Technical Note 57

Young Lives

Identifying Climate Shocks in Young Lives Communities: Estimating Weather Conditions Using a Global Gridded Time Series

Gerald McQuade and Marta Favara

www.younglives.org.uk

MARCH 2024

Identifying Climate Shocks in Young Lives Communities: Estimating Weather Conditions Using a Global Gridded Time Series

Gerald McQuade and Marta Favara

First published by Young Lives in March 2024

Printed on FSC-certified paper from traceable and sustainable sources.

About Young Lives

Young Lives is an international study of poverty and inequality, following the lives of 12,000 children in four countries (Ethiopia, India, Peru and Vietnam) since 2001. **www.younglives.org.uk**

Young Lives is funded by UK aid from the UK Government.

The views expressed are those of the author(s). They are not necessarily those of, or endorsed by, Young Lives, the University of Oxford, the UK Foreign, Commonwealth & Development Office or other funders.

Young Lives, Oxford Department of International Development (ODID), University of Oxford,
Queen Elizabeth House, 3 Mansfield Road, Oxford OX1 3TB, UK

Tel: +44 (0)1865 281751 • Email: younglives@qeh.ox.ac.uk

Contents

The authors

Gerald McQuade is a PhD candidate in Economics at Lancaster university. He has previously worked for Young Lives as a Research Consultant and was a Visiting Researcher in 2023. His research interests are in early childhood development and skill formation, the socio-economic impacts of climate and environmental shocks, and the interaction between these two processes in low- and middle-income contexts.

Marta Favara is the Director of Research at Young Lives. She joined the Young Lives team in 2015 and since then has led the quantitative research team. Marta is a development economist whose main research interests include labour economics, education economics and behavioural economics.

Acknowledgements

This study was funded by the UK's Foreign, Commonwealth and Development Office (FCDO) (Grant no. GB-GOV-1-301108). We are grateful to Georgina Fenson (former Young Lives Data Manager) and Maria de los Angeles Molina (current Young Lives Data Manager) for their support in accessing the data and for their feedback, and Jennifer Lopez (GRADE), Catherine Porter, Alan Sanchez and participants at the 'Data matching day' in Oxford in 2023 for providing very useful comments.

Summary

This technical note describes the process of selecting, preparing and matching external climate data to Young Lives respondents' locations to derive ex-post estimates of climate conditions at a community level in regions which traditionally have poor climate data collection capacity. The note also details how to use these matched data, specifically precipitation records, to assess experiences of anomalous conditions relative to historical mean conditions for each community. Finally, it briefly details the structure and content of the publicly archived dataset, and how this dataset may be used for further research.

1. Introduction

A growing body of research examines how climatic variables such as precipitation, temperature and natural disasters (droughts, floods and storms) can influence economic and social outcomes. Work has identified links between climate shocks and health (Deschenes 2014), conflict (Burke, Hsiang and Miguel 2015), agricultural productivity (Auffhammer and Schlenker 2014) and economic growth.[1] Additionally, the impact of climate change on communities and individuals is becoming increasingly important to quantify for social researchers (Baylis, Paulson and Piras 2011), particularly in low- and middle-income contexts where vulnerability to climate change is often higher, the ability to adapt is limited (Ford, Berrang-Ford, and Paterson 2011; Ravindranath and Sathaye 2002), and the effects are heterogenous across dimensions such as wealth or gender. Of particular interest is a large body of literature that considers the link between exposure to adverse climate conditions during key stages in early life and an individual's development and human capital formation, with lasting impacts observed in outcomes in later life. Many of these studies exploit climate anomalies as plausibly exogenous shocks to the early life period.[2]

Young Lives is a longitudinal study of 12,000 young people and their families across four low- and middle-income countries (Ethiopia, India, Peru and Vietnam) that examines the causes and consequences of poverty. There have been five rounds of in-person data collection to date, with the most recent data collection occurring between 2015 and 2016. In addition, a five-call phone survey was administered in 2020–21 to gather information about the impacts of the COVID-19 pandemic on the index children when they were young adults (Favara et al. 2022). The survey follows two cohorts – a Younger Cohort of approximately 8,000 children (2,000 per country) born in 2000–01 and tracked from age 6–18 months, and an Older Cohort of approximately 4,000 children (1,000 per country) born in 1994–95, first tracked when they were between 7.5 and 8.5 years old. Understanding respondents' experiences of climate conditions within their community, and across their life, is needed to address research needs and to enhance the policy impact of research in this area, through providing greater utility to the rich demographic data collected by Young Lives.

This technical note describes the process of selecting, preparing and matching external climate data to Young Lives respondents' locations to derive ex-post estimates of conditions at a community level in regions which traditionally have poor climate data collection capacity. The note also details how to use these matched data, specifically precipitation records, to assess experiences of anomalous conditions relative to historical mean conditions for each community. Finally, it briefly details the structure and content of the publicly archived dataset.

[1] For reviews of these, see Carleton and Hsiang (2016), Castells-Quintana, Lopez-Uribe and McDermott (2018) and Dell, Jones and Olken (2014).

[2] For reviews of the broader literature on early life circumstances, see Almond and Currie (2011), Almond, Currie and Duque (2018) and Currie and Vogl (2013).

2. Identifying climate shocks

While Young Lives respondents provide self-reports of their exposure to a wide range of shocks (including climate shocks), these reports are likely subject to measurement error and recall bias (Bound, Brown, and Mathiowetz 2001), being highly dependent on the perception of the severity and impact of a shock. Therefore, respondents may systematically misreport shock frequency and intensity (Nguyen and Nguyen 2020). It is also difficult to verify and accurately determine the exact timing and intensity of self-reported shocks that happened in between survey rounds, and therefore recorded every three or four years. Hence, the validity of any such analysis is likely to be improved by using external, objective measurements on climate conditions, such as terrestrial station data.

However, suitable weather station data are limited. The ideal would be to have a well-maintained and reliable station available near all communities across all four Young Lives countries. However, weather station density in low- and middle-income countries (LMICs), particularly near the equator, is generally uneven and coverage from any one source of data is often poor (Dinku et al. 2008). Even where there is a station near a community in our sample, that station's data may be subject to a high number of missing values, often driven by poor maintenance and neglect, lack of funding, or damage, particularly in regions which have been subject to conflict or poor governance (Donaldson and Storeygard 2016). Lastly, many stations may be privately owned and the data considered proprietary (Dell, Jones and Olken 2014), or they do not have the capability to automatically upload or broadcast measurements to an online service or network, and so are likely not included in any publicly available archive which relies heavily on stations with auto-updates for near real-time observations.

There are several potential alternatives to infer the climate conditions at any point in time in the community of interest. One option is to use gridded terrestrial datasets, which employ algorithms to interpolate ground weather station data from several sources to provide point estimates or cell averages at regular latitude-longitude intervals across a region (Auffhammer et al. 2013). This is useful as it provides complete coverage for a region. However, if station density in a region is low, this can lead to a few data points being interpolated over a large area. This may not be concerning if considered at a highly aggregated geographic level, but can misrepresent the actual situation if inference is based on small geographic units (Dell, Jones and Olken 2014). Additionally, averaging over large areas will smooth out trends in climate variables, underestimating more extreme precipitation or temperature events compared to the actual values observed (Harris et al. 2020). In contrast to smoothing out both extremes, some gridded datasets may display bias in a specific direction (Ensor and Robeson 2008). Therefore, caution must be exercised in choosing a specific gridded dataset product.

Another option is to use satellite weather measurements. This provides an advantage over ground station data in areas with poor weather station density and can allow for high resolution and frequency analysis. However, this is accompanied by two major drawbacks. First, geostationary or near-polar weather satellite data only became available relatively recently, and early coverage was incomplete, which limits the range of historical data available for assessing trends. Second, satellites cannot provide actual measurements of weather data like ground stations, but must infer conditions from measurable changes to infrared wavelengths in the atmosphere or cloud coverage, which may give a limited picture of the 'ground truth' conditions (Dell, Jones and Olken 2014). Additionally, while satellite data are less susceptible to missing records as stations come in and out of service, sensors and orbits (particularly non-geostationary satellites) are subject to subtle changes over time which will require corrections to be made. While satellite products provide an excellent opportunity for observing a wide range of phenomenon, from weather to pollution, forest/habitat loss, night lights and economic development, the length of record of full coverage for climate data is relatively short and may make assessing relative changes to climate conditions in communities over time difficult. [3] In comparison, gridded datasets often provide complete time series in excess of 100 years, with particular care taken to digitise print and handwritten station readings. Given the potential issues outlined above, gridded products based on several rich data sources may provide a more accurate picture of the ground truth in some of our regions of interest. Therefore, it was decided to match a gridded terrestrial dataset.

3. University of Delaware global gridded monthly climate time series

The dataset selected for this data matching process is the Terrestrial Air Temperature and Precipitation Gridded Monthly Time Series (v5.01)[4] (Matsuura and Willmott 2018) from the University of Delaware (hereafter ‘UDel’). This dataset provides global land-based estimates of monthly total precipitation and average air temperature at regular 0.5°x0.5° intervals (roughly 50x50km at the equator) on a grid for the period 1901–2017. Data are compiled from several ground station dataset sources: the Global Historical Climatology Network (GHCN2 (precipitation) and GHCN3 (temperature), monthly versions derived from the GHCN-Daily dataset (Menne et al. 2012)); daily records from the National Centers for Environmental Information (NCEI) Global Surface Summary of the Day (GSOD); Sharon Nicholson’s African station archive (Nicholson 1979, and subsequent updates); Webber and Willmott’s South American station archive (1998); daily station records from India, derived from the National Center for Atmospheric Research (NCAR); Greenland station records from GC-Net (Steffen, Box and Abdalat 1996); and a few other sources.[5] Filters are applied to temperature and precipitation records to remove unrealistic and likely erroneous records. Monthly total precipitation and average air temperature fields are estimated using climatologically-aided interpolation (Willmott and Robeson 1995), which by using a background climatology can potentially improve the accuracy of spatial interpolated station data. Values are then interpolated into a regular grid based on a spherical version of Shepard’s inverse distance-

[3] For a review of the applications of satellite products in social and economic research, see Donaldson and Storeygard (2016).

[4] University of Delaware Terrestrial Precipitation data provided by the NOAA PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov/data/gridded/data.UDel_AirT_Precip.html: OR https://web.archive.org/web/20230331042450/http://climate.geog.udel.edu/~climate/html_pages/download.html [Wayback Machine Archive]

[5] See the README files for further details: https://web.archive.org/web/20210414125704/http://climate.geog.udel.edu/~climate/html_pages/Tropics_files/README.tropic_precip_ts.html [Wayback Machine Archive]

weighting algorithm (Shepard 1968; Willmott et al. 1985) to produce a balanced latitude-longitude grid of point estimates. The number of nearby stations which may influence a grid point is 20, increasing from 7 in early versions and improving the accuracy of grid estimates (Matsuura and Willmott 2018). Interpolation of temperature estimates is assisted by a digital elevation model which adjusts for the effect of the differing elevations of stations on temperature measures.

As mentioned previously, the accuracy of point estimates heavily depends on the spatial density of weather stations in a region. This is particularly true of precipitation, which can exhibit high spatial variation (Dell, Jones and Olken 2014). However, in the regions considered there are few alternatives. The UDel dataset performs relatively well alongside other gridded products in comparison to ground station data (Ahmed et al. 2019; Akinsanola et al. 2017; Harris et al. 2020; Nashwan and Shahid 2019), although comparison across the regions of interest is limited. One such comparison in Vietnam with the region-specific Vietnam Gridded Precipitation (VnGP) dataset did indicate that UDel underestimated precipitation patterns in north and south-central Vietnam, and performed relatively poorly in estimating high rainfall during the south-west monsoon period (Vu et al. 2018). Another limitation of the UDel dataset is that it only provides two climatic variables, monthly total precipitation and average air temperature, whereas other options such as the University of East Anglia Climatic Research Unit Time Series (UEA CRU TS) (New, Hulme and Jones 2000) can offer other derived variables such as number of wet days and cloud cover, although these are generally derived from the same base variables.

4. Young Lives community GPS data

Young Lives uses a multi-stage sentinel site sampling approach in which 20 sentinel sites in each study country were purposively selected, with households within each sentinel site with children of the correct age chosen randomly to provide around 100 Younger Cohort and 50 Older Cohort children. Sentinel sites were chosen to meet the study aims; therefore, poor areas were oversampled.[6] Generally, Young Lives country samples were not intended to be nationally representative. Nevertheless, analysis of sampling procedures shows that despite biases, Young Lives samples cover a diverse distribution of children in each country, providing an appropriate instrument for analysing causal relations and modelling the longitudinal dynamics of child welfare (Escobal and Flores 2008; Outes-León and Sánchez 2008).

Within each sentinel site (or ‘cluster’) communities are defined according to administrative areas. In practice, the boundary of a Young Lives community may be smaller, larger, or the same size as the sentinel sites and might include a whole village or a suburb in a city. Wherever possible, researchers trace the new location of children who have moved between rounds and visit them at their new address, and as such new communities may be defined in subsequent rounds. Starting from Round 2 in India and Peru, Round 3 in Ethiopia, and Round 4 in Vietnam, GPS coordinates were collected by enumerators for communities with

[6] For an overview of sampling practices, see Young Lives (2017a). Each country team used slightly different methods to deliver this semi-purposive sampling strategy. For further details, see the individual country sampling and design reports (Young Lives 2017b, 2018a, 2018b, 2018c).

three or more respondents. These are not publicly archived to protect the anonymity and confidentiality of the Young Lives respondents. The community latitude and longitude coordinates are defined as a central location or landmark, usually a main square or park, market, church, town hall or school, using handheld GPS devices. To simplify cleaning and reduce potential inconsistencies across rounds, one set of coordinates was selected for each community and applied across all rounds.

As these coordinates were provided from enumerators as simple text strings, there was significant heterogeneity in format, length and precision of coordinate, and additional characters, as well as mistakes likely arising during digital data entry. Extensive data cleaning and validation was therefore required. Using Stata 16, raw text strings were first cleaned of any non-numeric characters, and the length of numeric strings was standardised (those coordinates which were evidently incomplete or erroneous were separately queried with country data managers). Strings were then split into parts, converted to numeric and transformed to decimal degree (DD) format. Where formats were in degree-minute-second (DMS) format, this was achieved by dividing minutes by 60 and seconds by 3,600. Where coordinates were in other formats, such as degree-decimal minute, they were converted accordingly.

To validate the location of communities, further information on location was matched, including administrative region names (in English) for levels 1–3, and community name. Shapefiles for administrative region boundaries were obtained from the Database of Global Administrative Areas (GADM) and imported as separate layers to QGIS 3.20. Community GPS coordinates were imported as a points layer, and a spatial join was conducted to match points to the lowest level of administrative region for which information was available. The administrative region name as listed by Young Lives was then compared with the spatially matched region name using regular expression matching.[7] Discrepancies were identified and also queried with country data teams. Where coordinates could not be corrected, new coordinates were defined using a central landmark of the community, with the agreement of each country data manager.

A final sample of cleaned GPS coordinates was obtained for 314 communities in total across the four countries. This is an incomplete list of all community locations for several reasons. First, as GPS was only collected in Round 2 or later, some communities in Round 1 may have ceased to exist if all respondents moved. Additionally, in some cases, communities defined in Round 1 may have been split or merged prior to GPS data collection and so the community code ceased to exist. Second, respondents who moved from a Young Lives community, but not to another Young Lives community may have been designated as living in a Young Lives mini-community, depending on the number of other respondents also now living in that community, or otherwise may be missing a community code. While mini-communities were first designated from Round 2, GPS coordinates were not collected until Round 4. For confidentiality reasons, mini-communities were not included in this data matching exercise. Table 1 provides a summary of the coverage by country.

[7] Additionally, informal validation of community points was conducted by checking community names against an OpenStreetMap layer; however, this was incomplete as many communities were not named at village or town level, but instead at neighbourhood level.

Table 1. *Communities for which GPS coordinates were obtained*

	Main	GPS	Percentage
Ethiopia	28	26	92.9
India	101	96	95.0
Peru	172	158	91.9
Vietnam	36	34	94.4
Total	**337**	**314**	**93.2**

5. Estimating community rainfall and temperature conditions

To approximate rainfall and temperature experienced by the Young Lives respondents in the community where they were living, an inverse distance-weighted (IDW) interpolation algorithm is used. Generally, a community lies between any four grid points. Therefore, using GPS coordinates collected at the community level, the distance between the community centre (defined as a point of interest as discussed above) and the four nearest grid points is measured. For each point, p, a weighting, w, is calculated:[8]

$$w_p = \frac{distance_p^{-1}}{\sum_{p=1}^{4} distance_p^{-1}}$$

Where $w_p \in (0,1]$, such that the closest grid points have a greater influence on the community estimate, which is the weighted average of those four points. This provides an estimate of total rainfall at each community, c, for each month, m, of each year, y:

$$Prcp_{cmy} = w_1 Prcp_{p_1} + w_2 Prcp_{p_2} + w_3 Prcp_{p_3} + w_4 Prcp_{p_4}$$

And similarly for average temperature. Using this method, an unbroken monthly record of these climatic variables is derived for the community across the period 1901–2017.

In the following two subsections, we discuss the accuracy of the data estimates (Section 5.1.) and derive a relative measure of community rainfall to define weather anomalies (Section 5.2).

5.1 Comparison of estimates with station data

While deriving community estimates from a gridded product provides an uninterrupted time series with global coverage across the regions and time period of interest, there is some concern over accuracy – whether community estimated values reflect relatively well the actual underlying conditions in that community. Estimates are interpolated from grid points which are in turn interpolated from several underlying data sources. Where station density and temporal coverage is poor, grid points could be interpolated from stations far from the area of interest, which may provide an inaccurate picture. To briefly assess the accuracy of

[8] This was achieved in Stata using the -geonear- user-written command (Picard 2010).

community estimates, raw station data for precipitation were obtained from the Monthly Global Historical Climatology Network (GHCN-Monthly) version 2 (Peterson and Vose 1997). All stations with records of precipitation across the period between January 1994 and December 2003 were obtained for each region,[9] yielding 137 stations. As noted, temporal and spatial coverage of station data for our regions of interest in publicly accessible online archives is mixed, therefore a relatively short record of 10 years, covering the period in which most of the Young Lives sample were born, was selected for comparison. Additionally, the proportion of stations with many missing values was high. To maintain a subsample of relatively reliable stations, all stations with more than 50 per cent of monthly records missing were discarded, leaving 67 stations. The distance from each Young Lives community to the nearest station was then measured. To provide a comparable measure to the estimates interpolated from grid points, the sample was restricted to communities with stations within 50km (the furthest possible distance a community could be located from a single grid point in the UDel dataset), yielding a subsample of 24 communities: 14, 3 and 7 from Ethiopia, India and Peru respectively, and none for Vietnam (see below), with values proxied by records from six stations. Stations ranged from <1km to 50km away, with 50 per cent of communities lying within 12km of a station.

This exercise produces relatively few stations for which to proxy a small sub-sample of communities, with no stations found within 50km of a community in Vietnam. Therefore, additional precipitation station data are obtained for each of the four countries from the NCEI Global Summary of the Month (GSOM) dataset, ordered separately for each country using the NCEI's online data search.[10] This archive contains monthly summaries for a wide range of climatic variables, computed from stations included in the GHCN-Daily dataset (while this dataset is similar, it is a separate product to the GHCN-Monthly dataset with some differences, in particular records in the GHCN-Monthly dataset have been bias corrected). These data are subjected to the same restriction criteria as the above stations, with duplicates found in both datasets removed.

A final sample was obtained of 51 communities from all four countries (14, 9, 20 and 8 from Ethiopia, India, Peru and Vietnam, respectively) with precipitation values provided from 12 stations. UDel precipitation estimates for these communities were matched, with the average monthly precipitation across communities calculated for both time series using communities for which a station record was available in that month. Figure 1 gives the mean deviation in estimates and station values across all communities for each month. The absolute difference in mean estimates across the series is small across the whole period, with the mean difference between a community estimate and the nearby station record of 3.07mm.

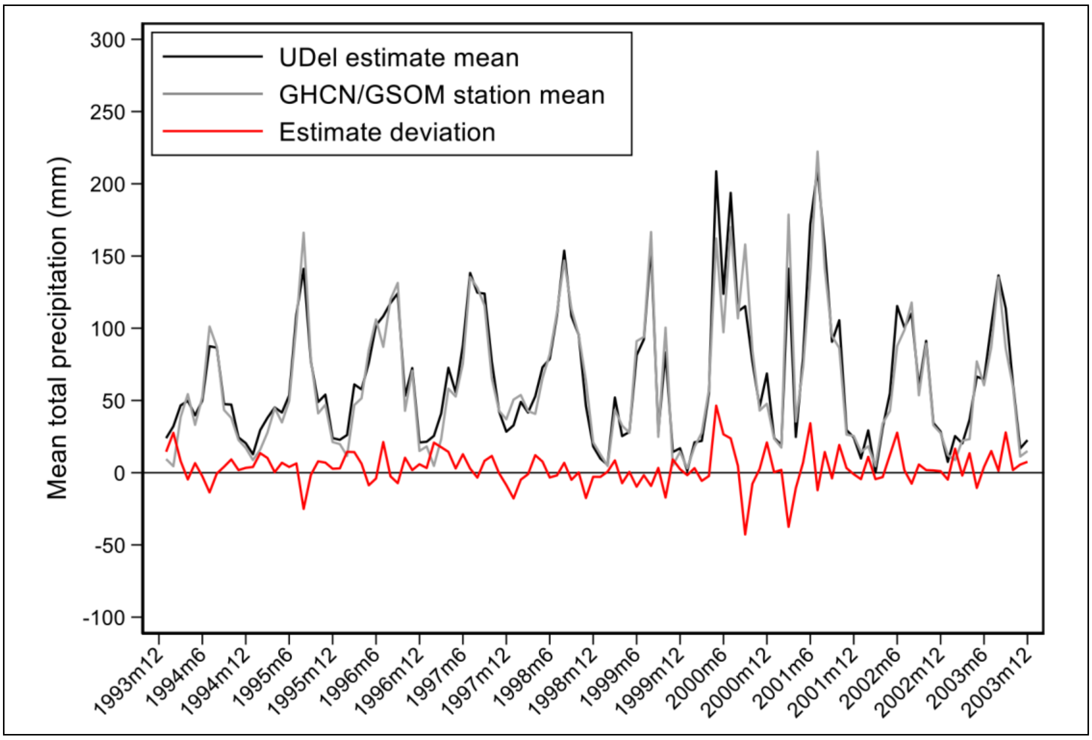
Figure 2 shows the kernel density plot of the difference in all community estimates from nearby station records. The distribution is highly focused around a zero difference, with a high level of kurtosis and a negative skew due to a small number of extreme negative differences in the far end of the tail (min/max. difference=-563.34/166.93, percentiles: 1st=-138.65, 99th=100.33). While it cannot be confirmed if all station records are fully accurate, the 10

[9] In some cases, the nearest station may be located across country borders. To allow for this, stations were first filtered within a box defined by latitude and longitude values rather than country borders. The bounds are as follows: Ethiopia: 16°N, 32°E and 2°N, 49°E; Peru: 1°N, 82°W and 19°S, 68°W; India (Andhra Pradesh and Telangana): 21°N, 76°E and 11°N, 86°E; Vietnam: 25°N, 101°E and 7°N, 110°E.

[10] The NCEI data search can be found at: https://www.ncei.noaa.gov/access/search/dataset-search.

largest underestimates (ranging -289.65 to -563.34) come from two stations: one located in northern Peru between December 1997 and March 1998, and another located in south-central Vietnam in November 1998. These periods coincide with extreme rainfall during the 1997–98 El Nino and 1998–2001 La Nina (and at the high point of local rainy season) events respectively, during which these two regions experienced extremely high rainfall over a very short period (Gobin et al. 2016; Ramírez and Briones 2017). While the differences are high, this reflects the extreme, sudden, short spikes in station values rather than extremely low estimates of community rainfall, and it cannot be ruled out that very high individual station values are simply erroneous or due to a measurement error.

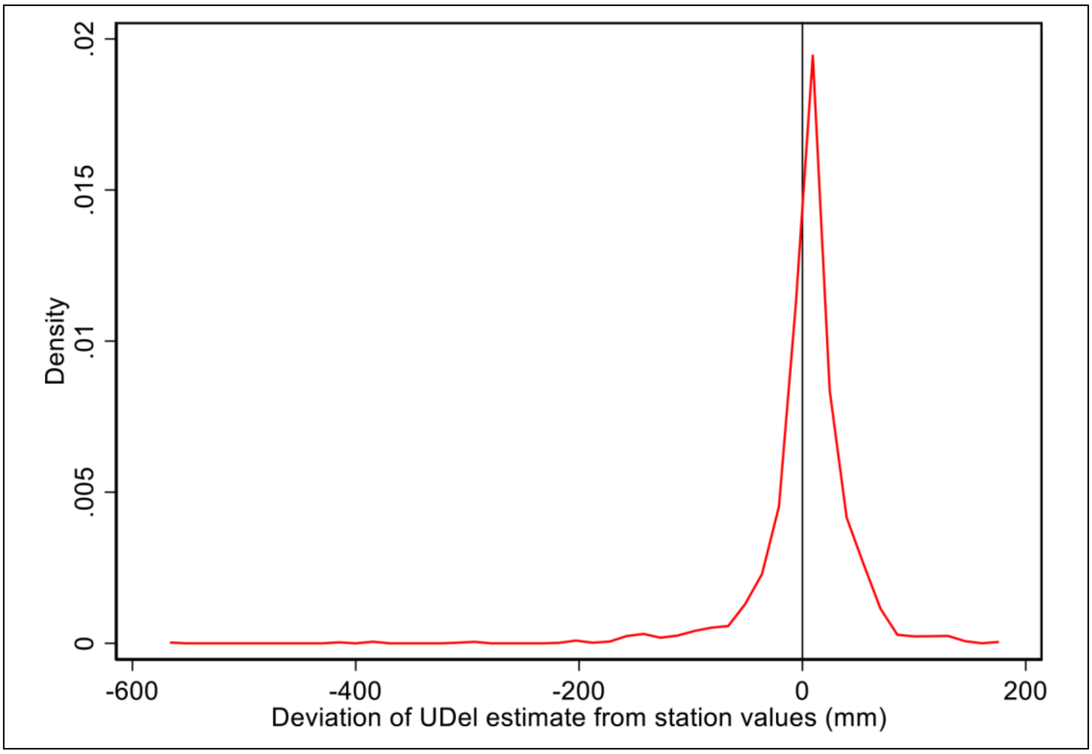
Figure 1. *Trends in mean precipitation, UDel versus GHCN/GSOM, non-missing station records*



Note: Mean rainfall across 51 communities, proxied by 12 stations from both GHCN-Monthly and GSOM datasets.

A limitation of this comparison is that both GHCN-Monthly and GSOM datasets are derived from the GHCN-Daily dataset, which is one of the station data sources for UDel. Therefore, it could be expected to see little difference in station values and estimates. However, as UDel includes other regional sources of station data, and grid estimates are influenced by a combination of nearby station values, it is unlikely to be identical. Overall, this comparison still shows that estimates derived as a weighted average of grid points, which were interpolated from underlying station data, do not on average differ significantly from real ground station data, therefore providing an at least equally good measure of ground conditions as the limited station data, while providing the advantage of an unbroken time series without missing values. Lastly, this comparison does not allow for the performance of the UDel dataset, and the subsequently derived community estimates, to be evaluated against the ‘ground truth’ conditions in areas where there are no stations. This reflects the common issue of analysis using observational data, that the counterfactual cannot be observed. Overall, these comparisons suggest that, for this subsample of communities at least, interpolated community estimates track well with nearby station data for most records – although some smoothing can be observed for short, sudden periods of extremely high precipitation, leading to underestimation of extreme rainfall events – suggesting that estimates provide a suitable measure in the relative absence of alternatives.

Figure 2. *Distribution of estimate differences, UDel vs GHCN/GSOM*



Note: Kernel = epanechnikov, bandwidth = 2.822.

5.2 Deriving a relative measure of community rainfall

Community rainfall is derived as an estimate of the absolute total rainfall for each month of each year, in each community. However, while absolute values may be suitable for assessing trends over time for a single community, absolute rainfall does not capture the relative dryness or wetness (that is, potential experience of drought or abnormally high rainfall), as what may represent abnormally low/high rainfall for one community could be perfectly acceptable for a community in a different region. As such, absolute values may not be accurate for representing anomalous weather across spatially different locations (Hayes et al. 2011), particularly across such diverse areas as the Young Lives study regions, and it would be optimal to derive a measure of conditions relative to each location. Additionally, to account for seasonality in precipitation across any given year, a relative measure should be specific to the month of the year. There are several potential relative measures of rainfall used in climatology literature and there is no consensus on a specific framework (Mishra and Singh 2010). Therefore, it was considered beyond the scope of this exercise to provide a specific relative measure, to allow as much flexibility for future users.

This section discusses one measure which may be derived by users, the Standardised Precipitation Index (SPI) (McKee, Doesken and Kleist 1993), which provides a simple measure of conditions relative to a long-term mean. An advantage of this measure is that it requires only precipitation to calculate and is computationally simplistic, unlike other measures such as the Palmer Drought Index (LLoyd-Hughes and Saunders 2002). The SPI derives a value for a month's rainfall in terms of standard deviations from the long-term mean of the transformed standardised normal distribution for that specific month of the year and community. This is preferred as, unlike a deviation from the simple long-term average, the non-negative and positively skewed nature of rainfall is accounted for prior to normalisation. The SPI is computed by fitting a suitable probability density function to the frequency distribution of precipitation summed over a time period of interest, with the probability density function then transformed into the standardised normal distribution. This is conducted for each month or period of the year at each location separately, providing a period-community specific measure of rainfall anomalies relative to long-run conditions.

As an example, we will derive a month-community specific one-month SPI, Z, by fitting a precipitation record over a period of n years to a mixed-distribution function, following a gamma distribution for non-zero precipitation. This variable will not be included in the publicly available dataset given the range of factors and modelling choices which may influence the accuracy of results for any one region (discussed below); however, example code for implementation of this worked example using Stata is provided in the Appendix.[11] Following Lloyd-Hughes and Saunders (2002), we can model a time series of non-zero precipitation using a gamma distribution with probability density function:

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

For $x > 0$ as non-zero precipitation, with shape parameter $\alpha > 0$, scale parameter $\beta > 0$, and gamma function $\Gamma(\alpha)$. In the example code snippet in the Appendix, the Stata user-written command -gammafit- (Cox and Jenkins 2003) is used to derive estimates of the parameters, $\hat{\alpha}$ and $\hat{\beta}$, using maximum likelihood. For instances which do not converge, $\hat{\alpha}$ and $\hat{\beta}$ can be approximated following Thom (1958):

$$\hat{\alpha} = \frac{1}{4A}\left(1 + \sqrt{1 + \frac{4A}{3}}\right)$$

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}$$

Where, for n observations:

$$A = \ln(\bar{x}) - \frac{\sum \ln(x)}{n}$$

Data can then be fit to the gamma distribution $g(x)$ using the -gammaden- command. To account for zero precipitation values, a mixed distribution is defined, given by:

$$H(x) = q + (1 - q)G(x)$$

[11] The SPI can be derived easily for a range of fitting distributions using the SPEI package in R (Vicente-Serrano, Beguería and López-Moreno 2010).

With $q = P(x = 0) > 0$ being the probability of zero rainfall and where integrating $g(x)$ with respect to x gives the incomplete cumulative probability function, $G(x)$:

$$G(x) = \int_0^x g(x)\,dx = \frac{1}{\hat{\beta}^{\hat{\alpha}}\Gamma(\hat{\alpha})}\int_0^x x^{\hat{\alpha}} e^{-\frac{x}{\hat{\beta}}} dx$$

Which is calculated in the worked example using the -gammap- command. An SPI, Z, can be computed by transforming the mixed distribution, $H(x)$, to the standard normal distribution. A practical method for computing SPIs for a large number of points is given by Edwards and McKee (1997) using the approximate conversions listed in Abramowitz and Stegun (1964):[12]

$$Z = \begin{cases} -\left(t - \dfrac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) & for \quad 0 < H(x) \leq 0.5 \\ +\left(t - \dfrac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) & for \quad 0.5 < H(x) < 1 \end{cases}$$

Where:

$$t \begin{cases} \sqrt{ln\left[\dfrac{1}{(H(x))^2}\right]} & for \quad 0 < H(x) \leq 0.5 \\ \sqrt{ln\left[\dfrac{1}{(1 - H(x))^2}\right]} & for \quad 0.5 < H(x) < 1 \end{cases}$$

And:

$$c_0 = 2.515517,\ c_1 = 0.802853,\ c_2 = 0.010328,\ d_1 = 1.432788,\ d_2 = 0.189269 d_3 = 0.001308$$

This is conducted for each month of the year at each location to produce a specific index for that community-month grouping. In the Stata example, this is achieved by specifying the above process as a programme and running over each grouping with the -runby- command (Picard and Schechter 2017). The derived SPI allows precipitation for each month to be expressed in terms of the number of standard deviations from the long-run mean rainfall of the standardised distribution. As such, simple indicators of relative dryness or wetness can be defined – following McKee, Doesken and Kleist (1993) and LLoyd-Hughes and Saunders (2002) – as shown in Table 2.

Table 2. *SPI categories*

SPI value	Category	Probability (%)
2.00 or more	Extremely wet	2.3
1.50 to 1.99	Severely wet	4.4
1.00 to 1.49	Moderately wet	9.2
0.99 to 0	Mildly wet	34.1
0 to -0.99	Mild drought	34.1
-1.00 to -1.49	Moderate drought	9.2
-1.50 to -1.99	Severe drought	4.4
-2.00 or less	Extreme drought	2.3

[12] See Lloyd-Hughes and Saunders (2002) for more details.

However, several factors or modelling choices can have a strong impact on the derived results of a calculated SPI. It must be assumed that the theoretical probability function chosen is appropriate for modelling monthly precipitation for the locations of interest. The above example employs a two-parameter gamma distribution as recommended by McKee, Doesken and Kleist (1993); other potential distributions include log-normal, Pearson type III, and exponential distributions (Mishra and Singh 2010). Relatedly, the goodness-of-fit of the data is directly related to the length of record used (number of month-specific observations, i.e. the number of years). Investigating the impact of different precipitation records for SPI calculation using gamma distributions, Wu et al. (2005) find significant differences stemming from differences in the estimated shape and scale parameters, affecting the shape of the gamma distribution to be fitted. McKee, Doesken and Kleist (1993) recommend using at least 30 years of high-quality observations, with an additional concern being that low-quality observations will lead to a different distribution being fitted than the true underlying theoretical distribution. Additionally, while the above example calculates one-month SPIs, when considering longer timescales, particularly in excess of six months, central limit theorem suggests that the observed probability distribution will shift towards normal; therefore, it may be more computationally efficient to model longer timescale SPIs as approximately normal (LLoyd-Hughes and Saunders 2002). Lastly, for relatively dry climates for which precipitation is strongly seasonal and zero values are common, short time frame SPIs may not be normally distributed due to a highly skewed underlying frequency distribution and the limitation of the fitted gamma distribution, potentially leading to large errors when using short precipitation records (Mishra and Singh 2010). Given these considerations, a predefined SPI is not provided in the dataset.

6. Matched dataset

Table 3 provides the variable name and description for all variables included in the publicly available dataset for each country. Climate variables are offered at the community level, with a panel data structure, such that the group variable is the anonymised unique community identifier (**COMMID**)[13] which can be used to match climate variables to Young Lives respondent data. The time component of the panel data is described uniquely by two variables, **YEAR** and **MONTH**. Hence, each community has a unique value of precipitation **PRCP** and temperature **TEMP**, for each year and month pairing. This allows the timing of a specific month's climate conditions to be identified exactly, both absolutely and relative to certain events (e.g. for the year prior to the date of interview in any round).

Table 3. *Data dictionary*

Variable	Description
COMMID	Unique anonymised Young Lives community identifier
PRCP	Monthly total precipitation estimated for the community centre point
TEMP	Monthly average air temperature estimated for the community centre point
YEAR	Year of record
MONTH	Month of record

[13] COMMID and PLACEID are the same variable and indicate the Young Lives community identifier.

7. Conclusion

This technical note details the process of selecting, preparing and matching external data for climate variables to the Young Lives respondent community locations, specifically monthly total precipitation and average air temperature, using the global gridded terrestrial time series dataset (UDel) (Matsuura and Willmott 2018). Community estimates were estimated as an inverse distance-weighted (IDW) average of the nearest four grid point estimates. Comparison of a subsample of community estimates with nearby station data, although limited, suggests that estimates derived from this interpolation, or the use of an interpolated gridded dataset, do not systematically differ from records observed by nearby ground stations, when available. This data matching widens the utility of an already rich demographic dataset, allowing the impacts of climate on children and young adults across key stages of development to be quantified.

References

Abramowitz, M., and I.A. Stegun (eds) (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Volume 55), Washington, DC: US Government Printing Office.

Ahmed, K., S. Shahid, X. Wang, N. Nawaz, and K. Najeebullah (2019) 'Evaluation of Gridded Precipitation Datasets over Arid Regions of Pakistan', *Water* 11.2: 210.

Akinsanola, A.A., K.O. Ogunjobi, V.O. Ajayi, E.A. Adefisan, J.A. Omotosho, and S. Sanogo (2017) 'Comparison of Five Gridded Precipitation Products at Climatological Scales Over West Africa', *Meteorology and Atmospheric Physics* 129.6: 669–689.

Almond, D., and J. Currie (2011) 'Killing Me Softly: The Fetal Origins Hypothesis', *Journal of Economic Perspectives* 25.3: 153–172.

Almond, D., J. Currie, and V. Duque (2018) 'Childhood Circumstances and Adult Outcomes: Act II', *Journal of Economic Literature* 56.4: 1360–1446.

Auffhammer, M., and W. Schlenker (2014) 'Empirical Studies on Agricultural Impacts and Adaptation', *Energy Economics* 46*:* 555–561.

Auffhammer, M., S.M. Hsiang, W. Schlenker, and A. Sobel (2013) 'Using Weather Data and Climate Model Output in Economic Analyses of Climate Change', *Review of Environmental Economics and Policy* 7.2: 181–198.

Baylis, K., N.D. Paulson, and G. Piras (2011) 'Spatial Approaches to Panel Data in Agricultural Economics: A Climate Change Application.',*Journal of Agricultural and Applied Economics* 43.3: 325–338.

Bound, J., C. Brown, and N. Mathiowetz (2001) 'Measurement Error in Survey Data' in *Handbook of Econometrics*, Volume 5, 3705–3843, Amsterdam: Elsevier.

Burke, M., S.M. Hsiang, and E. Miguel (2015) 'Climate and Conflict', *Annual Review of Economics* 7.1: 577–617.

Carleton, T.A., and S.M. Hsiang (2016) 'Social and Economic Impacts of Climate', *Science* 353*:* 6304.

Castells-Quintana, D., M. del P. Lopez-Uribe, and T.K.J. McDermott (2018) 'Adaptation to Climate Change: A Review Through a Development Economics Lens', *World Development* 104: 183–196.

Cox, N.J., and S.P. Jenkins (2003) 'GAMMAFIT: Stata Module to Fit a Two-parameter Gamma Distribution', Statistical Software Components, S435301.

Currie, J., and T. Vogl (2013) 'Early-Life Health and Adult Circumstance in Developing Countries', *Annual Review of Economics* 2013.5: 1–36.

Dell, M., B.F. Jones, and B.A. Olken (2014) 'What Do We Learn from the Weather? The New Climate-economy Literature', *Journal of Economic Literature* 52.3: 740–798.

Deschenes, O. (2014) 'Temperature, Human Health, and Adaptation: A Review of the Empirical Literature', *Energy Economics* 46: 606–619.

Dinku, T., S.J. Connor, P. Ceccato, and C.F. Ropelewski (2008) 'Comparison of Global Gridded Precipitation Products Over a Mountainous Region of Africa', *International Journal of Climatology* 28.12: 1627–1638.

Donaldson, D., and A. Storeygard (2016) 'The View from Above: Applications of Satellite Data in Economics', *Journal of Economic Perspectives* 30.4: 171–198.

Edwards, D.C., and T.B. McKee (1997). 'Characteristics of 20th Century Drought in the United States at Multiple Scales', *Atmospheric Science* 634: 1–30.

Ensor, L.A., and S.M. Robeson (2008) 'Statistical Characteristics of Daily Precipitation: Comparisons of Gridded and Point Datasets', *Journal of Applied Meteorology and Climatology* 47.9: 2468–2476.

Escobal, J., and E. Flores (2008) *An Assessment of the Young Lives Sampling Approach in Peru*, Young Lives Technical Note 3, Oxford: Young Lives. www.younglives.org.uk/publications (accessed 2 February 2024).

Favara, M., G. Crivello, M. Penny, C. Porter, E. Revathi, A. Sánchez, D. Scott, L.T. Duc, T. Woldehanna, and A. McKay (2022) 'Cohort Profile Update: The Young Lives Study', *International Journal of Epidemiology* 50.6: 1784–1785e.

Ford, J.D., L. Berrang-Ford, and J. Paterson (2011) 'A Systematic Review of Observed Climate Change Adaptation in Developed Nations', *Climatic Change* 106.2: 327–336.

Gobin, A., H.T. Nguyen, V.Q. Pham, and H.T.T. Pham (2016) 'Heavy Rainfall Patterns in Vietnam and their Relation with ENSO Cycles', *International Journal of Climatology* 36.4: 1686–1699.

Harris, I., T.J. Osborn, P. Jones, and D. Lister (2020) 'Version 4 of the CRU TS Monthly High-resolution Gridded Multivariate Climate Dataset', *Scientific Data* 7.1: 1–18.

Hayes, M., M. Svoboda, N. Wall, and M. Widhalm (2011) 'The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended', *Bulletin of the American Meteorological Society* 92.4: 485–488.

LLoyd-Hughes, B., and M. Saunders (2002) 'A Drought Climatology for Europe', *International Journal of Climatology* 22: 1571–1592.

Matsuura, K., and C.J. Willmott (2018) 'Terrestrial Air Temperature and Precipitation: 1900–2017 Gridded Monthly Time Series (V 5.01)', https://psl.noaa.gov/data/gridded/data.UDel_AirT_Precip.html (accessed 2 February 2024).

McKee, T.B., N. Doesken, and J. Kleist (1993) 'The Relationship of Drought Frequency and Duration to Time Scales', *8th Conference on Applied Climatology* 179–184.

Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston (2012) 'An Overview of the Global Historical Climatology Network-Daily Database', *Journal of Atmospheric and Oceanic Technology* 29.7: 897–910.

Mishra, A.K., and V.P. Singh (2010) 'A Review of Drought Concepts', *Journal of Hydrology* 391.1–2:, 202–216.

Nashwan, M.S., and S. Shahid (2019) 'Symmetrical Uncertainty and Random Forest for the Evaluation of Gridded Precipitation and Temperature Data', *Atmospheric Research* 230: 104632.

New, M., M. Hulme, and P. Jones (2000) 'Representing Twentieth-century Space-time Climate Variability. Part II: Development of 1901–96 Monthly Grids of Terrestrial Surface Climate', *Journal of Climate* 13.13: 2217–2238.

Nguyen, G., and T.T. Nguyen (2020) 'Exposure to Weather Shocks: A Comparison Between Self-reported Record and Extreme Weather Data', *Economic Analysis and Policy* 65: 117–138.

Nicholson, S.E. (1979) 'Revised Rainfall Series for the West African Subtropics', *Monthly Weather Review* 107: 620–623.

Outes-Leon, I., and A. Sánchez (2008) 'An Assessment of the Young Lives Sampling Approach in Ethiopia', Oxford: Young Lives.

Peterson, T.C., and R.S. Vose (1997) 'An Overview of the Global Historical Climatology Network Temperature Database', *Bulletin of the American Meteorological Society* 78.12: 2837–2849.

Picard, R. (2010) 'GEONEAR: Stata Module to Find Nearest Neighbors using Geodetic Distances', Statistical Software Components, S457146.

Picard, R., and C. Schechter (2017) 'RUNBY: Stata Module to Run a User's Program on By-groups of Observations', Statistical Software Components, S458413.

Ramírez, I.J., and F. Briones (2017) 'Understanding the El Niño Costero of 2017: The Definition Problem and Challenges of Climate Forecasting and Disaster Responses', *International Journal of Disaster Risk Science* 8.4: 489–492.

Ravindranath, N.H., and J.A. Sathaye (2002) *Climate Change and Developing Countries,* Berlin: Springer.

Shepard, D. (1968) 'A Two-dimensional Interpolation Function for Irregularly-spaced Data', *Proceedings of the 1968 23rd ACM National Conference*, 517–524.

Steffen, K., J.E. Box, and W. Abdalati (1996) 'Greenland Climate Network: GC-Net' in S.C. Colbeck (ed.) *Special Report 96-27 on Glaciers, Ice Sheets and Volcanoes, Tribute to Mark F. Meier*, 98–103, CRREL.

Thom, H.C.S. (1958) 'A Note on the Gamma Distribution', *Monthly Weather Review* 86.4: 117–122.

Vicente-Serrano, S.M., S. Beguería, and J.I. López-Moreno (2010) 'A Multi-scalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index – SPEI', *Journal of Climate* 23.7: 1696–1718.

Vu, T.M., S.V. Raghavan, S.Y. Liong, and A.K. Mishra (2018) 'Uncertainties of Gridded Precipitation Observations in Characterizing Spatio-temporal Drought and Wetness Over Vietnam', *International Journal of Climatology* 38.4: 2067–2081.

Webber, S.R., and C.J. Willmott (1998) 'South American Precipitation: 1960–1990 Gridded Monthly Time Series (Version 1.02)', Center for Climatic Research, Department of Geography, University of Delaware, Newark, Delaware.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe (1985) 'Statistics for the Evaluation and Comparison of Models', *Journal of Geophysical Research: Oceans* 90.5: 8995–9005.

Willmott, C.J., and S.M. Robeson (1995) 'Climatologically Aided Interpolation (CAI) of Terrestrial Air Temperature', *International Journal of Climatology* 15.2: 221–229.

Wu, H., M.J. Hayes, D.A. Wilhite, and M.D. Svoboda (2005) 'The Effect of the Length of Record on the Standardized Precipitation Index Calculation', *International Journal of Climatology* 25.4: 505–520.

Young Lives (2017a) 'A Guide to Young Lives Research: Section 5: Sampling', https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch-S5-Sampling.pdf (accessed 2 February 2024)

Young Lives (2017b) 'Young Lives Survey Design and Sampling (Round 5): United Andhra Pradesh', http://younglives.org.uk/sites/www.younglives.org.uk/files/INDIA-SurveyDesign-Factsheet-Oct17_0.pdf (accessed 2 February 2024).

Young Lives (2018a) 'Young Lives Survey Design and Sampling (Round 5) Factsheet: Ethiopia', https://www.younglives.org.uk/sites/default/files/migrated/ETHIOPIA-SurveyDesign-Factsheet-Jan18_0.pdf (accessed 2 February 2024).

Young Lives (2018b) 'Young Lives Survey Design and Sampling (Round 5) Factsheet: Peru', http://www.younglives.org.uk/sites/www.younglives.org.uk/files/PERU-SurveyDesign-Factsheet-Jan18_0.pdf (accessed 2 February 2024).

Young Lives (2018c) 'Young Lives Survey Design and Sampling (Round 5) Factsheet: Viet Nam', https://www.younglives.org.uk/sites/default/files/migrated/VIETNAM-SurveyDesign-Factsheet-Jan18_0.pdf (accessed 2 February 2024).

Appendix: SPI example Stata code snippet

First install the required packages:

```
ssc install runby, replace
ssc install gammafit, replace
```

Then specify the following program (assuming the precipitation variable in your dataset is named **PRCP**):

```
clear all
program define SPI

    //Step 1a: estimate parameters alpha and beta for monthly PRCP>0
    *****************************************************************
    capture noisily gammafit PRCP if PRCP>0, iter(2000)
    //capture required as some will not converge, causing an an error and
    //leading to them being discarded by the runby command otherwise.
    local alpha_hat = e(alpha)
    local beta_hat = e(beta)
    sort PRCP
    //sort required for -gammaden- and -gammap- commands below

    //Step 1b: for non-convergence, approx. alpha & beta by Thom (1958)
    *******************************************************************
    if `alpha_hat' == . {
        egen total = total(ln(PRCP)) if PRCP>0 //elicit sum of ln(x)
        quietly summarize total
        local total = r(max)
        //find parameter A
        quietly summarize PRCP if PRCP>0
        local A = ln(r(mean))-(`total'/r(N))
        //thom equations to approximate alpha_hat and beta_hat
        local alpha_hat = (1/(4*`A'))*(1+(1+4*`A'/3)^(1/2))
        local beta_hat = r(mean)/`alpha_hat'
        }

    //Step 2: Using these parameters, fit gamma density function to data
    ********************************************************************
    generate gammaden = gammaden(`alpha_hat',`beta_hat',0,PRCP) if PRCP>0

    //Step 3: Compute for cummulative density G(x) (excluding zeros)
    ****************************************************************
    generate gx = gammap(`alpha_hat', PRCP/`beta_hat')

    //Step 4: account for zeros and adjust cummulative density
    **********************************************************
    //want H(x) = q+(1-q)G(x), where q = P(x=0)>0 => n/N:
    //number of total observations
    quietly summarize PRCP
    local N = r(N)
    //number of zero observations
```

```
        count if PRCP == 0
        local n = r(N)
        //set H(x) to q = n/N if x = 0
        generate hx = `n'/`N' if PRCP == 0
        //otherwise we have H(x)=q+(1-q)G(x)
        replace hx = `n'/`N'+(1-(`n'/`N'))*gx if PRCP>0

        //Step 5: Compute SPI using approximation:
        ******************************************
        generate t = (ln(1/(hx^2)))^(1/2) if hx>0 & hx<=0.5
        replace t = (ln(1/((1-hx)^2)))^(1/2) if hx>0.5 & hx<1
        //generate SPI based on values given by Abramowitz and Stegun (1964)
        local c0 = 2.515517
        local c1 = 0.802853
        local c2 = 0.010328
        local d1 = 1.432788
        local d2 = 0.189269
        local d3 = 0.001308
        local exprs (`c0'+`c1'*t+`c2'*t^2)/(1+`d1'*t+`d2'*t^2+`d3'*t^3)
        generate SPI = -(t-`exprs') if hx>0 & hx<=0.5
        replace SPI = t-`exprs' if hx>0.5 & hx<1

    end
```

This will produce an approximately standard normal index which measures the deviation of precipitation relative to long-term mean of precipitation for that specific **MONTH** in that specific location (**COMMID**). Opening the rainfall dataset (replace **"DATA.dta"** as required), with panel data structure $N \times T$, where N is **COMMID** and T is described by **YEAR** and **MONTH**, we can specify the period over which we wish to define our long-term distribution of rainfall by setting the values of locals **`end**' and **`start**' based on the **YEAR** variable:

```
    use "DATA.dta", clear

    local end= XXXX
    local start= XXXX

    //set period to fit distribution over
    keep if inrange(YEAR,`start',`end')
```

The above **SPI** program can then be passed to the **runby** command, which iterates over the record for each month of the year in each community based on the length of record chosen.

```
    //loop for each commid-month grouping
    runby SPI, by(COMMID MONTH) status

    // drop intermediate variables
    capture drop gammaden gx hx t total
```

Once finished, we will have created a new variable **SPI** which provides values for every month-year period at each community; however, this is only for observations within the start and end years chosen above.

Young Lives

An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of poverty and inequality, following the lives of 12,000 children in four countries (Ethiopia, India, Peru and Vietnam). Young Lives is a collaborative research programme led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the four study countries.

Through researching different aspects of children's lives across time, we seek to improve policies and programmes for children and young people.

Young Lives Research and Policy Partners

Ethiopia

- *Policy Studies Institute*
- *Pankhurst Development Research and Consulting plc*

India (Andhra Pradesh and Telangana)

- *Centre for Economic and Social Studies, Hyderabad (CESS)*
- *Sri Padmavati Mahila Visvavidyalam (Women's University), Tirupati (SPMVV)*

Peru

- *Grupo de Análisis para el Desarollo (GRADE)*
- *Instituto de Investigación Nutricional (IIN)*

Vietnam

- *Centre for Analysis and Forecast, Viet Nam Academy of Social Sciences (CAF-VASS)*
- *General Statistics Office of Viet Nam (GSO)*

UKaid
from the British people

Contact:
Young Lives
Oxford Department of
International Development,
University of Oxford,
3 Mansfield Road,
Oxford OX1 3TB, UK
Tel: +44 (0)1865 281751
Email: younglives@qeh.ox.ac.uk
Website: www.younglives.org.uk

Young Lives