, J. P. (1993). Outdoor air temperature and mortality in The Netherlands: A time-series analysis. *American Journal of Epidemiology*, *137*(3), 331–341. https://doi.org/10.1093/oxfordjournals.aje.a116680

Lampos, V., Miller, A. C., Crossan, S., & Stefansen, C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, *5*(1), 12760. https://doi.org/10.1038/srep12760

Li, T., Ding, F., Sun, Q., Zhang, Y., & Kinney, P. L. (2016). Heat stroke internet searches can be a new heatwave health warning surveillance indicator. *Scientific Reports*, *6*(May), 1–6. https://doi.org/10.1038/srep37294

Li, Y., Ren, T., Kinney, P. L., Joyner, A., & Zhang, W. (2018). Projecting future climate change impacts on heat-related mortality in large urban areas in China. *Environmental Research*, *163*(November 2017), 171–185. https://doi.org/10.1016/j.envres.2018.01.047

Lockwood, J. G. (1993). Impact of global warming on evapotranspiration. *Weather*, *48*(9), 291–299. https://doi.org/10.1002/j.1477-8696.1993.tb05914.x

Matthews, T. K. R., Wilby, R. L., & Murphy, C. (2017). Communicating the deadly consequences of global warming for human heat stress. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(15), 3861–3866. https://doi.org/10.1073/pnas.1617526114

Miller, H. J., & Goodchild, M. F. (2014). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Nayak, S. G., Shrestha, S., Kinney, P. L., Ross, Z., Sheridan, S. C., Pantea, C. I., et al. (2018). Development of a heat vulnerability index for New York State. *Public Health*, *161*, 127–137. https://doi.org/10.1016/j.puhe.2017.09.006

Ng, K. H., Gan, Y. S., Cheng, C. K., Liu, K. H., & Liong, S. T. (2020). Integration of machine learning-based prediction for enhanced Model's generalization: Application in photocatalytic polishing of palm oil mill effluent (POME). *Environmental Pollution*, *267*, 115500. https://doi.org/10.1016/j.envpol.2020.115500

Pan, W.-H., Li, L.-A., & Tsai, M.-J. (1995). Temperature extremes and mortality from coronary heart disease and cerebral infarction in elderly Chinese Temperature. *The Lancet*, *345*(8946), 353–355. https://doi.org/10.1016/s0140-6736(95)90341-0

Robine, J. M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J. P., & Herrmann, F. R. (2008). Death toll exceeded 70, 000 in Europe during the summer of 2003. *Comptes Rendus Biologies*, *331*(2), 171–178. https://doi.org/10.1016/j.crvi.2007.12.001

Rohat, G., Flacke, J., Dosio, A., Dao, H., & Maarseveen, M. (2019). Projections of human exposure to dangerous heat in African cities under multiple socioeconomic and climate scenarios. *Earth's Future*, *7*(5), 528–546. https://doi.org/10.1029/2018EF001020

Salimi, F., Morgan, G., Rolfe, M., Samoli, E., Cowie, C. T., Hanigan, I., et al. (2018). Long-term exposure to low concentrations of air pollutants and hospitalisation for respiratory diseases: A prospective cohort study in Australia. *Environment International*, *121*(April), 415–420. https://doi.org/10.1016/j.envint.2018.08.050

Sato, T., Kusaka, H., & Hino, H. (2020). Quantitative assessment of the contribution of meteorological variables to the prediction of the number of heat stroke patients for Tokyo. *Scientific Online Letters on the Atmosphere*, *16*(0), 104–108. https://doi.org/10.2151/SOLA.2020-018

Schaffer, A., Muscatello, D., Broome, R., Corbett, S., & Smith, W. (2012). Emergency department visits, ambulance calls, and mortality associated with an exceptional heat wave in Sydney, Australia, 2011: A time-series analysis. *Environmental Health*, *11*(1), 3. https://doi.org/10.1186/1476-069x-11-3

Vicedo-Cabrera, A. M., Scovronick, N., Sera, F., Royé, D., Schneider, R., Tobias, A., et al. (2021). The burden of heat-related mortality attributable to recent human-induced climate change. *Nature Climate Change*, *11*(6), 492–500. https://doi.org/10.1038/s41558-021-01058-x

Wang, D., Lau, K. K. L., Ren, C., Goggins, W. B., Shi, Y., Ho, H. C., et al. (2019). The impact of extremely hot weather events on all-cause mortality in a highly urbanized and densely populated subtropical city: A 10-year time-series study (2006–2015). *Science of the Total Environment*, *690*, 923–931. https://doi.org/10.1016/j.scitotenv.2019.07.039

Wang, Y., Song, Q., Du, Y., Wang, J., Zhou, J., Du, Z., & Li, T. (2019). A random forest model to predict heatstroke occurrence for heatwave in China. *Science of the Total Environment*, *650*, 3048–3053. https://doi.org/10.1016/j.scitotenv.2018.09.369

Watts, N., Adger, W. N., Agnolucci, P., Blackstock, J., Byass, P., Cai, W., et al. (2015). Health and climate change: Policy responses to protect public health. *The Lancet*, *386*(10006), 1861–1914. https://doi.org/10.1016/S0140-6736(15)60854-6

Wu, Y., Wang, X., Wu, J., Wang, R., & Yang, S. (2020). Performance of heat-health warning systems in Shanghai evaluated by using local heat-related illness data. *Science of the Total Environment*, *715*(19), 136883. https://doi.org/10.1016/j.scitotenv.2020.136883

Yang, Y., Tang, R., Qiu, H., Lai, P. C., Wong, P., Thach, T. Q., et al. (2018). Long term exposure to air pollution and mortality in an elderly cohort in Hong Kong. *Environment International*, *117*(October 2017), 99–106. https://doi.org/10.1016/j.envint.2018.04.034

Zhang, B., Li, G., Ma, Y., & Pan, X. (2018). Projection of temperature-related mortality due to cardiovascular disease in Beijing under different climate change, population, and adaptation scenarios. *Environmental Research*, *162*(August 2017), 152–159. https://doi.org/10.1016/j.envres.2017.12.027

Zhang, K., Li, Y., Schwartz, J. D., & O'Neill, M. S. (2014). What weather variables are important in predicting heat-related mortality? A new application of statistical learning methods. *Environmental Research*, *132*, 350–359. https://doi.org/10.1016/j.envres.2014.04.004

Zhao, S., Li, J., Chen, C., Yan, B., Tao, J., & Chen, G. (2021). Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. *Journal of Cleaner Production*, *316*(May), 128244. https://doi.org/10.1016/j.jclepro.2021.128244

Zhu, X., Li, Y., & Wang, X. (2019). Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. *Bioresource Technology*, *288*(April), 121527. https://doi.org/10.1016/j.biortech.2019.121527