AREAdata: a worldwide climate dataset averaged across spatial units at different scales through time

Thomas P. Smith^{1,*}, Michael Stemkovski², Austin Koontz³, and William D. Pearse¹

¹The Georgina Mace Centre for the Living Planet, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, SL5 7PY, UK

²Department of Biology & Ecology Center, Utah State University, 5305 Old Main Hill, Logan, UT 84322, USA

³Center for Tree Science, Morton Arboretum, 4100 IL-53, Lisle, IL 60532, USA

*corresponding author: Thomas P. Smith (thomas.smith1@imperial.ac.uk)

ABSTRACT

In an era of increasingly cross-discipline collaborative science, it is imperative to produce data resources which can be quickly and easily utilised by non-specialists. In particular, climate data often require heavy processing before they can be used for analyses. Here we describe AREAdata, a free-to-use online global climate dataset, pre-processed to provide the averages of various climate variables across differing administrative units (*e.g.*, countries, states). These are daily estimates, based on the Copernicus Climate Data Store's ERA-5 data, regularly updated to the near-present and provided as direct downloads from our website (<https://pearselab.github.io/areadata/>). The daily climate estimates from AREAdata are consistent with other openly available data, but at much finer-grained spatial and temporal scales than available elsewhere. AREAdata complements the existing suite of climate resources by providing these data in a form more readily usable by researchers unfamiliar with GIS data-processing methods, and we anticipate these resources being of particular use to environmental and epidemiological researchers.

Background & Summary

High quality open-source weather and climate datasets are essential for researchers across the whole spectrum of scientific fields. Various sources of these data exist, such as the Copernicus Climate Data Store (CDS)¹, Berkeley Earth², WorldClim³, CHELSA⁴, and NASA's Earth Data (<https://earthdata.nasa.gov/>). These data are generally available in a gridded (raster) format, with each climate variable estimated for a grid of latitudes and longitudes at a given spatial scale. In order to use these data, extensive processing is often required to extract climate variables for the region(s) of interest.

To aid researchers, static datasets of climate variables averaged across spatial units (*e.g.*, countries) over a particular time period have been released (*e.g.* CCKP⁵) based on these gridded datasets. Such data are useful for many applications, but they are rarely regularly updated for up-to-the-minute modelling (*e.g.*, in the COVID-19 response), do not include comparable forecast data (*e.g.*, for climate change prediction), and are in spatial units not optimised for social science (*e.g.*, countries or regions). In particular, the COVID-19 pandemic has demonstrated a need for easily accessible up-to-date climate estimates, as researchers have scrambled to assemble datasets to investigate the potential seasonality of SARS-CoV-2. The need to process these data leads to a duplication of effort as many researchers used GIS methods to quantify climate parameters across the same spatial units, using the same original data sources. This is particularly clear when considering the number of studies which have independently processed the same CDS ERA-5 data across similar (although often the same) spatial units to investigate the effects of environment on SARS-CoV-2 transmission⁶⁻¹². Such duplication of effort is not just a barrier to reproducibility, but also to the efficient allocation of resources (particularly in the face of urgent crises like COVID-19).

We developed a pipeline to aggregate state-level climate variables in the USA across time as part of our work investigating SARS-CoV-2 seasonality⁹. Here, we extend this pipeline to cover additional climate variables and spatial units, which we release as the open-source resource AREAdata: "Administrative Region Environmental Averages data". AREAdata provides global daily estimates of temperature, specific humidity, relative humidity, ultraviolet (UV) radiation, and precipitation at three different levels of spatial organisation, equivalent to countries, states and counties, from the beginning of 2020 to the near-present. Additionally, AREAdata provides population densities for the same spatial units, as well as an array of forecasted future annual mean temperatures from nine global climate models under four shared socio-economic pathway scenarios, based on the CMIP6 future climate projections¹³. These are provided as downloadable 'RDS' files for direct use in the R statistical programming language¹⁴, and as compressed tab-delimited text files for other applications.

Methods

To produce the daily climate estimates provided in AREAdata, we gather gridded rasters describing daily climate data and average these climate variables across the geographic areas of spatial units at different levels of administrative organisation.

Below, all software packages given in *italics* are R packages (version 4.1.0)¹⁴ unless otherwise specified. The code to fully reproduce this pipeline is freely available under a GPL v3.0 license and can be acquired from our *GitHub* repository (<https://github.com/pearcelab/areadata>). Continual updates of the output files as new climate data becomes available can be found on our *GitHub* project website (<https://pearcelab.github.io/areadata/>). These continual updates are automatically released monthly, however the underlying code to run these updates locally is also shared so that users can update these data to-the-day when necessary. Output files for the county-level estimates are large (>100MB), and so are released on figshare (<https://doi.org/10.6084/m9.figshare.16587311>). Data on either platform are version-controlled with dates of submission recorded and past versions archived. To ensure that those who process and release the raw data going into AREAdata are properly acknowledged, a condition of use of AREAdata is the citation of the raw data, and this information is provided on the website.

Data collection

We acquire shapefiles for worldwide administrative areas from the Global Administrative Areas (GADM) database¹⁵ at three different spatial scales: GID 0, GID 1, and GID 2. GID 0 is equivalent to countries, and (in the USA) GID 1 and GID 2 are equivalent to states and counties respectively.

We collect hourly estimates of climatic variables for the ERA-5 reanalysis from the Copernicus Climate Change Service's Climate Data Store (CDS). Temperature (K), specific humidity (kg kg^{-1} ; mass of water vapour per kilogram of moist air), and relative humidity (%) (water vapour pressure as a percentage of the air saturation value) are acquired from the pressure-levels dataset¹⁶ at 1000 hPa (*i.e.*, surface atmospheric pressure). Estimates of ultraviolet (UV) levels (J m^{-2} ; the amount of UV radiation reaching the surface) and precipitation (m; total precipitation, the accumulated liquid and frozen water falling to the Earth's surface as measured in metres of water equivalent) are acquired from the surface-level dataset¹⁷.

Global population density data are acquired from the Gridded Population of the World collection, version 4, revision 11¹⁸. These data consist of population density estimates based on national and sub-national censuses and population registers. They use a gridding algorithm to assign population densities to grid cells, and these data are provided as rasters at different scales. Here we use the 15 arc-minute resolution for consistency with the resolution of the ERA5 climate data.

Downscaled CMIP6 future climate projections are acquired from WorldClim³. CMIP6 is the 6th phase of a global climate model (GCM) inter-comparison project, coordinating the design and distribution of global climate model simulations¹³. These model simulations are typically numerically complex and thus to facilitate fast computation, the world is divided into coarse grid cells. This is not ideal for studies investigating phenomena at higher spatial scales, and thus WorldClim provides downscaled versions of future predictions from GCM outputs, at higher spatial resolutions, based on WorldClim v2.1 as baseline climate. WorldClim provides these downscaled data for nine GCMs: BCC-CSM2-MR, CNRM-CM6-1, CNRM-ESM2-1, CanESM5, GFDL-ESM4, IPSL-CM6A-LR, MIROC-ES2L, MIROC6, MRI-ESM2-0, and for four Shared Socio-economic Pathways (SSPs): 126, 245, 370 and 585.

Climate averaging pipeline

We use the Climate Data Operators program¹⁹ to compute daily means from the hourly data for each of the climate variables acquired from the CDS. We then calculate the mean value of each environmental variable across the administrative units given in each of our acquired shapefiles (*i.e.* countries, states, etc.), using the *exactextractr* R package. Specifically, we compute the mean of all grid cells fully or partially covered by the administrative unit polygon, weighted by the fraction of each cell covered by the polygon. When new climate data becomes available, these are appended to the previously extracted data to produce a single, live, updated output file for each administrative level and environmental variable combination. The data produced are simple files containing the daily climate estimates by spatial unit, *e.g.* country and by date, which we output as .RDS files for use in R and as zipped tab-delimited text files for other applications. We use an automated pipeline to produce new estimates on a monthly basis, which updates these files and automatically publishes new versions to *GitHub* and figshare (the links for which remain constant). An example of the input data for temperature from the 1st January, 2020, and the output following application of our methodology for country-level spatial units is given in Figure 1.

We use the same methods to process the gridded population density data, which we provide similarly with a single population density estimate for each spatial unit. We process annual mean temperatures from the climate forecast data, and again provide estimates by spatial unit for each combination of GCM and SSP. The population density and temperature forecast output files are static (not continually updated). Our website provides an easy interface to download these data; however, users can also run the provided code locally to make adjustments to the calculations and generate their own files.

Data Records

AREAdata can be accessed at our *GitHub* site (<https://pearselab.github.io/areadata>), which contains download links to each data file.

These are distributed both as .RDS files for use in the R statistical programming environment and as zipped tab-delimited files for other uses. The daily climate files consist of a matrix of point estimates of an environmental variable (either temperature, specific humidity, relative humidity, UV or precipitation), with rows representing each spatial unit that the variable was averaged across and columns representing the date. The population density files consist of a matrix with a single column of population density point estimates, with rows for each spatial unit. The climate forecast files consist of a matrix of point estimates for annual mean temperatures, with rows representing each spatial unit, and columns representing the combination of global climate model (GCM) and shared socio-economic pathway (SSP), and the year range of the projection. Column headers for the forecasting files follow the labelling convention <GCM>_<SSP>_<XXXX-YYYY>, where XXXX-YYYY specifies the date range of the forecast. These files are all distributed by the level of spatial organisation that the data have been averaged across (*i.e.* separate files for countries, states, counties). In the initial release, AREAdata provided daily climate estimates from 2020-01-01 to 2021-08-31.

A visualisation of the daily temperature estimates for different levels of spatial organisation is shown in Figure 2.

Technical Validation

All of the underlying daily climate data contributing to the main product of AREAdata were acquired from the CDS ERA-5 reanalysis. The reanalysis is a data assimilation technique which combines a wide range of meteorological observations with a prediction model, to produce an improved gridded observational dataset²⁰. The benefits of using the ERA-5 reanalysis in particular include its frequent updates (five days into the past) and high temporal resolution (1hr)²⁰. The data underlying the daily climate estimates in AREAdata have thus been heavily validated and are already widely used, *e.g.*^{7,8}.

The CDS ERA-5 climate data is available as gridded values at the 15 arc-minute (0.25°) scale, equating to roughly 27x27 km grid cells at the equator. Very few countries have areas smaller than this, however here we investigate whether these data are granular enough for use with the smaller spatial units. When looking at the distribution of unique grid cells spanned by each spatial unit (*i.e.*, the number of climate values that the spatial polygon covers and is thus averaged across) we find that the vast majority of countries and states span many grid cells (Figure 3). However, the county-level (GID 2) administrative units are generally much smaller with most spanning fewer than 10 unique grid cells, including the 10% of counties represented by only one climate estimate (Figure 3). This may lead to climate estimates being duplicated for the smallest areas if there are instances where two or more areas are entirely encompassed by the same grid cell, thus returning the same climate estimate. To address this, we tested whether any regions return the same climate value through time. We found no instances of this duplication within our datasets – each GID 2 region has a unique climate profile. We therefore conclude that the underlying data, although somewhat coarse, are granular enough to return distinct climate estimates at the GID-2 scale. The code to run this analysis is available in our *GitHub* repository.

Few publicly available datasets exist to provide climate estimates for spatial regions through time to validate our daily outputs against. However, the Climate Change Knowledge Portal (CCKP) maintains an open dataset of historical monthly climate estimates at the country level, which includes mean temperature⁵. The CCKP data is not based on the CDS ERA-5 climate reanalysis, but derived from the “Climatic Research Unit gridded Time Series” (CRU-TS) dataset; a monthly gridded climate dataset at a 0.5° latitude/longitude scale²¹. Although these climate estimates are derived from a more coarse underlying dataset than AREAdata, both spatially (0.5° rather than our 0.25° grid) and temporally (monthly, as opposed to our daily estimates), this does provide a potentially useful comparison. Here, we aggregate our daily temperature estimates for countries to monthly means, and compare these to the CCKP estimates for 2020. When comparing monthly estimates of mean temperature for each month in 2020 from 182 countries within the CCKP data and AREAdata outputs (*i.e.* 2,184 point estimates), we find an overall Pearson’s correlation coefficient of 0.935. There are systematic differences in the AREAdata and CCKP outputs, with AREAdata consistently producing marginally higher temperature estimates than the CCKP estimates, especially in lower temperature countries (Figure 4). The CRU-TS dataset (upon which the CCKP estimates are based) is observational, whereas the ERA-5 data (upon which AREAdata is based) is a reanalysis combining forecasting model estimates with observations to provide spatially and temporally finer-grained climate estimates. Previous work has found similar overall correlations between the CRU-TS and ERA-5 data (approx 0.9)²², however these datasets can vary significantly at particular times and in particular places^{22,23}. In particular, there is likely a warm bias in the ERA-5 reanalysis in recent years due to lesser representation of snow on top of sea ice²⁴, which may explain the deviations we observe here. Therefore, despite the expected deviations due to differences in underlying datasets, the strong correlations we observe between AREAdata and the CCKP validates our methodology by showing that our country-level estimates are largely consistent with another source of country-level climate estimates through time.

To ensure validity of the downloaded climate estimate files, we provide a file with MD5 checksum values with each monthly data release. This is a hash function computed on a file, and if two files have the same value then it is very likely that they are identical since a single change to a file will radically alter its checksum. A script to automatically check the published MD5 checksum values against the downloaded files and note any discrepancies is provided in our *GitHub* repository.

Usage Notes

The AREAdata point estimates are distributed as .RDS files for use in the R statistical programming language, which can be imported with the `readRDS()` command. For example, to import daily temperature data for countries from the main *GitHub* directory after cloning the repository: `readRDS("output/temp-dailymean-countries-cleaned.RDS")`. Alternatively, the data could be loaded directly 'live' from the website itself: `readRDS("https://github.com/pearselab/areadata/raw/main/output/temp-dailymean-countries-cleaned.RDS")`, for example as part of a script that made use of the latest daily temperatures for near-real-time modelling and forecasting. The data objects are two-dimensional matrices with rows corresponding to administrative regions and columns corresponding to dates. The values of the matrix are in the units of the environmental variable in the given file. Each administrative region is described by the codes used by the GADM, however we also supply a metadata file linking these codes to place names ("data/name-matching.csv"). Place names from this metadata file could be added to the AREAdata objects in R by transforming them into dataframes and then using the `merge()` function.

We also distribute these data as zipped tab-delimited .txt files for use in other applications. However, these could also be directly read into R in a single line if required, e.g. `read.table(unz("output/temp-dailymean-countries-cleaned.zip", "temp-dailymean-countries-cleaned.txt"), check.names = FALSE, row.names = 1)`, or using a URL (as in the example above) if desired. Note that the zip compression is less efficient than the RDS compression, resulting in larger files, so for R applications the .RDS files are recommended.

Code availability

All code underlying AREAdata is available from our team's *GitHub* repository (<https://github.com/pearselab/areadata>).

References

1. Raoult, B., Bergeron, C., Lopez Alos, A., Thepaut, J.-N. & Dee, D. Climate service develops user-friendly data store. *ECMWF Newsl.* 22–27, [10.21957/p3c285](https://doi.org/10.21957/p3c285) (2017).
2. Rohde, R. & Hausfather, Z. The Berkeley Earth Land/Ocean Temperature Record. *Earth Syst. Sci. Data Discuss.* **4**, 1–16, [10.5194/essd-2019-259](https://doi.org/10.5194/essd-2019-259) (2020).
3. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315, [10.1002/joc.5086](https://doi.org/10.1002/joc.5086) (2017).
4. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 1–20, [10.1038/sdata.2017.122](https://doi.org/10.1038/sdata.2017.122) (2017). [1607.00217](https://doi.org/10.1038/sdata.2017.122).
5. World Bank Group. Climate Change Knowledge Portal.
6. Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E. & Grenfell, B. T. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. *Science* [10.1126/science.abc2535](https://doi.org/10.1126/science.abc2535) (2020).
7. Poirier, C. *et al.* The Role of Environmental Factors on Transmission Rates of the COVID-19 Outbreak: An Initial Assessment in Two Spatial Scales. *Sci. Reports* **10**, 17002, [10.1038/s41598-020-74089-7](https://doi.org/10.1038/s41598-020-74089-7) (2020).
8. Sajadi, M. M. *et al.* Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19). *JAMA network open* **3**, e2011834, [10.1001/jamanetworkopen.2020.11834](https://doi.org/10.1001/jamanetworkopen.2020.11834) (2020).
9. Smith, T. P. *et al.* Temperature and population density influence SARS-CoV-2 transmission in the absence of non-pharmaceutical interventions. *Proc. Natl. Acad. Sci. United States Am.* **118**, e2019284118, [10.1073/pnas.2019284118](https://doi.org/10.1073/pnas.2019284118) (2021).
10. Ma, Y., Pei, S., Shaman, J., Dubrow, R. & Chen, K. Role of meteorological factors in the transmission of SARS-CoV-2 in the United States. *Nat. Commun.* **12**, 1–9, [10.1038/s41467-021-23866-7](https://doi.org/10.1038/s41467-021-23866-7) (2021).
11. Nottmeyer, L. N. & Sera, F. Influence of temperature, and of relative and absolute humidity on COVID-19 incidence in England - A multi-city time-series study. *Environ. Res.* **196**, 110977, [10.1016/j.envres.2021.110977](https://doi.org/10.1016/j.envres.2021.110977) (2021).
12. Bekbulat, B. *et al.* Changes in criteria air pollution levels in the US before, during, and after Covid-19 stay-at-home orders: Evidence from regulatory monitors. *Sci. Total. Environ.* **769**, 144693, [10.1016/j.scitotenv.2020.144693](https://doi.org/10.1016/j.scitotenv.2020.144693) (2021).

13. Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model. Dev.* **9**, 1937–1958, [10.5194/gmd-9-1937-2016](https://doi.org/10.5194/gmd-9-1937-2016) (2016).
14. R Core Team. R: A language and environment for statistical computing. (2021).
15. Global Administrative Areas. GADM database of Global Administrative Areas, version 3.6 (2020).
16. Hersbach, H. *et al.* ERA5 hourly data on pressure levels from 1979 to present, [10.24381/cds.bd0915c6](https://doi.org/10.24381/cds.bd0915c6) (2018).
17. Hersbach, H. *et al.* ERA5 hourly data on single levels from 1979 to present, [10.24381/cds.adbb2d47](https://doi.org/10.24381/cds.adbb2d47) (2018).
18. Center for International Earth Science Information Network (CIESIN). Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11, <https://doi.org/10.7927/H49C6VHW> (2018).
19. Schulzweida, U. CDO User Guide (Version 1.9.8), [10.5281/zenodo.3539275](https://doi.org/10.5281/zenodo.3539275) (2019).
20. Hersbach, H. *et al.* The ERA5 global reanalysis. *Q. J. Royal Meteorol. Soc.* **146**, 1999–2049, [10.1002/qj.3803](https://doi.org/10.1002/qj.3803) (2020).
21. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data* **7**, 1–18, [10.1038/s41597-020-0453-3](https://doi.org/10.1038/s41597-020-0453-3) (2020).
22. Dergunov, A. V. & Yakubailik, O. E. Comparative analysis of data on air temperature based on current weather data sets for 2007–2019. *IOP Conf. Series: Earth Environ. Sci.* **548**, [10.1088/1755-1315/548/3/032034](https://doi.org/10.1088/1755-1315/548/3/032034) (2020).
23. Gleixner, S., Demissie, T. & Diro, G. T. Did ERA5 improve temperature and precipitation reanalysis over East Africa? *Atmosphere* **11**, 1–19, [10.3390/atmos11090996](https://doi.org/10.3390/atmos11090996) (2020).
24. Batrak, Y. & Müller, M. On the warm bias in atmospheric reanalyses induced by the missing snow over Arctic sea-ice. *Nat. Commun.* **10**, [10.1038/s41467-019-11975-3](https://doi.org/10.1038/s41467-019-11975-3) (2019).

Acknowledgements

This work was funded by Natural Environment Research Council Grant NE/V009710/1.

Author contributions statement

WDP and TPS conceived the work, WDP, TPS and MS conducted preliminary work, TPS and AK tested the code, TPS finalised the results, and TPS wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

We declare no competing interests.

Figures & Tables

Figure 1. An overview of the processing pipeline of AREAdata and its resulting output. (A) Climate data rasters are acquired. The example figure shows daily mean temperature (°C) for 2020-01-01. These are hourly estimates that are averaged to daily estimates at a spatial resolution of 0.25x0.25°. (B) Shapefiles from GADM are used to demarcate the boundaries of different administrative units, in this case countries (GID 0). (C) The climate variable is averaged (mean) across each spatial unit. This averaging is weighted according to the relative amount of each grid cell within a region, which is of particular importance for smaller units (*e.g.*, GADM level 2 regions). (D) This is the result of applying the same methodology to county-level (GID 2) regions. Here, for visualisation, land areas with no GID 2 demarcations are shown in grey, and border lines are not shown. The output more closely resembles the original data (panel A) at this finer scale. The data are released in .RDS and zipped .txt formats every month.

Figure 2. Visualisation of output daily temperature estimates for the USA, spanning 18 months from 2020-01-01 to 2021-06-30. (A). Grey lines are the county-level (GID 2) estimates for each of the 63 counties in New York State, blue line shows the state-level (GID 1) estimates for New York State. (B). Grey lines show the state-level (GID 1) estimates for each of the 51 states of the USA (50 states plus Washington DC), red line shows the country-level (GID 0) estimate across the whole of the USA. As expected, there is much greater variation in temperature between states (B) than between counties within the same state (A).

Figure 3. Variation in the number of distinct grid cells fully or partially covered by spatial area polygons across scales in (A) countries (GID 0), (B) states (GID 1), and (C) counties (GID 2). Across all scales, most administrative areas are large enough that their climate estimates are based on multiple grid cells of the climate rasters. This shows that the underlying data are largely granular enough for our usage; however, we further tested this by ensuring regions have distinct climate profiles (see Technical Validation).

Figure 4. Comparison between country-level monthly mean temperature estimates obtained from the Climate Change Knowledge Portal (CCKP)⁵, versus the daily estimates from AREAdata aggregated into monthly means. The CCKP estimates are derived not from the CDS ERA-5 data, but from the Climatic Research Unit gridded Time Series (CRU-TS). Despite the differences in underlying data and methods, we find strong correlations between AREAdata and the CCKP temperature estimates (all $r > 0.84$). This validates our work by showing that our methodology produces sensible temperature estimates in line with an alternative dataset.