
Climate Impact Assessment Requires Weighting: Introducing the *Weighted Climate Dataset*

Marco Gortan*

Faculty of Business and Economics
University of Basel
marco.gortan@unibas.ch

Lorenzo Testa*

Department of Statistics and Data Science
Carnegie Mellon University
ltesta@stat.cmu.edu

Giorgio Fagiolo

Institute of Economics and L'EMbeDS
Sant'Anna School of Advanced Studies
g.fagiolo@santannapisa.it

Francesco Lamperti

Institute of Economics and L'EMbeDS
Sant'Anna School of Advanced Studies
f.lamperti@santannapisa.it

Abstract

High-resolution gridded climate data are readily available from multiple sources, yet climate research and decision-making increasingly require country and region-specific climate information weighted by socio-economic factors. Moreover, the current landscape of disparate data sources and inconsistent weighting methodologies exacerbates the reproducibility crisis and undermines scientific integrity. To address these issues, we have developed a globally comprehensive dataset at both country (GADM0) and region (GADM1) levels, encompassing various climate indicators (precipitation, temperature, SPEI, wind gust). Our methodology involves weighting gridded climate data by population density, night-time light intensity, cropland area, and concurrent population count – all proxies for socio-economic activity – before aggregation. We process data from multiple sources, offering daily, monthly, and annual climate variables spanning from 1900 to 2023. A unified framework streamlines our preprocessing steps, and rigorous validation against leading climate impact studies ensures data reliability. The resulting *Weighted Climate Dataset* is publicly accessible through an online dashboard at <https://weightedclimatedata.streamlit.app/>.

1 Background & Motivation

Climate change is a major driver of economic and social challenges [1, 2]. Although research has documented its impacts on agriculture [3], economic production [4–6], conflict [7], income inequality [8], mortality [9], and energy consumption [10], a critical data gap persists. Climate data are often available at high resolutions (daily or hourly) and gridded formats, while socio-economic data are typically collected annually and associated with administrative units like regions or countries.

To bridge this gap, researchers typically aggregate weather data to match the lower resolution of socio-economic data. However, this aggregation can be misleading if not done with care. Simply averaging weather data across a region can mask important spatial heterogeneity. For example, average summer temperatures in the Mojave Desert may be much higher than Los Angeles, but socio-economic activity is concentrated in Los Angeles. So, labor productivity in California is likely more affected by temperatures in Los Angeles than in the desert. Therefore, researchers are increasingly using spatially weighted climate data to account for the varying distribution of socio-economic activities within a

*These authors contributed equally.

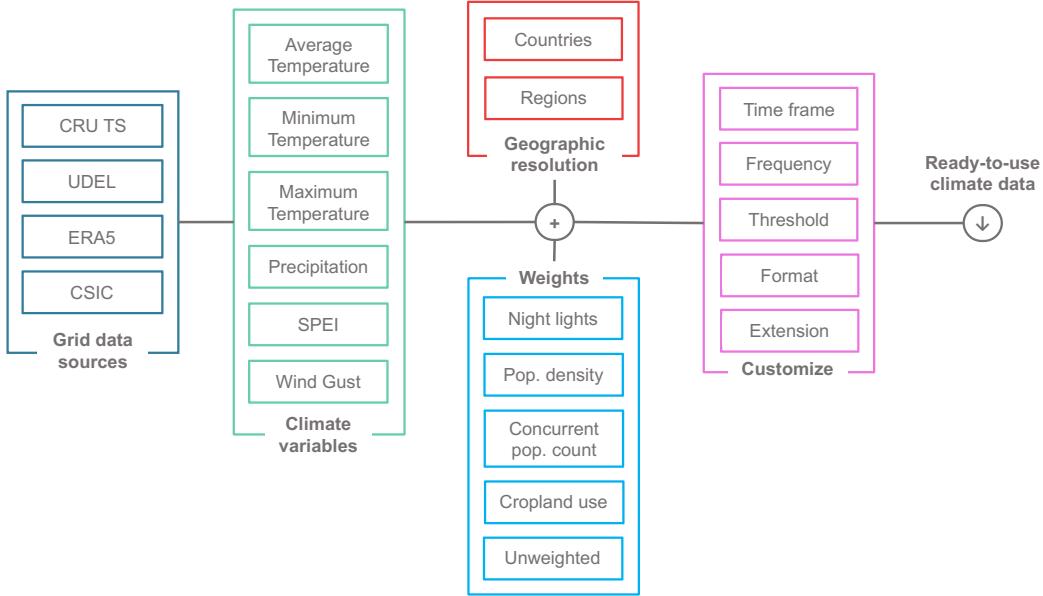


Figure 1: The *Weighted Climate Dataset* workflow combines climate data, socio-economic indicators, and administrative boundaries to produce region-specific climate variables *weighted* by socio-economic activity.

region. Weighting schemes such as population density can be used to prioritize areas with higher socio-economic activity when averaging weather data.

Despite the benefits, replicating studies using weighted climate data can be difficult. The specific weighting procedures used are often poorly documented or absent entirely in existing research. This lack of a transparent, standardized, documented, and open access source for spatially weighted climate data can lead to biased results, as the choice of weighting methodology can significantly impact analysis, hindering accurate and robust estimation of climate impacts [11]. This paper, by extending a previously published version [12], addresses this gap by introducing the *Weighted Climate Dataset*, a unified source of data that preprocesses and weights gridded climate data. We also provide a user-friendly interface at <https://weightedclimatedata.streamlit.app/> in order to explore and download ready-to-use, spatially weighted climate variables at national and sub-national levels for the period 1900-2023. This promotes:

- **Replicability:** Improved documentation and transparency in data processing;
- **Efficiency:** Saves researchers time and resources by providing pre-processed data;
- **Robustness:** Encourages testing the sensitivity of estimates to weighting choices.

By facilitating access to a standardized dataset, we aim to enhance the climate impact assessment community’s work and encourage robust and replicable research.

2 The Weighted Climate Dataset

Figure 1 outlines the construction of our dataset. We integrate gridded climate data from multiple open-access sources with gridded socio-economic indicators and administrative boundaries. Our methodology enables users to customize the weighting of climate data by various socio-economic factors to suit their specific research needs.

2.1 Sources

Climate Data. We incorporate raw gridded climate data from four established datasets commonly employed in climate impact research: Climate Research Unit Time-Series [13] (CRU TS v4.07),

available from 1901 to 2022; Consejo Superior de Investigaciones Científicas [14] (CSIC v2.7), spanning 1901 to 2020; ECMWF Reanalysis v5 [15] (ERA5), covering 1940 to 2023; and University of Delaware [16] (UDEL v5.01), ranging from 1900 to 2017. CRU TS, UDEL, and CSIC offer monthly data at a spatial resolution of $0.5^\circ \times 0.5^\circ$, whereas ERA5 provides data at a higher resolution of $0.25^\circ \times 0.25^\circ$. All sources provide *monthly* records of average temperature (in Celsius degrees, $^\circ C$) and total precipitation (in millimeters, mm), with the exception of CSIC, which uniquely offers an additional monthly climate variable: the Standardized Precipitation-Evapotranspiration Index [17], also known as SPEI (unit free). In addition to monthly data, ERA5 also provides records at the temporal resolution of *hours*. We aggregate these hourly data into *daily* values. Specifically, we calculate daily total precipitation (by taking the sum of hourly measurements), and average, minimum, and maximum temperature. Finally, ERA5 also provides 10-meter hourly wind speed measurements (measured in meters per second, m/s), which we aggregate into daily maximum wind gust.

Socio-economic Data. To capture the spatial distribution of economic and human activity, we incorporate five weighting schemes based on socio-economic variables. First, population density data are available from Columbia University’s Gridded Population of the World v4 (GPWv4) [18], and measured at 0.25° and 0.5° spatial resolutions. To accurately represent population size, we multiply population density by grid area. Second, night-time light data [19], measured in digital number values (DN) and originally available at a 30 arc-second spatial resolution (0.0083°), are incorporated after averaging to 0.25° and 0.5° resolutions. Third, cropland data are available from the History Database of the Global Environment (HYDE) [20], version 3.2, and are measured in square kilometers. These data are recorded at a spatial resolution of 5 arc-minutes (0.083°), and are incorporated after averaging to 0.25° and 0.5° resolutions. Users can select population density, night-time light, or cropland as weighting factors, with base years of 2000, 2005, 2010, or 2015. Fourth, climate data can be weighted solely by grid cell area, an option referred to as *unweighted*. Finally, for a dynamic approach, we provide a *concurrent* weighting strategy where climate variables are weighted by population counts from the beginning of the respective decade. For instance, temperatures in 1907 are weighted using population data from 1900. Population counts are derived from HYDE v3.2 [20] for 1900-2010 and GPW for 2020.

Administrative Boundaries. We utilize two administrative levels from the Global Administrative Areas Database (GADM v4.1, released July 16, 2022) [21]. The coarser GADM0 level corresponds to country boundaries, while the finer GADM1 level represents the largest subnational administrative units (e.g., states for the US).

A summary of the data sources and further information about our preprocessing pipelines can be found in the Supplementary Material.

2.2 Aggregation

We employ a general weighted aggregation scheme to calculate climate variables at the administrative unit level. The weighted value, $y_{i,t,w,T}$, for climate variable y in geographic unit i (at a specified GADM resolution) at time t using weight w measured in base year $T \in \{2000, 2005, 2010, 2015\}$ is computed as

$$y_{i,t,w,T} = \frac{\sum_{j \in J_i} a_j f_{i,j} w_{j,T} x_{j,t}}{\sum_{j \in J_i} a_j f_{i,j} w_{j,T}}, \quad (1)$$

where J_i represents the set of grid cells intersecting unit i , $f_{i,j}$ is the fraction of grid j within unit i , a_j is the area of grid j , and $x_{j,t}$ is the raw gridded climate value. Typically, the base year T is fixed, aligning with common practices in the literature [4, 22]. However, the *concurrent* weighting scheme uses a dynamic base year, setting T as the start of the decade containing t . In the *unweighted* case, $w_{j,T}$ equals 1 for all j and T . For cropland and concurrent weighting, the grid area term a_j is set to 1.

As an example of the aggregation strategy, Figure 2 shows three panels. The left and center panels display raw gridded data for night-time light intensity in 2015, and ERA5 average annual temperatures in 2015 for the contiguous US, respectively. Night-time lights are chosen as the weighting variable in this example. Figure 2, right panel, displays the resulting aggregation at GADM1 resolution and illustrates the output that users can retrieve from our dataset.

In the Supplementary Material, aiming to support the reliability and effectiveness of our approach, we validate the quality of our dataset by comparing it with those employed in two influential climate econometric exercises [4, 6].

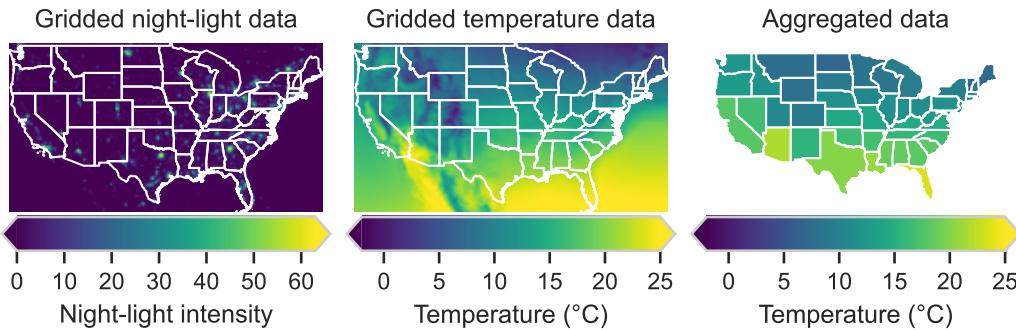


Figure 2: Climate data weighting example for the US in 2015. The left panel shows raw gridded night-time light data. The middle panel displays raw gridded average temperature. The right panel presents average temperature aggregated to the GADM1 level, *weighted* by night-time light intensity.

2.3 Customization

The *Weighted Climate Dataset* dashboard provides several ways to customize the aggregated data. Users can select specific time intervals and choose between *daily*, *monthly*, or *annual* frequencies. Notice that annual data are derived by aggregating daily or monthly values (e.g., summing monthly total precipitation). The SPEI is an exception, available only monthly due to its complex calculation.

To analyze extreme weather events, users can also define *absolute*, *relative*, and *cumulative* thresholds. The dashboard, by exploiting daily data provided by ERA5, then calculates the frequency of days exceeding these thresholds for each geographic unit, month, or year. For relative thresholds, percentiles are calculated based on the historical distribution for each region or country. For cumulative thresholds, instead of computing the frequency, the dashboard provides the sum of the residual values that exceed the threshold.

Data can be downloaded from the dashboard in two different formats: the *wide* format has geographical units as keys and values of a climate variable in different years as attributes; the *long* format has geographical units and years as keys, and the value of a climate variable as only attribute. Data can also be downloaded in three different extensions (csv, json, and parquet).

3 Discussion

The *Weighted Climate Dataset* offers a valuable resource for the climate impact assessment community, providing a standardized framework for evaluating and refining models and methodologies. By democratizing access to a comprehensive suite of datasets, we aim to propel research on the complex interplay between climate change and socio-economic dynamics. This resource allows researchers to further investigate the mechanisms through which climate variability and extremes influence economic growth, inequality, and other critical social factors. Ultimately, by enhancing our understanding of these relationships, the *Weighted Climate Dataset* can contribute to more informed decision-making, effective policy development, and public engagement in addressing the challenges posed by climate change.

Acknowledgments and Disclosure of Funding

This work is partially supported by Italian Ministry of Research, PRIN 2022 project “ECLIPTIC”. All analyses were performed on the computational resources of L’EMbeDS Department at Sant’Anna School of Advanced Studies.

References

- [1] Melissa Dell, Benjamin F Jones, and Benjamin A Olken. What do we learn from the weather? The new climate-economy literature. *Journal of Economic literature*, 52(3):740–798, 2014.

- [2] Tamara A Carleton and Solomon M Hsiang. Social and economic impacts of climate. *Science*, 353(6304):aad9837, 2016.
- [3] Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009.
- [4] Marshall Burke, Solomon M Hsiang, and Edward Miguel. Global non-linear effect of temperature on economic production. *Nature*, 527(7577):235–239, 2015.
- [5] Matthias Kalkuhl and Leonie Wenz. The impact of climate conditions on economic production. Evidence from a global panel of regions. *Journal of Environmental Economics and Management*, 103:102360, 2020.
- [6] Maximilian Kotz, Anders Levermann, and Leonie Wenz. The effect of rainfall changes on economic production. *Nature*, 601(7892):223–227, 2022.
- [7] Guy J Abel, Michael Brottrager, Jesus Crespo Cuaresma, and Raya Muttarak. Climate, conflict and forced migration. *Global environmental change*, 54:239–249, 2019.
- [8] Elisa Palagi, Matteo Coronese, Francesco Lamperti, and Andrea Roventini. Climate change and the nonlinear impact of precipitation anomalies on income inequality. *Proceedings of the National Academy of Sciences*, 119(43):e2203595119, 2022.
- [9] Tamara Carleton, Amir Jina, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Robert E Kopp, Kelly E McCusker, Ishan Nath, James Rising, Ashwin Rode, Hee Kwon Seo, Arvid Viaene, Jiacan Yuan, and Alice Tianbo Zhang. Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits. *The Quarterly Journal of Economics*, 137(4):2037–2105, 04 2022.
- [10] Maximilian Auffhammer and Erin T Mansur. Measuring climatic impacts on energy consumption: A review of the empirical literature. *Energy Economics*, 46:522–530, 2014.
- [11] Rui Wei, Yuxin Li, Jun Yin, and Xieyao Ma. Comparison of weighted/unweighted and interpolated grid data at regional and global scales. *Atmosphere*, 13(12):2071, 2022.
- [12] Marco Gortan, Lorenzo Testa, Giorgio Fagiolo, and Francesco Lamperti. A unified dataset for pre-processed climate indicators weighted by gridded economic activity. *Scientific Data*, 11(1):533, 2024.
- [13] Ian Harris, Timothy J Osborn, Phil Jones, and David Lister. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific data*, 7(1):109, 2020.
- [14] Sergio M Vicente-Serrano, Santiago Beguería, Juan I López-Moreno, Marta Angulo, and Ahmed El Kenawy. A new global 0.5 gridded dataset (1901–2006) of a multiscalar drought index: Comparison with current drought index datasets based on the Palmer drought severity index. *Journal of Hydrometeorology*, 11(4):1033–1043, 2010.
- [15] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(700):1999–2049, 2020.
- [16] Cort J Willmott and Kenji Matsuura. Terrestrial air temperature and precipitation: Monthly and annual time series (1950-1996). <https://climate.geog.udel.edu/>, 2000.
- [17] Sergio M Vicente-Serrano, Santiago Beguería, and Juan I López-Moreno. A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, 23(7):1696–1718, 2010.
- [18] Erin Doxsey-Whitfield, Kytt MacManus, Susana B Adamo, Linda Pistolesi, John Squires, Olena Borkovska, and Sandra R Baptista. Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3):226–234, 2015.

- [19] Xuecao Li, Yuyu Zhou, Min Zhao, and Xia Zhao. A harmonized global nighttime light dataset 1992–2018. *Scientific data*, 7(1):168, 2020.
- [20] Kees Klein Goldewijk, Arthur Beusen, Jonathan Doelman, and Elke Stehfest. Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth System Science Data*, 9(2):927–953, 2017.
- [21] GADM. Gadm maps and data. <https://gadm.org/>. Accessed: 2022-07-16.
- [22] Maximilian Auffhammer. Quantifying economic damages from climate change. *Journal of Economic Perspectives*, 32(4):33–52, 2018.
- [23] Thomas C Peterson and Russell S Vose. An overview of the global historical climatology network temperature database. *Bulletin of the American Meteorological Society*, 78(12):2837–2850, 1997.
- [24] Jay H Lawrimore, Matthew J Menne, Byron E Gleason, Claude N Williams, David B Wuertz, Russell S Vose, and Jared Rennie. An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116(D19), 2011.
- [25] Matthew J Menne, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. An overview of the global historical climatology network-daily database. *Journal of atmospheric and oceanic technology*, 29(7):897–910, 2012.
- [26] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.
- [27] H Hersbach et al. Technical report, 2023. ERA5 monthly averaged data on single levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. DOI: <https://doi.org/10.24381/cds.f17050d7>. Accessed: 2024-03-06.

A Additional information on sources

A.1 Climate data

CRU TS utilizes raw meteorological data collected from a vast network of weather stations. Monthly deviations from average climatic conditions, known as anomalies, are calculated and subsequently interpolated using a method called Angular-Distance Weighting (ADW) [13]. ADW is employed to account for the varying size of grid cells on the spherical Earth, specifically by considering the cosine of the latitude for each grid cell. This cosine value serves as a proxy for the change in grid cell area with respect to latitude. It is recognized that grid cells near the equator encompass larger areas compared to those closer to the poles.

CSIC leverages the CRU TS dataset to generate the Standardized Precipitation Evapotranspiration Index (SPEI), a drought index that comprehensively assesses both the severity and duration of drought conditions by incorporating information on precipitation and evapotranspiration. This index represents a standardized adaptation of the widely used Palmer Drought Severity Index (PDSI), which considers the combined influences of precipitation and temperature on water availability. Due to its multi-scalar nature, the SPEI can distinguish between different types of drought; for the present study, we concentrate on the one-month timescale to examine fluctuations in headwater levels.

ERA5 is a climate dataset that incorporates data collected by radiosondes, battery-powered instruments carried aloft by weather balloons to measure atmospheric parameters such as temperature, wind speed, and humidity. The data transmitted from these radiosondes to ground stations are integrated into the ERA5 dataset alongside observations from satellites and surface-based instruments through the utilization of numerical weather models, resulting in a comprehensive representation of the Earth’s climate system [15]. As a unique *reanalysis* product, ERA5 combines climate models with historical observations to deliver (i) consistent data over time and (ii) enhanced accuracy in grid areas without measurement stations.

Finally, UDEL provides gridded climate estimates primarily based on station records obtained from various publicly accessible sources (e.g., Global Historical Climatology Network dataset [23],

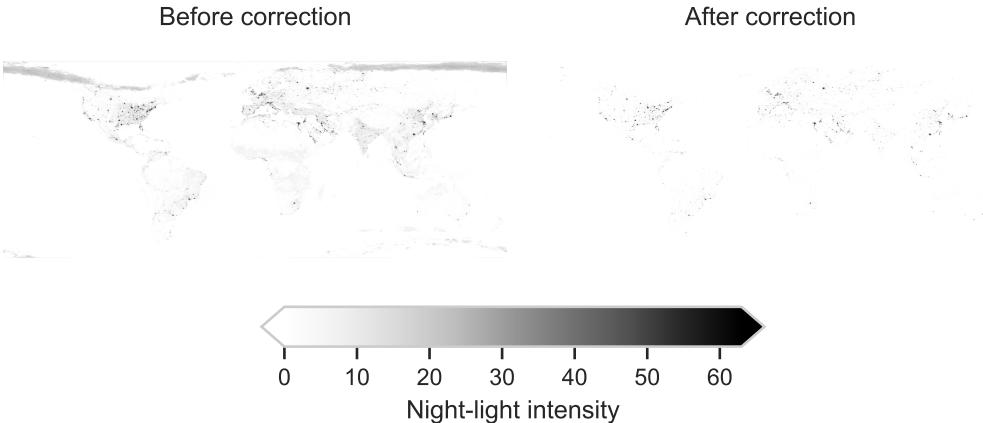


Figure 3: Correction of auroras and other noise sources in the night-time light data for the year 2015. The left plot shows night-time light data *before* correction; the right plot shows the same data *after* correction, which consists of setting to 0 the values in the grids whose digital number values are less than 30.

Global Historical Climatology Network Monthly dataset [24], Daily Global Historical Climatology Network archive [25]). The interpolation process is carried out using the Shepard spatial interpolation algorithm [26], modified to accommodate the near-spherical shape of the Earth.

A.2 Socio-economic data

Night-time light and cropland data require preprocessing prior to analysis. Night-time light data consists of digital number (DN) values, a standardized measure of pixel brightness ranging from 0 to 63. Initially available at a high-resolution of 30 arc-seconds (0.0083°), these data were aggregated to match the coarser resolutions of our gridded climate data (0.25° and 0.5°). This was accomplished by calculating the mean DN value for blocks of 900 (30×30) and 3600 (60×60) most upper-left cells in our coordinate system. We then iterate this procedure with the adjacent blocks of cells to obtain all gridded values of the night-light data for the coarser resolution. It is important to note that the harmonized VIIRS-DMSP dataset, particularly for the year 2015, contained noise from sources such as auroras and transient phenomena (e.g., boat lights, fires) as illustrated in Figure 3 (left panel). To address this, following the approach of [19], DN values below 30 were set to zero prior to aggregation. The impact of this correction is visualized in Figure 3 (right panel) for the year 2015.

Cropland data quantify the area, in square kilometers, dedicated to arable land and permanent crops within each grid cell. Originally recorded at a spatial resolution of 5 arc-minutes (0.083°), these data were aligned with the coarser resolutions of our climate datasets through an aggregation process identical to that applied to the night-time light data.

A summary of the main characteristics of the original data sources is provided in Table 1.

B Further details on aggregation

We observed inconsistencies in grid resolution across different datasets. ERA5 data are provided in a NetCDF file with a 721×1440 grid, spanning from 180.125°W to 179.875°E and 90.125°S to 90.125°N , at a 15 arc-minute resolution. In contrast, socio-economic data at 0.25° utilize a 720×1440 grid, extending from 180°W to 180°E and 90°S to 90°N . To align ERA5 data with socio-economic variables for weighting and analysis, we employed bilinear interpolation to resample the ERA5 weight grids. This process is visually illustrated in Figure 4, which compares the grid structures of the two data sources. This resampling procedure was applied consistently when combining ERA5 climate variables with population density, night-time light, cropland, and concurrent population count data.

Table 1: Summary of the main features of the employed data sources.

Source	Reanalysis	Variables	Coverage period	Frequency	Resolution	Version
CRU TS [13]	No	Average temperature, precipitation	1901–2022	Monthly	0.5°	4.07
CSIC [14]	No	SPEI 1-month	1901–2020	Monthly	0.5°	2.7
ERA5[27]	Yes	Average temperature, minimum temperature, maximum temperature, precipitation, wind gust	1940–2023	Daily	0.25°	5
UDEL [16]	No	Average temperature, precipitation	1900–2017	Monthly	0.5°	5.01
GPW [18]		Population density	2000, 2005, 2010, 2015	Yearly	0.25°	4
Li et al. (2020) [19]		Night-light intensity	2000, 2005, 2010, 2015	Yearly	0.0083°	7
HYDE [20]		Cropland	2000, 2005, 2010, 2015	Yearly	0.083°	3.2
GPW [18]		Population count	2020	Yearly	0.25°	4
HYDE [20]		Population count	1900-2010	10 years	0.083°	3.2
GADM [21]		Administrative boundaries			4.1	

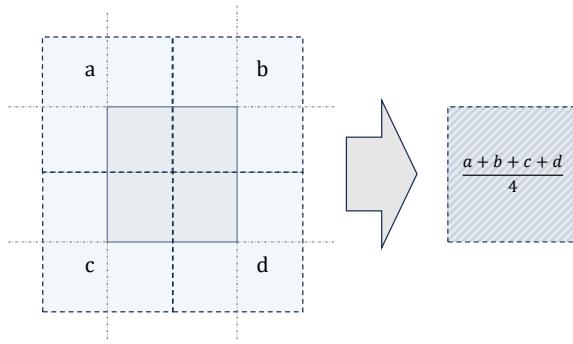


Figure 4: Stylized illustration of the resampling process applied to align weighting grids (population density, night-time light, cropland, and concurrent population count) with ERA5 climate variable grids. A slight spatial mismatch of 0.125° exists between the grids in both longitude and latitude. To ensure accurate weighting, the weighting grids were resampled using a simple averaging method of overlapping grid cell values.

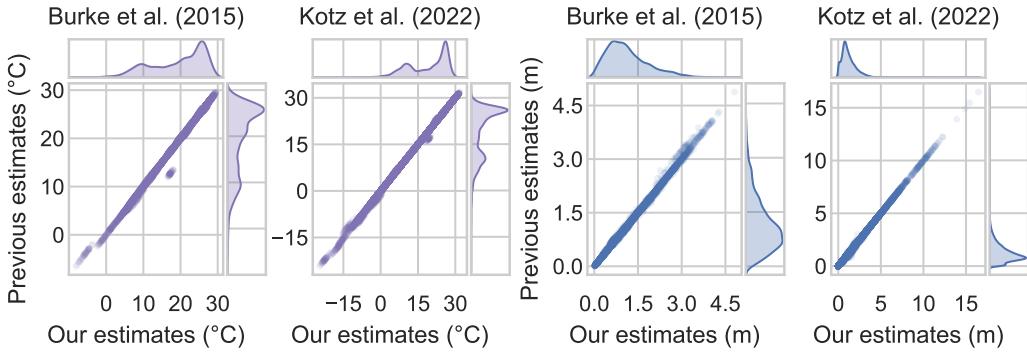


Figure 5: Comparison of weighted and/or aggregated average yearly temperature (degrees Celsius) and annual total precipitation (meters) variables in our datasets against data used in [4] (countries) and [6] (sub-national regions). Data points falling on the main diagonal indicate strong agreement between the estimates.

Both climate and socioeconomic datasets exhibited sporadic missing values. To maintain data integrity, we adopted a conservative approach. When accurate climate variable weighting was hindered due to factors such as zero-valued weights or missing climate data within specific geographic units, we opted to retain the missing values (NAs) rather than imputing estimates.

C Validation

This section validates our dataset against those used in influential climate econometric studies [6, 4]. By comparing our weighting procedures to theirs, we assess the reliability and effectiveness of our approach.

To ensure a rigorous validation, we replicated the exact data sources used in these studies, including older versions of climate and economic data. This enables us to verify the accuracy and robustness of our data processing pipeline and to objectively evaluate the quality and consistency of our estimates. [4] utilized UDEL v3.01 for precipitation and temperature data, and GPW v3 at 0.50° gridded population data for the year 2000. While the specific source and version of administrative boundaries are unspecified, their shapefiles are publicly accessible. In contrast, [6] employed 0.25° gridded ERA5 precipitation and temperature data without applying any weighting, using GADM1 v3.6 for spatial aggregation in their primary analysis.

Figure 5 presents scatterplots comparing our average temperature and rainfall estimates with those from the original studies by [4] and [6] (SPEI, wind gust, and other summary statistics of temperature data were not used in either study). Data points aligned along the main diagonal of these plots indicate strong agreement between the estimates. Our analysis reveals a high degree of concordance between our estimates and the weighted or aggregated data used by previous researchers across all study years. This supports the quality and reliability of our dataset and methodologies. However, minor discrepancies were observed. The first panel in Figure 5 highlights two key differences between our estimates and those of [4]. Lower-left quadrant points represent negative temperatures in Greenland, where our estimates are exceeded by the ones of [4]. Conversely, Bhutan exhibits slightly higher estimates in our data compared to [4]. These discrepancies primarily stem from the weighting scheme, particularly the concentrated population density in specific regions of Greenland and Bhutan.