

UC Irvine

UC Irvine Previously Published Works

Title

Online expressions of heat risk: social media posting and social vulnerability in California

Permalink

<https://escholarship.org/uc/item/3vc894p5>

Journal

Environmental Research Communications, 8(8)

ISSN

2515-7620

Authors

Garcia, Nicolas Rodriguez

Berrocal, Veronica J

Li, Chen

et al.

Publication Date

2026-08-01

DOI

10.1088/2515-7620/ae9359

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

PAPER • OPEN ACCESS

Online expressions of heat risk: social media posting and social vulnerability in California

To cite this article: Nicolas Rodriguez Garcia *et al* 2026 *Environ. Res. Commun.* **8** 085013

View the [article online](#) for updates and enhancements.

You may also like

- [Online signals of heat vulnerability: revealing city-level response to extreme heat using Google Trends](#)
Jangho Lee
- [Extreme heat in focus: comparing media narratives from the Philippines and Hong Kong](#)
Johanne Rei R Castro and Laurence L Delina
- [Heat alerts and information-seeking behavior: evidence from heat-related internet searches in the United States](#)
Quinn H Adams, Chad W Milando and Gregory A Wellenius

Environmental Research Communications



PAPER

OPEN ACCESS

RECEIVED

30 January 2026

REVISED

24 July 2026

ACCEPTED FOR PUBLICATION

31 July 2026

PUBLISHED

11 August 2026

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Online expressions of heat risk: social media posting and social vulnerability in California

Nicolas Rodriguez Garcia¹ , Veronica J Berrocal², Chen Li³ and Suellen Hopfer^{1,*}

¹ Department of Health, Society & Behavior, Joe C. Wen School of Public Health, University of California, Irvine, CA, United States of America

² Department of Statistics, Donald Bren School of Information and Computer Science, University of California, Irvine, CA, United States of America

³ Department of Computer Science, Donald Bren School of Information and Computer Science, University of California, Irvine, CA, United States of America

* Author to whom any correspondence should be addressed.

E-mail: shopfer@hs.uci.edu

Keywords: social media, heat discourse, Healthy Places Index, heat mitigation, heat interventions

Abstract

Heat waves in California are occurring with increasing frequency, duration, and intensity, largely due to climate change. Furthermore, the burden of extreme heat disproportionately affects various communities both geographically and socioeconomically. As a result, understanding how heat is discussed on social media is critical for informing the design of public health interventions. This study involved a spatiotemporal analysis of heat-related tweets ($N = 24,598$) posted across California between April and October from 2016 to 2022. Poisson regression models were used to examine county-level counts of heat tweets, with population size included as an offset term in the model while logistic regression models estimated the probability that a heat tweet reflected coping or life impact (LI) themes. A different set of logistic regression models were fit to examine the effect of socio-demographic and housing variables on the likelihood of heat tweets expressing positive and negative sentiment, and on the likelihood of expressing coping or LIs. Explanatory variables were derived mostly from the American Community Survey and the Healthy Places Index. Sensitive housing (percentage of households living in an RV, van, or boat) and income predicted posting of heat tweets about coping, whereas income and percentage of active commuters predicted posting of heat tweets about LIs. Percentage of active commuters, tree canopy, and sensitive housing predicted negative heat tweets. Counties with greater proportion of Hispanic population in California also appeared vulnerable, showing a modest tendency toward greater LI and fewer coping-related heat tweets. These findings suggest that public health interventions should prioritize outreach to individuals living in sensitive housing, with an emphasis on indoor heat mitigation strategies. Overall, results demonstrate the value of social media heat discourse as a complementary tool for identifying vulnerable communities and informing targeted extreme heat awareness campaigns.

1. Introduction

Extreme heat is a major cause of weather-related deaths in the United States and a growing public health threat, surpassing mortality from hurricanes, floods, and tornadoes [1]. Increases in temperature due to climate change have become a worldwide public health issue, given that the last 10 years have been the warmest years in recorded history [2]. The World Health Organization has acknowledged heat as an environmental and occupational health hazard that can generate weather-related deaths and worsen underlying illness in the population [3]. In the United States, rapid urbanization contributions to climate change will increase the probability of extreme heat events and risk of human exposure to deadly heat conditions [4–6]. Increasingly frequent and intense bouts of high temperatures are shaped by environmental phenomena such as seasonality, ocean, and land interactions. This complexity in increasing

heat events occurring globally is easier to comprehend by disaggregating factors shaping heat events regionally.

California's diverse geography—including coastal zones, inland deserts, and mountainous regions with wide elevation ranges—makes patterns of extreme heat difficult to characterize [7]. In addition to terrestrial factors, marine heat waves in the Pacific Ocean significantly influence high temperatures across the state, with consequences for population health, wildlife, water resources, and energy demands [8]. Climate change further intensifies these challenges by increasing the frequency, duration, and severity of heat waves. Beyond the traditional dry heat of inland deserts, California is experiencing longer and more humid heat waves in coastal regions, where dense populations and infrastructure are less acclimated to extreme heat [7]. As a result, escalating heat exposure has become a critical concern for organizations and researchers seeking to mitigate impacts on public health, productivity, and ecosystems.

Given the constant risk of heat waves in California, substantial efforts have been made to mitigate heat-related impacts. The state has developed a comprehensive extreme heat resilience plan centered on four strategies: public awareness and early notification, strengthened community services and response, increased resilience of the built environment, and the use of nature-based solutions [9]. Research also has underscored the importance of designing effective, targeted communication to reach vulnerable populations [10, 11]. Accordingly, understanding Californians' knowledge, behaviors, and literacy related to extreme heat resilience is critical. Prior studies using focus groups in San Diego found a need to diversify communication channels, clarify messaging content, and improve campaign coordination [11]. Additional research involving water-scarce households showed that resource vulnerability is associated with heightened perceptions of climate change risks [12]. More recently, community awareness has been examined through social media analyses, particularly text-based platforms such as X (formerly Twitter) [13–15].

This study examines the spatiotemporal distribution of heat-related discourse on social media in California using posts on X between 2016 and 2022. Explanatory variables from the Healthy Places Index (HPI) social vulnerability indicators are assessed for their ability to predict heat-related discourse and sentiment with heat-related tweets focused on coping strategies and life impacts (LIs). These outcomes were selected for their relevance to informing public health communication interventions for extreme heat mitigation.

By disaggregating social media heat discourse, this study identifies vulnerable communities in California that may benefit from targeted heat-risk awareness campaigns. The study framework also provides a transferable approach for analyzing heat-related social media discourse in other regions experiencing extreme heat and socioeconomically diverse populations.

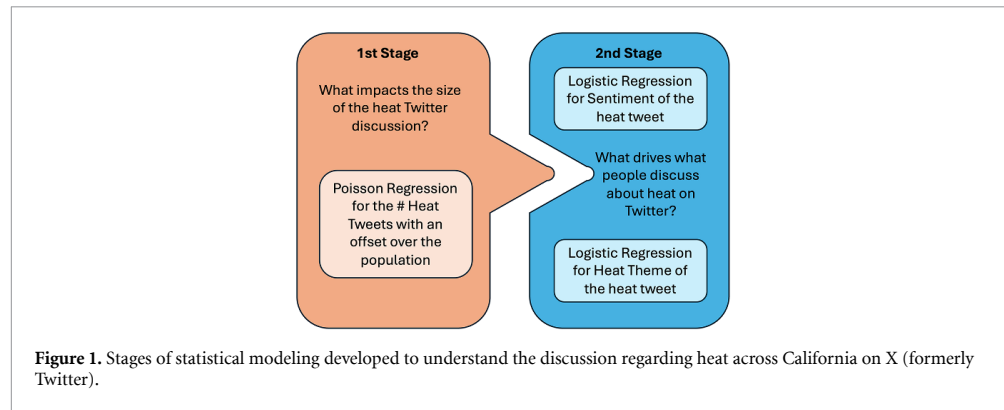
2. Methodology

2.1. Data extraction

Social media data (from Twitter, now X) were obtained from previous work by Sriperumbudur *et al* [13]. The original dataset was composed of a private archive of historical tweets from November 2016 through December 2022 collected through Twitter's public API with a random sample of 5% tweets. Each tweet contains the text, the time it was created, information about the user, and a bounding box indicating where the tweet may have been created geographically. A final set of 24 598 heat-related tweets was used for analyses. These data reflected original tweets and no retweets. Duplicates were removed. Manual inspection of heat tweets for bot activity revealed no suspicious accounts. This involved checking accounts for random or nonsensical text, excessive tweet frequency, multiple accounts posting exact same text, or tweets packed with excessive, irrelevant or made up hashtags.

The filtering process used to identify heat related tweets included four keywords heat wave, heatwave, extreme heat, high temperature. While the initial list used 25 keywords describing heat by place (car, home, office, school), period of time (summer, June/July), adjectives (unbearable, dry), adaptive capacity (AC use), and black and brown out (energy impacts) the terms black and brown out introduced too much noise in the data. The final four keywords yielded the most relevant tweets with the least noise. The complete filtering process is described in Sriperumbudur *et al* [13].

County-level HPI information was retrieved from the HPI API developed by the UCLA Luskin Center. The county shapefiles used to assign each tweet to a county were retrieved from the 2019 TIGER/Line shapefile for California counties on the federal government's open data site. Selected sociodemographic variables were extracted from the American Community Survey from the U.S. Census Bureau website.



2.2. Heat discourse classification

The classification process of the tweet context was performed using large language modeling (LLM) with in-context demonstration (few-shot approximation) of the three predefined heat discourse categories: coping, LI, and other heat related discourse topics. The heat discourse code definition represented the three main factors derived from the climate framework for studying climate change vulnerability [16]. The *coping mechanism* code corresponded to the adaptive capacity factor, the *life impact* code corresponded to the sensitivity factor, and the *other* code encompassed the exposure factor [13]. The *other* code collapsed the remaining heat discourse codes we identified in prior work—e.g. perceiving heat, community action, and linking climate change to heat [13]. The task given to the LLM was a multitask classification with exclusion of the themes, therefore the model had to prioritize the recognition of coping mechanisms (CMs) over the LIs or other heat themes to highlight heat mitigation awareness. The Generative Pre-Trained Transformer 4.1 mini (GPT-4.1 mini) was used, providing inclusion and exclusion criteria and examples for each heat-related code.

In the current validation dataset of 211 heat tweets, two coders independently coded tweets manually for coping, LIs or other and coded exclusively for one code. Annotation was discussed among the two coders establishing agreed upon exclusion and inclusion criteria. Coping tweets included people actively employing strategies to adapt to heat like staying hydrated, spending time in A/C, traveling to beach or pool, talking about cost of A/C, indoor heat mitigation, adapting to heat e.g. getting up early for a run, changing daily diet, etc. Exclusion criteria for CMs included messaging about purchasing a service or food to protect against a heatwave but in fact was an advertisement or expressing a desire for a change in weather, longing for weather in a different place but not executing the action. Life impact inclusion criteria ranged from high temperature impacts on pets, house plants, or personal life—health effects, food safety concerns, leisure activities canceled, negative mental state, businesses closed, work hours changing, work environments changing i.e. in-person meetings canceled or changed to zoom, and impact of heat on personal income and bills. The coding protocol relied on these decision criteria where the category of CMs was given priority over LI when a heat tweet expressed both. Using Krippendorff's α for interrater reliability between two coders, showed an $\alpha = 1.0$ for CM and $\alpha = 1.0$ for LI.

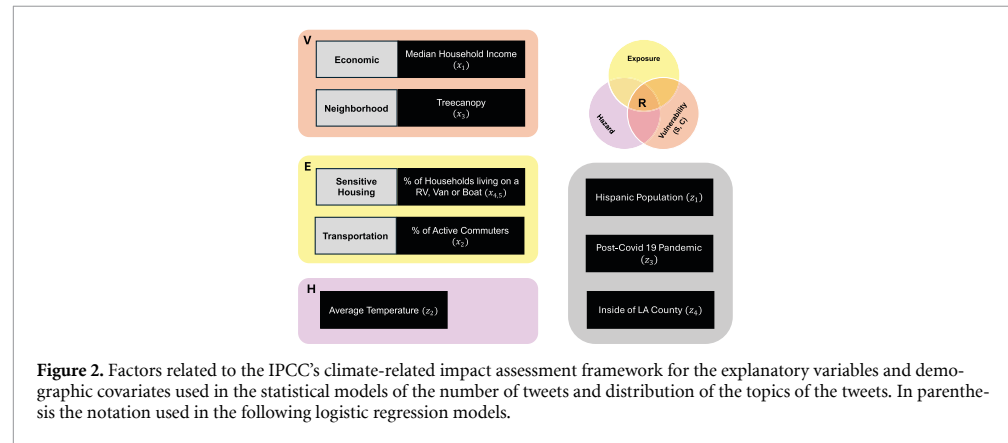
2.3. Statistical analysis

Figure 1 presents the framework for the statistical analysis carried out to address two research questions. We fit two regression models: the first model identified which HPI domain variables affected the monthly number of tweets posted in a county (orange piece). The second model determined which HPI domain variables affected the likelihood of a particular sentiment or heat theme being expressed in heat tweets in a county (blue piece).

The regression models comprised: a model where the outcome variable was the monthly number of tweets in a county (y_{ij}) with a given population of interest (POI). A second model compared categories in the distribution of sentiment and heat discourse codes. The monthly number of tweets in each county was modeled using a Poisson distribution, and it was linked to predictor variables through a Poisson regression model that included as an offset term the size of the POI (see table A in Appendix for goodness of fit indices). We defined the POI to consist of 15- to 55 year-olds as a proxy for the size of the Twitter user population [17, 18]. The equation of the Poisson regression model was as follows.

Table 1. Notation and description of the response variables for the statistical models.

Notation	Description	Unit	Variable
y_{ij}	Number of tweets in county i on month j	Count	Response
λ_{ij}	Rate of tweets per person in county i on month j	Decimal	Rate
P_i	Size of population of interest, consisting of adults between 15 and 55 years old in county i	Individual	Offset
x_{1i}	Median household income for county i	\$10,000 USD	Income
x_{2i}	Average percentage of active commuters (walking, cycling, public transportation) in county i during observation period	%	Active commuters
x_{3i}	Average population-weighted percentage of census tract area with tree canopy in county i	%	Tree canopy
x_{4i}	1 = county i belongs to middle tier of percentage of households living in an RV, van, or boat; 0 = does not satisfy the condition	Binary	Sensitive housing (mid)
x_{5i}	1 = county i belongs to upper tier of percentage of households living in an RV, van, or boat; 0 = does not satisfy the condition	Binary	Sensitive housing (high)
z_{1i}	Size of Hispanic or Latino population in county i	100 000 people	Hispanic
z_{2ij}	Average temperature in county i on month j	°F	Temperature
z_{3j}	1 = month j is post-Covid-19 pandemic (2020–2023); 0 = month j is pre-Covid-19 pandemic (2016–2019)	Binary	COVID-19
z_{4i}	1 = county i is inside Los Angeles County; 0 = county i is outside Los Angeles County	Binary	LA County



$$y_{ij} \sim \text{Poisson}(\lambda_{ij}) \log(\lambda_{ij}) = \log(P_i) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i}. \quad (1)$$

Table 1 presents the descriptions, units, and references for the predictor and outcome variables used in the model. These variables were also used in the logistic regression model shown in equation (2).

Predictor variables were of two types: (a) explanatory variables extracted from the HPI and (b) demographic covariates (size of Hispanic population, tweeting before and after the COVID-19 pandemic, tweeting from within or outside of Los Angeles County) that previous research acknowledged they can affect the outcome variable [19]. The selection of the explanatory variables was based on the Intergovernmental Panel on Climate Change (IPCC) climate-related impact assessment framework that defines community risk as the interaction of vulnerability, exposure, and hazard (figure 2) [20]. The variables selected from the HPI as explanatory variables represented one of the index's domains and could be assigned to one of the factors relevant to determining climate change risks. The first variable was median household income, representing the economic domain; the second was the percentage of active commuters, representing the transportation domain; the third was population-weighted percentage of census tract area with tree canopy, representing the neighborhood domain; and the fourth variable was the percentage of households living in an RV, van, or boat, representing the sensitive housing domain.

In figure 2, the study shows how the eight predictor variables (explanatory variables and demographic covariates) were categorized into one of the three factors used to understand heat risk under

Table 2. Notation and description of the parameter in the logistic regression models used to analyze categories of sentiment and heat discourse themes on social media.

Notation	Description	Unit
p_{ij}^{pos}	Probability that a heat tweet expresses a <i>positive</i> sentiment in county i in month j	%
p_{ij}^{neg}	Probability that a heat tweet expresses a <i>negative</i> sentiment in county i in month j	%
p_{ij}^{CM}	Probability that a heat tweet expresses a <i>coping mechanism</i> in county i during month j	%
p_{ij}^{LI}	Probability that a heat tweet expresses a <i>life impact</i> in county i during month j	%

Table 3. Heat discourse code classification metrics obtained by manually coded batch of heat tweets.

Code	Precision	Recall	F1
Coping	.72	.77	74.3%
Life impacts	.92	.83	87.6%
Other	.89	.89	89.6%

the climate change vulnerability framework developed by the IPCC [20]. The vulnerability factor (V in figure 2) encompasses human adaptive or coping capacity (represented by median household income) and sensitivity (represented by tree canopy in a neighborhood). The exposure factor (E in figure 2) is represented by sensitive housing and percentage of active commuters (walking, cycling, and/or using public transportation). The hazard factor (H in figure 2) is represented by the average temperature with respect to extreme heat.

The gray box represents proxy demographic covariates that capture factors that have been found in previous analyses and literature to be relevant for social media heat discourse: number of Hispanic Twitter users, difference in number of people using Twitter before and after COVID-19, and the number of Los Angeles County Twitter users that was not accounted for by the POI rate [18, 21].

All variables were used in their original scale with one exception: percentage of households living in an RV, van, or boat. This variable was transformed into a categorical variable with three levels ('Low', 'Mid', and 'High'), given its restricted range and right-tailed distribution. We defined and refer to this transformed variable as sensitive housing. The thresholds for each category were determined by the 33rd and 66th percentile split.

The second model was focused on understanding the probability of belonging to a specific sentiment or heat discourse theme compared to the 'other' category. The 'other' category was defined as an aggregate of the remaining heat discourse codes reported in a prior study [13] and included heat warnings, perceived heat, community action, venting, and expressing relief from extreme heat when it cools down for each county in California [13]. Four logistic regression models were developed, two modeling the probability that the heat tweet expressed a given sentiment and two for modeling the probability that the tweet expressed a given heat discourse theme (see table 2 for notation describing c). The general equation of the four logistic regression models was:

$$\text{logit}(p_{ij}^c) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \gamma_1 z_{1i} + \gamma_2 z_{2ij} + \gamma_3 z_{3j} + \gamma_4 z_{4i} \quad (2)$$

$$c \in \{\text{pos, neg, CM, LI}\}$$

3. Results

3.1. Heat discourse code classification

Table 3 shows the LLM performance metrics of classifying heat discourse tweets into one of the three predetermined codes: 'coping', 'life impact', or 'other' (an aggregate of the remaining seven heat discourse codes identified and reported in prior work) [13]. Precision answers, when the model predicts positive, how often is it correct. Precision is defined as true positive (TP)/TP + FP (true positives + false positives). High precision means few false positives. Recall (sensitivity) answers: Of all the true positives, how many did the model find. Recall is defined as TP/ TP + FN (true positives + false negatives). High recall means few misses. F1 score balances precision and recall using their harmonic mean: $F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$.

The LLMs demonstrated good precision and recall across the three heat discourse codes, with F1 ranging from 74% to 90% (see table 3).

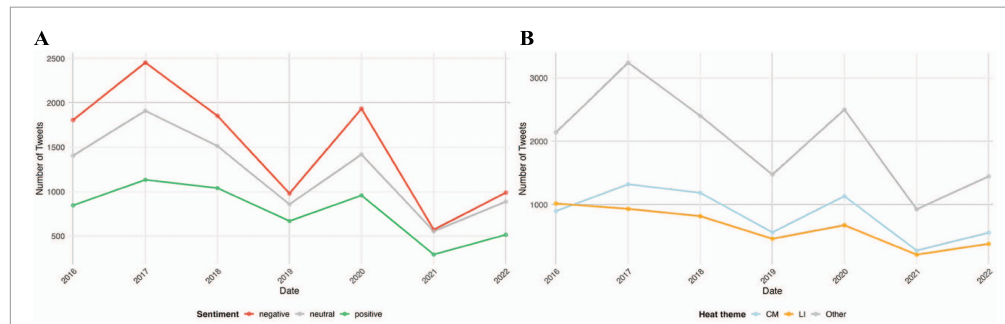


Figure 3. Heat tweets in California on a yearly aggregation between 2016 and 2022. (A) Yearly number of tweets grouped by sentiment: red represents negative tweets, gray represents neutral tweets, and green represents positive tweets. (B) Yearly number of tweets grouped by heat discourse themes. In (B), gray represents other tweets, blue represents coping mechanisms tweets, and yellow represents life impacts tweets. In the analysis, we focused on life impacts and coping; we aggregated other heat discourse categories for the purpose of informing public health interventions.

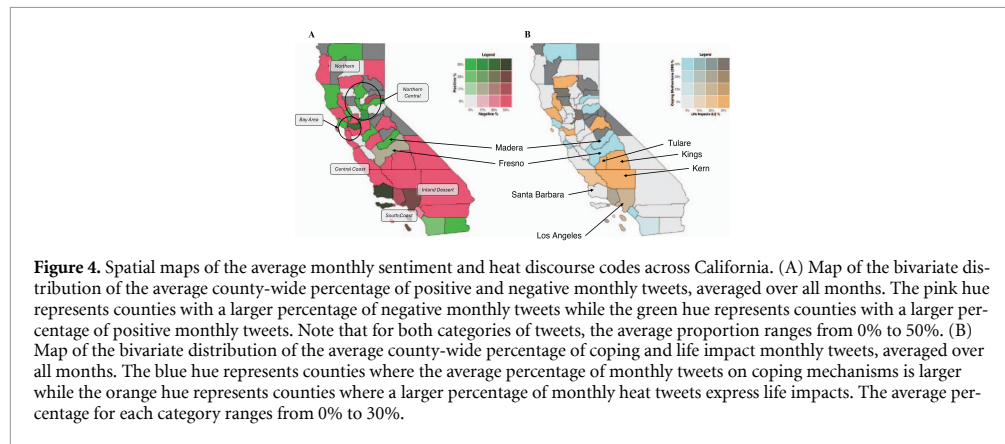


Figure 4. Spatial maps of the average monthly sentiment and heat discourse codes across California. (A) Map of the bivariate distribution of the average county-wide percentage of positive and negative monthly tweets, averaged over all months. The pink hue represents counties with a larger percentage of negative monthly tweets while the green hue represents counties with a larger percentage of positive monthly tweets. Note that for both categories of tweets, the average proportion ranges from 0% to 50%. (B) Map of the bivariate distribution of the average county-wide percentage of coping and life impact monthly tweets, averaged over all months. The blue hue represents counties where the average percentage of monthly tweets on coping mechanisms is larger while the orange hue represents counties where a larger percentage of monthly heat tweets express life impacts. The average percentage for each category ranges from 0% to 30%.

3.2. Descriptive statistics

Figure 3, panel A (sentiment) and panel B (coping, LI, and other heat tweets) show the volume of tweets on a yearly aggregation across the observation period (7 years). The year with the most tweets was 2017, with 5 499 heat tweets, whereas the year with the fewest tweets was 2021, with 1 419 heat tweets. The timeline shows a declining trend in the number of tweets related to heat in California during the 7 year study period. Regarding the sentiment distribution (figure 3, Panel A), the study shows that the negative sentiment had the highest proportion of tweets whereas the positive sentiment had the lowest proportion of tweets. With respect to heat theme distribution, the study found that the ‘other’ heat theme had the highest proportion of tweets while the LIs theme had the lowest proportion of tweets during most of the observation period. The only exception occurred in 2016, when the lowest proportion of tweets belonged to the CMs theme.

The spatial distribution of sentiment and heat discourse codes (coping and LIs) is shown in figure 4. The figure shows the average percentage of monthly tweets expressing a given sentiment and a certain heat code, averaged over the various months of the observation period. For the sentiment distribution (Map A), negative sentiment tweets in red hue dominated overall, with the Central, South Coast, and Inland Deserts regions heavily dominated by negative sentiment. Los Angeles and Santa Barbara counties (two darker counties in Southwest California) expressed a mix of positive and negative heat tweets. Among Northern county regions including the Bay Area and North Central California, there were equally distributed proportion of tweets belonging to each sentiment category (positive, negative, neutral).

On the heat tweet theme distribution (Map B), there was a dominance by the ‘other’ theme throughout California, with some regions tending to have more LIs (orange) or CMs (blue) tweets. In the Central region, there was a dominance by CMs (blue) in counties like Madera and Fresno, whereas LIs (orange) dominated in other counties like Tulare, Kings, and Kern.

Table 4. Estimates (EV) and standard errors (SE) of the regression coefficients in the Poisson regression model for the monthly number of heat tweets in each county, along with the rate ratio to show the relative magnitude.

Regression coefficient	Variable	EV	SE	Rate ratio
β_0	Intercept		8.445×10^{-2}	
β_1	Income	$1.726 \times 10^{-1*}$	3.857×10^{-7}	
β_2	Active commuters	-1.575^a	2.457×10^{-1}	2.07×10^{-1}
β_3	Tree canopy	-3.025^a	1.805×10^{-1}	4.85×10^{-2}
β_4	SH(Mid)	$2.398 \times 10^{-1*}$	1.662×10^{-2}	1.27
β_5	SH (High)	$1.441 \times 10^{-1*}$	2.75×10^{-2}	1.15
γ_1	Hispanic	$-1.552 \times 10^{-2*}$	9.002×10^{-9}	9.84×10^{-1}
γ_2	Temperature	$1.145 \times 10^{-1*}$		1.12
γ_3	COVID-19	$-2.792 \times 10^{-1*}$	1.184×10^{-2}	7.56×10^{-1}
γ_4	LA County	$7.2 \times 10^{-1*}$	3.477×10^{-2}	

^a $p < .0001$.

SH = Sensitive housing (%Households in RV, van or boat).

3.3. Estimates

Table 4 shows the estimates and standard errors of the regression coefficients for the Poisson regression model on tweet counts related to heat, aggregated at a monthly timeframe in each county. The table also reports rate ratios for each regression coefficient, to indicate its relative contribution.

Four coefficients, percentage of active commuters, tree canopy, Hispanic population size, and post-COVID-19 tweets, are estimated to be negative. The other five coefficients, median income household, mid- and high-level sensitive housing, average monthly temperature, and tweeting from Los Angeles County, were by contrast estimated as positive.

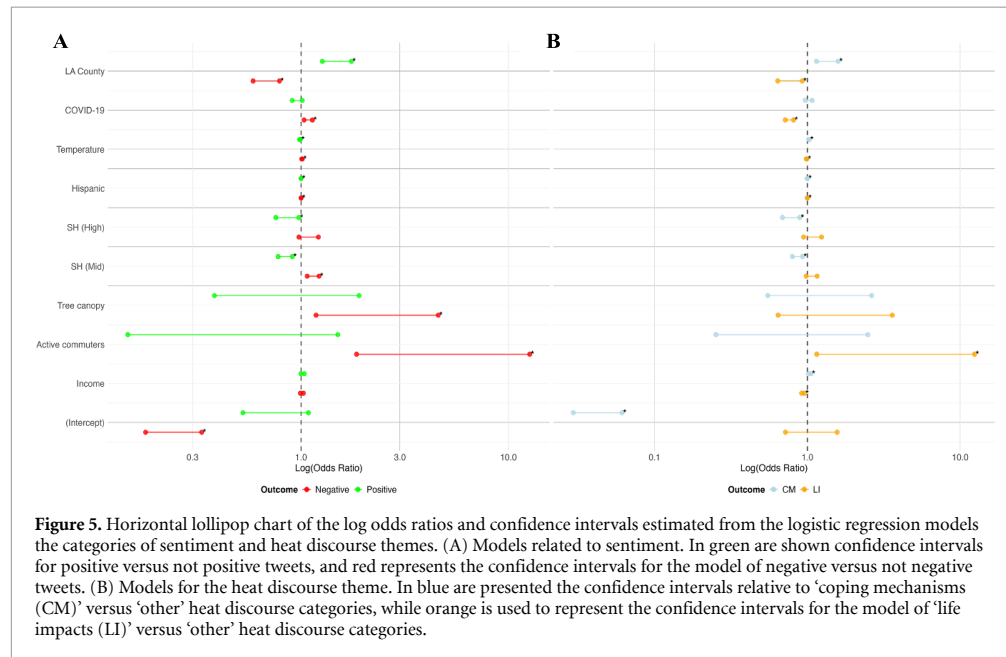
Results from the logistic regression models are expressed in terms of odds ratio for each category of sentiment and heat discourse vs the ‘other’ category which, in both cases, was taken as the reference category (see [appendix](#)).

All the regression coefficients in the Poisson regression model for the number of heat tweets are significantly different from zero. As shown in table 4, temperature is estimated to have a positive association with the number of monthly heat tweets: keeping all the other variables fixed, an increase of one degree in average temperature is associated with an increase in the rate of heat tweets per population size of 12% ($\hat{\gamma}_2 = 1.145 \times 10^{-1}$). On the other hand, the estimate of the coefficient of Hispanic population size indicates a negative association with heat tweets: keeping all the other variables fixed, an increase of 100 000 people in Hispanic population size is associated with a decrease in the rate of heat tweets by 2% ($\hat{\gamma}_1 = -1.552 \times 10^{-2}$). Thus, keeping all the other characteristics the same, counties with a larger Hispanic population have less heat-related tweets. In a comparable manner, we can see that the rate of heat tweets after the COVID-19 pandemic in a given county decreased by 25% compared with the rate in the same county prior to the pandemic (-2.792×10^{-1}), whereas Los Angeles County is estimated to have a rate of heat tweets that is about twice the rate of heat tweets of another county that is comparable to Los Angeles County in all the other aspects considered in the model ($\hat{\gamma}_4 = 7.2 \times 10^{-1}$).

Regarding the explanatory variables extracted from the HPI, table 4 shows that income has a positive association with the number of heat tweets (1.726×10^{-1} per \$10,000 USD increase), indicating that counties with larger median income are expected to have a greater number of heat tweets compared to counties with similar vulnerability, hazard, and exposure factors but a lower median income.

Figure 5 shows a horizontal lollipop chart with confidence intervals for the log odds ratios corresponding to the logistic regression models fitted to the different categories of sentiment and heat tweet codes. The vertical axis indicates the predictor variable whereas the horizontal axis contains the values of the log odds ratios. Statistical significance of a regression coefficient is indicated by the corresponding horizontal bar not crossing the vertical dashed black line.

In the sentiment analysis (Panel A), the logistic regression model for the positive sentiment yields five significant regression coefficients: tweeting from Los Angeles county, average monthly temperature, Hispanic population size, and mid- and high-level sensitive housing. For heat tweets expressing positive sentiment (versus not positive), tweeting from Los Angeles County significantly increases the likelihood of a heat tweet being positive, whereas increases in average monthly temperature, Hispanic population



size, and mid- and high-level sensitive housing are associated with a decrease in the likelihood of a heat tweet being positive.

The logistic regression for the negative versus not negative classification yields eight significant regression coefficients corresponding to the following predictors: tweeting after the COVID-19 pandemic, average monthly temperature, low- and high-level sensitive housing, Hispanic population size, percentage of active commuters, tree canopy, and tweeting from Los Angeles County.

Six of the eight predictors are estimated to have a positive association with the odds of negative tweeting, respectively: tweeting after COVID-19, high-level sensitive housing, Hispanic population size, percentage of active commuters, average monthly temperature, and tree canopy. The other two predictor variables are estimated to have a negative association with negative heat tweets: the mid-level of sensitive housing and tweeting from Los Angeles County.

For models classifying heat discourse themes (Panel B), the logistic regression model for LIs versus other categories indicated significant association between the odds of a LI heat tweet and six predictor variables. Specifically, increases in the percentage of active commuters and Hispanic population size are associated with an increase in the odds of a heat tweet expressing LIs. The other four predictors are associated with a decrease in the odds that a heat tweet expresses LIs; they are: tweeting from Los Angeles County, tweeting after COVID-19, average monthly temperature, and median household income of the county.

The logistic regression model for CMs versus other categories yields seven significant regression coefficients. Three of them are significantly positive and they are the coefficients of: tweeting from Los Angeles County, average monthly temperature, and median household income of the county. The other four predictors are found to have a negative association with the odds that a heat tweet expresses a CM: Hispanic population size and sensitive housing i.e. percentage of households living in an RV, van, or boat (each of the three sensitive housing levels including low-, mid- and high-level).

Table 5 shows the estimated values of the odds ratios obtained from the estimated coefficients of the logistic regression models for the sentiment and heat discourse theme analysis.

The logistic regression models fitted to classify the categories of interest in both sentiment and heat discourse theme indicate significant association with some explanatory variables and demographic covariates. Table 5 shows that three demographic covariates and two explanatory variables have a significant association with the odds that a heat tweet expresses a positive sentiment. Specifically, tweeting from Los Angeles County, is associated with a 48.8% increase in the odds that the heat tweet expresses a positive sentiment compared to tweeting from outside of Los Angeles County. In counties that had a

Table 5. Odds ratios (OR) estimated from the logistic regression models fit to data regarding the sentiment and heat discourse themes in monthly tweets.

Regression coefficient	Variable	OR (Neg)	OR (Pos)	OR (LI)	OR (CM)
β_0	Intercept	0.243 ^a	0.754	1.060	0.042 ^a
β_1	Income	1.009	1.017	0.936 ^a	1.036 ^a
β_2	Active commuters	4.842 ^a	0.468	3.781 ^a	0.792
β_3	Tree canopy	2.327 ^a	0.853	1.521	1.204
β_4	SH ^a (mid)	1.142 ^a	0.838 ^a	1.065	0.861 ^a
β_5	SH ^a (high)	1.087	0.858 ^a	1.082	0.782 ^a
γ_1	Hispanic	1.001 ^a	0.999 ^a	1.001 ^a	0.999 ^a
γ_2	Temperature	1.010 ^a	0.987 ^a	0.986 ^a	1.023 ^a
γ_3	COVID-19	1.083 ^a	0.958	0.762 ^a	1.022
γ_4	LA County	0.679 ^a	1.488 ^a	0.769 ^a	1.352 ^a

^a $p < .0001$.

SH = sensitive housing (%households living in RV, van or boat).

LI = Life impact; CM = Coping mechanism.

larger Hispanic population size, the odds of posting a positive heat tweet is estimated to decreased by 1 % per 100 000 individuals compared to counties that match them on all the other variables. Regarding temperature, keeping all the other variables fixed, the odds of posting a positive heat tweet is estimated to decrease by 2% per each additional Fahrenheit degree. Counties with high-level sensitive housing, that is, with a high percentage of households living in an RV, van, or boat, are associated with 15% lower odds of positive heat tweets than counties with a low-level of sensitive housing. Counties where the sensitive housing variable reaches a mid-level are associated with 17% lower odds of positive heat tweet than counties with a low-level of sensitive housing.

Table 5 shows that four demographic covariates and four explanatory variables have a significant association with the odds of posting a negative heat tweet. Heat tweets from Los Angeles County, are estimated to have 33% lower odds of being negative heat tweet than heat tweets from outside Los Angeles County. Heat tweets after COVID-19, are estimated to have 8.3% greater odds to be negative heat tweets than tweets posted before the pandemic. As for average monthly temperature, keeping all the other variables fixed, heat tweets from counties with higher average temperature are estimated to have 1% greater odds to be negative than tweets from counties with average monthly temperature one degree Fahrenheit lower. Additionally, the odds that a heat tweet is negative are expected to be 0.1 % times larger for every 100 000 additional Hispanic people living in the county. In counties where sensitive housing is mid-level, the odds that a heat tweet is negative is 14.2% greater than in counties with low-level of sensitive housing, keeping all the other variables fixed. Finally, keeping all the other variables fixed, counties with greater tree canopy coverage are associated with 2.3 times greater odds per additional percent coverage than counties with lower tree canopy coverage. The odds of a heat tweet being negative are 4.84 times larger per percentage point increase in counties with greater percentage of active commuters than in counties with lower percentage.

Table 5 shows that three demographic covariates are significant in explaining the odds of posting a coping tweet: tweeting from Los Angeles County, average monthly temperature, and the Hispanic population size. Tweeting from Los Angeles County, is associated with 35% greater odds that the heat tweet expresses a CM compared to heat tweets posted from outside the county. Considering average monthly temperature, the odds that a heat tweet expresses a CM is 2.3% larger for each additional Fahrenheit degree in counties with higher average monthly temperature. As for population size, counties with larger Hispanic population size are estimated to have lower odds of expressing a coping tweet by 1 % per 100 000 people compared to counties with the same values for all other variables but lower number of Hispanic people.

Three explanatory variables are significantly associated with coping heat tweets: high- and mid-level sensitive housing and median household income. Sensitive housing is associated with a decrease in the odds that a heat tweet expresses a CM: the decrease is by a factor of 22% for counties with high level of sensitive housing compared to counties with low level of sensitive housing, and by 14% for mid-level sensitive housing. The median household income variable increases the odds of expressing a coping tweet by % per \$10 000 U.S. Dollar.

Table 5 shows that four demographic covariates (tweeting from Los Angeles County, tweeting after COVID-19, average monthly temperature, and the Hispanic population) and two explanatory variables

(median household income and percentage of active commuters) are significantly associated with the odds that a heat tweet expresses a LI.

Heat tweets from Los Angeles County have a 23% greater odds of expressing a LI than heat tweets posted from outside of Los Angeles County. Heat tweets posted after COVID-19 have 24% lower odds of expressing a LI than tweets posted before COVID-19. On the other hand, tweets from counties with larger average monthly temperature are estimated to have 2% lower odds of expressing a LI heat per degree Fahrenheit than tweets from counties with lower average monthly temperature. Keeping all the other variables fixed, the results indicate that the likelihood of posting a LI tweet increases by 0.1% per 100 000 Hispanic individuals per county.

Finally, with respect to the explanatory variables and other variables fixed, the odds that a heat tweet expresses a LI decreases by 7% per additional \$10 000 USD in median income, whereas the odds increase by a factor of 3.7 as the percentage of active commuter per unit increase.

4. Discussion

Extreme heat events impose disproportionate health burdens across California communities [22]. This study examines the geographical distribution of heat-related social media posts across California to identify potentially vulnerable communities, structural, and individual heat risk factors, and develops a more nuanced understanding of heat wave impacts. Consistent with prior research [22], we found that increases in average monthly temperatures are associated with higher volumes of heat-related tweets. Patterns in social media heat discourse may provide public health officials with an additional tool to monitor heat impacts in near real time, identify vulnerable subpopulations, and gain insight into both indoor and outdoor heat risks.

4.1. Social media as complementary public health surveillance tool

Social media can serve as a proxy for human perception and knowledge about climate change and heat-related topics. Because social media has broad societal reach and allows individuals to express personal opinions, heat-related discourse on these platforms provides valuable insight into the social impacts of heat events [22–24]. Prior studies have shown that social media activity related to heat increases substantially during extreme heat events, underscoring the relevance of analyzing this discourse [25, 26]. Consistent with these findings, our results show that the rate of heat-related tweets per POI increases with rising average monthly temperature across counties. We also found that counties with higher median incomes generate greater volumes of heat-related tweets, likely reflecting greater access to resources and capacity to participate in Twitter conversations. Last, higher levels of sensitive housing (mid and high; the percentage of households living in RV, vans, or boats) were associated with increased rates of heat-related life-impacts tweets by POI, suggesting that populations vulnerable due to housing conditions are more likely to express concerns during periods of high temperatures.

4.2. Sentiment analysis

Previous studies have shown that the sentiment of heat-related tweets varies with average temperature, with more negative sentiment expressed during extreme heat events [22]. Research has also demonstrated that areas with higher HPI scores, which are associated with better quality of life, tend to exhibit a greater proportion of positive sentiment in heat-related tweets in California [13, 14]. In our study, disaggregating HPI domains relevant to heat awareness showed that a 1% increase in the percentage of active commuters is associated with a 4.8-fold increase in the odds of a negative heat tweet. Active commuters are recognized as a heat-sensitive population due to their direct exposure during travel, a finding that aligns with prior research [27]. These results highlight the importance of considering populations that rely on public transportation, walking, or biking in urban planning efforts aimed at implementing structural heat mitigation strategies [28].

4.3. Predictors of coping and life impact heat tweet distributions and suggestions for public health interventions

Study findings highlight how sensitive housing and income might be used to predict coping heat tweets, whereas percentage of active commuters and income are good predictors of LI heat tweets.

Higher levels of sensitive housing—defined as households living in RVs, vans, or boats—are associated with a lower likelihood of posting coping-related heat tweets, with reductions of 14% for mid-level sensitive housing and 22% for high-level sensitive housing. These findings identify Californians living in

RVs, vans, or boats as a critical subgroup for indoor heat mitigation efforts. Such housing types typically have poor insulation, placing residents at risk of life-threatening indoor temperatures during extreme heat events [29]. The limited discussion of heat coping strategies in this population highlights an important public health gap and underscores the need for targeted interventions to improve heat awareness and literacy for these vulnerable subgroups [30].

In counties with a larger Hispanic population, there was a significant 1% decrease in the odds of posting coping heat tweets and a 1% increase in the odds of posting LI tweets. Previous findings confirm that members of the Hispanic population tend to live in urban heat islands and a greater proportion have agricultural occupations, both of which may increase their susceptibility to health impacts due to high temperatures [31]. Something to consider is how the Hispanic population across California suffers from a triple burden involving the multiplicative effects from co-occurring extreme heat, pollution size, and health inequities [32].

Lastly, the vulnerability of the active commuter population, as indicated by fewer expected coping heat tweets, also emerges in the LI tweets. In table 5, the regression coefficient for the active commuters variable indicates an increased odds of a negative heat tweet: the odds of a negative tweet is estimated to be 3.7 times larger for every percentage point increase in the population that uses public transportation, bikes, or walking for their daily commute [27, 28], keeping all the other variables fixed. These results suggest a need for urban planning to consider populations that are vulnerable to extreme heat as a structural mitigation and public health intervention, with the goal of reducing heat-related illnesses.

4.4. Validating the IPCC climate change risk framework

The study applied the IPCC's core climate change risk framework grounded in vulnerability, exposure, and hazard, demonstrating its methodological utility for analyzing heat-related impacts [20, 33]. The findings indicate that both vulnerability (income) and exposure (housing) predicted coping related heat tweets whereas exposure factors (percentage of active commuters and living in a county with a higher proportion of Hispanics) predicted LI heat tweets [34]. Overall, these findings underscore the value of applying an actionable climate risk framework to understand and address heat-related communication and vulnerability.

4.5. Limitations

One limitation of this study is the use of counties as the primary geographic unit of analysis, which may have masked substantial heterogeneity in socioeconomic conditions. In highly populated counties such as Los Angeles and San Francisco, prior research by Yu *et al* [34] demonstrated that complex social phenomena are distributed unevenly across cities and neighborhoods, limiting the interpretability of county-level statistical analyses in California. Notably, the methodological framework proposed in this study is well suited for future research at finer spatial scales (e.g. city-level or census tract analyses), allowing for more granular examination of heat-related discourse and community-level variation.

The study relies on the former Twitter API's 5% sampled stream, which limits representativeness of heat related tweet estimates. Given that heat tweets constitute a small fraction of overall tweets, the effective sample size is substantially reduced, increasing sampling variability including in spatiotemporal analyses across California. These limitations are pronounced in geographically heterogeneous settings. In less populated and lower activity regions like inland and rural California, the sampled stream may yield sparse or zero observations despite ongoing heat related discourse. This introduces potential spatial misclassification where low tweet volume reflects sampling sparsity rather than true absence of engagement, complicating the interpretation of geographic variation.

Uncertainty in the Twitter sampling mechanism further limits inference. If the 5% stream disproportionately captures higher engagement content or more active users, estimates may be biased toward more visible discourse, underrepresenting less amplified experiences during extreme heat events. These issues compound known demographic biases in Twitter use, which overrepresents younger, urban coastal regions, and more digitally connected populations. Accordingly, results should be interpreted as relative indicators of observed social media discourse, rather than comprehensive measures of population level awareness of response to heat exposure.

Relatedly, another limitation of this study concerns bias with respect to representativeness of Twitter users, who are disproportionately younger adults, with reliable internet access, and a greater propensity to publicly express their views [35]. As a result, perceptions of heat waves captured through Twitter data may not adequately represent older adults, rural residents, lower-income populations, or non-English-speaking communities in California, all of whom are underrepresented on this platform [21]. Geographic variation in heat tweets across California reflects spatial differences in platform use and expressive behaviors about heat. Accordingly, geographic variation in this study should be interpreted

as differences in expressed heat-related discourse among Twitter users rather than direct measures of population-level exposure or vulnerability.

Additionally, it is noteworthy that we interpret low signal areas (i.e. low tweet volumes) cautiously. Low twitter use may indicate low expression norms and not necessarily low heat burden. High risk neighborhoods due to high heat burden but whose community has lower socioeconomic status, less digital access, or higher proportion of older adults may appear as silent in the data. In California, heat vulnerability is strongly patterned by housing, income, and occupation (e.g. agricultural or construction workers).

Twitter data reflect real-time public discourse, not population prevalence. Social media data represent traces of communication behavior expressed voluntarily and in the moment or context of the immediate environment typically (e.g. extreme heat event) distinct from survey data. Large scale Twitter data can approximate aggregate patterns even with bias. Aggregate communication signals of heat correlate with real-world increases in temperature.

5. Conclusions

This study demonstrates the value of disaggregating the HPI into heat-relevant domains to understand patterns in social media discourse related to heat coping strategies, LIs, and sentiment. Disaggregation acknowledges heterogeneity across communities and can enhance the effectiveness of targeted heat awareness campaigns. For example, communities with a relatively high proportion of residents living in sensitive housing (e.g. RVs, vans, or boats) require tailored messaging that emphasizes indoor heat coping strategies, given the elevated health risks during extreme heat events. Additionally, identifying counties with a high proportion of active commuters highlights areas where discussions of daily life disruptions are more prevalent, underscoring the importance of heat-conscious urban design and transportation planning. Finally, counties with proportionally larger Hispanic populations have greater engagement in discussions about heat-related LIs but comparatively less discussion about coping strategies, pointing to a critical opportunity for targeted heat awareness and intervention efforts to address their heightened vulnerability.

This research focused on examining how HPI domain-level drivers of population heat adaptive capacity can help explain patterns of heat-related discussion on social media. By bridging social media monitoring with automated language processing tools, the proposed framework establishes a clear pathway for extracting insights into sentiment and thematic content in heat discourse, serving as a proxy for population-level heat awareness. Future research can build on this methodology to identify communities in need of targeted messaging when developing heat warnings, actionable guidance, and community-level mitigation strategies for extreme heat events.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Author contributions

Nicolas Rodriguez Garcia  0009-0002-0797-4396

Data curation (lead), Formal analysis (lead), Investigation (equal), Methodology (lead), Validation (equal), Visualization (lead), Writing – original draft (equal)

Veronica J Berrocal

Investigation (equal), Methodology (equal), Supervision (equal), Writing – review & editing (equal)

Chen Li  0000-0001-8015-6870

Data curation (lead), Resources (lead), Writing – review & editing (supporting)

Suellen Hopfer  0000-0003-3232-9743

Conceptualization (lead), Investigation (lead), Supervision (lead), Validation (equal), Writing – original draft (lead), Writing – review & editing (lead)

Appendix

Table A. Goodness of fit index relative to the Poisson regression model performing a sentiment analysis and modeling the type of heat discourse.

Classification	Outcome variable	Residual deviance	Degrees of freedom	AIC
Sentiment	Positive	1534.4	1054	3760
	Negative	1761	1054	4316
Heat code	CM	1492	1054	3643.6
	LI	1633.1	1054	3788.3

References

- [1] Ebi K L, Hess J J, Stuart-smith R F, Vicedo-cabrera A M, Woodward A and Haines A 2025 Considerations for improving the relevance, use, and robustness of projections of the health risks of climate change *Lancet Planet. Health* **9** e442–7
- [2] World Meteorological Organization 2025 WMO confirms 2024 as warmest year on record at about 1.55 °C above pre-industrial level (available at: [wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level](https://www.wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level)) (Accessed 3 December 2025)
- [3] World Health Organization 2025 Heat and health (available at: www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health) (Accessed 4 December 2025)
- [4] Krayenhoff E S, Moustauou M, Broadbent A M, Gupta V and Georgescu M 2018 Diurnal interaction between urban expansion, climate change and adaptation in US cities *Nat. Clim. Change* **8** 1097
- [5] Mora C et al 2017 Global risk of deadly heat *Nat. Clim. Change* **7** 501
- [6] Trenberth K E, Fasullo J T and Shepherd T G 2015 Attribution of climate extreme events *Nat. Clim. Change* **5** 725–30
- [7] Gershunov A and Guirguis K 2012 California heat waves in the present and future *Geophys. Res. Lett.* **39** L18710
- [8] Scripps Institute of Oceanography. 2025 FAQ: California's marine heat waves (available at: scripps.ucsd.edu/research/climate-change-resources/californias-marine-heatwaves) (Accessed 4 December 2025)
- [9] Newsom G 2022 *Protecting Californians from Extreme Heat: A State Action Plan to Build Community Resilience* (California Natural Resources Agency)
- [10] Dunlap L 2024 *Communicating Heat Risk: A Guide to Inclusive, Effective, and Coordinated Public Information Campaigns* (UCLA Luskin Center for Innovation)
- [11] VanderMolen K, Kimutis N and Hatchett B J 2022 Recommendations for increasing the reach and effectiveness of heat risk education and warning messaging *Int. J. Disaster Risk Reduct.* **82** 103288
- [12] Dobbin K B, Fencl A L, Pierce G, Beresford M, Gonzalez S and Jepson W 2023 Understanding perceived climate risks to household water supply and their implications for adaptation: evidence from California *Clim. Change* **176** 40
- [13] Sriperumbudur G, Fan Y, Liu X, Ma M D, Wang W, Li C and Hopfer S 2025 How Californians tweet about extreme heat events on social media: a health equity perspective *Weather Clim. Soc.* **17** 669–81
- [14] Elkafi S and Tounsi A 2024 Examining public perceptions and concerns about the impact of heatwaves on health outcomes using Twitter data *J. Clim. Chang. Health* **17** 100320
- [15] Kirilenko A P, Molodtsova T and Stepchenkova S O 2015 People as sensors: mass media and local temperature influence climate change discussion on Twitter *Glob. Environ. Change* **30** 92–100
- [16] Manangan A P, Ck U, Sahal S, Schramm P J, Marinucci G D and Brown C L. Climate and health technical report series
- [17] Yildiz D, Munson J, Vitali A, Tinati R and Holland J A 2017 Using Twitter data for demographic research *Demogr. Res.* **37** 1477–514
- [18] Sloan L, Morgan J, Burnap P, Williams M and Preis T 2015 Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data *PLoS One* **10** e0115545
- [19] Maizlish N, Delaney T, Dowling H, Chapman D, Sabo R T, Woolf S, Orndahl C, Hill L and Snellings L 2019 California healthy places index: frames matter *Public Health Rep.* **134** 354–62
- [20] Estoque R C, Ishtiaque A, Parajuli J, Athukorala D, Rabby Y W and Ooba M 2023 Has the IPCC's revised vulnerability concept been well adapted? *Ambio* **52** 376–89
- [21] Mislove A, Lehmann S, Ahn Y Y, Onnela J P and Rosenquist J 2011 Understanding the demographics of Twitter users *Proc. Int. AAAI Conf. Weblogs Soc. Media* **5** 554–7
- [22] Young J C, Arthur R, Spruce M and Williams H T P 2021 Social sensing of heatwaves *Sensors* **21** 3717
- [23] Jung J and Uejio C 2017 Social media responses to heat waves *Int. J. Biometeorol.* **61** 1247–60
- [24] Jung J, Uejio C K, Duclos C and Jordan M 2019 Using web data to improve surveillance for heat sensitive health outcomes *Environ. Health* **18** 59
- [25] Minor K, Moro E and Obradovich N 2025 Worse weather amplifies social media activity *Psychol. Sci.* **36** 35–54
- [26] Jiang J, Murrugara-Ilerena N, Bos M W, Liu Y, Shah N, Neves L and Barbieri F 2022 Sunshine with a chance of smiles: how does weather impact sentiment on social media? *Proc. Int. AAAI Conf. Weblogs Soc. Media* **16** 393–404

- [27] Karner A, Hondula D M and Vanos J K 2015 Heat exposure during non-motorized travel: implications for transportation policy under climate change *J. Transp. Health* **2** 451–9
- [28] Belloc I, Gimenez-nadal J I and Molina J A 2025 Weather conditions and daily commuting *J. Reg. Sci.* **65** 818–42
- [29] Weitz C A 2024 Coping with extreme heat: current exposure and implications for the future *Evol. Med. Public Health* **12** 156–68
- [30] Palinkas L A *et al* 2022 Vulnerable, resilient, or both? A qualitative study of adaptation resources and behaviors to heat waves and health outcomes of low-income residents of urban heat islands *Int. J. Environ. Res. Public Health* **19** 11090
- [31] Lanza K, Jones J, Acuna F, Coudert M, Bixler R P, Kamath H and Niyogi D 2023 Heat vulnerability of Latino and Black residents in a low-income community and their recommended adaptation strategies: a qualitative study *Urban Clim.* **51** 101656
- [32] UCLA Latino Policy & Politics Institute. 2025 Latino neighborhoods across California face ‘triple burden’ of extreme heat, pollution, and health inequities (available at: latino.ucla.edu/press/climate-health-disparities-ca-latino-neighborhoods) (Accessed 19 December 2025)
- [33] Intergovernmental Panel on Climate Change 2021 Summary for policymakers
- [34] Yu J, Castellani K, Forsysinski K, Gustafson P, Lu J, Peterson E, Tran M, Yao A, Zhao J and Brauer M 2021 Geospatial indicators of exposure, sensitivity, and adaptive capacity to assess neighbourhood variation in vulnerability to climate change-related health hazards *Environ. Health* **20** 31
- [35] Steinert-threlkeld Z C 2018 *Twitter as Data* (Cambridge University Press)