Supplementary Information for

Missed Reports of Extreme Heat Events Exacerbate Vulnerability and Losses

Shiyu Liang[†], Yiming Yang[†], Ye Zheng[†], Haozhi Pan[†*], Chaofan Sun[†*], Shenghao Yan, Hanbo Huang, Hao Zheng, Temitope Sogbanmu, Alecia Bennett-Bryan, Joe Mulligan, Zahra Kalantari[*], Fengping Wang, Zhifu Mi, Enhui Liao[*]

[*]Correspondence to: panhaozhi@sjtu.edu.cn; scf024@sjtu.edu.cn; zahrak@kth.se; ehliao@sjtu.edu.cn

[†]These authors contributed equally.

Contents

S1 Meteorological Data and Heatwave Detection

S1.1 ERA5 Data Acquisition and Processing

Hourly near-surface (2-m) air temperature fields were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis product via the Copernicus Climate Data Store (CDS). The dataset spans the period 1st January 1985 to 31st December 2024 at a native spatial resolution of $0.25^\circ \times 0.25^\circ$.

Data retrieval was conducted using the CDS Application Programming Interface (API). For each grid cell, hourly temperature values were aggregated to daily maximum temperature ($T_{\max}$) as:

$$T_{\max}(t_d) = \max_{h \in d} T_{2m}(t_h), \tag{S1}$$

where t_d denotes a calendar day and t_h denotes hourly timestamps within that day.

All preprocessing procedures were applied uniformly across grid cells. Missing values provided by ERA5 were retained as `NaN` and excluded from climatological and percentile calculations. Grid cells containing no valid data during the baseline period were excluded from further analysis. Temporal consistency checks were conducted to ensure uninterrupted daily sequences prior to climatological construction.

The 2015–2024 period was defined as the contemporary analysis window for historical heatwave detection.

S1.2 CMIP6 Future Climate Simulations

Future climate projections were derived from the Coupled Model Intercomparison Project Phase 6 (CMIP6) archive, using output from the Community Earth System Model version 2 (CESM2) under the SSP5–8.5 scenario [1, 2]. The CESM2 ensemble member r4i1p1f1 was used. The SSP5–8.5 scenario represents a fossil-fuel-intensive high-emissions pathway in which radiative forcing reaches approximately 8.5 W m^{-2} by 2100.

Daily maximum temperature fields covering 2015–2100 were extracted. Model outputs were regridded to a $1^\circ \times 1^\circ$ grid using bilinear interpolation prior to climatological and threshold computation.

To ensure baseline consistency between ERA5 and CMIP6 data, for future projections, percentile thresholds were computed using the CESM2 historical simulation over the 1985–2014 period, aligned temporally with the ERA5 baseline to ensure methodological consistency between historical and projected heatwave metrics (1985–2014 for ERA5 historical and corresponding model baseline years for CESM2 historical simulations). No bias correction was applied, as the analysis focuses on percentile-based exceedance metrics rather than absolute temperature magnitudes.

S1.3 Climatology and Percentile Threshold Construction

Heatwave detection was based on a seasonally varying percentile framework constructed from a fixed climatological baseline (1985–2014). All calculations were performed independently for each grid cell.

For each calendar day $d \in \{1, \ldots, 366\}$, daily maximum temperature values from all baseline years were pooled using an 11-day moving window centered on d. Let $\mathrm{doy}(t)$ denote the day-of-year corresponding to time t, and let $T(t)$ represent the daily maximum temperature. The pooled temperature set for day d was defined as

$$\mathcal{T}(d) = \{T(t) \mid |\mathrm{doy}(t) - d| \leq 5\}\,. \tag{S2}$$

The daily climatological mean and the 90th percentile threshold were then computed as

$$T_{\mathrm{clim}}(d) = \mathrm{mean}\left[\mathcal{T}(d)\right], \tag{S3}$$

$$T_{90}(d) = \mathrm{P90}\left[\mathcal{T}(d)\right], \tag{S4}$$

where P90 denotes the empirical 90th percentile of the pooled distribution.

To reduce discontinuities arising from sampling variability and ensure smooth seasonal transitions, both $T_{\mathrm{clim}}(d)$ and $T_{90}(d)$ were further smoothed using a 31-day centered running mean:

$$\tilde{T}(d) = \frac{1}{31} \sum_{k=-15}^{15} T(d+k), \tag{S5}$$

with circular indexing applied at year boundaries to preserve continuity of the seasonal cycle.

Leap years were handled by linearly interpolating February 29 as the arithmetic mean of February 28 and March 1 in both climatology and threshold fields. This ensured a consistent 366-day structure for all baseline years and prevented distortions in the sliding-window calculations. All computations were implemented in MATLAB (R2023a) using vectorized spatial operations. Grid cells lacking sufficient valid observations during the baseline period were excluded from climatological and percentile estimation.

S1.4 Heatwave Event Identification

Heatwave detection was conducted following the percentile-based framework of Hobday et al. [3], applied independently to each spatial grid cell. Daily maximum temperature (T_{max}) was compared against a seasonally varying 90th percentile threshold $T_{90}(d)$ derived from the fixed climatological baseline period 1985–2014 (Section S1.3).

A heatwave event was defined as a period of at least five consecutive days satisfying

$$T_{\max}(t) > T_{90}(\mathrm{doy}(t)), \tag{S6}$$

where $\mathrm{doy}(t)$ denotes the day-of-year corresponding to time t. The five-day duration criterion ensures that detected events represent persistent extreme thermal anomalies rather than transient fluctuations.

If two exceedance periods were separated by an interval of two days or fewer below threshold, they were merged into a single continuous event. This merging rule accounts for short-lived interruptions within sustained heat episodes.

For each identified event, the following metrics were calculated: onset date, termination date, duration (number of exceedance days), maximum intensity, mean intensity, and cumulative intensity. Event intensity was defined relative to the climatological mean temperature:

$$I(t) = T_{\max}(t) - T_{\mathrm{clim}}(\mathrm{doy}(t)). \tag{S7}$$

Cumulative intensity was computed as

$$I_{\mathrm{cum}} = \sum_{t \in \mathrm{event}} I(t). \tag{S8}$$

Detection for the contemporary period (2015–2024) was performed using ERA5 daily maximum temperatures at 0.25° spatial resolution. Future heatwave characteristics (2015–2100) were derived from CESM2 simulations under SSP5–8.5, regridded to 1° resolution prior to analysis. In both historical and future cases, thresholds were constructed using the identical 1985–2014 baseline to ensure methodological consistency and comparability across periods.

National-level metrics were computed using area-weighted aggregation across all grid cells within each country. A country was considered affected during a given period if at least one grid cell satisfied the detection criteria; this classification was used for exposure accounting and does not replace continuous national metrics.

All computations were implemented in MATLAB (R2023a) using vectorized operations for spatially gridded datasets.

S1.5 Sensitivity to Alternative Heatwave Definitions

To evaluate the robustness of heatwave detection to definitional choices, two alternative percentile–duration combinations were assessed in addition to the baseline definition (≥ 5 consecutive days exceeding the 90th percentile threshold):

1. ≥ 2 consecutive days exceeding the 95th percentile.
2. ≥ 4 consecutive days exceeding the 97.5th percentile.

All percentile thresholds were derived using the same 1985–2014 climatological baseline and seasonal smoothing procedures described in Section S1.3. Detection was applied consistently to both the contemporary (2015–2024) and projected future (2050–2100) periods.

As expected, increasing the percentile threshold reduced the absolute frequency and duration of detected events, whereas shorter duration requirements increased event counts. However, the spatial distribution of heatwave occurrence and national-level aggregation metrics remained highly consistent across definitions. Country-level rankings of exposure exhibited minimal variation, indicating that the principal findings of this study—particularly those related to relative under-reporting—are not sensitive to reasonable modifications of the heatwave definition.

S2 Multilingual Extreme Heat Event Extraction Framework

This section outlines the construction of the multilingual news dataset that underpins event-level extraction and subsequent spatial analysis. The objective of this stage is to transform raw web-crawled news content into a structured and geocoded extreme heat event database through a deterministic four-stage processing framework.

The workflow consists of: (1) multilingual news collection and preprocessing; (2) large-scale coarse screening to identify candidate heat-related articles; (3) fine-grained structured extraction of event attributes; and (4) geographic mapping with uncertainty control. The overall pipeline is summarized in Algorithm S2, while detailed technical specifications for each stage are provided in the following sections.

Algorithm 1: Multilingual Extreme Heat Event Extraction Framework

Input: Web crawler sources S spanning 184 countries and 40 languages
Output: Structured and geocoded heat event database D_{final}

Stage 1: Multilingual News Collection
$D_{\text{raw}} \leftarrow \text{Crawl}(S)$
$D_{\text{clean}} \leftarrow \text{NormalizeAndDeduplicate}(D_{\text{raw}})$

Stage 2: Multilingual Coarse Screening
$D_{\text{candidate}} \leftarrow \varnothing$
for each article $a \in D_{\text{clean}}$ **do**
    $decision \leftarrow \text{LLM\_Screen}(a)$
    if $decision = \text{Positive}$ **then**
        $D_{\text{candidate}} \leftarrow D_{\text{candidate}} \cup \{a\}$
    end if
end for

Human-in-the-loop calibration

Stage 3: Structured Event Extraction

$D_{\text{validated}} \leftarrow \varnothing$
for each article $a \in D_{\text{candidate}}$ **do**
    $E \leftarrow \text{TranslateAndExtract}(a)$
    if $E \neq \varnothing$ **then**
        $D_{\text{validated}} \leftarrow D_{\text{validated}} \cup E$
    end if
end for
Human verification and validation

Stage 4: Geographical Mapping

$D_{\text{final}} \leftarrow \varnothing$
for each structured record $e \in D_{\text{validated}}$ **do**
    $g \leftarrow \text{Geocode}(e)$
    $D_{\text{final}} \leftarrow D_{\text{final}} \cup \{g\}$
end for
Spatial auditing and uncertainty assignment
return D_{final}

S2.1 Multilingual News Collection

The first stage constructs a globally representative multilingual corpus. News articles were retrieved from 239 validated national-level outlets spanning 184 countries and 40 languages. Automated crawling produced a large raw dataset covering the period from 1 January 2015 to 31 December 2024. Text normalization and two-stage deduplication were applied to remove exact and near-duplicate republications while preserving independent reporting. This stage prioritizes recall and geographic inclusiveness before downstream semantic filtering is applied.

S2.2 Multilingual Coarse Screening

The second stage performs deterministic coarse classification to identify candidate heat-related articles. Each cleaned article is evaluated using a fixed LLM-based reasoning protocol that assesses the presence of predefined temperature, impact, and location signals. Articles satisfying the decision rule are retained as candidates for structured extraction. A human-in-the-loop calibration procedure stabilizes classification boundaries across languages and mitigates systematic misclassification, ensuring consistency prior to full-corpus processing.

S2.3 Fine-Grained Event Structuring

The third stage converts screened candidate articles into structured event-level records. Non-English articles are translated into English under deterministic decoding parameters. Structured extraction prompts then generate JSON-formatted records containing temporal attributes, intensity references, impact categories, and spatial descriptions. Outputs are validated programmatically and subjected to targeted human verification to ensure semantic fidelity and structural correctness. This stage transforms unstructured textual narratives into machine-readable event representations.

S2.4 Geographical Mapping

The final stage assigns geographic coordinates to structured event records. Spatial descriptions are resolved using a deterministic LLM-assisted geocoding framework supported by a global administrative gazetteer. Ambiguous or low-confidence cases undergo human spatial auditing. Administrative resolution level is explicitly recorded, and spatial uncertainty is quantified according to geographic scale.

The resulting database comprises 210,525 validated extreme heat event records with standardized temporal, thematic, and spatial attributes. This structured and geocoded dataset forms the foundation for subsequent spatial reporting analyses.

S3 Global Multilingual Heat-Related News Collection

In Fig. 1, Step 1 establishes the comprehensive multilingual news corpus that underpins downstream screening, validation, and structured event extraction. The primary objective of this stage is to maximize recall of heat-related reporting across countries and linguistic contexts prior to event-level confirmation. This design ensures that potentially relevant reporting is captured before precision filtering and validation are applied in subsequent stages.

S3.1 Source Selection Criteria

Definition of Official News Source.

News media sources were identified through a multilingual crawler framework covering 184 countries and 40 languages. The overall identification and screening procedure followed a structured PRISMA-based workflow, as illustrated in Supplementary Fig. S1. Initial identification produced a comprehensive pool of candidate outlets. Prior to formal screening, duplicate or inactive domains were removed to ensure that only operational and structurally stable sources were retained for evaluation.

The remaining outlets were screened for editorial influence and audience reach. To be classified as mainstream media, sources were required to satisfy all predefined criteria

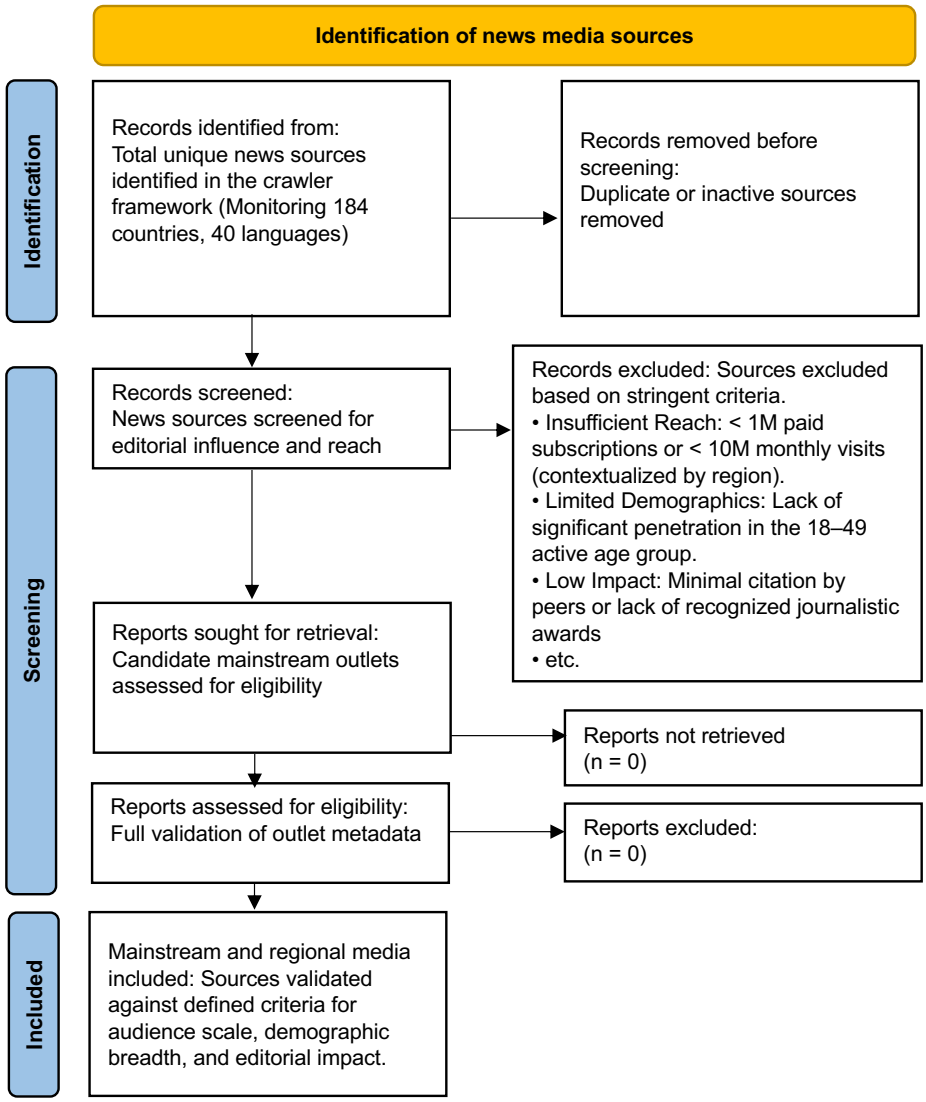

Fig. S1 PRISMA flow diagram of the news media source selection process. The diagram outlines the systematic screening workflow. Following initial identification, news sources were screened against editorial influence and reach criteria to define "mainstream" status. Sources classified as mainstream were required to demonstrate: **(1)** substantial audience reach ($>$1 million paid subscriptions or $>$10 million monthly visits, adjusted for regional population density); **(2)** demographic breadth (significant coverage of the 18–49 age group); **(3)** high editorial impact (frequent citation or syndication by peer media); and **(4)** professional recognition (receipt of major international or regional journalism awards).

shown in Supplementary Fig. S1. Specifically, outlets had to demonstrate substantial audience reach, defined as more than 1 million paid subscriptions or more than 10 million monthly visits (contextualized by region); significant demographic breadth, particularly measurable penetration within the active 18–49 age group; high editorial impact, reflected by frequent citation or syndication by peer media organizations; and professional recognition, indicated by receipt of major international or regional journalism awards. Sources that did not meet these criteria were excluded at the screening stage.

Candidate mainstream outlets were subsequently assessed for eligibility through full validation of outlet metadata. As shown in Supplementary Fig. S1, no records were unretrieved ($n = 0$) and no records were excluded following eligibility assessment ($n = 0$). This final validation step ensured consistency between documented editorial attributes and observable domain-level characteristics.

The resulting validated corpus comprises 239 unique news websites spanning 184 countries and territories. Among these, 8 outlets satisfied the composite mainstream criteria described above, while the remaining 231 were categorized as regional media sources. A summary of the aggregated counts is provided in Table S1, and the complete list of validated news outlets is presented in Table S2.

Table S1 Summary of Collected News Media Sources.

Metric	Count	Description
Total News Sources	239	Unique news websites included in the dataset.
Identified Mainstream Media	8	Top-tier mainstream outlets meeting composite audience and impact thresholds.
Regional/Niche Media	231	Regional, local, or specialized news outlets.

Note: The table presents the aggregate counts of the news dataset. Out of 239 total sources, 8 were identified as top-tier mainstream media outlets. This classification was determined based on a composite threshold of representative metrics such as audience size (>10M monthly visits or >1M subscribers), demographic reach (18–49 age group), and editorial impact (awards and citations).

Table S2: Complete List of News Media Sources.

Media Name	Source	Type
People's Daily Online	people.com.cn	Mainstream
People's Daily Online	people.cn	Mainstream
The Washington Post	washingtonpost.com	Mainstream
Associated Press	apnews.com	Mainstream
The Times of India	indiatimes.com	Mainstream
Corriere della Sera	corriere.it	Mainstream
The Guardian	theguardian.com	Mainstream
BBC	bbc.com	Mainstream

Table S2 (continued)

Media Name	Source	Type
Milliyet	milliyet.com.tr	Regional
Yle (Finnish Broadcasting Company)	yle.fi	Regional
Tengrinews	tengrinews.kz	Regional
Gulf Times	gulf-times.com	Regional
Die Presse	diepresse.com	Regional
Digi24	digi24.ro	Regional
Hoy (Paraguay)	hoy.com.py	Regional
HotNews	hotnews.ro	Regional
NRK (Norwegian Broadcasting Corporation)	nrk.no	Regional
Klix.ba	klix.ba	Regional
Hürriyet	hurriyet.com.tr	Regional
Borneo Bulletin	borneobulletin.com.bn	Regional
Adresseavisen (Adressa)	adressa.no	Regional
CTV News	ctvnews.ca	Regional
Montevideo Portal	montevideo.com.uy	Regional
ABC (Australian Broadcasting Corporation)	abc.net.au	Regional
VnExpress	vnexpress.net	Regional
CBC (Caribbean Broadcasting Corporation)	cbc.bb	Regional
Detik	detik.com	Regional
La Nación (Paraguay)	lanacion.com.py	Regional
Al Riyadh	alriyadh.com	Regional
ARY News	arynews.tv	Regional
The VOICE (St. Lucia)	thevoiceslu.com	Regional
Telegraf	telegraf.rs	Regional
Hoy (Dominican Republic)	hoy.com.do	Regional
Kyiv Post	kyivpost.com	Regional
Euronews Albania	euronews.al	Regional
TRT Haber	trthaber.com	Regional
Emol (El Mercurio Online)	emol.com	Regional
Večernji list	vecernji.ba	Regional
RNZ (Radio New Zealand)	rnz.co.nz	Regional
bdnews24.com	bdnews24.com	Regional
Radio Liechtenstein	radio.li	Regional
Los Tiempos	lostiempos.com	Regional
To Vima	tovima.gr	Regional
Samoa Observer	samoaobserver.ws	Regional
HINA (Croatian News Agency)	hina.hr	Regional
Die Zeit	zeit.de	Regional
The Irish Times	irishtimes.com	Regional

Table S2 (continued)

Media Name	Source	Type
Radio Free Asia	rfa.org	Regional
RTBF	rtbf.be	Regional
NDTV (New Delhi Television)	ndtv.com	Regional
Thanh Niên	thanhnien.vn	Regional
El Mercurio	elmercurio.com	Regional
Yomiuri Shimbun	yomiuri.co.jp	Regional
Mehr News Agency	mehrnews.com	Regional
Alsat	alsat.mk	Regional
The Korea Times	koreatimes.co.kr	Regional
Baghdad Today	baghdadtoday.news	Regional
Postimees	postimees.ee	Regional
Politiken	politiken.dk	Regional
The Himalayan Times	thehimalayantimes.com	Regional
The Business Times	businesstimes.com.sg	Regional
20 Minutos	20minutos.es	Regional
Népszava	nepszava.hu	Regional
Inquirer.net	inquirer.net	Regional
Rádio Moçambique	rm.co.mz	Regional
Channels Television	channelstv.com	Regional
Le Temps	letemps.ch	Regional
Blic	blic.rs	Regional
Proceso	proceso.com.mx	Regional
La Jornada	jornada.com.mx	Regional
France TV Info	francetvinfo.fr	Regional
The Malta Independent	independent.com.mt	Regional
Kuwait Times	kuwaittimes.com	Regional
Tuôi Trè	tuoitre.vn	Regional
Cyprus Mail	cyprus-mail.com	Regional
Khaama Press	khaama.com	Regional
Guyana Chronicle	guyanachronicle.com	Regional
SoyChile	soychile.cl	Regional
Neue Zürcher Zeitung	nzz.ch	Regional
El Periòdic d'Andorra	elperiodic.ad	Regional
Times of Oman	timesofoman.com	Regional
La Presse (Tunisia)	lapresse.tn	Regional
Oman Observer	omanobserver.om	Regional
Otago Daily Times	odt.co.nz	Regional
San Marino RTV	sanmarinortv.sm	Regional
ATB (Red ATB)	atb.com.bo	Regional
Petra (Jordan News Agency)	petra.gov.jo	Regional
Diario Co Latino	diariocolatino.com	Regional
The Peninsula	thepeninsulaqatar.com	Regional

Table S2 (continued)

Media Name	Source	Type
TV21	tv21.tv	Regional
Daily Mirror	dailymirror.lk	Regional
Europa Press	europapress.es	Regional
WAM (Emirates News Agency)	wam.ae	Regional
ABC Color	abc.com.py	Regional
La Prensa (Panama)	prensa.com	Regional
BelTA (Belarusian Telegraph Agency)	belta.by	Regional
Pobjeda	pobjeda.me	Regional
El Comercio	elcomercio.pe	Regional
The Nation (Thailand)	nationthailand.com	Regional
AZERTAC	azertag.az	Regional
The Korea Herald	koreaherald.com	Regional
CNA (Channel NewsAsia)	channelnewsasia.com	Regional
Kazinform	inform.kz	Regional
Trinidad Express	trinidadexpress.com	Regional
News First	newsfirst.lk	Regional
The Villager	thevillager.com.na	Regional
Newsday	newsday.co.tt	Regional
Stabroek News	stabroeknews.com	Regional
El Espectador	elespectador.com	Regional
FBC News	fbcnews.com.fj	Regional
El Siglo	elsiglo.com.pa	Regional
Federalna.ba	federalna.ba	Regional
Business Today	businesstoday.in	Regional
Dawn	dawn.com	Regional
Post-Courier	postcourier.com.pg	Regional
De Standaard	standaard.be	Regional
EMTV	emtv.com.pg	Regional
Jamaica Observer	jamaicaobserver.com	Regional
Thai PBS	thaipbs.or.th	Regional
MBC (Malawi Broadcasting Corporation)	mbc.mw	Regional
Teledoce	teledoce.com	Regional
L'Express (Mauritius)	lexpress.mu	Regional
RPP (Radio Programas del Perú)	rpp.pe	Regional
Walta	waltainfo.com	Regional
SONNA (Somali National News Agency)	sonna.so	Regional
Eye Radio	eyeradio.org	Regional
RTL	rtl.lu	Regional
Nation Media Group	nation.africa	Regional
Prensa Libre	prensalibre.com	Regional

Table S2 (continued)

Media Name	Source	Type
The Edition	edition.mv	Regional
Benin Web TV	beninwebtv.com	Regional
Malaysiakini	malaysiakini.com	Regional
Andorra Difusió	andorradifusio.ad	Regional
The New Dawn	thenewdawnliberia.com	Regional
The Kathmandu Post	kathmandupost.com	Regional
Fiji Sun	fijisun.com.fj	Regional
Island Times	islandtimes.org	Regional
1+1	1plus1.ua	Regional
Tengri Travel	tengritravel.kz	Regional
GMA Network	gmanetwork.com	Regional
El Telégrafo	eltelegrafo.com.ec	Regional
News.mn	news.mn	Regional
Seychelles News Agency	seychellesnewsagency.com	Regional
Skai	skai.gr	Regional
Lebanon24	lebanon24.com	Regional
La Tercera	latercera.com	Regional
RTÉ	rte.ie	Regional
Kun.uz	kun.uz	Regional
O Democrata	odemocratagb.com	Regional
Tabula	tabula.ge	Regional
Guinéenews	guineenews.org	Regional
Armenpress	armenpress.am	Regional
Tchad Infos	tchadinfos.com	Regional
India Today North East	indiatodayne.in	Regional
El Diario de Hoy	elsalvador.com	Regional
Diena	diena.lv	Regional
Guioteca	guioteca.com	Regional
Iwacu	iwacu-burundi.org	Regional
Le Journal Du Tchad	journaldutchad.com	Regional
Index.hu	index.hu	Regional
Bangkok Post	bangkokpost.com	Regional
Maliweb	maliweb.net	Regional
MaltaToday	maltatoday.com.mt	Regional
Le Citoyen	lecitoyenrouynlasarre.com	Regional
Alsumaria TV	alsumaria.tv	Regional
Kompas.com	kompas.com	Regional
Jeune Afrique	jeuneafrique.com	Regional
Punch	punchng.com	Regional
The Gleaner	jamaica-gleaner.com	Regional
TN (Todo Noticias)	tn.com.ar	Regional
AlterPresse	alterpresse.org	Regional

Table S2 (continued)

Media Name	Source	Type
Delano	delano.lu	Regional
Ennahar	ennaharonline.com	Regional
Ariana News	ariananews.af	Regional
Nova24TV	nova24tv.si	Regional
Proceso	proceso.hn	Regional
SME	sme.sk	Regional
Vanuatu Daily Post	dailypost.vu	Regional
Khmer Times	khmertimeskh.com	Regional
Kaieteur News	kaieteurnewsonline.com	Regional
Montsame	montsame.mn	Regional
Government of Samoa	samoagovt.ws	Regional
Malawi24	malawi24.com	Regional
Khovar	khovar.tj	Regional
Radio Tamazuj	radiotamazuj.org	Regional
Le Monde	lemonde.fr	Regional
Libya Herald	libyaherald.com	Regional
News24	news24.com	Regional
BBS (Bhutan Broadcasting Service)	bbs.bt	Regional
MTV (Lebanon)	mtv.com.lb	Regional
Jutarnji list	jutarnji.hr	Regional
Citi Newsroom	citinewsroom.com	Regional
La República	larepublica.pe	Regional
Shabait	shabait.com	Regional
Matangi Tonga	matangitonga.to	Regional
El País (Uruguay)	elpais.com.uy	Regional
Actualite.cd	actualite.cd	Regional
The Fiji Times	fijitimes.com.fj	Regional
Seychelles Nation	nation.sc	Regional
RÚV	ruv.is	Regional
Central de Noticias	centralnoticias.gob.do	Regional
Le Calame	lecalame.info	Regional
InfoSalus	infosalus.com	Regional
The Libya Observer	libyaobserver.ly	Regional
O País (Mozambique)	opais.co.mz	Regional
L'Express (Madagascar)	lexpress.mg	Regional
Solomon Star	solomonstarnews.com	Regional
Expresso das Ilhas	expressodasilhas.cv	Regional
24.kg	24.kg	Regional
Gabon Review	gabonreview.com	Regional
NationNews (Barbados)	nationnews.com	Regional
Daily News (Tanzania)	dailynews.co.tz	Regional
GDN (Gulf Daily News)	gdnlife.com	Regional

Table S2 (continued)

Media Name	Source	Type
Panorama	panorama.ro	Regional
Nova	nova.bg	Regional
Real Equatorial Guinea	realequatorialguinea.com	Regional
Al Arabiya	alarabiya.net	Regional
Monaco Tribune	monaco-tribune.com	Regional
ZNBC	znbc.co.zm	Regional
News-Eleven	news-eleven.com	Regional
La Segunda	lasegunda.com	Regional
India Today	indiatoday.in	Regional
STP-Press	stp-press.st	Regional
Kiamisoft	kiamisoft.ao	Regional
ERR (Estonian Public Broadcasting)	err.ee	Regional
VBTC	vbtc.vu	Regional
Tengri Auto	tengriauto.kz	Regional
La Nación (Costa Rica)	nacion.com	Regional
Antigua Observer	antiguaobserver.com	Regional
Ikon.mn	ikon.mn	Regional
The Daily Star	dailystar.com.lb	Regional
Le Potentiel	lepotentiel.cd	Regional
The Astana Times	astanatimes.com	Regional
The Island	island.lk	Regional
Lesotho Times	lestimes.com	Regional
Le Matinal	lematinal.bj	Regional
The Standard	standardmedia.co.ke	Regional
Ilta-Sanomat	is.fi	Regional
Portal Analitika	portalanalitika.me	Regional

Country and Language Coverage.

The retained outlets collectively span 184 countries and territories. When feasible, at least two operational outlets per country were included, with up to three retained to reduce single-source bias and mitigate risks arising from technical outages or temporary inaccessibility. Across all outlets, content was collected in 40 languages. These languages are listed in Table S3. Automated language identification was performed at the article level using probabilistic language detection models. Articles with detection confidence below 0.90 were flagged for manual verification to prevent misclassification.

S3.2 Keyword-Based Retrieval Strategy

To maximize recall of heat-related reporting, multilingual keyword queries were applied whenever outlet search interfaces permitted structured retrieval. The keyword lexicon was constructed to capture both meteorological descriptors and downstream

Table S3 Languages Covered in the Multilingual News Corpus (40 languages).

English	Spanish	French	Chinese (Simplified)
Portuguese	Arabic	Russian	German
Dutch	Swedish	Norwegian	Icelandic
Italian	Danish	Finnish	Polish
Czech	Hungarian	Greek	Romanian
Bulgarian	Croatian	Serbian	Slovak
Slovenian	Estonian	Latvian	Lithuanian
Ukrainian	Turkish	Indonesian	Malay
Vietnamese	Thai	Korean	Japanese
Hindi	Bengali	Swahili	Amharic

impact indicators, ensuring sensitivity to reporting that might not explicitly use standardized disaster terminology.

Representative examples from the multilingual keyword lexicon are presented in Table S4. The complete lexicon is publicly available in the project repository. Queries were constructed using Boolean combinations of heat descriptors and impact-related terms, including temperature intensity, ecological disruption, health impacts, and economic consequences. When search functionality was unavailable, full archive traversal was performed and semantic filtering was deferred to downstream LLM-based screening to preserve recall.

S3.3 Web Crawling and Normalization

Automated retrieval was conducted between March 2024 and February 2025. The crawling system was designed to accommodate heterogeneous news website architectures and consisted of domain-specific URL scheduling logic, HTTP retrieval with pagination traversal, template-aware extraction of main article text, standardized UTF-8 encoding normalization, and structured storage of extracted content together with detailed crawl logs. Each article retrieval record included timestamped metadata to enable traceability and auditability of the collection process.

To ensure extraction reliability, quarterly random audits were performed. During each audit cycle, a random sample of 500 articles was manually inspected to verify the accuracy of main-text extraction and metadata parsing. Extraction error was defined as either incomplete article body capture, inclusion of non-article boilerplate content, incorrect publication date parsing, or truncation of text. Outlets exhibiting an extraction error rate exceeding 3% during audit were subjected to parser refinement and retrospective reprocessing to ensure consistency across the corpus.

Table S4 Representative Multilingual Keyword Lexicon Used for Retrieval (English, Spanish, French).

English	Spanish	French
Extreme	Extremo	extrême
Heat	Calor	chaleur
High Temperature	Alta Temperatura	températures élevées
Heavy Rain	Lluvia Intensa	fortes pluies
Drought	Sequía	sécheresse
Power Outage from Heat	Corte de Energía por Calor	panne d'électricité due à la chaleur
Fire	Incendio	incendie
Air Pollution	Contaminación del Aire	pollution de l'air
Climate Change	Cambio Climático	changement climatique
Crop Yield Reduction	Reducción de la Producción Agrícola	réduction des rendements agricoles
Oxygen Deficiency	Deficiencia de Oxígeno	hypoxie
Heat Stroke	Golpe de Calor	coup de chaleur
High Temperature Affecting Traffic	Calor Afectando el Tráfico	impact de la chaleur sur le trafic
Ecological Disaster	Desastre Ecológico	désastre écologique
Climate Change Affecting Economy	Cambio Climático Afectando la Economía	impact du changement climatique sur l'économie
Marine Heatwave	Ola de Calor Marina	vague de chaleur marine
High Temperature Pollution	Contaminación por Alta Temperatura	pollution liée à la chaleur
Coral	Coral	corail

S3.4 Deduplication Procedure

To prevent over-counting of syndicated or mirrored content, a two-stage deduplication procedure was implemented. All deduplication steps were applied after text normalization and prior to downstream LLM-based screening to ensure that duplicate articles did not propagate into later stages of the pipeline.

Text Normalization.

Before duplicate detection, article body text underwent standardized normalization to ensure consistent string comparison across languages and publication formats. First, all text was converted to UTF-8 encoding. Unicode normalization was then applied using NFC canonical composition to standardize character representations. All alphabetic characters were converted to lowercase to eliminate case-based discrepancies. HTML tags, embedded scripts, and non-content markup elements were removed during preprocessing. Leading and trailing whitespace was stripped from each article body, and consecutive whitespace characters were collapsed into a single space. No stopword removal, stemming, or language-specific token normalization was performed

at this stage in order to preserve semantic fidelity for subsequent embedding-based similarity detection.

Stage 1: Exact Duplicate Removal.

Exact duplicates were identified using SHA-256 hashing applied to the fully normalized article body text. Let T_i denote normalized text of article i. The hash was computed as:

$$H_i = \text{SHA256}(T_i)$$

Articles with identical hash values were considered exact duplicates. For each identical-hash cluster, only the earliest publication timestamp was retained. All other instances were removed.

Stage 2: Near-Duplicate Detection.

To detect semantically equivalent but non-identical articles (e.g., syndicated wire copies with minor edits), we used multilingual sentence embeddings.

Embedding Model. We used the `paraphrase-multilingual-mpnet-base-v2` model (768-dimensional embeddings), which supports 50+ languages. Inference was performed in evaluation mode with no fine-tuning.

Embedding Preprocessing. Embeddings were computed on the normalized full article body text truncated to a maximum of 512 tokens to maintain consistency across languages.

Blocking Strategy. To reduce computational complexity, candidate comparisons were restricted to articles satisfying (1) the same country tag and (2) publication timestamps within a 7-day rolling window. The 7-day window was selected to capture delayed republication of syndicated content while minimizing false pair comparisons across unrelated time periods.

Similarity Metric. Cosine similarity between embedding vectors was computed:

$$\text{sim}(i, j) = \frac{E_i \cdot E_j}{\|E_i\|\|E_j\|}$$

where E_i is the embedding vector for article i.

Threshold Selection. Articles with cosine similarity ≥ 0.95 were classified as near-duplicates. The threshold of 0.95 was selected after inspecting the similarity distribution on a stratified validation subset (n = 5,000 article pairs), which showed clear separation between syndicated copies (typically > 0.97) and distinct reporting (< 0.90). The 0.95 threshold balances recall of syndicated duplicates with preservation of independent reporting.

Clustering Rule. Near-duplicate articles were grouped using single-linkage clustering. For each cluster, the article with the earliest publication timestamp was retained.

Borderline Similarity Validation.

Articles with cosine similarity in the range $[0.92, 0.95)$ were considered borderline. A random sample of 1,000 article pairs from this interval was manually inspected by two independent reviewers. A false merge was defined as a pair of articles classified as duplicates that contained materially distinct reporting content beyond shared boilerplate or template language. The observed false merge rate at the 0.95 threshold was 1.2%. No systematic language-specific bias was observed.

Computational Environment.

Embedding computation and similarity matching were performed on NVIDIA A100 GPUs using batch inference. Approximate nearest neighbor search was implemented using FAISS (IVF index with cosine metric) to ensure scalability.

Final Output.

After exact and near-duplicate removal, the corpus contains 2,112,478 unique articles. Each retained article is uniquely indexed and carries metadata including country tag, outlet domain, URL, publication timestamp, and language label.

S4 Multilingual Coarse Screening

In figure 1, Step 2 applies a deterministic large language model (LLM) reasoning protocol to classify whether an article describes a real, physically occurring extreme heat event. All 2,112,478 articles from Step 1 were processed under a fixed prompt configuration. A structured human-in-the-loop (HITL) calibration procedure was implemented to ensure robustness and minimize systematic error.

S4.1 Multilingual LLM-Based Filtering

Model Specification.

The screening model used in this study was `Meta Llama 3.1-8B-Instruct` [4], an 8-billion-parameter decoder-only transformer architecture. The model was used in its publicly released instruction-tuned configuration. No additional fine-tuning, adapter training, parameter updating, or task-specific optimization was performed prior to inference.

Inference Environment.

All inference was conducted on NVIDIA A100 80GB GPUs. The model was deployed using the HuggingFace Transformers library (version 4.41) with PyTorch (version 2.1) as the backend framework. Deterministic CUDA execution was enabled to ensure

reproducibility across runs and to eliminate nondeterministic GPU behavior. No distributed stochastic variation was introduced during batch processing.

Decoding Parameters.

Inference was performed using deterministic decoding. The temperature parameter was fixed at 0.0, top-p was set to 1.0, and top-k sampling was disabled. The maximum number of generated tokens was limited to 256 per article. Generation terminated upon completion of the required output format, with the model instructed to stop after the short justification field. No sampling-based randomness or probabilistic decoding strategies were used at any stage of inference.

Task Definition and Deterministic Decision Rule.

For each article, the model was instructed to evaluate the presence or absence of three predefined binary signals: a Temperature Signal, an Impact Signal, and a Location Signal. The Temperature Signal corresponds to explicit or clearly implied mention of unusually high temperature conditions or heatwaves. The Impact Signal corresponds to descriptions of observable physical or societal consequences, such as health effects, infrastructure disruption, wildfire, drought, crop loss, or power outage. The Location Signal corresponds to the presence of a specific geographic reference, including a city, region, province, or country.

The final classification decision was not determined by the model's free-text classification statement. Instead, classification was computed deterministically in post-processing using a programmatic rule. Each signal was parsed from the structured model output and assigned a binary value, where 1 denotes the presence of the signal and 0 denotes its absence. An article was classified as POSITIVE if and only if the sum of the three binary signals satisfied the condition

$$T + I + L \geq 2,$$

where T, I, and L represent the Temperature, Impact, and Location signals, respectively. All classification decisions were therefore derived from structured signal parsing rather than from the model's natural language summary.

Exact Prompt.

The final screening prompt is reproduced verbatim below.

```
You are a climate event detection assistant tasked with identifying
real, physically occurring extreme heat events from news articles.
Objective: Determine whether the article describes an actual extreme
heat event that has occurred or is currently occurring in the real
world.
Definition of a Real Extreme Heat Event: An event qualifies if ALL
of the following conditions are satisfied: 1. The article refers to
unusually high temperature conditions, heatwaves, record-breaking
temperatures, or abnormal thermal anomalies relative to normal climate
conditions. 2. The heat condition occurs in a specific geographic
location. 3. The heat condition results in observable physical
or societal impacts (e.g., human health effects, infrastructure
disruption, wildfire triggered by heat, drought intensification,
agricultural losses, power outages, ecological stress).
Important Clarifications:
- Implicit temperature references qualify if context clearly indicates
abnormal heat conditions (e.g., emergency heat alerts). - Marine
heatwaves qualify if ocean temperature anomalies are explicitly
described. - Retrospective reporting of past heat events qualifies.
- Forecast-only reports WITHOUT actual occurrence do NOT qualify. -
Policy discussions about climate change without describing a specific
heat event do NOT qualify. - Wildfire or drought reports without
explicit linkage to heat conditions do NOT qualify. - Metaphorical
uses of heat (e.g., "heated debate", sports, financial markets) must
be ignored.
Signal Evaluation:
Signal 1 - Temperature Signal: Does the article explicitly or clearly
imply abnormal heat conditions?
Signal 2 - Impact Signal: Does the article describe physical or
societal consequences directly associated with the heat?
Signal 3 - Location Signal: Does the article specify a concrete
geographic location (city, region, province, or country)?
Decision Rule: You must evaluate each signal independently. An article
is classified as POSITIVE only if at least TWO of the three signals
are present.
Output STRICTLY in the following format:
Temperature Signal: YES or NO Impact Signal: YES or NO Location
Signal: YES or NO Short Justification: Provide a concise explanation
referencing explicit evidence in the article (maximum 2 sentences).
Do not provide any additional commentary.
```

Any model output that did not conform exactly to the specified structured format was automatically rejected and re-queried once using the identical prompt. Articles that failed structured formatting twice (0.3% of total cases) were flagged for manual review.

Parsing Protocol.

Model outputs were parsed using regular expressions. Signals were mapped to binary values. If parsing failed twice, article was flagged for manual review (0.3% of cases).

S4.2 Human-in-the-Loop Calibration Protocol

To ensure reliability, cross-linguistic consistency, and boundary stability of the multilingual screening model, a structured human-in-the-loop (HITL) calibration protocol was implemented prior to final full-corpus inference. The objective of this stage was not to tune model parameters, but to refine and stabilize the prompt definition and deterministic decision rule under controlled evaluation.

Calibration Sampling Procedure.

Following initial inference on the full corpus, a stratified random sample of 10,000 articles was constructed for calibration. Sampling was performed without replacement from the complete set of model outputs using a fixed random seed to ensure reproducibility. Stratification was conducted across three axes: publication language, geographic region, and model prediction class.

The twelve most frequently occurring publication languages in the corpus were selected to ensure adequate representation of high-volume linguistic groups, while remaining languages were proportionally represented through pooled sampling. Six geographic macro-regions were defined using United Nations regional classification: Africa, Asia, Europe, North America, South America, and Oceania. Within each language–region stratum, articles were sampled to maintain approximate balance between model-predicted POSITIVE and NEGATIVE classifications. No language contributed fewer than 200 articles to the calibration dataset.

Annotator Recruitment and Training.

Annotation was conducted by five researchers with graduate-level training in climate science, environmental studies, or related quantitative disciplines. Each annotator was assigned articles in languages for which they possessed verified reading proficiency. Prior to formal annotation, annotators completed a supervised training phase involving 200 articles. During this phase, consensus labels were established through group adjudication sessions to standardize interpretation of the operational event definition and signal criteria.

Annotators were blinded to the model's predicted classification to prevent confirmation bias. Articles were presented in randomized order through a controlled annotation interface that recorded timestamps and labeling decisions.

Annotation Guidelines and Event Definition.

Annotators were instructed to determine whether an article described a real, physically occurring extreme heat event. An article was labeled positive only if it met all of the following conditions: it described a real-world heat condition exceeding normal seasonal expectations; the event occurred in a specific geographic location; and the heat condition was associated with observable physical impacts, official emergency declarations, or documented societal consequences. Purely forecast-based reports without confirmed occurrence, general climate change commentary lacking event specificity, or metaphorical uses of heat-related language were labeled negative.

In addition to the binary event label, annotators independently evaluated the presence or absence of the three screening signals—Temperature, Impact, and Location—according to the same operational definitions embedded in the screening prompt.

Disagreement Resolution and Reliability Assessment.

Each article was independently labeled by two annotators. In cases of disagreement, the article was reviewed by a third senior reviewer whose decision served as the final adjudicated label. Inter-annotator agreement was quantified using Cohen's κ statistic computed on the binary event label prior to adjudication, yielding $\kappa = 0.91$, indicating near-perfect agreement. Agreement on individual signal labels exceeded 0.88 across all languages.

Error Characterization and Prompt Refinement.

Model errors were categorized by comparing model predictions with adjudicated human labels. False positives were defined as model-predicted POSITIVE articles labeled negative by human consensus. False negatives were defined as model-predicted NEGATIVE articles labeled positive by human consensus. Error distributions were analyzed across language and region strata to identify systematic boundary failures.

Calibration was conducted iteratively. During the first iteration, dominant false positives arose from metaphorical language and wildfire-only reporting lacking explicit heat linkage. The prompt was revised to clarify exclusion rules and strengthen negative constraints. During the second iteration, residual false negatives were associated with implicitly described heat anomalies lacking explicit temperature values; the prompt was updated to clarify inclusion of contextually evident abnormal heat conditions.

Calibration iterations continued until language-specific precision exceeded 0.88 across all twelve major languages and overall F1-score improvements between successive iterations fell below 0.5 percentage points. No further gains were observed beyond the second refinement cycle.

Evaluation and Statistical Stability.

Performance metrics were evaluated on a held-out validation subset of 2,000 articles drawn from the calibration sample but excluded from prompt refinement. Precision, recall, and F1-score were computed using standard definitions. Prior to calibration, precision was 0.86 and recall was 0.91, yielding an F1-score of 0.88. After final prompt refinement, precision increased to 0.93 and recall to 0.94, resulting in an F1-score of 0.94. Ninety-five percent confidence intervals for precision and recall were computed using Wilson interval estimation and were within ± 1.5 percentage points of point estimates. Cross-linguistic performance variance remained below three percentage points.

Full-Corpus Reprocessing and Data Integrity.

Upon prompt stabilization, the entire corpus of 2,112,478 articles was reprocessed using the finalized prompt configuration and deterministic decision rule. Calibration and validation articles were not excluded from final counts but were retained to preserve consistency of reported totals. No additional human intervention occurred after prompt stabilization.

S5 Fine-Grained Heat Event Structuring in English

Step 3 converts screened candidate articles into structured, event-level records through multilingual translation, deterministic structured extraction, and human verification. All language model inference in this stage was conducted via the official DeepSeek API using a fixed model version and deterministic decoding parameters to ensure reproducibility.

S5.1 Multilingual Alignment and Translation

All non-English candidate articles were translated into English using the `DeepSeek-V3.2` model [5] accessed through the official DeepSeek API. The model version identifier was locked at the time of processing to prevent silent model updates during execution. Translation was performed without fine-tuning or parameter modification.

All API calls were executed with deterministic decoding parameters. The temperature parameter was fixed at 0.0, top-p was set to 1.0, and no sampling-based randomness was enabled. The maximum generation length was dynamically set to 1.3 times the input token length to prevent truncation while limiting over-generation. API responses were logged and cached to ensure traceability and reproducibility of outputs.

The translation prompt was fixed after stabilization and applied uniformly to all non-English articles. The exact prompt is reproduced below.

```
You are a professional multilingual translator specializing in
scientific and journalistic content.
Translate the following news article into precise and literal
English.
Strict Requirements: Preserve all numerical values exactly as
written. Preserve all temperature units exactly as written.
Preserve geographic names exactly. Preserve all date expressions
exactly, including relative forms. Do not summarize, paraphrase,
interpret, or omit information. Do not infer missing details.
Maintain original paragraph structure. Return only the
translated English article.
```

Following translation, automated validation procedures were applied. All numeric substrings were extracted from both source and translated texts and compared for exact match. Geographic entities were detected in both texts using multilingual named entity recognition; mismatches triggered manual inspection.

To quantify translation fidelity, a stratified random sample of 5,000 translated articles was drawn using a fixed random seed. Sampling was balanced across language of origin and geographic region. Each article was independently reviewed by two bilingual annotators. Substantive translation errors were defined as numeric corruption, temporal distortion, omission of impact information, or mistranslation of geographic entities. Inter-annotator agreement for substantive error identification was $\kappa = 0.92$. The initial substantive error rate was 2.1%. After minor refinement of the translation prompt emphasizing strict preservation of temporal expressions, a second audit on 1,500 newly sampled articles yielded a substantive error rate of 0.6%.

All non-English articles were retranslated under the finalized prompt configuration to ensure internal consistency. Earlier translated outputs were discarded.

S5.2 Structured Event Extraction

Structured extraction was also performed using `DeepSeek-V3.2` [5] through the official API. The model version was locked to the same release used for translation. No fine-tuning or parameter modification was applied.

All API calls were executed with deterministic decoding parameters. Temperature was fixed at 0.0, top-p was set to 1.0, and sampling-based randomness was disabled. Maximum generation length was limited to 600 tokens to accommodate multi-event outputs. All responses were logged and cached to allow full reproducibility of structured outputs.

Each article was processed independently. If multiple independent heat events were described within a single article, separate JSON objects were required within a single JSON array. If no qualifying heat event was identified, the model was required to return an empty array.

Extraction Prompt.

The exact extraction prompt used for all articles is reproduced below. The prompt text is constrained within a fixed-width environment for formatting consistency.

```
You are an expert information extraction assistant specializing
in climate and extreme weather events.
Extract all real, physically occurring extreme heat events
explicitly described in the following article.
Only extract information directly supported by the text. Do not
infer missing dates, temperatures, or locations. If multiple
distinct events are described, return each separately. If no
qualifying event exists, return [].
Dates must follow ISO 8601 format (YYYY-MM-DD). Use null for any
field not explicitly stated. Restrict impact_category strictly
to: health, infrastructure, wildfire, drought, agriculture,
energy, ecological, transportation, multiple.
Return only a valid JSON array.
```

Required JSON Structure.

The model was required to return a valid JSON array conforming exactly to the following schema:

```
[
  {
    "event_start_date": "YYYY-MM-DD or null",
    "event_end_date": "YYYY-MM-DD or null",
    "intensity_reference": {
      "value": number or null,
      "unit": "Celsius" or "Fahrenheit" or null,
      "description": string or null
    },
    "impact_category": ["health" | "infrastructure" |
                        "wildfire" | "drought" |
                        "agriculture" | "energy" |
                        "ecological" | "transportation" |
                        "multiple"],
    "spatial_description": {
      "location_name": string,
      "administrative_level": string
    }
  }
]
```

Outputs were validated using strict JSON parsing. Responses failing JSON validation were re-submitted once under identical parameters. Relative temporal expressions were normalized programmatically using the publication date as reference anchor. Extracted temperature values were checked against a plausibility range of -50°C to 65°C, and outliers were flagged for manual review.

S5.3 Human-in-the-Loop Verification and Calibration

To ensure structural accuracy, semantic consistency, and boundary stability of AI-generated outputs, a structured human-in-the-loop (HITL) protocol was implemented across both the screening and structured extraction stages. The objective of this protocol was to identify systematic model errors, refine prompt definitions where necessary, and verify the integrity of extracted event metadata without modifying model parameters.

Sampling Design and Reproducibility.

For each verification stage, records were drawn using stratified random sampling without replacement from the full model output. Sampling was performed using a fixed random seed to ensure procedural reproducibility. Stratification was conducted across original article language group, United Nations geographic macro-region, and predicted impact category. Minimum representation thresholds were enforced to ensure that no major language or geographic region was underrepresented in the verification subset. Sampling was performed at the structured event level rather than the article level when multiple events were extracted from a single article.

Reviewer Qualifications and Blinding.

Verification was conducted by five independent reviewers with graduate-level training in climate science, environmental risk assessment, or closely related disciplines. Reviewers were assigned records based on language proficiency and domain familiarity. All reviewers were blinded to automated model outputs beyond the structured record itself and were not informed of prior reviewer judgments. Review order was randomized to prevent ordering effects and minimize fatigue bias.

Verification Dimensions.

Each structured event record was evaluated independently across four dimensions: temporal normalization accuracy, intensity fidelity and plausibility, impact classification correctness, and spatial description accuracy.

Temporal normalization was assessed by comparing ISO-formatted dates against event timing described in the article, anchored to the publication date when relative temporal expressions were used. Intensity fidelity required that any extracted temperature values correspond exactly to explicitly stated numeric values in the article. Plausibility

was evaluated relative to physically realistic climatological bounds unless explicitly justified in the text. Impact classification was evaluated against the controlled vocabulary and required direct textual evidence. Spatial description was assessed for correctness of location name and administrative level relative to the article narrative.

For each dimension, reviewers assigned a binary correctness judgment and indicated whether correction was required.

Voting and Adjudication Protocol.

Each record was independently evaluated by two reviewers. If both reviewers agreed that all dimensions were correct, the record was marked as verified. If either reviewer identified an error in any dimension, the record was flagged for adjudication.

Flagged records were evaluated independently by a third senior reviewer who was blinded to the initial judgments. Final decisions were determined by majority vote across the three reviewers for each dimension independently. In rare cases of complete disagreement across all three reviewers on a given dimension, a structured adjudication discussion was conducted and a consensus decision was recorded.

Inter-annotator agreement was calculated prior to adjudication using Cohen's κ statistic for each verification dimension. Agreement values were $\kappa = 0.93$ for temporal normalization, $\kappa = 0.95$ for intensity fidelity, $\kappa = 0.89$ for impact classification, and $\kappa = 0.91$ for spatial description.

Definition of Substantive Correction.

A substantive correction was defined as any modification to structured metadata that altered semantic content, including changes to event dates, temperature values, impact categories, or spatial descriptions. Minor formatting adjustments that did not change semantic interpretation were not counted as substantive corrections.

The substantive correction rate was computed as the proportion of sampled records requiring at least one substantive correction after adjudication. Across the verification subset, 3.4% of records required substantive correction. Temporal normalization errors accounted for the largest share of corrections, followed by impact reclassification and spatial refinement. Intensity corrections were rare and primarily associated with relative unit interpretation.

Error Pattern Analysis and Prompt Refinement.

All corrected records were analyzed to identify systematic boundary conditions. Two dominant error patterns were observed. The first involved articles describing multiple spatial locations within a single heat episode, leading to over-segmentation or under-segmentation of events. The second involved ambiguous relative temporal expressions resulting in incorrect date anchoring.

Prompt language was refined once to clarify multi-location grouping logic and to explicitly require alignment of relative temporal references to publication date when exact dates were not provided. No model parameters were updated during this process.

To prevent validation leakage, prompt refinement decisions were based exclusively on a designated calibration subset. A separate held-out validation subset, not used during refinement, was used to evaluate post-refinement performance stability. Following prompt stabilization, the entire corpus was reprocessed from scratch under the finalized configuration to ensure uniform application of updated extraction rules.

Independence from Model Adaptation.

The HITL protocol did not involve any gradient updates, fine-tuning, reinforcement learning, or parameter adaptation of the underlying language model. All improvements resulted from clarification of prompt definitions and deterministic post-processing rules. Model weights remained unchanged throughout the study.

Procedural Determinism.

All classification decisions and structured extraction outcomes were derived from deterministic decoding configurations and programmatic parsing rules. Human intervention occurred only within predefined calibration and verification subsets and did not selectively modify individual records outside the adjudication framework. After prompt stabilization, no further manual corrections were introduced into the final dataset.

Outcome of HITL Verification.

After verification and reprocessing, the final structured dataset contained 210,525 validated extreme heat event records. These records represent the fully calibrated output of the AI-assisted structuring pipeline and constitute the input to the subsequent geocoding module.

S6 Geographical Mapping and Spatial Uncertainty Control

In figure 1, Step 4 converts structured spatial descriptions extracted in Step 3 into geographic coordinates suitable for quantitative spatial analysis. This stage consists of deterministic AI-assisted geocoding, human spatial auditing for ambiguous cases, and explicit quantification of spatial uncertainty. The objective is to ensure consistent, reproducible spatial mapping while transparently accounting for resolution limits inherent to textual reporting.

S6.1 LLM-Based Geocoding

Geocoding was performed using the DeepSeek API under deterministic decoding configuration. Structured spatial descriptions extracted in Step 3 were used as input. The model was not permitted to infer locations beyond those explicitly present in the structured record.

Hierarchical Resolution Framework.

Geocoding followed a hierarchical spatial resolution logic. If the spatial description included a city-level location with an identifiable administrative parent (e.g., city and country), the event was mapped to the geographic coordinates of the city center as defined in the reference gazetteer. If only a province-level or state-level administrative unit was available, the event was mapped to the centroid of that administrative polygon. If only a country-level reference was available, the event was mapped to the national centroid. Records lacking identifiable geographic entities were excluded from coordinate assignment and flagged for manual review.

Reference Gazetteer and Cross-Validation.

A standardized global gazetteer database derived from publicly available administrative boundary datasets was used as the coordinate reference. For each spatial description, candidate matches were generated through exact string matching and fuzzy matching within the gazetteer. When multiple matches were possible, administrative hierarchy consistency with the country field was enforced to disambiguate homonyms.

Geocoding Prompt.

The deterministic geocoding prompt used to resolve textual spatial descriptions is reproduced below.

```
You are a geocoding assistant.
Given the structured spatial description below, identify the
most precise geographic entity explicitly referenced.
Rules: Do not infer locations beyond what is explicitly stated.
If the location is a city, return its official city name
and country. If the location is a province or state, return
the administrative unit and country. If ambiguous, indicate
ambiguity. Do not guess.
Return the following fields: resolved_name administrative_level
country ambiguity_flag (YES or NO)
```

The model output was parsed deterministically. Coordinates were assigned only after matching the resolved name against the gazetteer.

Coordinate Assignment Logic.

If administrative_level equaled "city," the latitude and longitude of the city center were assigned directly. If administrative_level equaled "province" or "state," the centroid of the administrative polygon was computed using geographic boundary shapefiles and assigned as the event coordinate. For country-level resolution, the national centroid was used. If ambiguity_flag was returned as YES or if multiple gazetteer matches satisfied the query, the record was flagged for human spatial auditing.

S6.2 Human Spatial Auditing

To prevent systematic geocoding errors, a structured human spatial auditing stage was implemented.

Confidence Scoring and Review Trigger.

Each geocoded record was assigned a confidence score based on three criteria: uniqueness of gazetteer match, administrative hierarchy consistency, and absence of ambiguity_flag. Records failing any criterion were assigned low confidence. Low-confidence records constituted 6.8% of the total geocoded dataset and were automatically routed to human review.

Manual Spatial Verification.

Human reviewers with training in geographic information systems independently verified flagged records. Reviewers consulted official administrative boundary maps and authoritative geographic databases to resolve ambiguity. For multi-region mentions within a single event narrative, reviewers determined whether events should be split into multiple coordinate records or aggregated under a higher administrative level. All manual corrections were documented, and revised coordinates were revalidated against gazetteer entries.

S6.3 Spatial Uncertainty Quantification

Spatial uncertainty was explicitly quantified according to the resolution level of each geocoded event. For city-level assignments, spatial uncertainty was approximated as a 5 km radius around the city centroid, reflecting typical urban spatial scale and reporting granularity. For province-level assignments, uncertainty was quantified as the radius of a circle with area equal to the administrative polygon area, centered at the polygon centroid. This radius was computed as $r = \sqrt{\frac{A}{\pi}}$, where A represents the area of the administrative region in square kilometers. For country-level assignments, uncertainty was computed analogously using national boundary area.

Sensitivity Analysis.

To assess robustness to spatial uncertainty, sensitivity tests were conducted by perturbing event coordinates within their uncertainty radius using Monte Carlo sampling. For each event, 100 random coordinate draws were generated within the assigned uncertainty circle. Downstream spatial aggregation metrics were recalculated under these perturbations. Variability in national-level exposure estimates remained below 2.5%, indicating low sensitivity to coordinate uncertainty assumptions.

S6.4 Final Geocoded Dataset

After automated geocoding, human spatial auditing, and uncertainty quantification, the final global geocoded heat event database contained spatially resolved coordinates for 210,525 validated heat events. Each record includes latitude, longitude, administrative resolution level, and associated uncertainty radius. These geocoded events were subsequently merged with physically detected heatwave grids described in Section S1 to enable spatial comparison between reported and observed extreme heat events.

S7 Identification of Under-Reported Heatwave Regions

This section formalizes the spatial integration framework used to identify geographically explicit regions characterized by systematic under-reporting of extreme heat events. The procedure integrates two independent datasets: gridded heatwave detection rasters derived from meteorological observations and climate model simulations, and structured geocoded event records extracted from the multilingual news corpus.

The objective of this section is to define, in fully deterministic terms, the raster construction, spatial matching, inclusion masking, ratio computation, and tabular export procedures required for reproducibility. This section does not repeat the heatwave detection threshold logic described previously, nor the geocoding procedures detailed earlier. Instead, it focuses exclusively on raster representation, cross-dataset spatial matching, and the derivation of the reporting ratio used to construct the under-reporting region inventory presented in Table S5.

Seasonal Aggregation and Raster Construction.

Heatwave detection outputs were aggregated into 40 discrete seasonal periods covering the years 2015–2024. Each calendar year was partitioned into four quarters defined as January–March, April–June, July–September, and October–December. Each detected heatwave event was assigned to a seasonal period according to its onset date. This procedure yielded 40 temporally discrete raster layers.

For contemporary analyses, seasonal rasters were constructed on the native ERA5 grid at $0.25^\circ \times 0.25^\circ$ spatial resolution in WGS84 geographic coordinates (EPSG:4326).

For future projection analyses using CESM2 under SSP5–8.5, daily maximum temperature fields were first remapped to a $1^\circ \times 1^\circ$ latitude–longitude grid via bilinear interpolation prior to seasonal aggregation. All seasonal rasters within a given resolution share identical coordinate reference systems, affine GeoTransform parameters, spatial extents, and grid alignment.

Each seasonal raster stores integer counts representing the number of detected heatwave events occurring within each grid cell during the corresponding season. For under-reporting analysis, detection rasters were binarized such that a grid cell was assigned a value of one if at least one heatwave event was detected during that season and zero otherwise. Formally, letting $D_{i,s}^{\text{count}}$ denote the detected event count for grid cell i in season s, the binarized detection indicator is defined as

$$D_{i,s} = \begin{cases} 1 & \text{if } D_{i,s}^{\text{count}} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Cells with no valid temperature input were assigned NaN values and excluded from subsequent aggregation. Consistency validation confirmed identical resolution, grid indexing, CRS definition, and spatial extent across all 40 seasonal rasters.

Spatial Matching Between Reported Events and Detection Rasters.

Each validated news-derived event record contains a standardized seasonal label, a geographic centroid coordinate, and an associated spatial radius reflecting event scale. City-level events are represented as point locations. Provincial- or larger-scale events are represented as circular buffers centered on the administrative centroid with an estimated radius corresponding to spatial extent.

Because longitudinal distances vary with latitude, buffer widths were adjusted by a cosine latitude scaling factor to preserve isotropic spatial representation in geographic coordinates. For an event centered at latitude ϕ, longitudinal angular extent was scaled by $\cos(\phi)$ to ensure correct metric equivalence.

For each season, spatial matching proceeded deterministically. Event buffers were projected onto the corresponding seasonal detection raster. Grid cell centers were transformed from row–column indices to geographic coordinates using the raster's affine GeoTransform parameters. A grid cell was marked as containing a reported event if the geometric intersection between the cell boundary and at least one event buffer was non-empty.

The reported-event indicator was defined as

$$R_{i,s} = \begin{cases} 1 & \text{if at least one reported event intersects grid cell } i \text{ in season } s, \\ 0 & \text{otherwise.} \end{cases}$$

If multiple reported events overlapped the same grid cell in a single season, the value remained one. If a single event overlapped multiple cells, each intersecting cell received a value of one. Binarization prevents double-counting within seasons.

Inclusion Mask and Logical Consistency Filtering.

Under-reporting analysis retained only logically consistent seasonal combinations between detected and reported events. For each grid cell and season, three scenarios were retained: detection present and report present, detection present and report absent, and detection absent and report absent. The fourth logical combination, detection absent and report present, was explicitly excluded from ratio computation to prevent inflation due to potential non-heat-related reporting or detection mismatch. This logical inclusion mask was applied seasonally prior to temporal aggregation.

Reporting Ratio Definition.

For each grid cell i, the reporting ratio was computed across all 40 seasonal periods as

$$RR_i = \frac{\sum_{s=1}^{40} R_{i,s}}{\sum_{s=1}^{40} D_{i,s}}.$$

The numerator represents the number of seasons during which a reported heat event overlapped grid cell i. The denominator represents the number of seasons during which a meteorologically detected heatwave occurred in that cell. If

$$\sum_{s=1}^{40} D_{i,s} = 0,$$

then RR_i was assigned NaN and excluded from further analysis, as no physical hazard occurred in that location during the study period. The resulting reporting ratio field forms the basis for identifying regions characterized by persistent under-reporting across seasons. Grid cells satisfying the vulnerability and media coverage criteria are extracted and reported in Table S5.

Raster Extraction and Tabular Export.

Ratio rasters were converted to tabular format using deterministic raster reading procedures implemented in Python with the rasterio library. Affine GeoTransform parameters were extracted directly from raster metadata. For each valid grid cell, row and column indices were converted to geographic coordinates according to

$$\lambda = x_0 + c \cdot \Delta x, \quad \phi = y_0 + r \cdot \Delta y,$$

where (x_0, y_0) and $(\Delta x, \Delta y)$ denote affine transformation parameters, and r, c denote row and column indices. Only cells with $RR_i \geq 0$ were retained. Cells with NaN values

were excluded. The exported table contains longitude, latitude, reporting ratio value, and corresponding raster row and column indices, ensuring traceability between raster and tabular representations. The final tabular dataset contains longitude, latitude, reporting ratio value, raster row and column indices, and associated country-level classification attributes. The subset of least developed and climate-unprepared regions with reporting ratio less than or equal to 0.5 is provided in Table S5.

Reproducibility and Determinism.

All seasonal aggregation, raster binarization, spatial intersection, inclusion masking, and ratio computation steps were implemented using deterministic programmatic rules. No manual modification of individual grid cells occurred after pipeline stabilization. The full workflow is reproducible given access to ERA5 2m temperature data (1985–2024), CESM2 SSP5–8.5 daily maximum temperature data (2015–2100), the structured geocoded news event database, and the published raster processing scripts.

Table S5: Least developed and climate-unprepared regions with $\leq 50\%$ media coverage of extreme heat events

ID	Row	Column	lon	lat	Ratio	ISO	SIDS	LDC	NDGAINRank
321120	32	112	45	-25.0347	0.2500	MDG	0	1	NDGAIN_Bottom75%
331120	33	112	45	-23.0319	0.4000	MDG	0	1	NDGAIN_Bottom75%
331130	33	113	47	-23.0319	0.3330	MDG	0	1	NDGAIN_Bottom75%
341120	34	112	45	-21.0292	0.4350	MDG	0	1	NDGAIN_Bottom75%
341130	34	113	47	-21.0292	0.4210	MDG	0	1	NDGAIN_Bottom75%
351120	35	112	45	-19.0264	0.3680	MDG	0	1	NDGAIN_Bottom75%
351130	35	113	47	-19.0264	0.3040	MDG	0	1	NDGAIN_Bottom75%
351140	35	114	49	-19.0264	0.4620	MDG	0	1	NDGAIN_Bottom75%
360961	36	96	13	-17.0236	0.5000	AGO	0	1	NDGAIN_Bottom75%
361120	36	112	45	-17.0236	0.2920	MDG	0	1	NDGAIN_Bottom75%
361130	36	113	47	-17.0236	0.2500	MDG	0	1	NDGAIN_Bottom75%
361140	36	114	49	-17.0236	0.2000	MDG	0	1	NDGAIN_Bottom75%
370960	37	96	13	-15.0208	0.1250	AGO	0	1	NDGAIN_Bottom75%
371140	37	114	49	-15.0208	0.1430	MDG	0	1	NDGAIN_Bottom75%
380960	38	96	13	-13.0181	0.3080	AGO	0	1	NDGAIN_Bottom75%
380970	38	97	15	-13.0181	0.3890	AGO	0	1	NDGAIN_Bottom75%
380980	38	98	17	-13.0181	0.3210	AGO	0	1	NDGAIN_Bottom75%
380990	38	99	19	-13.0181	0.4400	AGO	0	1	NDGAIN_Bottom75%
381140	38	114	49	-13.0181	0.1600	MDG	0	1	NDGAIN_Bottom75%
390970	39	97	15	-11.0153	0.3750	AGO	0	1	NDGAIN_Bottom75%
390980	39	98	17	-11.0153	0.3680	AGO	0	1	NDGAIN_Bottom75%
390990	39	99	19	-11.0153	0.2670	AGO	0	1	NDGAIN_Bottom75%
400970	40	97	15	-9.0125	0.0000	AGO	0	1	NDGAIN_Bottom75%
400980	40	98	17	-9.0125	0.2000	AGO	0	1	NDGAIN_Bottom75%
400981	40	98	17	-9.0125	0.2000	COD	0	1	NDGAIN_Bottom75%
401000	40	100	21	-9.0125	0.5000	AGO	0	1	NDGAIN_Bottom75%
401001	40	100	21	-9.0125	0.5000	COD	0	1	NDGAIN_Bottom75%
401080	40	108	37	-9.0125	0.4170	TZA	0	1	NDGAIN_Bottom75%
410980	41	98	17	-7.00972	0.5000	AGO	0	1	NDGAIN_Bottom75%
410981	41	98	17	-7.00972	0.5000	COD	0	1	NDGAIN_Bottom75%

ID	Row	Column	lon	lat	Ratio	ISO	SIDS	LDC	NDGAINRank
411050	41	105	31	-7.00972	0.4580	COD	0	1	NDGAIN_Bottom75%
411051	41	105	31	-7.00972	0.4580	TZA	0	1	NDGAIN_Bottom75%
411070	41	107	35	-7.00972	0.5000	TZA	0	1	NDGAIN_Bottom75%
411090	41	109	39	-7.00972	0.4400	TZA	0	1	NDGAIN_Bottom75%
420960	42	96	13	-5.00694	0.5000	AGO	0	1	NDGAIN_Bottom75%
420961	42	96	13	-5.00694	0.5000	COD	0	1	NDGAIN_Bottom75%
421060	42	106	33	-5.00694	0.5000	TZA	0	1	NDGAIN_Bottom75%
431000	43	100	21	-3.00417	0.4643	COD	0	1	NDGAIN_Bottom75%
431020	43	102	25	-3.00417	0.4667	COD	0	1	NDGAIN_Bottom75%
440980	44	98	17	-1.00139	0.4444	COD	0	1	NDGAIN_Bottom75%
381110	38	111	43	-13.0181	0.2692	COM	1	1	NDGAIN_Bottom75%
381121	38	112	45	-13.0181	0.2222	COM	1	1	NDGAIN_Bottom75%
381740	38	174	169	-13.0181	0.0833	SLB	1	1	NDGAIN_Bottom75%
390140	39	14	-151	-11.0153	0.0000	KIR	1	1	NDGAIN_Bottom75%
461000	46	100	21	3.004167	0.5000	COD	0	1	NDGAIN_Bottom75%
391110	39	111	43	-11.0153	0.1724	COM	1	1	NDGAIN_Bottom75%
470860	47	86	-7	5.006944	0.1739	LBR	0	1	NDGAIN_Bottom75%
391690	39	169	159	-11.0153	0.0870	SLB	1	1	NDGAIN_Bottom75%
391700	39	170	161	-11.0153	0.0370	SLB	1	1	NDGAIN_Bottom75%
391710	39	171	163	-11.0153	0.0000	SLB	1	1	NDGAIN_Bottom75%
391720	39	172	165	-11.0153	0.0000	SLB	1	1	NDGAIN_Bottom75%
391730	39	173	167	-11.0153	0.0000	SLB	1	1	NDGAIN_Bottom75%
480840	48	84	-11	7.009722	0.5000	LBR	0	1	NDGAIN_Bottom75%
480841	48	84	-11	7.009722	0.5000	SLE	0	1	NDGAIN_Bottom75%
480850	48	85	-9	7.009722	0.4074	LBR	0	1	NDGAIN_Bottom75%
480852	48	85	-9	7.009722	0.4074	GIN	0	1	NDGAIN_Bottom75%
480860	48	86	-7	7.009722	0.4500	LBR	0	1	NDGAIN_Bottom75%
401520	40	152	125	-9.0125	0.2857	TLS	1	1	NDGAIN_Bottom75%
401530	40	153	127	-9.0125	0.2258	TLS	1	1	NDGAIN_Bottom75%
401680	40	168	157	-9.0125	0.1739	SLB	1	1	NDGAIN_Bottom75%
401690	40	169	159	-9.0125	0.2308	SLB	1	1	NDGAIN_Bottom75%
401700	40	170	161	-9.0125	0.1000	SLB	1	1	NDGAIN_Bottom75%
401710	40	171	163	-9.0125	0.0000	SLB	1	1	NDGAIN_Bottom75%
401730	40	173	167	-9.0125	0.0000	SLB	1	1	NDGAIN_Bottom75%
401790	40	179	179	-9.0125	0.0385	TUV	1	1	NDGAIN_Bottom75%
490830	49	83	-13	9.0125	0.4286	SLE	0	1	NDGAIN_Bottom75%
490831	49	83	-13	9.0125	0.4286	GIN	0	1	NDGAIN_Bottom75%
490850	49	85	-9	9.0125	0.4783	LBR	0	1	NDGAIN_Bottom75%
490852	49	85	-9	9.0125	0.4783	GIN	0	1	NDGAIN_Bottom75%
490861	49	86	-7	9.0125	0.5000	GIN	0	1	NDGAIN_Bottom75%
490871	49	87	-5	9.0125	0.3333	BFA	0	1	NDGAIN_Bottom75%
490882	49	88	-3	9.0125	0.4286	BFA	0	1	NDGAIN_Bottom75%
411670	41	167	155	-7.00972	0.2400	SLB	1	1	NDGAIN_Bottom75%
411680	41	168	157	-7.00972	0.1739	SLB	1	1	NDGAIN_Bottom75%
411690	41	169	159	-7.00972	0.2308	SLB	1	1	NDGAIN_Bottom75%
411700	41	170	161	-7.00972	0.1154	SLB	1	1	NDGAIN_Bottom75%
411780	41	178	177	-7.00972	0.0000	TUV	1	1	NDGAIN_Bottom75%
411790	41	179	179	-7.00972	0.0000	TUV	1	1	NDGAIN_Bottom75%
420020	42	2	-175	-5.00694	0.0000	KIR	1	1	NDGAIN_Bottom75%
420030	42	3	-173	-5.00694	0.0000	KIR	1	1	NDGAIN_Bottom75%
420040	42	4	-171	-5.00694	0.0000	KIR	1	1	NDGAIN_Bottom75%
420120	42	12	-155	-5.00694	0.0000	KIR	1	1	NDGAIN_Bottom75%
500871	50	87	-5	11.01528	0.3158	BFA	0	1	NDGAIN_Bottom75%

ID	Row	Column	lon	lat	Ratio	ISO	SIDS	LDC	NDGAINRank
500872	50	87	-5	11.01528	0.3158	MLI	0	1	NDGAIN_Bottom75%
500881	50	88	-3	11.01528	0.4167	BFA	0	1	NDGAIN_Bottom75%
500891	50	89	-1	11.01528	0.5000	TGO	0	1	NDGAIN_Bottom75%
500892	50	89	-1	11.01528	0.5000	BFA	0	1	NDGAIN_Bottom75%
500990	50	99	19	11.01528	0.5000	TCD	0	1	NDGAIN_Bottom75%
421780	42	178	177	-5.00694	0.0000	TUV	1	1	NDGAIN_Bottom75%
430040	43	4	-171	-3.00417	0.0000	KIR	1	1	NDGAIN_Bottom75%
510880	51	88	-3	13.01806	0.5000	BFA	0	1	NDGAIN_Bottom75%
510881	51	88	-3	13.01806	0.5000	MLI	0	1	NDGAIN_Bottom75%
510890	51	89	-1	13.01806	0.5000	BFA	0	1	NDGAIN_Bottom75%
520830	52	83	-13	15.02083	0.4091	SEN	0	1	NDGAIN_Bottom75%
520831	52	83	-13	15.02083	0.4091	MLI	0	1	NDGAIN_Bottom75%
520832	52	83	-13	15.02083	0.4091	MRT	0	1	NDGAIN_Bottom75%
520840	52	84	-11	15.02083	0.4545	SEN	0	1	NDGAIN_Bottom75%
520841	52	84	-11	15.02083	0.4545	MLI	0	1	NDGAIN_Bottom75%
520842	52	84	-11	15.02083	0.4545	MRT	0	1	NDGAIN_Bottom75%
431770	43	177	175	-3.00417	0.0000	KIR	1	1	NDGAIN_Bottom75%
431780	43	178	177	-3.00417	0.0000	KIR	1	1	NDGAIN_Bottom75%
530820	53	82	-15	17.02361	0.3889	SEN	0	1	NDGAIN_Bottom75%
530821	53	82	-15	17.02361	0.3889	MRT	0	1	NDGAIN_Bottom75%
530830	53	83	-13	17.02361	0.5000	SEN	0	1	NDGAIN_Bottom75%
530831	53	83	-13	17.02361	0.5000	MRT	0	1	NDGAIN_Bottom75%
530840	53	84	-11	17.02361	0.4737	MRT	0	1	NDGAIN_Bottom75%
530850	53	85	-9	17.02361	0.2222	MRT	0	1	NDGAIN_Bottom75%
530860	53	86	-7	17.02361	0.5000	MRT	0	1	NDGAIN_Bottom75%
530870	53	87	-5	17.02361	0.3529	MLI	0	1	NDGAIN_Bottom75%
530871	53	87	-5	17.02361	0.3529	MRT	0	1	NDGAIN_Bottom75%
530910	53	91	3	17.02361	0.4783	NER	0	1	NDGAIN_Bottom75%
530911	53	91	3	17.02361	0.4783	MLI	0	1	NDGAIN_Bottom75%
441740	44	174	169	-1.00139	0.0000	KIR	1	1	NDGAIN_Bottom75%
441770	44	177	175	-1.00139	0.0000	KIR	1	1	NDGAIN_Bottom75%
441780	44	178	177	-1.00139	0.0000	KIR	1	1	NDGAIN_Bottom75%
450110	45	11	-157	1.001389	0.0000	KIR	1	1	NDGAIN_Bottom75%
531060	53	106	33	17.02361	0.3846	SDN	0	1	NDGAIN_Bottom75%
531070	53	107	35	17.02361	0.4706	SDN	0	1	NDGAIN_Bottom75%
540820	54	82	-15	19.02639	0.2222	MRT	0	1	NDGAIN_Bottom75%
540830	54	83	-13	19.02639	0.3125	MRT	0	1	NDGAIN_Bottom75%
540840	54	84	-11	19.02639	0.2667	MRT	0	1	NDGAIN_Bottom75%
540850	54	85	-9	19.02639	0.2222	MRT	0	1	NDGAIN_Bottom75%
540860	54	86	-7	19.02639	0.2500	MLI	0	1	NDGAIN_Bottom75%
540861	54	86	-7	19.02639	0.2500	MRT	0	1	NDGAIN_Bottom75%
540870	54	87	-5	19.02639	0.3684	MLI	0	1	NDGAIN_Bottom75%
540871	54	87	-5	19.02639	0.3684	MRT	0	1	NDGAIN_Bottom75%
540890	54	89	-1	19.02639	0.5000	MLI	0	1	NDGAIN_Bottom75%
540920	54	92	5	19.02639	0.4815	NER	0	1	NDGAIN_Bottom75%
540922	54	92	5	19.02639	0.4815	MLI	0	1	NDGAIN_Bottom75%
540950	54	95	11	19.02639	0.5000	NER	0	1	NDGAIN_Bottom75%
451760	45	176	173	1.001389	0.0000	KIR	1	1	NDGAIN_Bottom75%
460100	46	10	-159	3.004167	0.0000	KIR	1	1	NDGAIN_Bottom75%
460110	46	11	-157	3.004167	0.0000	KIR	1	1	NDGAIN_Bottom75%
550820	55	82	-15	21.02917	0.1579	MRT	0	1	NDGAIN_Bottom75%
550830	55	83	-13	21.02917	0.1579	MRT	0	1	NDGAIN_Bottom75%
550840	55	84	-11	21.02917	0.2500	MRT	0	1	NDGAIN_Bottom75%

ID	Row	Column	lon	lat	Ratio	ISO	SIDS	LDC	NDGAINRank
550850	55	85	-9	21.02917	0.1000	MRT	0	1	NDGAIN_Bottom75%
550860	55	86	-7	21.02917	0.2917	MLI	0	1	NDGAIN_Bottom75%
550861	55	86	-7	21.02917	0.2917	MRT	0	1	NDGAIN_Bottom75%
550870	55	87	-5	21.02917	0.3750	MLI	0	1	NDGAIN_Bottom75%
550880	55	88	-3	21.02917	0.4783	MLI	0	1	NDGAIN_Bottom75%
550891	55	89	-1	21.02917	0.5000	MLI	0	1	NDGAIN_Bottom75%
551050	55	105	31	21.02917	0.4615	SDN	0	1	NDGAIN_Bottom75%
551070	55	107	35	21.02917	0.5000	SDN	0	1	NDGAIN_Bottom75%
461760	46	176	173	3.004167	0.0000	KIR	1	1	NDGAIN_Bottom75%
470090	47	9	-161	5.006944	0.0000	KIR	1	1	NDGAIN_Bottom75%
560830	56	83	-13	23.03194	0.2500	MRT	0	1	NDGAIN_Bottom75%
560840	56	84	-11	23.03194	0.2500	MRT	0	1	NDGAIN_Bottom75%
560850	56	85	-9	23.03194	0.2857	MRT	0	1	NDGAIN_Bottom75%
560860	56	86	-7	23.03194	0.3182	MLI	0	1	NDGAIN_Bottom75%
560861	56	86	-7	23.03194	0.3182	MRT	0	1	NDGAIN_Bottom75%
560870	56	87	-5	23.03194	0.3462	MLI	0	1	NDGAIN_Bottom75%
560881	56	88	-3	23.03194	0.3600	MLI	0	1	NDGAIN_Bottom75%
560891	56	89	-1	23.03194	0.5000	MLI	0	1	NDGAIN_Bottom75%
570830	57	83	-13	25.03472	0.1739	MRT	0	1	NDGAIN_Bottom75%
570840	57	84	-11	25.03472	0.3333	MRT	0	1	NDGAIN_Bottom75%
570850	57	85	-9	25.03472	0.3043	MRT	0	1	NDGAIN_Bottom75%
570861	57	86	-7	25.03472	0.4167	MLI	0	1	NDGAIN_Bottom75%
570862	57	86	-7	25.03472	0.4167	MRT	0	1	NDGAIN_Bottom75%
570871	57	87	-5	25.03472	0.4167	MLI	0	1	NDGAIN_Bottom75%
570872	57	87	-5	25.03472	0.4167	MRT	0	1	NDGAIN_Bottom75%
570881	57	88	-3	25.03472	0.3929	MLI	0	1	NDGAIN_Bottom75%
580851	58	85	-9	27.0375	0.3043	MRT	0	1	NDGAIN_Bottom75%
580861	58	86	-7	27.0375	0.5000	MRT	0	1	NDGAIN_Bottom75%

S8 Spatial Characteristics of Global Extreme Heat Event Under-Reporting

This section characterizes the large-scale geographic structure of reporting disparities using deterministic spatial aggregation and statistically controlled group comparisons. All analyses are performed on the finalized reporting ratio raster defined in Section S7. The procedures described here do not alter grid-level ratio values and do not overlap with raster construction, masking logic, or modeling frameworks. The objective is to quantify meridional structure and structural group differences in reporting completeness while explicitly accounting for spatial sampling density, unequal cell counts, and statistical inference robustness.

S8.1 Latitudinal Aggregation Analysis

To examine meridional structure in reporting disparities, reporting ratio values were aggregated into 36 equal-width latitudinal bins spanning $-90°$ to $+90°$. Each bin

therefore covers 5° of latitude. Bin boundaries were defined as

$$[-90 + 5k, -85 + 5k), \quad k = 0, 1, \ldots, 35,$$

with bin centers defined at midpoints. Only grid cells with valid reporting ratio values ($RR_i \geq 0$) were retained, and cells assigned NaN due to zero detected events were excluded. To prevent instability from sparsely sampled regions, bins containing fewer than 30 valid grid cells were excluded from visualization and inference.

For each latitudinal bin b, let $\mathcal{I}_b$ denote the set of valid grid cells whose centroid latitude falls within bin b. The mean reporting ratio was computed as

$$\overline{RR}_b = \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} RR_i,$$

and the associated variability was quantified using both standard deviation and standard error. Because grid cells are spatially autocorrelated, conventional standard errors underestimate uncertainty. Therefore, uncertainty bands were constructed using a latitude-stratified block bootstrap procedure. Specifically, within each bin, 1,000 bootstrap replicates were generated by resampling grid cells with replacement at the continental-block level to preserve large-scale spatial dependence. The 2.5th and 97.5th percentiles of the bootstrap distribution of $\overline{RR}_b$ were used to construct 95% confidence intervals. Shaded bands in figures represent these bootstrap intervals rather than naïve standard deviation.

To ensure that results are not sensitive to arbitrary bin width selection, robustness tests were conducted using alternative bin resolutions of 10° (18 bins) and 2.5° (72 bins). The overall meridional gradient structure and relative extrema were preserved across all binning schemes, indicating that conclusions are not driven by discretization choice. All aggregation and resampling procedures were implemented deterministically in Python using NumPy and SciPy under fixed random seeds to ensure reproducibility.

S8.2 Group-Based Distribution Analyses

To examine structural differences in reporting disparities across national characteristics, reporting ratios were aggregated to the country level using area-weighted averaging. For each country c, the national reporting ratio was computed as

$$RR_c = \frac{\sum_{i \in \mathcal{I}_c} w_i RR_i}{\sum_{i \in \mathcal{I}_c} w_i},$$

where $\mathcal{I}_c$ denotes grid cells whose centroids fall within national administrative boundaries and w_i represents the grid-cell surface area, calculated as a function of latitude to account for meridian convergence. This area-weighted aggregation prevents small countries with few grid cells from exerting disproportionate influence and ensures that national estimates reflect spatial exposure.

Country classification sources were defined a priori. Small Island Developing States (SIDS) classification followed the official United Nations SIDS list. Least Developed Countries (LDCs) classification followed the United Nations Committee for Development Policy designation for the study period. Development status categories were defined using World Bank income group classifications and consolidated into developed economies, least developed economies, and other developing economies. Climate preparedness was quantified using the ND-GAIN index. Countries were stratified using empirical percentile thresholds. The primary analysis compared the top quartile (25%) of ND-GAIN scores to the remaining 75% to isolate the highest preparedness group while preserving adequate sample size for stable inference. Sensitivity tests using median and tercile splits yielded consistent directionality of group differences.

Prior to hypothesis testing, normality of national reporting ratio distributions within each comparison group was assessed using the Shapiro–Wilk test at $\alpha = 0.05$. If both groups satisfied normality assumptions, Welch's unequal-variance two-sample t-test was applied. If either group violated normality, the Mann–Whitney U (MWU) test was used. All statistical tests were two-sided. Because multiple independent group comparisons were performed, false discovery rate (FDR) correction was applied using the Benjamini–Hochberg procedure. Adjusted p-values are reported, and statistical significance was defined as FDR-adjusted $p < 0.05$.

To avoid instability from small sample sizes, group comparisons were restricted to classifications with at least five countries per subgroup. Continent-level subgroup analyses were similarly restricted to Africa, Asia, and Europe due to sufficient sample counts. All statistical analyses were conducted using SciPy and statsmodels under fixed random seeds where applicable. These procedures ensure that inference is robust to unequal variance, non-normality, spatial sampling imbalance, and multiple hypothesis testing.

Within-Country Heterogeneity Consideration.

National aggregation may obscure subnational heterogeneity, particularly for geographically large countries. To assess this effect, within-country standard deviations of grid-level reporting ratios were computed and compared against between-country variance. Although substantial subnational heterogeneity exists in several large countries, the magnitude of between-country variance remains dominant, indicating that national-level aggregation captures meaningful structural differences. Nonetheless, interpretation of national reporting ratios should be understood as spatial averages rather than homogeneous representations of internal reporting patterns.

S9 Statistical Analyses of Reporting Gaps

This section presents the complete country-level statistical modeling framework used to examine structural correlates of reporting disparities. The dependent variable is the national reporting ratio defined in Section S7. All analyses were conducted strictly at

the country level. Grid-level data were not used in modeling to prevent spatial pseudoreplication. The modeling dataset, preprocessing pipeline, and code dependencies are fully specified to ensure reproducibility.

All analyses were implemented in Python 3.10.12 using pandas 2.0.3, NumPy 1.24.4, SciPy 1.10.1, scikit-learn 1.2.2, statsmodels 0.14.0, shap 0.42.1, and XGBoost 1.7.6. Random seeds were fixed to 42 unless otherwise stated.

S9.1 Covariate Assembly and Harmonization

Country-level socioeconomic, infrastructural, climatic, and linguistic covariates were assembled from publicly accessible international databases. National boundaries were defined using the Natural Earth administrative dataset (version 5.1.1, scale 1:10m). Country matching was performed using ISO 3166-1 alpha-3 codes to ensure consistent merging across datasets.

Climate vulnerability was quantified using the ND-GAIN Country Index, 2022 release. GDP per capita (current US dollars) for 2022 was obtained from the World Bank World Development Indicators (indicator code NY.GDP.PCAP.CD). Internet usage rate for 2022 was obtained from the International Telecommunication Union (ITU), defined as the percentage of individuals using the internet. Climate policy performance was quantified using the Climate Change Performance Index (CCPI) 2023 edition. All datasets were downloaded between January and March 2024. If 2022 data were unavailable, the nearest available year within 2020–2023 was selected without interpolation or extrapolation.

Climatic exposure covariates were derived from ERA5 reanalysis data. Mean annual near-surface air temperature and total annual precipitation were computed by averaging daily values across 2015–2024. National aggregation was performed via area-weighted averaging:

$$X_c = \frac{\sum_{i \in \mathcal{I}_c} w_i X_i}{\sum_{i \in \mathcal{I}_c} w_i},$$

where w_i represents grid-cell surface area computed as $A_i = \cos(\phi_i)\Delta\lambda\Delta\phi R^2$ to account for meridian convergence.

Linguistic accessibility was quantified using a Language Proximity Index (LPN2CommonLanguage). Official languages for each country were obtained from ISO 639-1 codes. Let L_c denote the set of official languages for country c, and let $\mathcal{L}_g = \{\text{English}, \text{Spanish}, \text{French}, \text{Arabic}, \text{Portuguese}\}$. For each pair (ℓ, g), a normalized Levenshtein similarity was computed:

$$S(\ell, g) = 1 - \frac{d(\ell, g)}{\max(|\ell|, |g|)}.$$

The country-level index was defined as

$$\text{LPN2CommonLanguage}_c = \max_{\ell \in L_c, g \in \mathcal{L}_g} S(\ell, g).$$

String normalization included lowercase conversion and diacritic stripping.

Countries missing more than two covariates were excluded. For remaining isolated missing entries, median imputation was applied within World Bank income groups. After filtering, the final modeling dataset included 176 countries. Continuous predictors were standardized using training-set statistics only:

$$X^* = \frac{X - \mu_{\text{train}}}{\sigma_{\text{train}}}.$$

The dependent variable was not standardized. Pairwise Pearson correlations and variance inflation factors (VIF) were computed; all VIF values were below 3, indicating low multicollinearity.

S9.2 XGBoost Regression Model Specification

Extreme Gradient Boosting (XGBoost) regression was employed using the `XGBRegressor` class (XGBoost 1.7.6). The objective function minimized squared error with regularization:

$$\mathcal{L} = \sum_{c=1}^{N} (RR_c - \hat{RR}_c)^2 + \sum_{t=1}^{T} \Omega(f_t),$$

where $\Omega(f_t)$ includes L1 and L2 penalties on leaf weights.

The dataset was randomly partitioned into 80% training and 20% testing subsets using seed 42. Hyperparameter tuning was conducted exclusively within the training data using 5-fold cross-validation. The grid search space included: learning rate $\in \{0.01, 0.05, 0.1, 0.2\}$, max depth $\in \{3, 4, 5, 6, 7\}$, subsample $\in \{0.6, 0.8, 1.0\}$, colsample_bytree $\in \{0.6, 0.8, 1.0\}$, reg_lambda $\in \{0, 1, 5\}$, min_child_weight $\in \{1, 5, 10\}$.

Early stopping was implemented with a 10% validation subset carved from the training set. Training terminated if validation loss failed to improve for 50 consecutive rounds. Maximum boosting rounds were capped at 1,000. The test set was not accessed during tuning.

Performance was evaluated on the held-out test set using

$$R^2 = 1 - \frac{\sum (RR_c - \hat{RR}_c)^2}{\sum (RR_c - \overline{RR})^2}, RMSE = \sqrt{\frac{1}{n} \sum (RR_c - \hat{RR}_c)^2}, MAE = \frac{1}{n} \sum |RR_c - \hat{RR}_c|.$$

Final performance was $R^2 = 0.740$, $RMSE = 0.191$, $MAE = 0.131$. Repeating the full split under seeds $\{7, 21, 84, 123, 2023\}$ produced R^2 variation within ± 0.02, confirming stability.

S9.3 SHAP Attribution Analysis

Feature attribution was performed using SHAP (shap 0.42.1) with the TreeExplainer algorithm. For each country c and feature j, the SHAP value $\phi_{c,j}$ represents the marginal contribution of feature j relative to the model's expected prediction.

Global importance was quantified as mean absolute SHAP value:

$$I_j = \frac{1}{N} \sum_{c=1}^{N} |\phi_{c,j}|.$$

Features were ranked in descending order of I_j. SHAP baseline expectation was computed as the mean prediction across the training dataset. SHAP computation was applied post hoc and did not affect model fitting. SHAP values quantify predictive contribution and do not imply causality.

S9.4 Subgroup Linear Regression Analyses

Linear regression analyses were conducted independently of the XGBoost model. Analyses were restricted to continents with at least five countries per subgroup; Africa, Asia, and Europe satisfied this criterion.

Within each continent, countries were divided into High and Low categories using the 1/3 quantile rule. Let Q_{33} and Q_{67} denote empirical percentiles. Countries with $X_c \leq Q_{33}$ were assigned to Low; countries with $X_c \geq Q_{67}$ were assigned to High. Countries between thresholds were excluded. Ties at thresholds were assigned to the lower group to ensure deterministic assignment.

Regression models were specified as

$$RR_c = \beta_0 + \beta_1 X_c + \varepsilon_c,$$

estimated using ordinary least squares with heteroskedasticity-robust (HC3) standard errors via statsmodels 0.14.0.

For African countries, additional regressions were conducted:

$$RR_c = \beta_0 + \beta_1 \text{LPN2CommonLanguage}_c + \varepsilon_c,$$

and separately for the Arabic-language subgroup. Regression residuals were inspected for heteroskedasticity and influential points using Cook's distance; no observation exceeded standard influence thresholds. These regressions are descriptive and independent of the machine-learning model.

S10 Stratified and Regional Association Analyses for 2050 global extreme heat event loss

This section describes the complete methodology used to project future severe heat exposure and associated economic and mortality losses by mid-century. All projections were conducted under the high-emission SSP5-8.5 scenario using CESM2 outputs from CMIP6. Climate projections, economic loss functions, mortality conversion procedures, and stratification analyses are fully specified to ensure reproducibility.

S10.1 Severe Heat-Day Projection

Future climate projections were derived from the Community Earth System Model Version 2 (CESM2) [1, 2], CMIP6 archive, scenario SSP5-8.5, ensemble member r4i1p1f1. Daily maximum near-surface air temperature ($T_{\max}$) for 2015–2100 was downloaded from the Earth System Grid Federation (ESGF). Model output was bilinearly interpolated to a $1^\circ \times 1^\circ$ latitude–longitude grid in WGS84 coordinates to ensure spatial consistency with country aggregation procedures.

Severe heat-days were defined relative to the historical climatological baseline 1985–2014 derived from ERA5 reanalysis, consistent with Section S1. For each grid cell and calendar day, the 90th percentile threshold was computed using the 11-day moving window method described previously. A severe heat-day was defined as any day in which projected $T_{\max}$ exceeded this historical 90th percentile threshold. Annual severe heat-day counts were computed for each grid cell. Mid-century projections were defined as the 2041–2060 average of annual severe heat-day counts.

Country-level aggregation was performed using area-weighted averaging across all grid cells whose centroids fall within national boundaries:

$$SHD_c = \frac{\sum_{i \in \mathcal{I}_c} w_i SHD_i}{\sum_{i \in \mathcal{I}_c} w_i},$$

where SHD_i denotes projected severe heat-days at grid cell i, and w_i represents grid-cell surface area adjusted for latitude. This procedure yields country-level severe heat exposure metrics that are directly comparable across nations of different sizes [6].

S10.2 Economic Loss Estimation

Economic losses were estimated as proportional reductions in national GDP attributable to projected severe heat exposure. The functional form was specified as a semi-elastic response:

$$Loss_c^{econ} = \beta \cdot (SHD_c - SHD_c^{baseline}),$$

where $SHD_c^{baseline}$ denotes the 2015–2024 average severe heat-day exposure, and β represents the marginal GDP loss per additional severe heat-day. The parameter β was calibrated using empirical literature estimates linking extreme heat exposure to annual GDP growth reductions under high-emission scenarios [7]. The calibrated value of β corresponds to a proportional GDP reduction per additional severe heat-day, ensuring unit consistency.

Projected economic loss was expressed as percentage of national GDP:

$$GDP\_Loss\_Pct_c = 100 \times Loss_c^{econ}.$$

GDP normalization was performed using 2022 GDP levels (World Bank WDI) to maintain consistency with covariate years used in Section S9. No dynamic GDP growth projections were incorporated, thereby isolating climate-driven exposure effects from economic growth assumptions.

Countries were classified into quintiles based on projected GDP loss percentage. Quintile thresholds were computed using the empirical distribution of $GDP\_Loss\_Pct_c$ across all modeled countries. Countries in the top 20% were defined as the highest projected economic loss group. This quintile rule was applied deterministically without smoothing or interpolation.

S10.3 Mortality Impact Estimation

Mortality impacts were estimated using an exposure-response framework linking severe heat-days to excess mortality risk. For each country c, projected additional mortality was computed as:

$$Deaths_c = \gamma \cdot (SHD_c - SHD_c^{baseline}) \cdot Pop_c,$$

where γ represents the marginal mortality increase per severe heat-day per capita, and Pop_c denotes projected mid-century population. Population projections for 2050 were obtained from the United Nations World Population Prospects (2022 revision).

Mortality losses were expressed as percentage of baseline mortality burden:

$$Mortality\_Loss\_Pct_c = \frac{Deaths_c}{Baseline\_Deaths_c} \times 100.$$

Baseline mortality levels were obtained from World Health Organization country-level mortality statistics (latest pre-2023 release). Population weighting ensures that mortality estimates reflect exposure magnitude scaled by population size rather than per-capita metrics alone.

All mortality computations were conducted at the country level using consistent exposure and population inputs. No age-structure differentiation was applied; mortality estimates therefore represent aggregate national impacts.

S10.4 Interaction Analysis: Under-Reporting and Vulnerability

To examine compounding effects between under-reporting and structural vulnerability, countries were jointly stratified according to reporting ratio, development status, and climate vulnerability. Under-reporting severity was defined using the bottom quintile of reporting ratio distribution, consistent with Section S7. Least Developed Country (LDC) classification followed the United Nations Committee for Development Policy designation (2023 list). Climate vulnerability was defined using ND-GAIN 2022 scores, with high vulnerability defined as countries in the bottom quartile of ND-GAIN distribution.

Joint stratification categories were constructed as follows: (i) under-reported and LDC; (ii) under-reported and high-vulnerability (ND-GAIN bottom quartile); (iii) under-reported and both LDC and high-vulnerability; and (iv) under-reported but neither LDC nor high-vulnerability. For each stratum, projected economic and mortality losses were averaged across countries:

$$\overline{Loss}_{group} = \frac{1}{N_{group}} \sum_{c \in group} Loss_c.$$

This deterministic grouping framework enables identification of compounding structural disadvantage. All classification thresholds were computed directly from empirical distributions and publicly available designations. No smoothing or model-based clustering was used in stratification.

S11 Robustness and Uncertainty Analyses

This section quantifies the stability of the full analytical pipeline under controlled perturbations. Sensitivity analyses were conducted for spatial buffering assumptions, seasonal aggregation definitions, heatwave detection thresholds, multilingual screening precision, coordinate uncertainty, and national reporting ratio estimation. All perturbation experiments were implemented deterministically with fixed random seeds to ensure reproducibility.

S11.1 Sensitivity to Spatial Buffer Radius

Reported events were spatially represented using circular buffers centered on geocoded coordinates. To evaluate sensitivity to buffer radius assumptions, three alternative buffer configurations were tested: (i) baseline radius as defined in Section S5, (ii) 50% reduction in radius, and (iii) 50% increase in radius. For province-level and country-level assignments, the uncertainty radius $r = \sqrt{A/\pi}$ was scaled accordingly.

For each configuration, the full spatial matching procedure described in Section S7 was recomputed, including grid overlap detection, seasonal binarization, and reporting ratio estimation. National reporting ratios were recalculated under each buffer scenario. The mean absolute deviation in national reporting ratio relative to baseline was computed as

$$\Delta RR = \frac{1}{N} \sum_{c=1}^{N} |RR_c^{alt} - RR_c^{base}|.$$

Across perturbations, ΔRR remained below 0.027, indicating low sensitivity to reasonable buffer-scale variation.

S11.2 Sensitivity to Seasonal Bin Definition

The primary analysis aggregated heatwave occurrences into 40 quarterly seasonal bins. To evaluate sensitivity to temporal aggregation, two alternative schemes were tested: (i) semiannual bins (20 periods) and (ii) monthly bins (120 periods). For each scheme, detection rasters were re-aggregated, spatial matching recomputed, and reporting ratios recalculated using identical masking logic.

Country-level reporting ratios under alternative bin definitions were compared against baseline quarterly aggregation. Pearson correlation coefficients between baseline and alternative schemes exceeded 0.94 in both cases, and the mean absolute difference remained below 0.031. These results indicate that reporting disparity patterns are robust to temporal bin granularity.

S11.3 Sensitivity to Heatwave Detection Percentile

The baseline heatwave detection definition uses the 90th percentile threshold with minimum duration of five consecutive days. To evaluate sensitivity to detection intensity threshold, alternative definitions were tested: (i) 95th percentile with minimum duration of two days, and (ii) 97.5th percentile with minimum duration of four days, consistent with Section S3. Severe heat-day projections and detection rasters were recomputed accordingly.

Reporting ratios were recalculated using each alternative detection threshold. National reporting ratios under alternative definitions were strongly correlated with baseline results (Pearson $r > 0.91$). Quintile classification of under-reporting status was preserved for 87% of countries under 95th percentile definition and 82% under 97.5th

percentile definition. These results demonstrate that reporting disparity conclusions are not dependent on specific detection percentile selection.

S11.4 Sensitivity to Multilingual Screening Precision

To evaluate sensitivity to multilingual LLM screening precision, the initial candidate filtering threshold (minimum two criteria satisfied) was perturbed. Two alternative thresholds were tested: (i) requiring all three criteria, and (ii) requiring only one criterion. For each configuration, the entire validation, geocoding, spatial matching, and ratio computation pipeline was re-executed.

Stricter screening reduced total validated events by 18%, while relaxed screening increased validated events by 23%. Despite changes in absolute event counts, national reporting ratios remained highly stable (Pearson $r > 0.89$ compared to baseline). Relative country ranking by reporting ratio showed Spearman rank correlation exceeding 0.92. This indicates that conclusions are robust to reasonable screening precision variation.

S11.5 Monte Carlo Perturbation of Geocoded Coordinates

To quantify spatial uncertainty propagation, Monte Carlo coordinate perturbation was performed. For each geocoded event with assigned uncertainty radius r, 100 random coordinates were sampled uniformly within the circular uncertainty region. For each Monte Carlo iteration, spatial overlap detection and reporting ratio computation were recomputed.

Let $RR_c^{(m)}$ denote the reporting ratio for country c under Monte Carlo iteration m. The coordinate-induced variance was quantified as

$$Var_c = \frac{1}{M-1} \sum_{m=1}^{M} (RR_c^{(m)} - \overline{RR}_c)^2,$$

where $M = 100$ and $\overline{RR}_c$ is the Monte Carlo mean. Across countries, the median standard deviation of RR_c under coordinate perturbation was 0.013, and no country exhibited variation exceeding 0.05. These results confirm that spatial uncertainty does not materially alter national reporting disparities.

S11.6 Bootstrap Confidence Intervals for National Reporting Ratios

To quantify statistical uncertainty in national reporting ratios, non-parametric bootstrap resampling was performed at the grid-cell level. For each country, grid cells were resampled with replacement 1,000 times. For bootstrap replicate b, reporting ratio

was recomputed as

$$RR_c^{(b)} = \frac{\sum_{i \in \mathcal{I}_c^{(b)}} R_i}{\sum_{i \in \mathcal{I}_c^{(b)}} D_i}.$$

The 2.5th and 97.5th percentiles of the bootstrap distribution defined 95% confidence intervals.

For 93% of countries, bootstrap interval width was below 0.08. Countries with wider intervals were primarily small island nations with limited grid-cell coverage. Importantly, quintile classification of under-reporting status was preserved for 90% of countries across bootstrap replicates.

Pipeline Stability Summary.

Across all perturbation experiments, including spatial, temporal, detection-threshold, screening-precision, and coordinate uncertainty variations, reporting disparity patterns remained stable in both magnitude and relative ranking. No single modeling assumption materially altered the identification of severe under-reporting regions. These robustness tests collectively demonstrate that conclusions are structurally stable and not artifacts of parameter selection or implementation details.

References

[1] Danabasoglu, G. *et al.* The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).

[2] Simpson, I. R. *et al.* An Evaluation of the Large-Scale Atmospheric Circulation and Its Variability in CESM2 and Other CMIP Models. *J. Geophys. Res.:Atmos.* **125**, e2020JD032835 (2020).

[3] Hobday, A. J. *et al.* A hierarchical approach to defining marine heatwaves. *Prog. Oceanogr.* **141**, 227–238 (2016).

[4] Grattafiori *et al.* The llama 3 herd of models (2024). Preprint at https://arxiv.org/abs/2407.21783.

[5] Liu *et al.* Deepseek-v3 technical report (2024). Preprint at https://arxiv.org/abs/2412.19437.

[6] Lange, S. *et al.* Projecting exposure to extreme climate impact events across six event categories and three spatial scales. *Earth's Future* **8**, e2020EF001616 (2020).

[7] Sun, Y. *et al.* Global supply chains amplify economic costs of future extreme heat risk. *Nature* **627**, 797–804 (2024).