UNHCR
The UN Refugee Agency

COMPLEX RISK ANALYTICS Fund

Appendix
Description of data

DATA AND DATA SOURCES

This project aims to develop a model that predicts the risk of forced displacement from confined geographical regions. Countries in East, Central, and West Africa were chosen for this project as case regions based on the availability of displacement data, current high levels of displacement, and the expected high risk of displacement in the future as well as the rapidly changing environmental conditions. The spatial unit that the analyses and models are based upon are 0.5° grid cells, of which there are 6,225 which have geo-coordinate points intersecting with at least some land within the 25 countries of interest. Figure A.2 maps the case countries along with the embedded 0.5° grid cells.

Although displacement data are located within the 0.5° grid cells, where possible, we collect feature variable data at the 0.1° grid cell level to account for how populations and events are distributed and interact at a more localized level within the 0.5° grid cells.[1] On the temporal dimension, our data are collected on a monthly basis starting in January 2000, though some variables only have data from 2010 onwards.

Figure A.2: 0.5° grid cells within project region of interest

1 Around the equator, a 0.5° grid cell corresponds to approximately 55 times 55 kilometers and 0.1° grid cell corresponds to 11 times 11 kilometers.

DISPLACEMENT DATA

PRIMES displacement data

This project relies on displacement data from UNHCR's PRIMES Database, a case management tool for refugees, first developed in 2002. PRIMES's primary purpose is to register refugees and provides a common source of information about individuals that is used by different work units to facilitate protection activities by the organization. PRIMES has evolved into a comprehensive system that currently houses records of approximately 18 million people across more than 130 countries. The database contains detailed individual-level information, with each record representing a unique displaced person. The displacement data used for this project are displaced individuals registered in PRIMES who arrived in their country of asylum since 2000.

CLIMATE AND ENVIRONMENTAL DATA

Temperature

Climate Hazards InfraRed Temperature with Stations

Temperature data are extracted from the Climate Hazards InfraRed Temperature with Stations (CHIRTS) product. CHIRTS data offer daily 2-meter maximum temperature in degrees Celsius at a native resolution of 0.05° x 0.05°. These data are available from 01 January 1980 to near present. These temperature data points are placed inside of the 0.1° grid cells defined by the project (approximately 4 per grid cell) and averaged by grid cell and day to find the mean daily maximum temperature for each 0.1° grid cell.

As these temperature data are daily and the standard temporal resolution for the project is monthly, several techniques are used for aggregating the daily temperature data to the monthly level. The first technique was to group the data by grid cell and month and take standard aggregate measurements such as mean, maximum, minimum and standard deviation. In order to account for natural differences in the temperature among the grid cells due to geography and climate, the baseline temperature of each grid cell for each month is also calculated. This baseline temperature is the average temperature for each month for each individual grid cell for the years 1980 - 1990. For example, for each individual grid cell, the monthly temperatures for the month of January are selected from the year 1980 – 1990. The average of these figures is the January baseline temperature. The same process is done for every other month. A variable is then created which calculates the difference between the mean temperature for each month and the baseline.

As higher temperatures can hinder vegetation growth, with daily temperatures above certain thresholds being referred to as 'killing days', the second technique was to set daily temperature thresholds of 30, 35 and 38 degrees Celsius, categorized as low-killing, medium-killing and high-killing, respectively. The number of days in each month where the maximum temperature exceeds these thresholds were then counted. Similar to the raw temperature data, the average low-, medium-, and high-killing days in each month in each grid cell for the years 1980 – 1990 were calculated. These figures were set as the baseline low-, medium-, and high-killing days for each grid cell in each month. Variables were then created recording the difference between the number of killing days for each month period and the baseline killing days for the grid cells in the same month.

In addition to the temperature and killing day metrics for each individual month, we calculate the mean of each of these variables for each grid cell over the preceding 3-, 6-, and 12-month windows, inclusive of the current month.

The temperature variables developed through CHIRTS data along with brief descriptions are listed below:

MEAN_TEMP	Mean temperature (°C) among daily temperature values for that 0.1° grid cell in given month
MEAN_TEMP_DIFF	Difference between mean_temp and mean baseline temperature for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
MEAN_TEMP_P3	Average mean_temp in the current month and previous 2 months
MEAN_TEMP_P3_DIFF	Difference between the average mean_temp in the current month and previous 2 months and baseline average during that period
MEAN_TEMP_P6	Average mean_temp in the current month and previous 5 months
MEAN_TEMP_P6_DIFF	Difference between the average mean_temp in the current month and previous 5 months and baseline average during that period
MEAN_TEMP_P12	Average mean_temp in the current month and previous 11 months
MEAN_TEMP_P12_DIFF	Difference between the average mean_temp in the current month and previous 11 months and baseline average during that period
MAX_TEMP	Maximum daily temperature (°C) among temperature values for that grid cell in given month
MAX_TEMP_DIFF	Difference between max_temp and mean baseline maximum temperature for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
MAX_TEMP_P3	Average max_temp in the current month and previous 2 months
MAX_TEMP_P3_DIFF	Difference between the average max_temp in the current month and previous 2 months and baseline average during that period
MAX_TEMP_P6	Average max_temp in the current month and previous 5 months
MAX_TEMP_P6_DIFF	Difference between the average max_temp in the current month and previous 5 months and baseline average during that period

MAX_TEMP_P12	Average max_temp in the current month and previous 11 months
MAX_TEMP_P12_DIFF	Difference between the average max_temp in the current month and previous 11 months and baseline average during that period
LOW_KILLING	Number of days grid cell experienced temperature high above 30°C in given month
LOW_KILLING_ DIFFERENCE	Difference between low_killing and mean baseline low killing days (days above 30°C) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
LOW_KILLING_P3	Average number of monthly low_killing days in the current month and previous 2 months
LOW_KILLING_DIFF_P3	Difference between the average killing_killing in the current month and previous 2 months and baseline average during that period
LOW_KILLING_P6	Average number of monthly low_killing days in the current month and previous 5 months
LOW_KILLING_DIFF_P6	Difference between the average killing_killing in the current month and previous 5 months and baseline average during that period
LOW_KILLING_P12	Average number of monthly low_killing days in the current month and previous 11 months
LOW_KILLING_DIFF_P12	Difference between the average low_killing in the current month and previous 11 months and baseline average during that period
MEDIUM_KILLING	Number of days grid cell experienced temperature high above 35°C in given month
MEDIUM_KILLING_ DIFFERENCE	Difference between medium_killing and mean baseline medium killing days (days above 35°C) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
MEDIUM_KILLING_P3	Average number of monthly medium_killing days in the current month and previous 2 months

MEDIUM_ KILLING_DIFF_P3	Difference between the average medium_killing in the current month and previous 2 months and baseline average during that period
MEDIUM_KILLING_P6	Average number of monthly medium_killing days in the current month and previous 6 months
MEDIUM_ KILLING_DIFF_P6	Difference between the average medium_killing in the current month and previous 6 months and baseline average during that period
MEDIUM_KILLING_P12	Average number of monthly medium_killing days in the current month and previous 12 months
MEDIUM_ KILLING_DIFF_P12	Difference between the average medium_killing in the current month and previous 12 months and baseline average during that period
HIGH_KILLING	Number of days grid cell experienced temperature high above 38°C in given month
HIGH_KILLING_ DIFFERENCE	Difference between high_killing and mean baseline high killing days (days above 38°C) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
HIGH_KILLING_P3	Average number of monthly high_killing days in the current month and previous 2 months
HIGH_KILLING_DIFF_P3	Difference between the average high_killing in the current month and previous 2 months and baseline average during that period
HIGH_KILLING_P6	Average number of monthly high_killing days in the current month and previous 6 months
HIGH_KILLING_DIFF_P6	Difference between the average high_killing in the current month and previous 6 months and baseline average during that period
HIGH_KILLING_P12	Average number of monthly high_killing days in the current month and previous 12 months
HIGH_KILLING_DIFF_P12	Difference between the average high_killing in the current month and previous 12 months and baseline average during that period

Copernicus ERA-5 post-processed daily-statistics

In addition to the CHIRTS data, ERA5-land post-processed daily-statistics from 1950 to near-present, available are available from Copernicus. These data are at 0.1° native resolution, so each point is placed into an individual 0.1° grid cell. The Copernicus daily temperature data allow for setting temperature thresholds from the baseline period of 1950 - 1980 to calculate monthly heatwave and high temperature variables for each grid cell. From these figures, a series of heatwave, high temperature, and high temperature threshold variables are developed.

Heatwave days are calculated as the number of days in current month for a grid cell where daily temperature high is (1) above 95th percentile of the daily maxima temperature of all days during baseline period (1950 – 1980); (2) above 35°C; and (3) one of at least 3 consecutive days meeting the criteria of 1 and 2. Based on this heatwave variable, a variable for heatwave current, which is the number of heatwave days in the current month and over the previous 11 months (1 year window); and heatwave accumulated, which is the number of heatwave days in the current month and over the previous 71 months (7-year period) are developed.

High temperature days are calculated as the number of days in current month for a grid cell where daily temperature high is (1) above 99th percentile of the daily maxima temperature of all days during the baseline period. A variable for high temperature current, which is the number of high temperature days in the current month and over the previous 11 months (1 year window); and high temperature accumulated, which is the number of high temperature days in the current month and over the previous 71 months (6 year window) are created.

High temperature threshold days are also calculated, this figure is similar to the high temperature variables, but, in addition to being above the 99th percentile of the daily maxima, are also above the threshold temperature of 35°C. Similar to the heatwave and high temperature variables, high temperature threshold variables include one for high temperature threshold current, which is the number of with temperatures meeting the criteria over the previous 1 year window and high temperature threshold accumulated, which is the number of days over the previous 6 year window.

The temperature variables developed through Copernicus ERA-5 data along with brief descriptions are listed below:

HEATWAVE_DAYS	Number of days in current month for grid cell where daily temperature high is (1) above 95th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; (2) above 35°C; and (3) one of at least 3 consecutive days meeting the criteria of 1 and 2.
HEATWAVE_CURRENT	Number of days over the previous 12 months for grid cell where daily temperature high is (1) above 95th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; (2) above 35°C; and (3) one of at least 3 consecutive days meeting the criteria of 1 and 2.

HEATWAVE_ACCUMULATED	Number of days over the previous 6 years (72 months) for grid cell where daily temperature high is (1) above 95th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; (2) above 35°C; and (3) one of at least 3 consecutive days meeting the criteria of 1 and 2.
HIGH_TEMPERATURE	Number of days in current month for grid cell where daily temperature high is above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980.
HIGH_ TEMPERATURE_CURRENT	Number of days over the previous 12 months for grid cell where daily temperature high is above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980.
HIGH_TEMPERATURE_ ACCUMULATED	Number of days over the previous 6 years (72 months) for grid cell where daily temperature high is above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980.
HIGH_ TEMPERATURE_THRESHOLD	Number of days in current month for grid cell where daily temperature high is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) above 35°C.
HIGH_TEMPERATURE_ THRESHOLD_ CURRENT	Number of days over the previous 12 months for grid cell where daily temperature high is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) above 35°C.
HIGH_TEMPERATURE_ THRESHOLD_ ACCUMULATED	Number of days over the previous 6 years (72 months) for grid cell where daily temperature high is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) above 35°C.

Berkeley Earth historical temperature

Historical temperature data from Berkley Earth was also utilized. These data offer monthly average temperature from January 1850 to the near present at a resolution of 0.25° x 0.25°. Using the coordinate point location as the centroid, a 0.25° polygon was created around each point and placed the 0.1° grid cells in these polygons to find the approximate monthly temperature of each 0.1° grid cell.

For each observation, the medium-term temperature is calculated as the average temperature for each grid cell in the current month and over the preceding 119 months (10-year period). The long-term temperature is calculated as the average temperature in the current month and over the preceding 359 months (30-year period). In order to account for baseline differences in temperature among the grid cells, the average temperature for each individual month over a 10-year period (medium-term) and 30-year period (long-term) during the baseline period of 1950 to 1980 were calculated and the difference between the medium- and long-term average temperatures in each observation and the respective baseline for that month is recorded.

The temperature variables developed through Berkely Earth data along with brief descriptions are listed below:

TEMP_MEDIUM_TERM	Mean temperature over previous 10 years (120-month window) for that grid cell
TEMP_MEDIUM_TERM_DIFF	Difference between temp_medium_term for current observation and baseline temperature over medium term (mean temperature for that grid cell in that month over previous 120 months for baseline years; 1950 – 1980)
TEMP_LONG_TERM	Mean temperature over previous 30 years (360-month window) for that grid cell
TEMP_LONG_TERM_DIFF	Difference between temp_long_term for current observation and baseline temperature over long term (mean temperature for that grid cell in that month over previous 360 months for baseline years; 1950 – 1980)

Precipitation

Climate Hazards InfraRed Precipitation with Stations

Precipitation data are extracted from the Climate Hazards InfraRed Precipitation with Stations (CHIRPS) product. CHIRPS data offer daily accumulated rainfall in millimeters at a native resolution of 0.05° x 0.05°. These data are available from 01 January 1981 to near present. These precipitation data points are placed inside of the 0.1° grid cells defined by the project and averaged to find the mean daily accumulated precipitation for each 0.1° grid cell.

Similar to the temperature data, precipitation data are daily so several techniques several techniques are employed for aggregating these data to the monthly level. The first technique is to group the data by grid cell and month and take standard aggregate measurements such as mean, maximum, minimum and standard deviation. The second technique was to set daily precipitation thresholds. The number of days in each month where the accumulated precipitation falls within these thresholds are recorded; greater than 1 and less than 10 mm (low rainfall), equal to or greater than 10 and less than 30 mm (medium rainfall), and equal to or greater than 30 mm (high rainfall). In order to account for natural differences in the precipitation among the grid cells due to geography and climate, the baseline precipitation of each grid cell for each month is also calculated. This baseline precipitation is the average precipitation for each month for each grid cell for the years 1981 - 1990. Thes figures allow for calculating the difference between the mean precipitation for each month period and the baseline precipitation for the grid cells in the same month.

In addition to the precipitation and raining day metrics for each individual month, we calculate the mean of each of these variables for each grid cell over the preceding 3-, 6-, and 12-month windows, inclusive of the current month.

The precipitation variables developed through CHIRPS data along with brief descriptions are listed below:

MEAN_PREC	Mean daily precipitation (mm) for that grid cell in given month.
MEAN_PREC_DIFF	Difference between mean_prec and mean baseline precipitation for that month (baseline is monthly average for that grid cell in that month for years 1981 - 1990)
MEAN_PREC_P3	Average mean_prec in the current month and previous 2 months
MEAN_PREC_P3_DIFF	Difference between the average mean_prec in the current month and previous 2 months and baseline average during that period
MEAN_PREC_P6	Average mean_prec in the current month and previous 5 months
MEAN_PREC_P6_DIFF	Difference between the average mean_prec in the current month and previous 5 months and baseline average during that period
MEAN_PREC_P12	Average mean_prec in the current month and previous 11 months
MEAN_PREC_P12_DIFF	Difference between the average mean_prec in the current month and previous 11 months and baseline average during that period
MAX_PREC	Maximum daily precipitation (mm) among precipitation values for that grid cell in given month

MAX_PREC_DIFF	Difference between max_prec and mean baseline maximum precipitation for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
MAX_PREC_P3	Average max_prec in the current month and previous 2 months
MAX_PREC_P3_DIFF	Difference between the average max_prec in the current month and previous 2 months and baseline average during that period
MAX_PREC_P6	Average max_prec in the current month and previous 5 months
MAX_PREC_P6_DIFF	Difference between the average max_prec in the current month and previous 5 months and baseline average during that period
MAX_PREC_P12	Average max_prec in the current month and previous 11 months
MAX_PREC_P12_DIFF	Difference between the average max_prec in the current month and previous 11 months and baseline average during that period
LOW_RAINING	Number of days grid cell experienced precipitation between 1 and 10 mm in given month
LOW_KILLING_ DIFFERENCE	Difference between low_raining and mean baseline low raining days (days between 1 and 10 mm of precipitation) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
MEDIUM_RAINING	Number of days grid cell experienced precipitation between 10 and 30 mm in given month
MEDIUM_RAINING_ DIFFERENCE	Difference between medium_raining and mean baseline medium raining days (days between 10 and 30 mm of precipitation) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)
HIGH_RAINING	Number of days grid cell experienced precipitation above 30 mm in given month
HIGH_KILLING_ DIFFERENCE	Difference between high_raining and mean baseline high raining days (days above 30 mm of precipitation) for that month (baseline is monthly average for that grid cell in that month for years 1980 - 1990)

Copernicus ERA5-land hourly precipitation

In addition to the CHIRPS data described above, ERA5-land hourly precipitation data from 1950 to near-present available from Copernicus are used. The data record the accumulated data precipitation at each hour at 0.1° native resolution. The accumulated precipitation level at the last hour of the day, which is the total daily accumulated precipitation at each point are extracted. The Copernicus precipitation data allow setting precipitation thresholds from the baseline period of 1950 – 1980 to calculate the heavy precipitation variables. Heavy precipitation days are calculated as the number of days in current month for a grid cell where accumulated precipitation is (1) above 99th percentile of the daily maxima accumulated precipitation for all days during the baseline period and (2) at least 1 mm of accumulated precipitation.

In addition to a monthly heavy precipitation variable, we created variables for heavy precipitation over the last 2-, 3- and 6- month windows; heavy precipitation current, which is the number of heavy precipitation days in the current month and over the previous 11 months (1 year period); and heavy precipitation accumulated, which is the number of heavy precipitation days in the current month and over the previous 71 months (6 year period).

The precipitation variables developed through Copernicus ERA-5 data along with brief descriptions are listed below

HEAVY_PRECIPITATION	Number of days in current month for grid cell where daily accumulated precipitation is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) at least 1 mm of precipitation
HEAVY_ PRECIPITATION_CURRENT	Number of days over the previous year (12 months) for grid cell where daily accumulated precipitation is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) at least 1 mm of precipitation
HEAVY_PRECIPITATION_ ACCUMULATED	Number of days over the previous 6 years (72 months) for grid cell where daily accumulated precipitation is (1) above 99th percentile of the daily maxima temperature of all days during baseline period of 1950 – 1980; and (2) at least 1 mm of precipitation

Normalized Difference Vegetation Index

We extract Normalized Difference Vegetation Index (NDVI) is an index used to approximate the level of vegetation at a point in time. NDVI is calculated from spectral band data from satellites, specifically red and near-infrared (NIR) spectral bands. The level of NDVI is equivalent to:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

NDVI values fall between −1 and 1, with values of 0 and less indicating no vegetative growth. Monthly NDVI measurements are extracted from NASA's Earthdata Database at a resolution of 0.05° ×0.05°, which are available from January 2000 to near present. These NDVI data points are placed inside the 0.1° grid cells defined by the project and averaged to find the mean monthly NDVI for each 0.1° grid cell.

Similar to temperature and precipitation, there is a need to account for baseline differences in vegetation, irrespective of climate change. Baseline NDVI is the average NDVI for each month for each grid cell for the years 2000 - 2010. These figures allow for calculating the difference between the NDVI for each month period and the baseline NDVI for the grid cells in the same month.

The NDVI variables developed along with brief descriptions are listed below:

NDVI	Normalized difference vegetation index (NDVI) is an index used to approximate the level of vegetation at a point in time. Calculated through satellite spectral band data: (near-infrared - red) / (near-infrared + red)
MEAN_NDVI_DIFF	Difference between ndvi and mean baseline ndvi for that month (baseline is monthly average for that grid cell in that month for years 2000 - 2010)

Peak precipitation and NDVI months

Using the average monthly precipitation from the baseline period of 1980 – 1990, months in which there was a peak in average precipitation were identified through SciPy's find_peaks function, which finds local maxima by comparing neighbouring values. A peak month represents a month with heightened levels of precipitation on average. The primary peak precipitation month for each grid cell was first identified. Secondary peak months were also identified but must have been at least 10% of the primary peak precipitation level, to avoid random deviations. With each month being labelled as a primary precipitation peak or secondary precipitation peak or not, each observation was recorded with the time in months since the last precipitation peak month. This exact process was used for finding NDVI primary and secondary peak months based on the baseline period of 2000 – 2010.

From these monthly peak identifier and time since last peak variables, a series of variables records the level of precipitation, heavy precipitation, and high temperature in the months relative to the last peak months. These variables carry the tag 'lag' and a number in their title, the number stands for the number of months prior to the relative peak month.

The peak variables along with brief descriptions are listed below:

PEAK_PREC_PRIM	Binary variable with 1 if that month is the month in which the baseline period experienced the highest average precipitation, and 0 otherwise.
PEAK_PREC_SEC	Binary variable with 1 if that month is the month in which the baseline period experienced a peak in average precipitation at least 10% of the primary precipitation peak month average, and 0 otherwise.
SINCE_PREC_PEAK	Number of months since either a peak_prec_prim or peak_prec_sec month.
SINCE_PREC_PRIM_PEAK	Number of months since a peak_prec_prim month.
PEAK_NDVI_PRIM	Binary variable with 1 if that month is the month in which the baseline period experienced the highest average NDVI, and 0 otherwise.
PEAK_NDVI_SEC	Binary variable with 1 if that month is the month in which the baseline period experienced a peak in average NDVI at least 10% of the primary NDVI peak month average, and 0 otherwise.
SINCE_NDVI_PEAK	Number of months since either a peak_ndvi_prim or peak_ndvi_sec month.
SINCE_NDVI_PRIM_PEAK	Number of months since a peak_ndvi_prim month.
PREC_LAST_PREC_PEAK_LAG0	Precipitation difference from the baseline (mean_prec_diff) in the last precipitation peak month
PREC_LAST_PREC_PEAK_LAG1	Precipitation difference from the baseline (mean_prec_diff) in the month prior to the last precipitation peak month
PREC_LAST_PREC_PEAK_LAG2	Precipitation difference from the baseline (mean_prec_diff) in the month two months prior to the last precipitation peak month
PREC_LAST_PREC_PEAK_LAG3	Precipitation difference from the baseline (mean_prec_diff) in the month three months prior to the last precipitation peak month
PREC_LAST_PREC_PEAK_LAG4	Precipitation difference from the baseline (mean_prec_diff) in the month four months prior to the last precipitation peak month

PREC_LAST_ PREC_PEAK_LAG5	Precipitation difference from the baseline (mean_prec_diff) in the month five months prior to the last precipitation peak month
HEAVY_ PRECIPITATION_LAST_ PREC_PEAK_LAG0	Number of heavy precipitation days in the last precipitation peak month
HIGH_ TEMPERATURE_LAST_ NDVI_PEAK_LAG1	Number of high temperature days in the month prior to the last NDVI peak month
HIGH_ TEMPERATURE_LAST_ NDVI_PEAK_LAG2	Number of high temperature days in the month two months prior to the last NDVI peak month
HIGH_ TEMPERATURE_LAST_ NDVI_PEAK_LAG3	Number of high temperature days in the month three months prior to the last NDVI peak month
HIGH_ TEMPERATURE_LAST_ NDVI_PEAK_LAG4	Number of high temperature days in the month four months prior to the last NDVI peak month
HIGH_ TEMPERATURE_LAST_ NDVI_PEAK_LAG5	Number of high temperature days in the month five months prior to the last NDVI peak month

Standardized Precipitation Evapotranspiration Index

Standardized Precipitation Evapotranspiration Index (SPEI) is an index used to approximate the level of dryness and drought at a point in time. SPEI levels below −1 signify a dry area and potential drought conditions. The SPEI Database offers historical data on SPEI levels at a resolution of 0.5° x 0.5° and available at monthly intervals with measurements of average SPEI over the previous 1 to 48 months. The historical data are available from January 1901 to December of the previous year. In other words, in 2025, monthly historical SPEI data is available up to December 2024.

In order to update the analyses and models on a monthly basis, this project uses data from the SPEI Global Drought Monitor, which offers data on monthly SPEI levels with a 1.0° spatial resolution at a 2-month delay. SPEI-03 is used, which is the average SPEI level of the previous 3-month rolling window at each location. Using the coordinate point location as the centroid, a 1.0° polygon was created around each point and placed the 0.1° grid cells in these polygons to find the approximate monthly SPEI of each 0.1° grid cell.

The SPEI/drought variables along with brief descriptions are listed below:

SPEI	This data is based on SPEI-03 (level of SPEI over previous 3 months)
DROUGHT	Binary variable, with 1 if spei is equal to or less than -1 and 0 if greater than -1
DROUGHT_LEVEL	spei * (-1 * drought)
DROUGHT_CURRENT	Sum of drought_level over previous 12-month window
DROUGHT_ACCUMULATED	Sum of drought_level over previous 7-year window (72 months)

RESOURCE AND GEOGRAPHIC DATA

Landcover

Land cover data is extracted at 0.05° x 0.05° resolution point locations from Copernicus. There are 38 unique land cover classifications in the raw data. In order to make these data more usable, the points are recategorized into one of 8 general categories of: tree cover, bare area, shrubland, cropland, grassland, urban, cover flooded and water. The individual points are placed inside of 0.1° grid cells. Each grid cell is labelled with the modal classification of the points that lie within it and also labelled with the proportion of each landcover classification of each of the points within it.

In order to account for urban areas covering a smaller geographical area, we extracted polygon coordinate data on the locations of urban zones from Natural Earth. Any 0.1° grid cell that had any overlap with an urban zone was labelled as 'urban'.

The landcover variables along with brief descriptions are listed below:

LC_MODAL	Classification of each 0.1° grid cell as: 'urban', 'shrubland', 'cropland', 'water', 'tree_cover', 'cover_flooded', 'grassland' or 'bare_areas' based on the model identification of the points within each grid cells. But, as urban areas are geographcially smaller, grid cells are identified as urban if any part of the grid cell's polygon overlaps with an urban area polygon.
BARE_AREA	Proportion of 0.1° grid cell which is designated bare_areas
COVER_FLOODED	Proportion of 0.1° grid cell which is designated cover_flooded
CROPLAND	Proportion of 0.1° grid cell which is designated cropland
GRASSLAND	Proportion of 0.1° grid cell which is designated grassland
SHRUBLAND	Proportion of 0.1° grid cell which is designated shrubland
TREE_COVER	Proportion of 0.1° grid cell which is designated tree_cover
URBAN	Proportion of 0.1° grid cell which is designated urban
WATER	Proportion of 0.1° grid cell which is designated water

Agro-ecological zone

Agro-ecological zone (AEZ) data is extracted from the International Food Policy Research Institute, which are coordinate polygons of the various zones. AEZ classifications for Africa have three dimensions: major climate zone (tropics or subtropics), moisture zones (water availability) and highland/lowland (warm or cool based on elevation). Each of the 0.1° grid cells are labelled with an AEZ categorization based on where the grid cell centroids lie.

Pastoral areas and populations are of particular interest for studying climate change and changing human mobility. As pastoral grazing generally takes place in the AEZ zones defined as tropic warm-arid and tropic warm-semiarid, we labelled each grid cell with 4 pastoral categories: pastoral grazing, if the AEZ categorization was tropic warm-arid or tropic warm-semiarid and the land cover classification was grassland or shrubland; pastoral cropland, if the AEZ categorization was tropic warm-arid or tropic warm-semiarid and the land cover classification was cropland; pastoral water, if the AEZ categorization was tropic warm-arid or tropic warm semiarid and the land cover classification was water; and pastoral urban, if the AEZ categorization was tropic warm-arid or tropic warm-semiarid and the land cover classification was urban.

The agro-ecological zone variables along with brief descriptions are listed below:

AGEC_ZONE	Classification of each 0.1° grid cell based on its agro-ecological zone: 'tropic_warm_semiarid', 'tropic_warm_arid', tropic_warm_subhumid', 'subtropic_warm_arid', 'tropic_warm_humid', 'tropic_cool_arid', 'tropic_cool_subhumid', 'tropic_cool_humid', 'subtropic_cool_arid', 'subtropic_warm_semiarid', 'tropic_cool_semiarid'
PASTORAL_GRAZING	Binary classifier of 0.1° grid cell. 1 if grassland or shrubland value for grid cell is above 0 (proportional landcover) and agec_zone is grid cell is tropic_warm_arid or tropic_warm_semiarid; 0 otherwise
PASTORAL_CROPLAND	Binary classifier of 0.1° grid cell. 1 if cropland value for grid cell is above 0 (proportional landcover) and agec_zone is grid cell is tropic_warm_arid or tropic_warm_semiarid; 0 otherwise
PASTORAL_WATER	Binary classifier of 0.1° grid cell. 1 if water value for grid cell is above 0 (proportional landcover) and agec_zone is grid cell is tropic_warm_arid or tropic_warm_semiarid; 0 otherwise

Elevation

Elevation data is taken from the HarvestChoice CELL5M Database. The data gives the elevation, in meters, at coordinate points at a resolution of 5 arc minutes (approximately 9.30 km). The coordinate points are placed inside of the 0.1° grid cells. If multiple elevation points lie within a single grid cell, the average is taken.

The elevation variable along with brief descriptions is listed below:

ELEVATION	Mean elevation of 0.1° grid cell in meters

River

Raw data from Natural Earth, which gives the linestring geometry of the locations of the centre of rivers. Binary variable for each 0.1° grid cell, 1 if the polygon geometry of the grid cell intersects with the linestring geometry of a river, and 0 if not.

The river variable along with brief descriptions is listed below:

RIVER	Binary classifier of 0.1° grid cell. 1 if grid cell polygon intersects with Natural Earth river linestring; 0 otherwise

Road

Data on the geographical locations of roads was developed by Humanitarian OpenStreetMap. Each 0.1° grid cell is labelled based on the presence of at least some of the grid cell area encompassing a paved road or unpaved road, if neither of these are present, then the grid cell is labelled as no-road.

The road variable along with brief descriptions is listed below:

ROAD_SURFACE	Categorical value for each 0.1° grid cell: ‘paved’ if at least part of the grid cell’s geometry overlaps with the paved road geometry; ‘unpaved’ is part of the grid cell overlaps with an unpaved road and no part overlaps with a paved road; ‘no_road’ is the grid cell does not overlap with either a paved or unpaved road.

Market access and subsistence indices

Using data from the International Food Policy Research Institute, each grid cell is labelled with the time in hours to the nearest market within towns of various size: 20,000, 50,000, 100,000, 250,000 and 500,000. Using these 5 variables, which show the time to a market town of various sizes, the 'market access index' (MAI) was derived. The variable sums the travel times to different town sizes (500k, 250k, etc.) weighted by the inverse of the population size (so closer big towns count more). The measure is then log-transformed to reduce skew, normalized to a 0–1 range, and inverted so that higher final mai values indicate better access to large population centres (shorter travel times to bigger towns). This market access index was used in combination with the 2019 neighbourhood wealth data developed for the paper, "Temporal neighborhood-level material wealth maps of Africa (1990-2019)" (Pettersson, 2023), data available from Harvard Dataverse. The MAI is multiplied by the 2019 neighbourhood wealth level, this value is then scaled and inverted. The resulting subsistence index has higher values for more remote and poor grid cell and lower values for more connected and wealth grid cells.

The market access and subsistence indices variables along with brief descriptions are listed below:

MAI	Market Access Index Index derived from the amount of time from each grid cell to the nearest market in towns of various population sizes
SUBSISTENCE_INDEX	Subsistence Index = 1 – (MAI * 2019 neighborhood wealth) Index uses the Market Access Index, derived above and multiplies the MAI by the mean Neighborhood Wealth of grid cell from 2019. This figure is scaled and subtracted from 1.

DEMOGRAPHIC DATA

Population density

Chapter 3.1 details the data and processes undertaken for deriving the prediction of monthly population within each 0.1° grid cell. The predictions are built on annual population predictions from LandScan, which have a native resolution of 1,000 meters. These points are placed inside of grid cells and summed to find the annual population with each 0.1° grid cell.

Monthly nightlight radiance data from the Defense Meteorological Satellite Program (DMSP) – Operational Linescan System (OLS) Nighttime Lights Time Series (2000 – 2013) and the Visible Infrared Imaging Radiometer Suite (VIIRS) (2014 – present) are then used to fill in the monthly population between the annual population data.

The population variable along with brief descriptions is listed below:

PREDICTED_POPULATION	Output of the population model (detailed in Chapter 3.1). These are the predicted monthly population counts within each 0.1° grid cell based on the annual LandScan values and average monthly nightlight trends.

ETHNICITY DATA

ETH Zurich

ETH Zurich offers data on with the geographic outlines of ethnic groups. Using the centroids of the 0.1° grid cells, each grid cell is placed inside of an ethnic area and thereby labelled with an ethnic group. Based on Euclidean distance, the distance from the centroid of each 0.1° grid cell to the nearest grid cell which is of a different ethnic group.

SWALIM Somali clan locations

As the Somali ethnic group is divided into clans, coordinate polygons of Somali clan locations through SWALIM data source were used to categorize Somali ethnic grid cells in Somalia, Kenya, and Ethiopia based on their clan identity.

The ethnic variables along with brief descriptions are listed below:

ETHNIC_GROUP	Classification of each 0.1° grid cell based on grid cell centroid point within ethnic defined polygons
DIST_ETHNIC	Euclidean distance from centroid of each grid cell to centroid of nearest grid cell with different ethnic_group value
CLAN_FAM	Clan family classification of grid cells for which ethnic_group = Somali. For grid cells where ethnic_group is not Somali, value is identical to ethnic_group
CLAN	Clan classification (sub-division of clan family) of grid cells for which ethnic_group = Somali. For grid cells where ethnic_group is not Somali, value is identical to ethnic_group
DIST_ETHNIC_SOMALI	Euclidean distance from centroid of each grid cell (where ethnic_group = Somali) to centroid of nearest grid cell (also with ethnic_group = Somali) with different clan_fam value. For grid cells where ethnic_group is not Somali, value is identical to dist_ethnic

FOOD SECURITY DATA

FEWS NET

As detailed in Chapter 3.2, this project predicts the food security situation within each grid cells on a monthly basis. These predictions build from the Famine Early Warning Systems Network (FEWS NET) historical food security data, which offers food security classifications of sub-national regions every 3 or 4 months from 2011 to the near present. For predictions the food security situation in each grid cell, the model uses feature variables on climate, conflict, grid cell land-use classification, annual national level data on fragility (Fragile States Index), health indicators, and food prices. The Monthly Global food price data are extracted from the Food and Agriculture Organization (FAO). These global data on the food price index, meat price index, dairy price index, cereals price index, oil price index, and sugar price index are used as feature variables in the food security model.

The food security variable along with brief descriptions is listed below:

CS	Predicted food security classification resulting from the food security model, which predicts the food security classification in each 0.1° grid cell from January 2009 to near present.

SOCIO-ECONOMIC AND WELLBEING DATA

Child health

Child health variables include infant mortality rate and child malnutrition rate and are extracted from the PRIO dataset. These variables are gridded to 0.5° resolution and stagnant, reflecting the values of the respective variables for the year 2000.

The child health variables along with brief descriptions is listed below:

IRM_MEAN	Infant mortality rate at 0.5° resolution, reflecting value in year 2000
CMR_MEAN	child malnutrition rate at 0.5° resolution, reflecting value in year 2000

Inequality

A stagnant Gini coefficient was calculated by gridding WorldPop data and VIIRS nightlight radiance data, averaged over the year 2020. These variables were placed inside 0.02° grid cell. The mean population within these 0.02° grid cells was used along with the 2020 nightlight intensity data to estimate nightlight per capita at the 0.02° grid cell-level. The nightlight per capita of each 0.02° grid cell is used to estimate the inequality (Gini index) with in each 0.1° grid cell.

The inequality variable along with brief descriptions is listed below:

GINI	Gini coefficient as an approximation of inequality within each 0.1° grid cell based on the distribution of nightlight per capita among 0.02° grid cells within each grid cell. Nightlight and (WorldPop) population data both from 2020.

MACRO POLITICAL VARIABLES

Fragile States Index

A series of indicators from the Fragile States Index are used. These indicators are annual and at the national level, so all grid cells in a country and in the same year carry the same value for each individual indicator.

The fragility variables along with brief descriptions are listed below:

Conflict data

Conflict data used for this project is extracted from Armed Conflict Location & Event Data (ACLED), which offers daily data on conflict events with geo-coordinate point locations of each conflict event as well as a description of each event. The descriptions of each event include information on the actors involved, number of fatalities, whether civilians were targeted, a categorization of each actor involved (as state forces, rebel group, political militia, identity militia, civilians, etc.), and the type of event which it was (battle, protest, riot, strategic development, etc.).

Events and fatalities

Based on the conflict location and date, we place each event inside of a 0.1° grid cell for the respective month. The sum of conflict events, which we consider all events except those identified as protest or riot, occurring in a grid cell in a given month is recorded with the variable events. From the conflict events that take place in a grid cell in a month, we also sum the fatalities caused by these events, which is recorded with the variable fatalities.

A linear relationship between conflict and displacement could be expected – that higher levels of conflict and more intense conflict, as measured through fatalities, would yield higher levels of displacement. But, from analyzing trend lines between the level and intensity of conflict events compared and displacement, this linear relationship was not evident. Rather, it seemed that the highest levels of displacement occurred at the onset of conflict events. This makes sense, as the initial shock of conflict breaking out may cause people, and especially those with the means to do so, to flee. While ongoing conflict events may become 'normalized' in a way and those without the means or ability to flee may remain, even as the number of conflict events rises. In an attempt to capture this relationship, new event variables were created. The variable new event in a binary variable, with a 1 indicating that there is at least one conflict event in that grid cell and it is the first month in at least 12 months with a conflict event, and 0 otherwise. For instance, if the event value for a grid cell is at least 1 in the month of August, the new event variable will only be 1 if the events variable values were all 0 from September of the previous year (12 months earlier). If there was a single month where the events variable was 1 or higher between September and August, then this observation will not count as having a new event. In order to account for the residual effect that a new event may have on the subsequent months for that grid cell, a variable call new event decay was also created. The formula for the values in this variable is 1 – (No. of months since new_event / 12). So the month where the new event took place will take a value of 1.00 and the next month will take 1 - (1/12), etc.

The events and fatalities variables along with brief descriptions are listed below

EVENTS	Number of (non-protest and non-riot) conflict events in each grid cell per month
FATALITIES	Number of (non-protest and non-riot) fatalities in each grid cell per month
NEW_EVENT	Binary variable with 1 indicating that the grid cell experienced a conflict event for the first time in at least 12 months, 0 otherwise
NEW_EVENT_DECAY	1 – (No. of months since new_event / 12). Values will be 1.000 for the month in which a new event occurs; 0.917 for the next month; 0.833 for two months since the new event; etc. In this way, the value 'decays' in the months since a new event.

Actor based conflict

The raw ACLED data feature a column which categorizes each actor involved in a conflict based on the type of group that it is. We were particularly interested in understanding the activities of certain groups, which are formed based on an identity, such as ethnicity, as well as rebel groups. Conflict events featuring such groups have the actor categorized as an identity militia or rebel group. So, we filtered the data for events featuring an identity militia or rebel group, placed the events into the 0.1° grid cells and summed the number of conflict events featuring an identity militia and the number of events featuring a rebel group for each month.

We were similarly interested in events featuring an extremist or separatist group, which are commonly involved in conflicts in certain areas of our study. ACLED does not specifically label actors as either extremist or separatist, so we filtered for events featuring an actor which is a known extremist or separatist group in the region. In identifying conflict events featuring an extremist group, we filtered for events in which an actor had one or more of the following key words/phrases in their name: Al Shabaab, ISIS, Al Qaeda, Islamist, Extremist, Katiba, Salafist, Ansar, Islam, Al Furqan, Jihad, Boko Haram, Sunni, Shiite, Al Mourabitoune Battalion, Emirate of the Sahara, Al-Sunna, Mouride Brotherhood Militia, Lord's Resistance Army.

We used the same logic for identifying conflict events featuring a separatist going; filtering for events in which an actor had one or more of the following words in their name: separatist, liberation, independence, or featuring an actor with one of the following names: MSA: Movement for Azawad Salvation, CSP-DPA: Permanent Strategic Framework - Azawad People Defense, MAA: Arab Movement of the Azawad (Platform), CMA: Coordination of Movements of the Azawad, HCUA: High Council for the Unity of Azawad, MNLA: National Movement for the Liberation of Azawad, MSA: Movement for Azawad Salvation (Sekou Bolly Faction), CJA: Congress for Justice in Azawad, MAA: Arab Movement of the Azawad (CMA), GMA: Mourabitounes Group of Azawad, MAA: Arab Movement of the Azawad, CPA: Azawad People's Coalition, High Council for the Azawad, MIA: Islamic Movement of Azawad, FNLA: National Liberation Front of Azawad, MPLA: Popular Movement for the Liberation of Azawad, MFDC: Movement of Democratic Forces of Casamance, IPOB: Indigenous Peoples of Biafra, BNL: Biafra Nations League Dragon Fighter Marine Faction, BNL: Biafra Nations League Black Marine Faction, BNL: Biafra Nations League, BNG: Biafran National Guard, Biafra Motherland Warriors, MASSOB: Movement for the Actualization of a Sovereign State of Biafra, MASSOB-Uwazuruike: Movement for the Actualization of a Sovereign State of Biafra (Uwazuruike Faction), Ambazonian Separatists (Cameroon), Ambazonian Separatists (Cameroon) Buffalos of Bali Nyonga Faction, Ambazonian Separatists (Cameroon) Restoration Forces, Ambazonian Separatists (Cameroon) Fako Black Tar Faction, ADF: The Ambazonia Defense Forces, Ambazonian Separatists (Cameroon) Red Dragons Faction, Ambazonian Separatists (Cameroon) Butabu Kwifor Faction, Ambazonian Separatists (Cameroon) Bafut 7 Karta Faction, SCACUF: Southern Cameroons Ambazonia Consortium United Front, FLEC-FAC: Front for the Liberation of the Enclave of Cabinda-Armed Forces of Cabinda, FLEC: Front for the Liberation of the Enclave of Cabinda, FLEC-PM: Front for the Liberation of the Enclave of Cabinda (Military Position), FLEC-Renouvada: Front for the Liberation of the Enclave of Cabinda (Renouvada Faction), Darfur Joint Forces/JSAMF: Joint Force of Armed Struggle Movement, Darfur Joint Forces, Darfur Joint Resistance Forces, Darfur Liberation Front, TPLF: Tigray People's Liberation Front, CRCT: Covenant for Radical Change in Tigray, TPDM: Tigray People's Democratic Movement, OLF-OLA: Oromo Liberation Army-Oromo Liberation Front, OLF-OLA/Central Oromia Zone: Oromo Liberation Army-Oromo Liberation Front-Central Oromia Zone Command, OLF: Oromo Liberation Front, Oromo Ethnic Militia, OLA: Oromo Liberation Army (Abba Torbee Splinter Faction), ODP: Oromo Democratic Party, ONLF: Ogaden National Liberation Front, ONLA: Ogaden National Liberation Army.

Civilian targeting

Conflict events featuring an actor categorized as an extremist or separatist group are placed into the 0.1° grid cells and summed the number of conflict events featuring an extremist group and the number of events featuring a separatist group for each month.

The actor-based conflict variables along with brief descriptions are listed below:

REBEL	Number of conflict events featuring a rebel militia in each grid cell per month
IDENTITY	Number of conflict events featuring an identity militia (ethnic or communal) in each grid cell per month
EXTREMIST	Number of conflict events featuring an extremist group in each grid cell per month
SEPARATIST	Number of conflict events featuring a separatist group in each grid cell per month

As each conflict event recorded by ACLED also notes whether civilians were targeted and civilian targeting would be a reason that people would become displaced, the data were filtered for events where civilians were targeted by different groups. ACLED data identifies whether a conflict event involved state forces, a political militia, a rebel group, or an identify group. So, the events were simply labelled where they featured one of these groups and civilians were targeted. Civilian targeting by extremist groups and separatist groups was also labelled based one of these groups being involved and civilians being targeted.

The civilian targeting variables along with brief descriptions are listed below:

STATE_CIV	Number of events in which state forces used violence on civilians in each grid cell per month
POLMIL_CIV	Number of events in which political militia used violence on civilians in each grid cell per month
REBEL_CIV	Number of events which featured a rebel militia and in which civilians were targeted
IDENTITY_CIV	Number of events which featured an identity militia (ethnic or communal) and in which civilians were targeted
EXT_CIV	Number of events which featured an extremist group and in which civilians were targeted
SEP_CIV	Number of events which featured a separatist group and in which civilians were targeted

Vicinity aggregation

As these conflict variables are aggregated at such high resolution, 0.1°, it is rare for a conflict event to take place within the confines of a given grid cell in each month. Additionally, people may be impacted in their decisions to flee not just based on their immediate surroundings, but also on conflict taking place in the wider areas. Each of the conflict variables listed above was therefore aggregated to the number of events, and level of intensity, within a 50 km radius of the grid cell in that month.

The conflict in surrounding area variables along with brief descriptions are listed below:

EVENTS_NEIGHBORS	Number of (non-protest and non-riot) conflict events in grid cells within 50 km radius per month
FATALITIES_NEIGHBORS	Number of (non-protest and non-riot) conflict fatalities in grid cells within 50 km radius per month
NEW_EVENT_NEIGHBORS	Sum of new_event values in grid cells within 50 km radius per month
NEW_EVENT_ DECAY_NEIGHBORS	Sum of new_event_decay values in grid cells within 50 km radius per month
REBEL_NEIGHBORS	Number of conflict events featuring a rebel militia in grid cells within 50 km radius per month
IDENTITY_NEIGHBORS	Number of conflict events featuring an identity militia in grid cells within 50 km radius per month
EXTREMIST_NEIGHBORS	Number of conflict events featuring an extremist group in grid cells within 50 km radius per month
SEPARATIST_NEIGHBORS	Number of conflict events featuring a separatist group in grid cells within 50 km radius per month
STATE_CIV_NEIGHBORS	Number of events in which state forces used violence on civilians in grid cells within 50 km radius per month
POLMIL_CIV_NEIGHBORS	Number of events in which political militia used violence on civilians in grid cells within 50 km radius per month
REBEL_CIV_NEIGHBORS	Number of events which featured a rebel militia and in which civilians were targeted in grid cells within 50 km radius per month
IDENTITY_ CIV_NEIGHBORS	Number of events which featured an identity militia (ethnic or communal) and in which civilians were targeted in grid cells within 50 km radius per month
EXT_CIV_NEIGHBORS	Number of events which featured an extremist group and in which civilians were targeted in grid cells within 50 km radius per month
SEP_CIV_NEIGHBORS	Number of events which featured a separatist group and in which civilians were targeted in grid cells within 50 km radius per month

Social tension

As mentioned above, ACLED data categorizes each event based on its type of event (battle, protest, riot, strategic development, etc.). The conflict variables listed above are based on observations of all event types, except protests and riots. The number of 'non-violent' events, protests and riots, were used along with V-Dem's liberal democracy index ('v2x_libdem') as an indicator of social tension.

The social tension variable along with brief descriptions are listed below:

SOCIAL_TENSION	Calculated as the logged number of demonstration events (protests and riots) in a given grid cell, multiplied with the V-Dem liberal democracy index on a reversed scale.

UNHCR
The UN Refugee Agency

COMPLEX
RISK
ANALYTICS
Fund