frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 22 July 2024
DOI 10.3389/fpubh.2024.1420608

OPEN ACCESS

EDITED BY
Wei-Chun Chou,
University of California, Riverside,
United States

REVIEWED BY
Ming Luo,
Sun Yat-sen University, China
Huasheng Tong,
General Hospital of Guangzhou Military Command, China

*CORRESPONDENCE
Yan Tang
✉ tangyan97_1017@sina.com
Aiqing Han
✉ aqhan@hotmail.com

†These authors have contributed equally to this work and share second authorship

RECEIVED 02 May 2024
ACCEPTED 08 July 2024
PUBLISHED 22 July 2024

CITATION
Xu H, Guo S, Shi X, Wu Y, Pan J, Gao H, Tang Y and Han A (2024) Machine learning-based analysis and prediction of meteorological factors and urban heatstroke diseases. *Front. Public Health* 12:1420608. doi: 10.3389/fpubh.2024.1420608

Machine learning-based analysis and prediction of meteorological factors and urban heatstroke diseases

Hui Xu[1], Shufang Guo[1†], Xiaojun Shi[1†], Yanzhen Wu[1], Junyi Pan[1], Han Gao[2], Yan Tang[1*] and Aiqing Han[1*]

[1]School of Management, Beijing University of Chinese Medicine, Beijing, China, [2]School of Humanities, Beijing University of Chinese Medicine, Beijing, China

Introduction: Heatstroke is a serious clinical condition caused by exposure to high temperature and high humidity environment, which leads to a rapid increase of the core temperature of the body to more than 40°C, accompanied by skin burning, consciousness disorders and other organ system damage. This study aims to analyze the effect of meteorological factors on the incidence of heatstroke using machine learning, and to construct a heatstroke forecasting model to provide reference for heatstroke prevention.

Methods: The data of heatstroke incidence and meteorological factors in a city in South China from May to September 2014–2019 were analyzed in this study. The lagged effect of meteorological factors on heatstroke incidence was analyzed based on the distributed lag non-linear model, and the prediction model was constructed by using regression decision tree, random forest, gradient boosting trees, linear SVRs, LSTMs, and ARIMA algorithm.

Results: The cumulative lagged effect found that heat index, dew-point temperature, daily maximum temperature and relative humidity had the greatest influence on heatstroke. When the heat index, dew-point temperature, and daily maximum temperature exceeded certain thresholds, the risk of heatstroke was significantly increased on the same day and within the following 5 days. The lagged effect of relative humidity on the occurrence of heatstroke was different with the change of relative humidity, and both excessively high and low environmental humidity levels exhibited a longer lagged effect on the occurrence of heatstroke. With regard to the prediction model, random forest model had the best performance of 5.28 on RMSE and dropped to 3.77 after being adjusted.

Discussion: The incidence of heatstroke in this city is significantly correlated with heat index, heatwave, dew-point temperature, air temperature and *zhongfu*, among which the heat index and dew-point temperature have a significant lagged effect on heatstroke incidence. Relevant departments need to closely monitor the data of the correlated factors, and adopt heat prevention measures before the temperature peaks, calling on citizens to reduce outdoor activities.

KEYWORDS

heatstroke, meteorological factor, machine learning, time series, DLNM

1 Introduction

Heatstroke is a series of clinical symptoms caused by fluid and electrolyte disorder, acid-base imbalance, and dysfunction of the thermoregulatory center and the cardiac and cerebral nerves due to prolonged body exposure to high temperature and heat radiation (1). Heatstroke may occur when the temperature exceeds 36°C and the relative humidity exceeds 58% (2). With increasing greenhouse gas emissions and El Nino events, the probable appearance of a year with extreme heat within the next 5 years is as high as 98 per cent (3), which may lead to a significant rise in the number of heatstroke victims. Heatstroke predisposes the heart to added burden, triggering neurological organ damage and systemic inflammatory response syndromes, which can lead to a dramatic increase in the risk of death (4, 5). Compared with the 1986–2005 average, Chinese people experienced 7.85 more heatwave days on average in 2021 (6), and the number of deaths associated with high-temperature heatwaves in China has risen rapidly since 1979 (7). Therefore, the analysis of the impact of meteorological factors on the incidence of heatstroke is crucial for preventing heatstroke and maintaining public health.

Previous studies exploring the effect of meteorological factors on heatstroke have focused on key variables such as temperature and humidity, and analyzed them with a single statistical method. Kumar et al. used simple statistical estimation methods to analyse the effect of heat exposure on human health in the Indian region (8). Wang et al. (9) used a random-effects Poisson regression model to estimate the relative risk (RR) of hospital admission for heatstroke in heatwave weather vs. non-heatwave weather, and had found that the more severe and prolonged the heatwave, the higher the RR value. Li et al. (10) used a zero-inflated Poisson regression model with a logistic distribution to analyze the influence of daily maximum temperature on the occurrence of heatstroke, considering factors such as gender, age, and the severity of heatstroke. In recent years, machine learning algorithms have been gradually applied to the environmental and public health fields. Compared with traditional statistical methods, machine learning algorithms have stronger data processing and model generalization capabilities, and are able to better capture complex non-linear relationships and interactions between multiple factors. The application of machine learning algorithms has made significant progress in the study of the relationship between heatstroke and meteorological factors. Han et al. used correlation analysis and random forest model to analyze the relationship between meteorological variables and heatstroke search index in 333 Chinese cities from 2013 to 2020 (2). Wang et al. used a random forest model to predict heatstroke occurrence for heatwave based on 3 years' data in typical cities with high temperatures in China, which had better performance than the traditional linear regression model. The results indicates that meteorological factors play the most significant role in the model's estimation of the parameters evaluated (11). In addition, some studies have found that the high temperature and high humidity of *sanfu* (the dog days of summer in China) is closely related to the occurrence and treatment of many diseases, such as heatstroke and asthma. Zhu et al. (12) conducted a study on the treatment of asthma by acupuncture, and came to the conclusion that the treatment of the disease is related to *sanfu* in China. Although a number of studies have examined the relationship between heatstroke and meteorological factors, relatively few studies have combined multiple meteorological factors to analyse heatstroke disease in a multidimensional manner.

Therefore, a variety of characteristic data, such as meteorological factors, comprehensive indicators and time series of *sanfu*, were incorporated in this study to reveal the influencing factors of the onset of heatstroke more comprehensively. Adopting a variety of machine-learning algorithms, this study tried to fully exploited the potential information of the data, and has selected the optimal model for making predictions by comparing the performance and prediction effects of different algorithms, so as to improve the accuracy and reliability of the predictions. In addition, to understand the lagged effect and non-linear relationship of heatstroke incidence in a deeper way, this study used the traditional statistical method of Distributed Lag non-linear Model (DLNM) for analysis, thus describing more accurately the relationship and pattern between meteorological factors and heatstroke incidence, which provided an important basis for formulating effective early warning strategies and constructing prediction models. This study has a positive effect on reducing the incidence of heatstroke and protecting public health.

2 Materials and methods

2.1 Data source and variables

This study used data from *Data on heatstroke incidence and meteorological factors in a southern city from May to September in 2014–2019* created by the Chinese Center for Disease Control and Prevention (CDC). The dataset was collected and filled in through the existing monitoring system, integrating data from multiple testing sources, and was released after review by experts, thus reliable data quality. The data contains 919 records and 11 features.

In order that the characteristic data could be better used for the prediction of heatstroke, we calculated the data of daily average air temperature, daily maximum air temperature and relative humidity, and obtained two commonly used comprehensive meteorological indicators, namely heat index and dew-point temperature.

Heat index, i.e., apparent temperature, taking into account the combined effect of both air temperature and relative humidity, refers to the fact that at high temperatures, when the relative humidity is increased, the temperature felt by the human body is higher than the actual temperature. Research has shown that at the same temperature, different relative humidity levels will give individuals different levels of comfort, which in turn will have different impacts on human health (13, 14). The formula for its calculation is as follows:

$$HI = c_1 + c_2T + c_3[RH] + c_4T[RH] + c_5T^2 + c_6[RH]^2 + c_7T^2[RH] + c_8T[RH]^2 + c_9T^2[RH]^2$$

Dew-point temperatu3ature at which the atmosphere is saturated with water vapor when it is cooled without changing its pressure or vapor content (15). When the dew-point temperature

is low, the air temperature or the relative humidity will also be low, either of which can facilitate effective heat dissipation by the human body, thereby reducing the risk of heatstroke. This study employed the Magnus formula to calculate dew-point temperature, utilizing values of $a = 17.27$ and $b = 237.7°C$.

$$T_d = \frac{b\gamma(T, RH)}{a - \gamma(T, RH)}$$

$$\gamma\ (T, RH) = \frac{aT}{b + T} + ln(RH/100)$$

High-temperature heatwaves were included in our study as features as well. A high-temperature heatwave is a complex atmospheric phenomenon that usually refers to a series of consecutive hot days (16). According to the criteria of China Meteorological Administration (CMA), a daily maximum temperature ≥35°C is considered as a "high-temperature day," and three or more consecutive high-temperature days are considered as a high-temperature heatwave. Based on this standard, the daily maximum temperature in the original data was converted, and those who were in a high-temperature heatwave were assigned a value of 1 and those who were not 0.

In addition, the effect of *sanfu* timing characteristics on the number of heatstroke victims was examined. According to the theory of TCM, *sanfu* refers to the three specific periods of the Chinese lunar year from July to August. Specifically, there are *toufu* (the beginning part of *sanfu*) and *mofu* (the ending part of *sanfu*), each lasting precisely 10 days, as well as *zhongfu* (the middle part of *sanfu*), which lasts either 10 or 20 days (17). *Sanfu* has typical climate characteristics such as high temperature, low air pressure, high humidity and low wind speed. The three variables, *toufu*, *zhongfu*, and *mofu*, were assigned 0 and 1 according to the dates of *sanfu* in each year. "0" means it is not in the corresponding period, while "1" means it is. The specific time of sanfu from 2014 to 2019 is shown in Supplementary Table 1.

The finalized dataset comprised primarily temporal variables such as the onset date, year, month, day, weekday, holiday status, and periods of the *sanfu*. It also encompassed meteorological variables including daily average temperature, daily maximum temperature, relative humidity, heat index, dew-point temperature, and high-temperature heatwaves. Additionally, it featured daily total counts of heatstroke incidents and the total population, amounting to a total of 17 variables. The individual variables and their descriptions are shown in Supplementary Table 2.

2.2 Method

Based on a number of meteorological characteristic data and time characteristic data, a distributed lag non-linear model was used to analyse the effects of meteorological factors, such as temperature, humidity and their integrated indicators, high-temperature heatwaves and the *sanfu* time series, on the incidence of heatstroke. The results of the analyses were combined to construct a heatstroke early warning model through machine learning models such as random forest, which provided a basis for preventing the occurrence of heatstroke.

2.2.1 Distributed lag non-linear models

Previous study has shown lag in effect of heat on heatstroke (18), and that the relationship between heat and mortality in the population was mostly non-linear with a "J" curve (19). Therefore, in this paper, a distributed lag nonlinear model (DLNM) was used to fit the relationship between the number of heatstroke occurrences and meteorological factors. The DLNM describes the distribution of the dependent variables in the independent and lagged dimensions by constructing a cross-base, and is now mostly used in analyses of the effects of meteorological factors (20, 21). The formula for its calculation is as follows:

$$\log E\left[Y_t\right] = \alpha + cb\left(x_i, lag\right) + ns\left(date, 10^*1\right) + dow + holiday$$

$E\left[Y_t\right]$ was the number of daily heatstroke occurrences on day t, α was the intercept, cb(x_i, lag) was the established cross-basis function. A 4th order polynomial function was used to specify the maximum number of lag days as 30, and xi was the heat index, dew-point temperature, daily maximum temperature and relative humidity respectively, which was used to illustrate the use of the natural spline function to control for long-term and seasonal trends. "Dow" and "holiday" were respectively week and holiday variables, used to remove confounding effects of week and holiday. The relative hazards were obtained and the lagged effects were visualized through the usage of the R language.

2.2.2 Early warning modeling of heatstroke

Heatstroke occurrence has obvious time-series characteristics such as seasonality and is influenced by multiple factors (e.g., temperature, relative humidity, etc.) (22). Therefore, an attempt was made in this study to predict the number of heatstroke using two time series models, ARIMA and LSTM, along with several machine learning models such as regression decision tree, gradient boosting tree, SVR and random forest, from which the optimal algorithms were selected to be used as the main prediction tool for heatstroke early warning.

Autoregressive Integrated Moving Average (ARIMA), or Autoregressive Sliding Average Model, is a classical statistical method widely used for time series modeling and forecasting (23). Long short-term memory (LSTM) is a special variant of recurrent neural networks with a "gate" structure, which allows the network to converge better and faster, and can effectively improve prediction accuracy (24, 25).

Random forest is a powerful and flexible integrated learning algorithm commonly used for classification and regression problems (26). It is built on decision trees and improves the performance and generalization of the overall model by combining multiple decision trees (27). The algorithm uses Bootstrap sampling technique to randomly select multiple subsamples from the original dataset, each of which is used to train an independent decision tree (28). Its prediction results are based on the integration of multiple decision trees (Supplementary Figure 1). For the regression task the predicted value of the random forest is the average of all the decision trees. Suppose there are B decision trees and the predicted value of the ith tree is fi (x), then the predicted value of the random

forest is:

$$\hat{Y}(x) = \frac{1}{B}\sum_{i-1}^{B} f_i(X)$$

The ARIMA and LSTM models were constructed by analyzing the time series of daily heatstroke occurrences, and both used rolling forecasts for better model predictions (29). The other machine learning models were trained with multiple features in mind and used static prediction in their forecasting. To facilitate the comparison of the models, each model used 2014–2018 data as the training set and 2019 data as the validation set. Parameter tuning of the models were performed by methods such as grid search to improve the generalization ability of the models.

3 Results

3.1 Descriptive statistics

This study analyzed the occurrence of heatstroke, related meteorological factors and comprehensive indicators from May to September in a southern city over a 6-year period from a variety of perspectives. The results showed that there were obvious seasonal fluctuations in the distribution of the number of heatstroke occurrences in this place, with the *sanfu* period being the high incidence time of heatstroke. The occurrence of heatstroke was mainly affected by the local temperature and relative humidity, and the calculated high-temperature heatwave and heat index had the strongest correlation with the number of heatstroke occurrences.

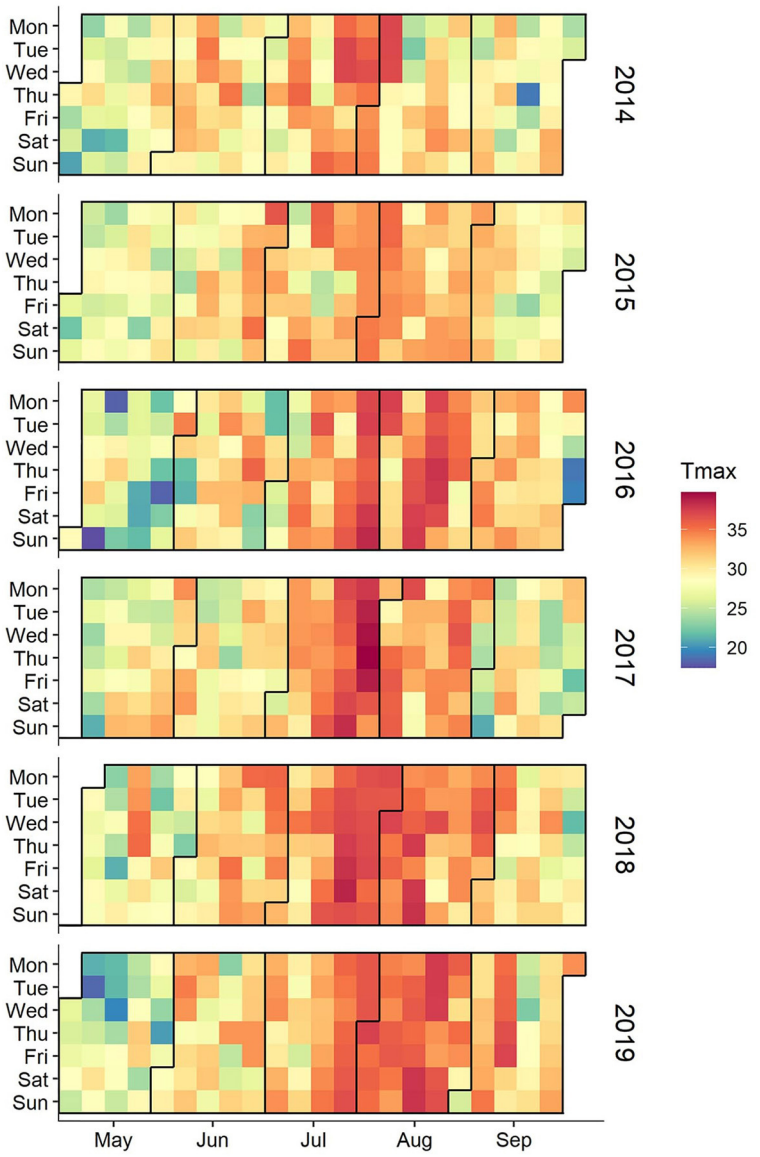

FIGURE 1
Date distribution of daily maximum temperature, May–September 2014–2019.

3.1.1 Distribution of heatstroke occurrences

Descriptive statistical analysis of the number of heatstroke occurrences showed that July and August were the peak periods for heatstroke. From 2014 to 2019, the number of heatstroke occurrences in July was 1,753, accounting for 59.14 per cent of the total number; the number of heatstroke occurrences in August was 966, making up 32.59 per cent of the total number; and the number of heatstroke occurrences in June was 178, constituting 6.01 per cent of the total number; the number of heatstroke occurrences in May and September was comparable, with 24 and 43 occurrences respectively. The number of heatstroke occurrences from 2014 to 2019 showed a more pronounced seasonal variation, with a general trend of increasing and then decreasing (Supplementary Figure 2).

Through visual analyses of daily maximum temperatures, the daily distribution of daily maximum temperatures from May to September 2014–2019 was obtained, which is shown in Figure 1. Maximum temperatures concentrated in the *sanfu* period, and the *sanfu* days were the peak time for the occurrence of heatstroke. The highest number of heatstroke occurrences in *zhongfu* was 1,630, accounting for 54.99% of the total number. The number of heatstroke occurrences in the toufu, *mofu*, and non-*sanfu* days

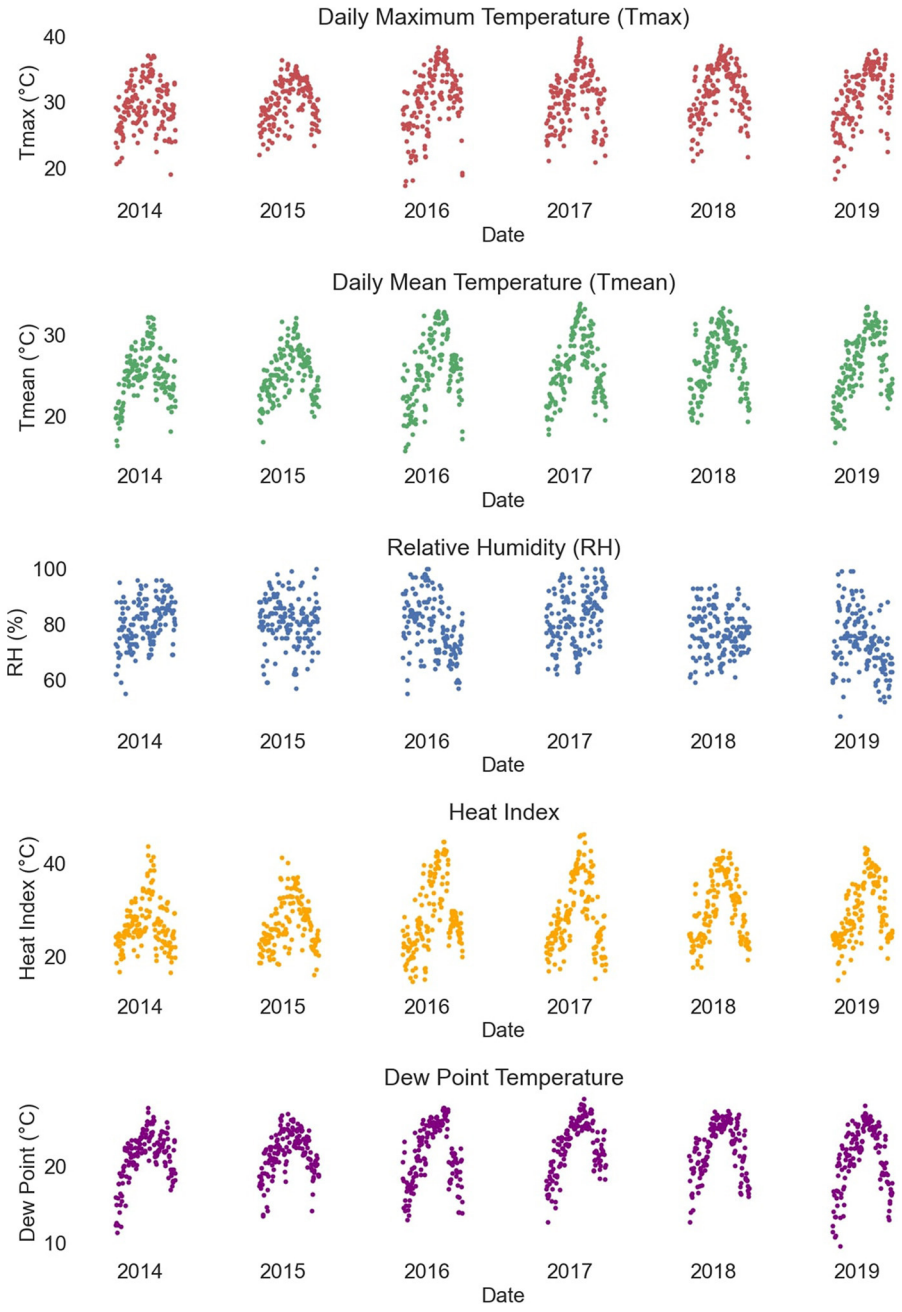

FIGURE 2
Time-series plot of meteorological data.

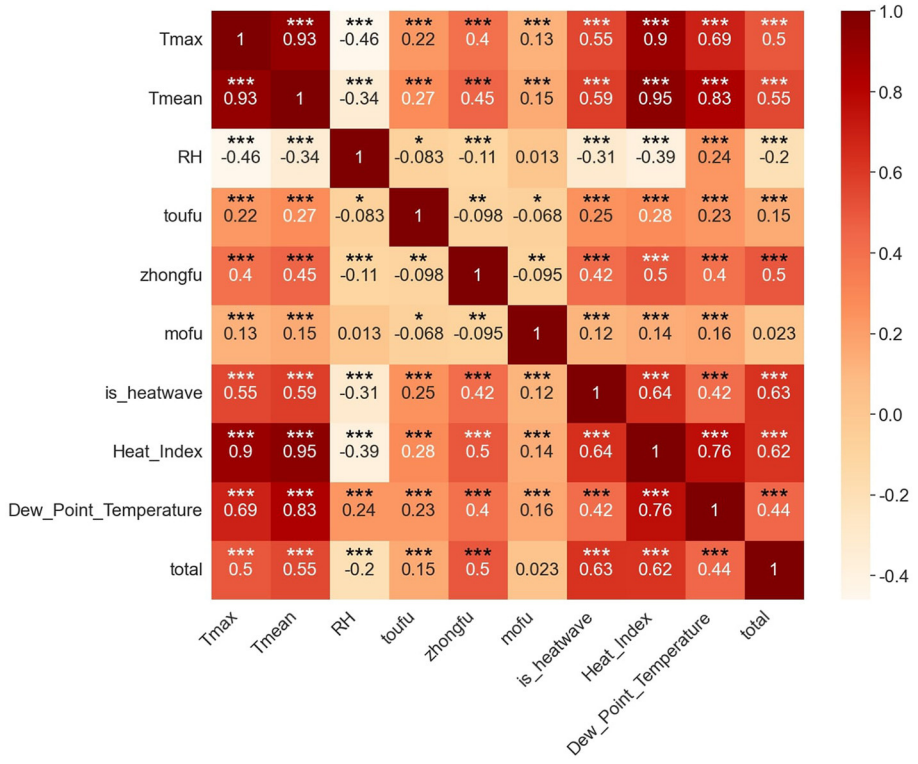

FIGURE 3
Heat map for correlation analysis. *represents statistical significance at the $p < 0.05$ level. **represents statistical significance at the $p < 0.01$ level. ***represents statistical significance at the $p < 0.001$ level.

was relatively small, at 488, 227, and 619 respectively, constituting 16.46%, 7.66%, and 20.88% of the total number.

3.1.2 Relevance analysis

As shown in Figure 2, the daily maximum air temperature, daily mean air temperature, heat index, and dew-point temperature all exhibit distinct seasonal fluctuation patterns over the annual cycle. Further visualization analysis shows a significant positive correlation between temperature and the number of daily heatstroke events, and a clear temporal correspondence between the peak in the number of heatstroke events and the highest point in temperature (Supplementary Figure 3). Relative humidity shows some negative correlation with the number of daily heatstroke occurrences and corresponds to the peak in the number of daily heatstroke occurrences when the relative humidity drops to certain low points (Supplementary Figure 4).

The correlation of the variables in the data was visualized and the heat map obtained is shown in Figure 3, which reveals that the number of daily heatstroke occurrences shows a statistically significant correlation with daily maximum temperature, daily average temperature, relative humidity, *toufu*, *zhongfu*, high-temperature heatwaves, heat index, and dew-point temperature, with all *P*-values <0.001, indicating a strong significance. The correlation coefficients between the number of daily heatstroke occurrences and the daily maximum temperature, daily average temperature are 0.5 and 0.55 respectively, indicating that the higher the temperature, the higher the likelihood of heatstroke occurrences. In addition, the correlation coefficient between the number of daily heatstroke occurrences and whether or not it is *zhongfu* is 0.5, indicating that the likelihood of heatstroke also increases during *zhongfu*. However, the correlation coefficient between the number of daily heatstroke occurrences and relative humidity is −0.2, indicating that the direct link between relative humidity and the number of heatstroke is not strong.

3.2 Cumulative and lagged effects

Figure 4 presents the visualization of the results from the DLNM analysis, encompassing contour plots illustrating the changes in lag time, relative risk (RR), and meteorological data, along with three-dimensional representations depicting various meteorological factors, lag days, and RR values. Within the two-dimensional graphs, regions are color-coded, with red and blue areas signifying where RR is >1 and <1, respectively. These graphical illustrations demonstrate the varying RR of heatstroke incidence in relation to shifts in heat index, dew-point temperature, maximum temperature, and relative humidity, indicating a non-linear association. The lagged effect of heat index, dew-point temperature and daily maximum air temperature on the number of heatstroke incidence was 0–5 days. When the daily heat index was >30°C, dew-point temperature was >23°C and maximum air temperature was >35°C, the risk of heatstroke increased

10.3389/fpubh.2024.1420608

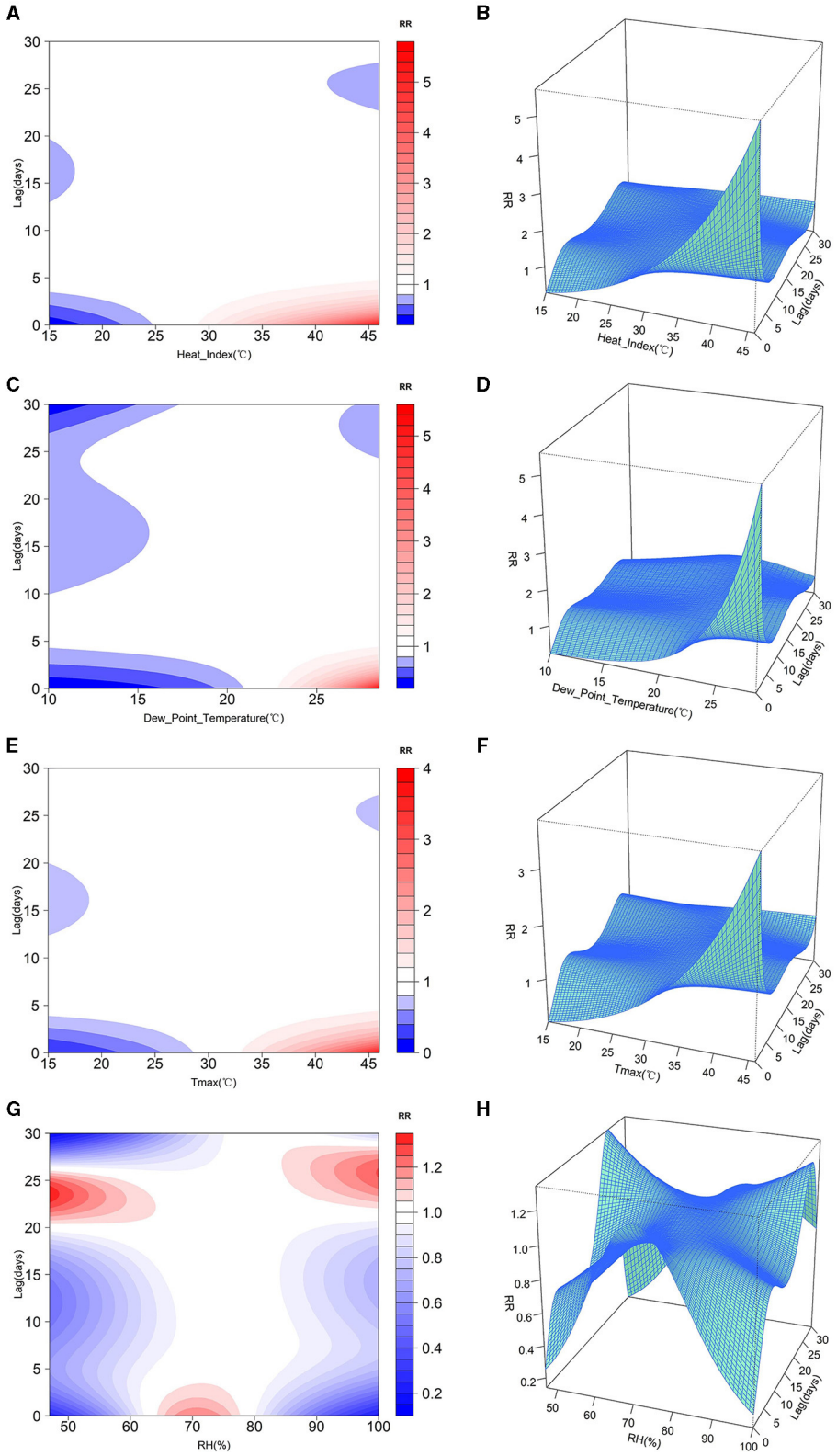

FIGURE 4
Relative risk of different meteorological factors and lagged days: **(A)** Heat Index 2D plot, **(B)** Heat Index 3D plot, **(C)** Dew-Point Temperature 2D plot, **(D)** Dew-Point Temperature 3D plot, **(E)** Daily Maximum Temperature 2D plot, **(F)** Daily Maximum Temperature 3D plot, **(G)** Relative Humidity 2D plot, **(H)** Relative humidity 3D plot.

significantly on the same day and within the following 5 days, and the risk of heatstroke decreased gradually with increasing lag time. Relative humidity had a lagged effect on the number of heatstroke occurrences and the effect varied with relative humidity. When the daily relative humidity was <65%, the relative risk of the lagged 20–25 days was >1, and the risk of heatstroke increased; when the daily relative humidity was between 65 and 78%, the relative risk of the same day and the lagged 5 days was >1 and the risk of heatstroke was relatively high; when the daily relative humidity was > 85%, the relative risk of the lagged 22–28 days was >1, which shows that high humidity has a longer lagged effect on the number of heatstroke incidence.

3.3 Results and comparison of models for predicting the number of heatstroke

3.3.1 Feature selection

In pursuit of enhancing the predictive accuracy of the model, this study initiated the process with the implementation of the Boruta Algorithm for feature selection. Boruta is a feature selection method grounded in random forests, which introduces randomized "shadow features" to compare against real features within an augmented feature matrix. The algorithm trains on this composite matrix and employs the importance scores of these shadow features as a reference baseline, thereby identifying a subset of real features that exhibit genuine relevance to the dependent variable. Given that the suggested depth for Boruta operates optimally with trees pruned to depths ranging from 3 to 7, our study configured each tree in the forest to a depth of 4, retaining default settings for all other parameters, including an estimator count set to "auto," perc at 100%, alpha at 0.05, a two-step approach enabled (two_step = True), and a maximum iteration limit of 100. The resultant analysis identified day, daily maximum temperature (T_{max}), daily mean temperature (T_{mean}), relative humidity (RH), *zhongfu*, high-temperature heatwaves (is_heatwave), heat index (Heat_Index), and dew-point temperature (Dew_Point_Temperature) as variables exerting significant influence on the target variable.

3.3.2 Model comparison

In order to compare the prediction ability of each model, three indicators, mean square error (MSE), root mean square error (RMSE) and coefficient of determination (R^2), were selected for model evaluation in this study. The calculation formulas are as follows:

$$MSE = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

n is the number of samples; y_i is the ith observation; $\hat{y}_i$ is the ith predicted value; and $\hat{y}$ is the mean of the observations.

TABLE 1 Comparison of evaluation indicators for different models.

Model name	MSE	RMSE	R^2
Regression decision tree	15.13	3.89	0.79
Random forest	12.74	3.57	0.82
Gradient boosting decision tree	14.64	3.82	0.80
Linear SVR	14.86	3.85	0.80
ARIMA	34.58	5.89	0.52
LSTM	44.86	6.70	0.38

Residuals are defined as the difference between observations and model predictions. Mean square error (MSE) measures the extent to which the residuals are dispersed, while root mean square error (RMSE) measures the magnitude of residual fluctuations. RMSE is the same scale as MSE, but being on the same order of magnitude as the data points makes it easier to visually compare with the raw data (30). The lower the MSE and RMSE of the model, the higher the quality of the fit. The coefficient of determination R^2 measures the strength of correlation between the predicted and actual values of the model, and its value tends to be closer to 1 indicates the stronger predictive ability the model has. The results of the evaluation of these indicators are shown in Table 1 and indicate that the random forest model stands out among all the compared models with its smallest MSE, RMSE, and R^2 value closest to 1, which suggests that it has higher accuracy in predicting the number of heatstroke victims per day. In order to visualize the effectiveness of the different model algorithms in predicting the number of daily heatstroke occurrences from May to September 2019, a line graph of the actual number of observations against the predictions of the four models was plotted, using time as the horizontal coordinate and the number of daily heatstroke occurrences as the vertical coordinate, as shown in Figure 5.

3.3.3 Random forest model

In this study, the random forest model was hyper-parametrically optimized by a grid search method to determine the optimal parameter configuration: the maximum depth of the decision tree was set to 4; the minimum number of samples to be included in each leaf node was set to 1; and the minimum number of samples required to split a node was 4; the model as a whole consisted of 13 decision trees.

Furthermore, this study calculated the SHAP values and SHAP interaction values for the finalized model and presented them visually, and the results obtained are exhibited in Figure 6. The results show that the thermal index received the highest feature importance score and played a key role in the model prediction process. This finding is consistent with thermodynamic principles and established medical a priori knowledge, the latter suggesting that the onset of heatstroke is closely related to high temperature and relative humidity. Daily mean temperature, daily maximum temperature, relative humidity, and dew-point temperature also exhibited notable significance in the model. The feature importance of the date in the time series was notably high, suggesting a potential periodicity in heatstroke cases throughout the months. Although

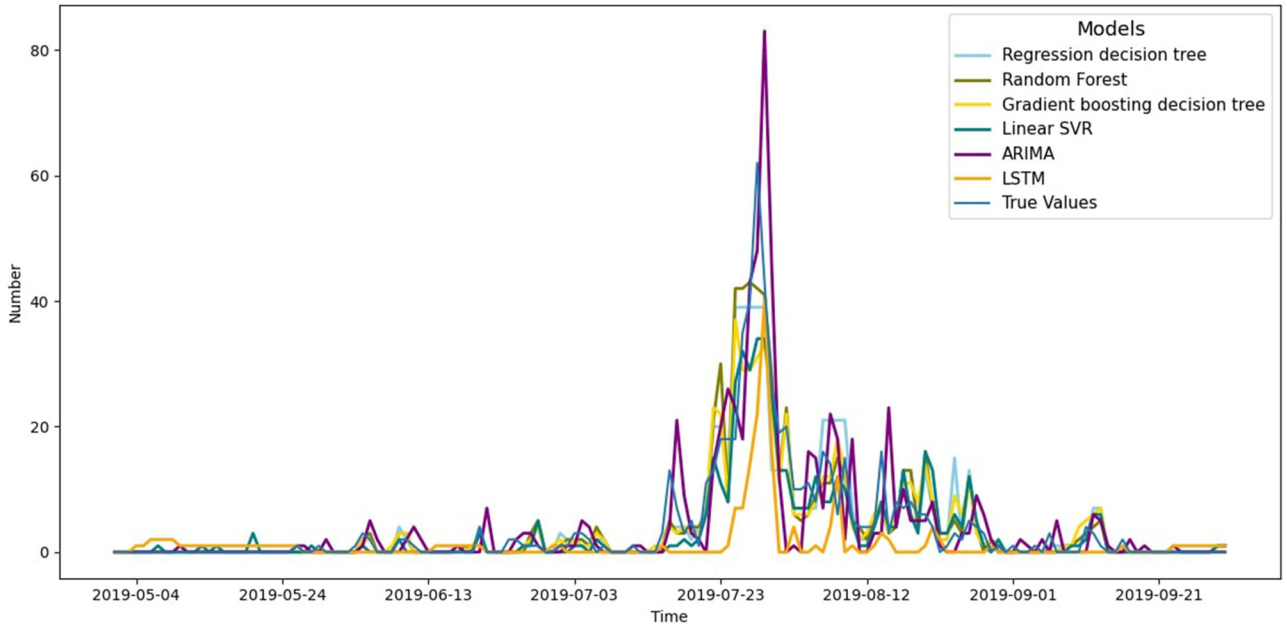

FIGURE 5
Forecasts of daily heatstroke numbers from May to September in 2019 by different model algorithms.

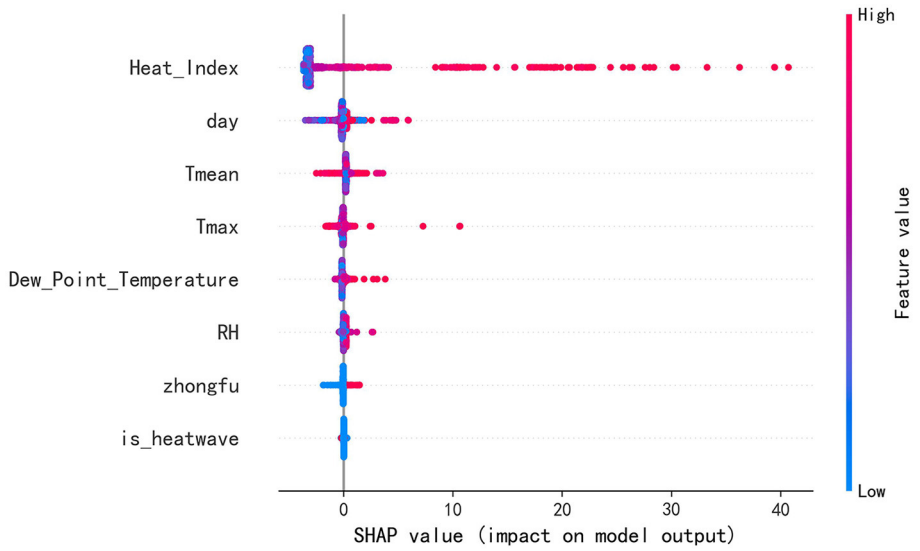

FIGURE 6
SHAP summary plot.

the feature importance score for the *zhongfu* period was relatively low, the graphical depiction clearly illustrated a pronounced positive impact of *zhongfu* on the escalation of heatstroke cases, aligning with the descriptive statistical findings of this research. By comparison, the feature related to high-temperature heatwaves had a lesser role in the model, yet it still contributed to the model's performance to some extent.

In order to assess the robustness of the constructed random forest regression model under different conditions, this study conducted a sensitivity analysis by adjusting the values of each feature individually and monitoring the possible effects of these adjustments on the model performance. By comparing the differences in model scores before and after adjusting the feature values, we assessed the specific impact of different features on model performance (Supplementary Figure 5). The results show that changes in the values of features such as heat index, daily mean temperature, daily maximum temperature, onset day, and dew-point temperature had some impact on the model, but the impacts

were all small, indicating that the random forest regression model developed in this study has good stability and robustness.

4 Discussion

In the context of rising global temperatures, the onset of high-temperature red alerts is occurring earlier, their durations are extending, the affected areas are broadening, their intensities are amplifying, and their extremities are enhancing. Consequently, the incidence of heatstroke, a meteorologically sensitive illness, is anticipated to rise. Therefore, the prediction and early warning of heatstroke are vital, which can enable the relevant departments and the public to get the relevant information in time, which is conducive to the adoption of protective measures in advance, to avoid health risk, and to reduce the damage to health. This study analyzed the cumulative and lagged effects of these factors on heatstroke by screening the influential features of heatstroke and constructing DLNM model analysis. Moreover, various machine learning methods were tried to construct a prediction model for the number of heatstroke victims, and after comparison, we found that the random forest model had the best prediction effect. Through the sensitivity analysis, the model showed high robustness, which indicated that the model would still be able to maintain a highly reliable prediction performance even in the face of some parameter variations or uncertainties.

4.1 Influence of meteorological factors on the number of heatstroke victims

Meteorological factors have a significant impact on summer heatstroke. This study has found that there existed a high correlation between the number of heatstroke cases and the following meteorological variables: high-temperature heatwaves, heat index, daily mean temperature, and daily maximum temperature, as evidenced by substantial correlation coefficient values, implying that they not only directly lead to discomfort, but also may cause high-risk health problems, especially in areas where extreme heat is infrequent (31). However, the feature importance of high-temperature heatwaves was not prominent in the random forest predictive model. Current research on the daytime, nighttime, and compound heatwaves suggests that nighttime heatwaves predominantly occur in low-latitude regions and are typically accompanied by high humidity conditions during nighttime, and that nighttime warmth may impose additional health risks (32). Furthermore, in recent years, heatwaves have exhibited trends of longer durations, greater spatial extents, and slower movement, with slow-moving heatwaves indicative of prolonged high temperatures, potentially having a substantial impact on heatstroke incidences (33). Consequently, future studies could incorporate a broader range of data related to heatwaves to enhance the precision of predictive early warning systems.

Regarding time series aspects, the total number of heatstroke cases during the *sanfu* periods accounted for 79.11% of the total heatstroke occurrences, closely aligning with the peak timing of heatstroke incidents. Particularly during the *zhongfu* phase, temperatures typically reached seasonal highs, and this temporal characteristic exhibited a significant correlation with the heatstroke incidence, reaching 0.5, further confirming high temperatures as a pivotal meteorological factor in heatstroke occurrences. Within the studied region, the effect of relative humidit on heatstroke was relatively minor. As the DLNM model analysis suggests, the impact of relative humidity on heatstroke morbidity was neither linear nor monotonous but an inverted-U shape. Relative humidity displayd a pronounced short-term lagged effect within the range of 65%−78%, while showing more evident long-term lagged effects when relative humidity was below 65% or above 85%, leading to a lower Pearson correlation coefficient. This inverted-U pattern might result from dehydration under low humidity in high temperatures and severe hindrance of heat dissipation under high humidity, both of which disrupt thermoregulation and elevate heatstroke risk over extended periods (34).

Moreover, the DLNM model analysis reveals a lagged effect of heat index and dew-point temperature on heatstroke incidence, with a marked increase in risk on the day of exposure and up to following 5 days once certain threshold values are surpassed (35). High-temperature heatwaves, heat index, daily mean temperatures, and daily maximum temperatures all exert noticeable lagged effects on heatstroke occurrences. In preventing and managing heatstroke, the delayed impacts of these meteorological factors must be fully considered, and appropriate protective measures should be implemented to mitigate heatstroke incidents.

4.2 Forecasting and early warning models

The frequency and duration of extreme temperature events are increasing (31), and the number of heatstroke victims is likely to show a continuous increase in the future (11). It is necessary to construct a forecasting and early warning model for the prediction of the number of heatstroke victims to provide early warning of heatstroke incidence (36). Based on the analysis results of DLNM, this study experimented with a variety of machine learning algorithms to construct a forecasting and warning model for heatstroke, and chose to adopt the random forest model, which is the most effective and robust model, to predict the number of people suffering from heatstroke per day by using multiple meteorological factors and time factors.

Comparing the structure of the decision trees within the final models, commonalities in the branching structure are visible at certain levels, which maps to a consistent understanding of the key predictors in the model. For example, the heat index and average daily temperature are prominent in most trees, implying that these two features contribute more to predicting the target variable in the overall model. At the same time, individual decision trees were observed to pay more attention to additional factors such as "day," "*zhongfu*," and "high-temperature heatwaves," signaling that the model has a certain degree of versatility and is able to make more adaptable predictions for different data patterns.

Synthesizing the analysis of the DLNM model and the results of the comparison of the prediction models, this study proposes a comprehensive early warning mechanism for heatstroke. The core strategy is to use the random forest model to make accurate heatstroke number predictions. Based on the predicted data and

the set warning thresholds, an early warning is implemented when the model predicts a high risk of heatstroke events, and further calculation of the relative humidity, heat index, and dew-point temperature is made. If one of them is identified as a driver of heatstroke occurrences beyond the warning thresholds, then a reinforced warning process for at least five consecutive days is started. At the same time, if the relative humidity is <65 per cent on that day, another warning is issued on the 20th to 25th day after that day, and if it is >85 per cent, another warning is issued on the 22nd to 28th day after that day.

The model could provide real-time early warning information to governments, medical institutions and other public health departments, and promote the development of appropriate preventive measures. By using the model, public health departments can intervene early to improve public health and safety emergency response capabilities, enhance group health protection, and reduce the incidence of heatstroke.

4.3 Research programmes and prospects

The scope of this paper is mainly limited to a specific region in southern China, and thus it is difficult to directly apply the conclusions obtained so far to cities in other geographical environments or climatic conditions. In terms of the selection of meteorological factors, this study covers a relatively limited number of indicators and does not take into account factors such as wind speed, air pressure and weather phenomena. Future research endeavors will broaden the scope of data collection to encompass climatic data from diverse regions, facilitating inter-regional comparative analyses. This enhanced dataset will incorporate a wider array of meteorological variables, including wind velocity, atmospheric pressure, nocturnal heatwaves, compound heatwaves, and the velocity of heatwave movement, aligning with the forefront of meteorological research domains. Such comprehensive data integration aims to enhance the precision of predictive early warning systems.

Exertional heatstroke poses a persistent threat to individuals exposed to high temperatures, with young, healthy individuals of higher body mass index exhibiting an elevated risk, as evidenced in recent literature (37). The diagnostic criteria for occupational heatstroke released by the Chinese Center for Disease Control and Prevention in 2019 highlighted outdoor occupations such as construction, engineering, agricultural labor, and sanitation work as prevalent causes during summer months. Furthermore, intense activities during summer, including sports competitions and military drills, significantly contribute to heatstroke incidents. Subsequent studies could, therefore, stratify participants based on occupation and duration of outdoor exposure, in addition to gender and age, to create detailed population profiles. Integrating these profiles with the city-specific meteorological factors would enable a holistic analysis and the targeted delivery of heatstroke warnings.

The aspiration is that these methodologies will augment the universality and accuracy of heatstroke alerts, thereby furnishing more scientifically grounded approaches for the prevention and control of urban heatstroke diseases.

5 Conclusion

In this study, a distributed lag nonlinear model was used to investigate the lagged and cumulative effects of various climatic factors on heatstroke, and a forecasting model for daily heatstroke occurrences was constructed using the random forest algorithm. The early warning strategy for heatstroke shows that exposure to hot and humid weather tends to increase the risk of heatstroke, and their effects are not limited to the day, but can last for days afterwards. These findings may inform government departments, medical institutions and other organizations of more accurate early warning of heatstroke risks, thus improving public health and safety emergency response capabilities, and reducing the damage to health caused by hot weather.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: http://www.ncmi.cn.

Author contributions

HX: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Supervision. SG: Data curation, Formal analysis, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. XS: Data curation, Formal analysis, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. YW: Investigation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. JP: Data curation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. HG: Writing – original draft, Writing – review & editing. YT: Conceptualization, Funding acquisition, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. AH: Conceptualization, Funding acquisition, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported in part by the China Ministry of Education Industry-University Cooperation Collaborative Education Project (220500643305240) and Horizontal Project of Beijing University of Chinese Medicine, Project No. BUCM-2021-JS-FW-024; Ministry of Education University-Industry Collaborative Education Program, Batch 2, 2022, Project No. 2205 0064 3305 240.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2024.1420608/full#supplementary-material

References

1. Hifumi T, Kondo Y, Shimizu K, Miyake Y. Heat stroke. *J Intensive Care.* (2018) 30:1–8. doi: 10.1186/s40560-018-0298-4

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

3. World Meteorological Organization. *WMO Global Annual to Decadal Climate Update (Target years: 2023-2027).* (2023). Available online at: https://www.un-ilibrary.org/content/books/9789210027939 (accessed March 23, 2024).

4. Kephart JL, Sánchez BN, Moore J, Schinasi L, Bakhtsiyarava M, Ju Y, et al. City-level impact of extreme temperatures and mortality in Latin America. *Nat Med.* (2022) 28:1700–5. doi: 10.1038/s41591-022-01872-6

5. Liu J, Varghese BM, Hansen A, Zhang Y, Driscoll T, Morgan G, et al. Heat exposure and cardiovascular health outcomes: a systematic review and meta-analysis. *Lancet Planet Health.* (2022) 6:e484–95. doi: 10.1016/S2542-5196(22)00117-6

6. Cai W, Zhang C, Zhang S, Bai Y, Callaghan M, Chang N, et al. The 2022 China report of the Lancet Countdown on health and climate change: leveraging climate actions for healthy ageing. *Lancet Public Health.* (2022) 7:e1073–90. doi: 10.1016/S2468-2667(22)00224-9

7. Chen H, Zhao L, Dong W, Cheng L, Cai W, Yang J, et al. Spatiotemporal variation of mortality burden attributable to heatwaves in China, 1979-2020. *Sci Bull.* (2022) 67:1340–4. doi: 10.1016/j.scib.2022.05.006

8. Kumar A, Singh DP. Heat stroke-related deaths in India: an analysis of natural causes of deaths, associated with the regional heatwave. *J Therm Biol.* (2021) 95:102792. doi: 10.1016/j.jtherbio.2020.102792

9. Wang Y, Bobb JF, Papi B, Wang Y, Kosheleva A, Di Q, et al. Heat stroke admissions during heatwaves in 1,916 US counties for the period from 1999 to 2010 and their effect modifiers. *Environ Health.* (2016) 15:83. doi: 10.1186/s12940-016-0167-3

10. Li Y, Li C, Luo S, He J, Cheng Y, Jin Y. Impacts of extremely high temperature and heatwave on heatstroke in Chongqing, China. *Environ Sci Pollut Res Int.* (2017) 24:8534–40. doi: 10.1007/s11356-017-8457-z

11. Wang Y, Song Q, Du Y, Wang J, Zhou J, Du Z, et al. A random forest model to predict heatstroke occurrence for heatwave in China. *Sci Total Environ.* (2019) 650:3048–53. doi: 10.1016/j.scitotenv.2018.09.369

12. Zhu L, Zhang W, Wong V, Eric Z, Lao L, Lo K, et al. Randomized trial of acupoints herbal patching in *Sanfu* Days for asthma in clinical remission stage. *Clin Transl Med.* (2016) 5:5. doi: 10.1186/s40169-016-0084-7

13. Anderson GB, Bell ML, Peng RD. Methods to calculate the heat index as an exposure metric in environmental health research. *Environ Health Perspect.* (2013) 121:1111–9. doi: 10.1289/ehp.1206273

14. Awasthi A, Vishwakarma K, Pattnayak KC. Retrospection of heatwave and heat index. *Theor Appl Climatol.* (2022) 147:589–604. doi: 10.1007/s00704-021-03854-z

15. Encyclopedia Britannica. *Dew point.* (2024). Available online at: https://www.britannica.com/science/dew-point-temperature (accessed March 23, 2024).

16. Wu X, Wang L, Yao R, Luo M, Wang S, Wang L. Quantitatively evaluating the effect of urbanization on heat waves in China. *Sci Total Environ.* (2020) 731:138857. doi: 10.1016/j.scitotenv.2020.138857

17. Zhou F, Yang D, Lu JY, Li YF, Gao KY, Zhou YJ, et al. Characteristics of clinical studies of summer acupoint herbal patching: a bibliometric analysis. *BMC Complement Altern Med.* (2015) 15:381. doi: 10.1186/s12906-015-0905-z

18. Lu WH, Gu SH, Sun SQ, Zhang CM, Zhu XC. Quantitative analysis of the lagged effects of heat-wave on heatstroke in Ningbo from 2013 to 2019. *J Meteorol Environ.* (2022) 38:106–12. doi: 10.3969/j.issn.1673-503X.2022.01.014

19. Chen K, Horton RM, Bader DA, Lesk C, Jiang L, Jones B, et al. Impact of climate change on heat-related mortality in Jiangsu Province, China. *Environ Pollut.* (2017) 224:317–25. doi: 10.1016/j.envpol.2017.02.011

20. Dong J, Chen Y, Zhang B, Zhou J, Wang S. distributed lag effects in the relationship between daily mean temperature and the incidence of stroke in Lanzhou. *Clim Change Res.* (2017) 13:366–74. doi: 10.12006/j.issn.1673-1719.2016.222

21. Gasparrini A. Distributed lag linear and non-linear models in R: the package dlnm. *J Stat Softw.* (2011) 43:1–20. doi: 10.18637/jss.v043.i08

22. Fujibe F, Matsumoto J, Suzuki H. Spatial and temporal features of heat stroke mortality in Japan and their relation to temperature variations, 1999-2014. *Geogr Rev Japan Ser B.* (2018) 91:17–27. doi: 10.4157/geogrevjapanb.91.17

23. Newbold P. ARIMA model building and the time series analysis approach to forecasting. *J Forecast.* (2010) 2:23–35. doi: 10.1002/for.3980020104

24. Hua Y, Zhao Z, Li R, Chen X, Liu Z, Zhang H. Deep Learning with Long short-term memory for time series prediction. *IEEE Commun Mag.* (2019) 57:114–9. doi: 10.1109/MCOM.2019.1800155

25. Iqbal M, Al-Obeidat F, Maqbool F, Razzaq S, Anwar S, Tubaishat A, et al. COVID-19 patient count prediction using LSTM. *IEEE Trans Comput Soc Syst.* (2021) 8:974–81. doi: 10.1109/TCSS.2021.3056769

26. Rigatti SJ. Random forest. *J Insur Med.* (2017) 47:31–9. doi: 10.17849/insm-47-01-31-39.1

27. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

28. Ljumović M, Klar M. Estimating expected error rates of random forest classifiers: a comparison of cross-validation and bootstrap. In: *2015 4th Mediterranean Conference on Embedded Computing (MECO)* (2015). p. 212–5. doi: 10.1109/MECO.2015.7181905

29. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice.* (2021). Available online at: https://otexts.com/fpp3 (accessed March 23, 2024).

30. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci.* (2021) 7:e623. doi: 10.7717/peerj-cs.623

31. Wainwright HM, Finsterle S, Jung Y, Zhou Q, Birkholzer JT. Making sense of global sensitivity analyses. *Comput Geosci.* (2014) 65:84–94. doi: 10.1016/j.cageo.2013.06.006

32. He C, Kim H, Hashizume M, Lee W, Honda Y, Kim SE, et al. The effects of night-time warming on mortality burden under future climate change scenarios: a modelling study. *Lancet Planet Health.* (2022) 6:e648–57. doi: 10.1016/S2542-5196(22)00139-5

33. Luo M, Wu S, Lau GN, Pei T, Liu Z, Wang X, et al. Anthropogenic forcing has increased the risk of longer-traveling and slower-moving large contiguous heatwaves. *Sci Adv.* (2024) 10:eadl1598. doi: 10.1126/sciadv.adl1598

34. Deng Q, Zhao J, Liu W, Li Y. Heatstroke at home: prediction by thermoregulation modeling. *Build Environ.* (2018) 137:147–56. doi: 10.1016/j.buildenv.2018.04.017

35. Du Y, Jing M, Lu C, Zong J, Wang L, Wang Q. Global population exposure to extreme temperatures and disease burden. *Int J Environ Res Public Health.* (2022) 19:13288. doi: 10.3390/ijerph192013288

36. Lowe D, Ebi KL, Forsberg B. Heatwave early warning systems and adaptation advice to reduce human health consequences of heatwaves. *Int J Environ Res Public Health.* (2011) 8:4623–48. doi: 10.3390/ijerph8124623

37. Giersch GEW, Taylor KM, Caldwell AR, Charkoudian N. Body mass index, but not sex, influences exertional heat stroke risk in young healthy men and women. *Am J Physiol.* (2023) 324:R15–9. doi: 10.1152/ajpregu.00168.2022