Gridded 20-year climate parameterization of Africa and South America for a stochastic weather generator (CLIGEN)

Andrew Fullhart, Guillermo E. Ponce-Campos, Menberu B. Meles, Ryan P. McGehee, Gerardo Armendariz, Paulo Tarso S. Oliveira, Cristiano Das Neves Almeida, José C. de Araújo, Werner Nel & David C. Goodrich

To cite this article: Andrew Fullhart, Guillermo E. Ponce-Campos, Menberu B. Meles, Ryan P. McGehee, Gerardo Armendariz, Paulo Tarso S. Oliveira, Cristiano Das Neves Almeida, José C. de Araújo, Werner Nel & David C. Goodrich (2022): Gridded 20-year climate parameterization of Africa and South America for a stochastic weather generator (CLIGEN), Big Earth Data, DOI: 10.1080/20964471.2022.2136610

To link to this article: https://doi.org/10.1080/20964471.2022.2136610

© 2022 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the International Research Center of Big Data for Sustainable Development Goals, and CASEarth Strategic Priority Research Programme.

Submit your article to this journal

View related articles

View Crossmark data

This article has been awarded the Centre for Open Science 'Open Data' badge.
Gridded 20-year climate parameterization of Africa and South America for a stochastic weather generator (CLIGEN)

Andrew Fullhart\textsuperscript{a,}*, Guillermo E. Ponce-Campos\textsuperscript{b}, Menberu B. Meles\textsuperscript{c,}*, Ryan P. McGehee\textsuperscript{d,}*, Gerardo Armendáriz\textsuperscript{e,}*, Paulo Tarso S. Oliveira\textsuperscript{e,}*, Cristiano Das Neves Almeida\textsuperscript{f,}*, José C. de Araújo\textsuperscript{g,}*, Werner Nel\textsuperscript{h,}*, and David C. Goodrich\textsuperscript{a,}*

\textsuperscript{a}Southwest Watershed Research Center USDA-ARS, Tucson, AZ, USA; \textsuperscript{b}School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, USA; \textsuperscript{c}Sustainable Agricultural Water Systems Unit USDA-ARS, Davis, CA, USA; \textsuperscript{d}Agricultural and Biological Engineering, Purdue University, West Lafayette, IN, USA; \textsuperscript{e}Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Campo Grande, MS, Brazil; \textsuperscript{f}Water Resources and Environmental Engineering Laboratory, Federal University of Paraíba, Joao Pessoa, PB, Brazil; \textsuperscript{g}Federal University of Fortaleza, Department of Agricultural Engineering, Fortaleza, Brazil; \textsuperscript{h}Department of Geography and Environmental Science, University of Fort Hare, Alice, South Africa

\textbf{ABSTRACT}

CLIGEN is a stochastic weather generator that creates statistically representative timeseries of daily and sub-daily point-scale weather variables from observed monthly statistics and other parameters. CLIGEN precipitation timeseries are used as climate input for various risk-assessment modelling applications as an alternative to observe long-term, high temporal resolution records. Here, we queried gridded global climate datasets (TerraClimate, ERA5, GPM-IMERG, and GLDAS) to estimate various 20-year climate statistics and obtain complete CLIGEN input parameter sets with coverage of the African and South American continents at 0.25 arc degree resolution. The estimation of CLIGEN precipitation parameters was informed by a ground-based dataset of >10,000 locations worldwide. The ground observations provided target values to fit regression models that downscale CLIGEN precipitation input parameters. Aside from precipitation parameters, CLIGEN’s parameters for temperature, solar radiation, etc. were in most cases directly calculated according to the original global datasets. Cross-validation for estimated precipitation parameters quantified errors that resulted from applying the estimation approach in a predictive fashion. Based on all training data, the RMSE was 2.23 mm for the estimated monthly average single-event accumulation and 4.70 mm/hr for monthly maximum 30-min intensity. This dataset facilitates exploration of hydrological and soil erosional hypotheses across Africa and South America.

\textbf{1. Introduction}

Precipitation and other climate timeseries serve as the primary forcing input for many hydrological and soil erosional modelling applications. Obtaining long-term, high resolution,
and high-quality precipitation timeseries is a common requirement and ongoing challenge. Therefore, the quality and availability of these data are often limiting factors for risk-assessments based on such applications. Unfortunately, difficulties in meeting these data requirements may preclude risk-assessment modelling from being done or may necessitate the use of nonideal climate data, subsequently leading to increased model uncertainty.

To address data limitations in climate timeseries, globally available gridded datasets may be used to derive the required timeseries. However, model applications commonly assume that climate timeseries are point-scale, i.e. that they represent observations at a single location. Grid-scale data may be statistically downscaled to meet the assumption of point-scale climate data by quantifying the effect of spatial averaging (which depends on the resolution of the data product), and accordingly, applying bias corrections (Yuan et al., 2015). The process of downscaling spatially distributed precipitation is important because these gridded datasets differ substantially from what would be expected at point-scale in terms of daily and sub-daily factors, such as precipitation intensity, storm frequency, and storm accumulation (Tan et al., 2018; Jiang et al., 2021; Rivoire et al., 2021). Climate model datasets may also have inherent systematic errors that vary spatially, temporally, and across specific processes within the climate model framework (Gleixner et al., 2020; Gosset et al., 2018). Furthermore, any precipitation dataset with coarse temporal resolution is affected by scaling bias due to time-averaging of the true intensity-time profile of a precipitation event, leading to less vigorous precipitation patterns (Pananos et al., 2016). Correcting these sources of error reduces uncertainty in model outcomes, as in runoff modelling, where the error in precipitation inputs can result in comparable uncertainty as the error in model structure and parameterization (Renard et al., 2011).

Several regions of the world have been identified in a systematic literature review as having minimal or absent coverage of information from runoff and soil erosion modelling applications (Borreli et al., 2021). For example, such studies are missing in large regions of Africa (Nordling, 2019; Le Coz and van de Giesen, 2020) and South America (Almagro et al., 2021; Fagundes et al., 2021) due in part to data limitations in available climate data. Stochastic weather generators such as CLImate GENerator (CLIGEN) Version 5.3 (Agricultural Research Service, 2022) provide an approach for downscaling global gridded climate data to produce long-term daily and sub-daily timeseries in places where these data are scarce or non-existent. Typically, input parameters for stochastic weather generators come from daily and sub-daily datasets aggregated to statistics that represent longer time windows, e.g. monthly statistics. As such, downscaling may be achieved by applying downscaling bias-adjustments to grid-scale values of stochastic weather generator input parameters (Mehan et al., 2017).

This paper describes a calibrated downscaling method to estimate CLIGEN precipitation parameters on a 0.25 arc deg. grid of African and South American continents. At each grid point, these parameter sets are part of complete CLIGEN input files that may be used to generate timeseries for downscaled precipitation and other basic weather parameters. CLIGEN has been validated in a variety of global climate types across extremes in precipitation, temperature, and climate dynamics. Daily-scale dynamics are generally accurately represented, while sub-daily dynamics show more uncertainty in some cases (Elliot and Arnold, 2001; Fan et al., 2013; Lobo et al., 2015). Among applications for CLIGEN, determining these precipitation factors is important for representing derivative metrics.
such as rainfall erosivity in the Universal Soil Loss Equation (Wischmeier and Smith, 1978), which describes the erosive potential of rainfall (Kinnell, 2019). As an example of generated CLIGEN timeseries being used as model input, CLIGEN was used to drive watershed simulations in a semi-arid 148-km² watershed in the southwest U.S. where it was shown how CLIGEN might be used for multi-site rainfall generation to better represent spatial variability (Zhao et al., 2021). Bayley et al. (2010) developed a Climate Assessment Tool for the Water Erosion Prediction Project (WEPP, Nearing et al., 1989) that can alter CLIGEN parameters based on regional or global climate model (GCM) projections including precipitation intensity for extreme events.

The primary goal of this paper is to develop a grid of 20-year average CLIGEN parameter sets at 0.25 arc deg. resolution that may be used to produce point-scale climate timeseries for numerous modelling applications in Africa and South America. Precipitation parameters were calculated from gridded climate datasets and were bias-adjusted using flexible machine learning regression methods that consider a number of predictors and effectively quantify biases by comparison to long-term ground network data from more than 10,000 locations worldwide. The ground data provided relatively few long-term parameter sets in Africa and South America, and therefore, a secondary goal of this work is to show that error introduced by transferring the applied statistical models to areas with poor spatial coverages of ground data has a minimal impact on the determined parameters. To assess uncertainties, cross-validation analyses included a leave-one-out cross-validation for 60 parameter sets in both studied continents. These and other validation results are discussed along with characteristics of the gridded CLIGEN parameter dataset.

2. Materials and methods

The product of this paper is a gridded dataset of 20-year CLIGEN input parameters with coverage of the continents of Africa and South America at 0.25 arc deg. resolution (Fullhart et al., 2022). The grid represents 40,936 point locations in Africa and 24,588 locations in South America for which complete CLIGEN parameter sets were created. In Table 1, four different global climate products and two globally distributed ground networks are listed that were involved in the development of the gridded parameter sets. The four gridded climate products were accessed using the Google Earth Engine cloud computing platform (Gorelick et al., 2017). Two ground networks, whose locations are plotted in Figures 1 and 2, provided observed ground-based long-term parameter sets at a total of 11,065 locations (described in more detail later). CLIGEN’s monthly parameters that aggregate daily-scale precipitation were taken from ERA5-Daily and were down-scaled from grid-scale. ERA5 is the fifth-generation reanalysis product of the European Centre for Medium Range Weather Forecasts (Hersbach et al., 2020). Twenty-year average monthly accumulations were set directly according to the higher resolution (1/24th arc deg.) TerraClimate product, which is a monthly climatic water balance dataset (Abatzoglou et al., 2018), and which showed better agreement to monthly accumulation from ground data than ERA5-Daily. Precipitation values aggregated over longer time periods such as months or years, though still affected by spatial averaging at the resolution of the model, are generally not impacted by spatial averaging of individual storms. Therefore, in this approach, monthly accumulation from TerraClimate was treated as
Table 1. Various gridded climate datasets and ground observation networks involved in development of gridded CLIGEN parameterizations. Google Earth Engine system versions are shown for gridded products.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Spatial resolution</th>
<th>Time interval</th>
<th>N point locations</th>
<th>Reference period</th>
<th>Version</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA5-Daily</td>
<td>Gridded</td>
<td>0.25 arc degree</td>
<td>Daily</td>
<td>40,938</td>
<td>2000–2020</td>
<td>1641990118708959</td>
<td>Hersbach et al., (2020)</td>
</tr>
<tr>
<td>GPM-IMERG</td>
<td>Gridded</td>
<td>0.1 arc degree</td>
<td>30 min</td>
<td>40,938</td>
<td>2000–2020</td>
<td>1645999532722295</td>
<td>Huffman et al., (2015)</td>
</tr>
<tr>
<td>GLDAS</td>
<td>Gridded</td>
<td>0.25 arc degree</td>
<td>3 hr</td>
<td>40,938</td>
<td>2001–2021</td>
<td>1645607236743494</td>
<td>Rui et al. (2021)</td>
</tr>
<tr>
<td>Intl. CLIGEN Dataset (GHCN-D)</td>
<td>Ground</td>
<td>Point-scale</td>
<td>Monthly</td>
<td>7,491</td>
<td>10, 20, and 30 yr</td>
<td>-</td>
<td>Fullhart et al. (2021)</td>
</tr>
</tbody>
</table>

Figure 1. International CLIGEN dataset stations ground data is considered for the latitude band of 40° N and southward. These stations provided point-scale observations of MEAN P, SDEV P, SKEW P, and RATIO. The number of stations is n=10,456.

a point-scale variable, separate from CLIGEN parameters. Parameter values for CLIGEN’s monthly 30-min intensity parameter were also downcaled from the Global Precipitation Measurement (GPM-IMERG) product at 30-min temporal resolution and 0.1 arc deg. spatial resolution (Huffman et al., 2015). Since GPM-IMERG has a finer resolution than the 0.25 arc deg. grid, only grid cells covering centre points of the 0.25 arc deg. grid were sampled (because the centre points were chosen to represent point-scale locations of each grid cell). Alternatively, spatial weighting of GPM-IMERG values based on the distance of a point to adjacent grid cell centres may possibly result in better estimates; however, this approach was not taken because of the computational expense that would be required by sampling every GPM-IMERG grid cell. Monthly temperature parameters were taken directly from the ERA5 temperature variable fields, while monthly solar radiation parameters were taken directly from the Global Land Data Assimilation System (GLDAS) variable fields (Rui et al., 2021). Additionally, CLIGEN’s time-to-peak
intensity cumulative probability distribution and various wind parameters that have little influence on other CLIGEN output factors were taken from representative observation locations using the same procedure described in 2021.

To inform downscaling statistical models, we used point-scale CLIGEN precipitation parameters calculated from two globally distributed ground networks (Figures 1 and 2). Parameter values derived from daily-scale precipitation data were taken from 7,491 stations in an international dataset of ground-based CLIGEN parameters. This international CLIGEN dataset was derived from sufficiently long, gap-free, records in the Global Historical Climate Daily Network (GHCN-D), consisting of available 30-year, 20-year, and 10-year records, with most records being 30-year (Fullhart et al. 2021). Record lengths precluded many stations in GHCN-D from being used to determine long-term parameter sets. This resulted in particularly poor coverage for Africa, partly because of a large portion of long record lengths in South America and Africa that consisted of 20th century periods. Therefore, a second, less strict querying of GHCN-Daily was done for any stations in Africa and South America with sufficiently long records to produce 20-year monthly averages and at least some data from the year 1990 onward, which brought the total number of stations to 10,456. These additional stations primarily exist in the countries of Brazil and South Africa and are not preferred for the validation analyses done in this paper of individual stations. The Automated Surface Observing System (ASOS), a high temporal resolution (1-min) network in the U.S. (Flanagan et al., 2020), was used to calculate observed monthly maximum 30-minute intensity parameters at 609 stations for the time period 2006–2019 (13 years) with some variability with this time window for the available stations.

Differences in the available reference periods and geographical coverages of the ground datasets are a potential source of uncertainty in the estimation approach for CLIGEN precipitation parameters. The 20-year records taken from each gridded dataset
were compared to collocated ground records with variable reference periods of 10–30 year durations. Therefore, for these comparisons to be made, an underlying assumption is that the reference periods are, in all cases, representative of long-term climate, and furthermore, offsets in reference periods do not affect the comparison. This assumption may lead to bias from factors such as long-term climate cycles, extremes in precipitation, non-stationary climate trends, etc. Secondly, the transferability of the applied downscaling models to areas with poor ground coverage is a source of uncertainty because of localized climate dynamics, localized systematic errors in each gridded dataset, and variation in the spatial area of grid cells across latitude. Only stations from the international CLIGEN dataset south of the +40 degree latitude band were selected to coincide with the latitudinal extent and range of climates in the target area (stations were also weighted in the downscaling models with respect to the extents of global climate types in the target areas, as discussed later). However, use of this latitude band was not imposed for the selection of ASOS stations because of the large number of stations outside of this region, leading to additional uncertainty in the maximum 30-minute intensity downscaling model. The uncertainties related to the downscaling approach were assessed in several ways, as discussed later. While not insignificant, these uncertainties are justifiable if ground observations are unavailable for an area of interest in South America or Africa.

2.1. Calculation of CLIGEN input parameters

2.1.1. Monthly aggregate daily precipitation parameters

The ERA5-Daily dataset was used to calculate five of CLIGEN’s monthly precipitation parameters that represent aggregate statistics of daily-scale precipitation data. These consist of monthly mean (MEAN $P$), standard deviation (SDEV $P$), and skewness (SKEW $P$) of accumulation on days with non-zero accumulation, and two-state Markov chain conditional probabilities of the transition to precipitation days, $P(W|D)$ and $P(W|W)$, which together define the conditional probability of a non-zero accumulation day, $P(W)$. The values for each parameter are calculated on a monthly basis as follows:

$$MEAN \ P = \frac{1}{n} \sum_{i=1}^{n} p_i,$$

$$SDEV \ P = \left( \frac{1}{n} \sum_{i=1}^{n} (p_i - \bar{p})^2 \right)^{0.5},$$

$$SKEW \ P = \frac{\sum_{i=1}^{n} (p_i - \bar{p})^3}{(n-1) \cdot (SDEV P)^3},$$

$$P(W|W) = \frac{N_{ww}}{N_{ww} + N_{wd}},$$

$$P(W|D) = \frac{N_{dw}}{N_{dw} + N_{dd}},$$

where $p$ is single-day accumulation on days with non-zero accumulation ($>0$), $n$ is the number of days with non-zero accumulation, $P(W|W)$ is the probability of a “wet” day
following a “wet” day, and $P(W|D)$ is the probability of a “wet” day following a “dry” day (with both frozen and liquid precipitation being considered “wet”). The conditional probabilities are determined from the number of days, $N$, in which each possible transition occurred, shown as subscripts to $N$ (e.g., $N_{dw}$ is the number of dry days that followed a wet day). The length unit is U.S. customary inches (in). Statistical models were developed to bias-adjust grid-scale $MEAN\ P$, $SDEV\ P$, and $SKEW\ P$ in order to downscale these values, as discussed in the following Section 2.2.

Monthly accumulation ($ACCUM$), while not a CLIGEN input parameter itself, may be calculated from input parameters as follows:

$$ACCUM = n_{days} \times MEAN\ P \times \left\{P(W|D)/\left[1 - P(W|W) + P(W|D)\right]\right\}, \quad (6)$$

where $n_{days}$ is the total number of days in a given month, and the bracketed term is equivalent to $P(W)$. In this case, $ACCUM$ taken directly from TerraClimate, along with downscaled $MEAN\ P$, were entered into Equation 6 allowing the $P(W)$ term to be solved for (i.e. the point-scale fraction of monthly precipitation days). Then, the $P(W)$ term was isolated to begin to solve for the required point-scale $P(W|W)$ and $P(W|D)$ values. To obtain enough information to determine these values, the point-scale ratio of $P(W|W)$ to $P(W|D)$ was treated as a variable as follows:

$$RATIO = P(W|W)/P(W|D). \quad (7)$$

A separate downscaling model using ERA5 data was used to estimate $RATIO$, which had the benefit of controlling the wet/dry transitions to yield a more realistic sequencing. Then, with $P(W)$ and $RATIO$ known, values of $P(W|W)$ and $P(W|D)$ were adjusted until the $P(W)$ term was balanced, thereby solving for point-scale $P(W|W)$ and $P(W|D)$. These values were iteratively adjusted by equal proportions in order to maintain the assumed $RATIO$ value. Since the tendency of grid-scale precipitation is to produce relatively small accumulation amounts distributed over a relatively large number of timesteps, values for $P(W|D)$ and $P(W|W)$ usually decreased from their grid-scale values and reduced the total number of wet days.

Using this parameter adjustment scheme, 20-year average values of $ACCUM$ taken from TerraClimate are obtained from the estimated CLIGEN parameter sets, though generated CLIGEN output may result in slight differences in long-term average precipitation from that given by TerraClimate. As discussed later, values of $ACCUM$ were also used as a predictor variable for downscaling precipitation models. The correlation of TerraClimate versus ground observations of $ACCUM$ is shown in Figure 3. The fact that 20-year average TerraClimate values were compared against long-term ground observations with variable reference periods and record lengths likely explains a significant part of the correlation variance. Furthermore, the slope of the fitted regression shows an apparent underestimation of 7% by TerraClimate and 2% underestimation based on percent bias, indicating that systematic bias exists in one or both datasets being compared. The decision to directly use TerraClimate values of $ACCUM$ in Equation 6 may help resolve uncertainties from issues like climate change and spatial variability by using a dataset that, itself, is well documented and has continuous spatial and temporal coverage. TerraClimate $ACCUM$ values were chosen for use over ERA5-Daily $ACCUM$ (which would have also been possible to use) because the comparison of ERA5-Daily to ground data
yielded slightly more variance and apparent underestimation bias than TerraClimate. This is likely due in part to the fact that TerraClimate has considerably higher spatial resolution.

2.1.2. Monthly maximum 30-minute intensity
The monthly maximum 30-minute intensity (MX.5P) parameter was determined using the 30-minute GPM-IMERG dataset. A downscaling model was developed for MX.5P that accounts for spatial and temporal averaging of MX.5P taken from GPM-IMERG (discussed in Section 2.2). The calculation of MX.5P is done as follows:

$$MX.5P = \frac{1}{k} \sum_{i=1}^{k} max_i,$$

where max is the maximum 30-minute intensity in a given month, and k is the number of monthly records for a given month (k = 20 for calculation of MX.5P from 20-year records). The resolution of GPM-IMERG is 0.1 arc deg., which differs from the 0.25 arc deg. resolution of ERA5, GLDAS, and the gridded product of this paper. The mismatch in resolutions meant that calculations of MX.5P were done using only the grid cells overlapping centre points of the 0.25 arc deg. grid.

2.1.3. Monthly temperature parameters
Variable fields from ERA5 were used to calculate CLIGEN temperature parameters. The obtained parameter values were accepted with no bias adjustments. The temperature parameters are determined on a monthly basis and listed as follows: average maximum daily temperature (TMAX AV), average minimum daily temperature (TMIN AV), standard deviation of maximum daily temperature (SD TMAX), standard deviation of minimum daily temperature (SD TMIN), and average daily dew point temperature (DEW PT). Units are in
U.S. customary degrees Fahrenheit. ERA5 temperature variables are validated for Africa
and other places in studies such as Gleixner et al. (2020).

2.1.4. Monthly solar radiation parameters
Solar radiation parameters were determined from the 0.25 arc deg. GLDAS land-surface
dataset with no bias adjustments. These were average daily incoming solar radiation (SOL.
RAD) and the standard deviation of daily incoming solar radiation (SD SOL). Units are in
Langleys (1 Langley = 41,840 J·m⁻²). A validation of GLDAS solar radiation was done
against selected stations from a ground network in Fullhart et al. 2021 that found good
agreement of parameter values. However, gaps in spatial coverage resulted in a small
percentage (<1%) of parameter sets with missing solar radiation parameters. SOL.RAD is
also used as a predictor variable in the downscaling models, and in this procedure, the
equivalent variable is taken from TerraClimate which does not have the gaps in available
spatial coverage. Correlation of ER5-Daily precipitation to average daily solar radiation
from GLDAS was poor overall, but as expected, a negative correlation was found. This was
also the case for correlations of ER5 precipitation to ER5 temperatures (in this case
mean daily temperature was used). For two major cities (São Paulo, Brazil and Nairobi,
Kenya), the strongest Pearson correlation coefficient of monthly correlations of precipita-
tion versus solar radiation was -0.52 for São Paulo and -0.48 for Nairobi. The strongest
Pearson correlation coefficient of precipitation versus temperature was -0.34 for São Paulo
and -0.46 for Nairobi. This suggests GLDAS solar radiation may have a slightly better
correlation to precipitation than the other non-precipitation parameters, which may have
implications for the downscaling models that all use various monthly parameters.

2.1.5. Representative parameter sets
Two CLIGEN parameter sets could not be estimated from available data because of high
data requirements for their calculation, being: time-to-peak precipitation intensity (Time
Pk) and various wind parameters. The Time Pk parameter values require high temporal
resolution data to determine and define interval values of the normalized time-to-peak
intensity cumulative probability distribution (and are therefore not monthly values like
other CLIGEN parameters). Representative Time Pk parameters were assumed based on
averages for Köppen-Geiger climate types according to the appendix table in
Fullhart et al. (2021). The Time Pk parameter has a small effect on other precipitation
parameters (Wang et al., 2018) and Time Pk values were found to vary across small ranges
within Köppen-Geiger climate types (Fullhart et al., 2021a).

Various and detailed wind parameters are required by CLIGEN that have no effect on
other CLIGEN output variables but are included for applications such as evaporation
modelling and wind erosion modelling. Data limitations precluded the necessary para-
eters from being known, so wind parameters were taken from representative locations
with the most similar precipitation and temperature parameter values, as determined
using the “international conversion program” tool (Agricultural Research Service, 2022).
This tool was used to query the U.S. CLIGEN ground network described in Srivastava et al.
(2019). Information about the query made by the international conversion program tool is
included at the bottom of each CLIGEN input file.
2.2. Gradient boosting regression downscaling models

Gradient boosting (GB) machine learning regression models were developed to down-scale \textit{MEAN P}, \textit{SDEV P}, \textit{SKEW P}, \textit{RATIO}, and \textit{MX.5P} using the Sci-kit Learn v 3.2 Python implementation of GB (Pedregosa et al., 2011). GB is an ensemble framework of decision trees that classify subsets of data for which individual regressions are done. Gradient boosting has been used in hydrological and climatological applications and usually out-performs multiple regression or random forest regression. In the present application, each GB model was fitted to observe point-scale values for respective target variables. The fitting was based on sets of collocated predictor variables that consisted primarily of grid-scale statistics. Essentially, the five accepted GB models provided a means of downscaling bias-adjustment for the African and South American precipitation grid. A consistent set of predictor variable inputs were used for all GB models that included grid-scale values of required CLIGEN parameters, including the grid-scale values of each point-scale parameter being estimated. As such, several predictor variables were already necessarily determined for the process of creating CLIGEN parameter sets. A flow chart of the GB approach is shown in Figure 4. In the training phase, GB models were fitted to known point-scale target values from the ground station distribution. A selection process for fitted GB models was based on a simple hyperparameter calibration, which decided the five models used in the prediction phase. The selection process, discussed in the following sections, further involved quantifying the degree of model fitting and the accuracy of fitting for the training data. However, model fitting error is commonly lower than prediction error, and two forms of cross-validation were done in this paper that reserve subsets or individual samples of data to be used to derive prediction errors that may be compared.

![Flowchart of the statistical downscaling approach using gradient boosting models.](image-url)
to the known fitting error. Importantly, a leave-one-out cross-validation provided a representative assessment of model error by fitting all ground data to the model except for single stations of interest so that model predictions for these stations can be compared to observations. Using cross-validations, prediction errors along trends of climate, geographic location and density of ground data could be understood.

2.2.1. Gradient boosting setup

To characterise aspects of climate, seasonality, and spatial variability, including trends across elevation and latitude, we used a total of 21 predictor variables (see Table 2). To use the information necessary to determine CLIGEN input sets, most predictor variables represented grid-scale CLIGEN parameters or were variables that could easily be determined from querying the selected datasets. In addition to previously given CLIGEN parameters, the predictor variables included annual rainfall (ANNUAL), the standard deviation of monthly accumulation (SDEV ACCUM), monthly surface air pressure (PRESS), elevation (ELEV), latitude (LAT), and the minimum, maximum and average grid-scale values of the point-scale target variables (TARGET MIN, TARGET MAX, TARGET AVG). The predictor variables represented a mixture of grid-scale monthly variables, grid-scale variables with single values (i.e. not monthly), and spatial variables, elevation and latitude. Predictor variables were checked for single-interaction cross-correlation using the variance inflation factor (VIF), defined as \((1-R^2)^{-1}\), which is a measure of the increase in variance caused by collinearity (Salmerón et al., 2018). Each VIF was found to be less than 6, while most values ranged between 0 and 2. In practice, interactions between variables and the importance of individual predictor variables differed for each of the five GB models.

The international distribution of 10,456 GHCN-D ground stations with 12 monthly parameters per station were pooled to create the training datasets for the \(\text{MEAN} P\), \(\text{SDEV} P\), \(\text{SKEW} P\), and \(\text{RATIO}\) models, while a second set of data from 609 ASOS stations

<table>
<thead>
<tr>
<th>Target variable</th>
<th>Predictor variables</th>
<th>Predictor data table size used for training</th>
<th>Ground network</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{MEAN} P^a); (\text{SDEV} P^a); (\text{SKEW} P^a); (\text{RATIO}^a)</td>
<td>(\text{MEAN} P^1); (\text{SDEV} P^3); (\text{SKEW} P^1); (\text{P(W</td>
<td>W)}^1); (\text{P(W</td>
<td>D)}^2); (\text{ACCUM}^2); (\text{MX.SP}^3); (\text{TMAX AV}^1); (\text{TMIN AV}^1); (\text{SD TMAX}^1); (\text{SD TMIN}^1); (\text{TDW}^1); (\text{SOL.RAD}^2); (\text{PRESS}^1); (\text{TARGET MIN}^3); (\text{TARGET MAX}^1); (\text{TARGET AV}^1); (\text{ANNUAL}^2); (\text{SDEV ACCUM}^2); (\text{ELEV}^\theta); (\text{LAT}^\theta); (\text{AV}^\theta); (\text{MEAN}^\theta); (\text{SDEV}^\theta); (\text{SKEW}^\theta); (\text{RATIO}^\theta)</td>
</tr>
<tr>
<td>(\text{MX.SP}^a)</td>
<td>(\text{MEAN} P^1); (\text{SDEV} P^1); (\text{SKEW} P^1); (\text{P(W</td>
<td>W)}^1); (\text{P(W</td>
<td>D)}^1); (\text{ACCUM}^2); (\text{MX.SP}^3); (\text{TMAX AV}^1); (\text{TMIN AV}^1); (\text{SD TMAX}^1); (\text{SD TMIN}^1); (\text{TDW}^1); (\text{SOL.RAD}^2); (\text{PRESS}^1); (\text{TARGET MIN}^3); (\text{TARGET MAX}^1); (\text{TARGET AV}^1); (\text{ANNUAL}^2); (\text{SDEV ACCUM}^2); (\text{ELEV}^\theta); (\text{LAT}^\theta);</td>
</tr>
</tbody>
</table>

\(^a\)denotes variables that are point-scale.
\(^d\)denotes variables that are grid-scale.
\(^@\)denotes variables that have the same value for all months for a given station or point (i.e. not a monthly value).
\(^^\wedge\)denotes absolute value (used for latitude in the MX.SP GB model).
\(^1\)denotes ERA5-derived variable.
\(^2\)denotes TerraClimate-derived variable.
\(^3\)denotes GPM-derived variable.
was used for the \textit{MX.SP} model. For the \textit{MEAN P}, \textit{SDEV P}, \textit{SKEW P}, and \textit{RATIO} models, this resulted in a data table of predictors with 125,472 rows, equivalent to the number of observed monthly values for each parameter, and 21 columns for each predictor variable. Using ASOS stations distributed only across the U.S., the training data table for \textit{MX.SP} had 7,200 rows and 21 columns.

For the \textit{MEAN P}, \textit{SDEV P}, \textit{SKEW P}, and \textit{RATIO} models, which have globally distributed training data, a sample weighting scheme was applied to each pair of monthly target and predictor variable value sets during the training phase that considered the degree to which global climate types are represented by the ground distributions compared to the spatial representation of the same climate types in the African and South American Continents. This was done to account for the fact that certain climates are over or under-sampled by the ground distribution used for training relative to the spatial extents of global climate types in Africa and South America. The weighting scheme used the Beck et al. (2018) Köppen-Geiger climate type map to define global climate types. From this, the proportion of ground stations in each climate type and the proportion of each climate type present in the combined spatial extents of Africa and South America were determined. The applied sample weight was the ratio of the latter quantity over the former quantity based on the climate type of a given station. This weighting scheme was not applied to the GB model for \textit{MX.SP} because the ASOS dataset was largely distributed over climate types only present inside the U.S.

\subsection*{2.2.2. Gradient boosting model selection and fitting}

When hyperparameters are tuned, the complexity of the GB structure changes, resulting in varying degrees of both overfitting and underfitting such that prediction error differs from the error of the fitted datapoints. Therefore, a model selection process involved tuning a single hyperparameter and ascertaining both a training and prediction error for each resulting candidate model that was tested. The goal was not to develop models with equal training and prediction errors because of the high computational requirements necessary for precise tuning. Rather, each accepted model was allowed to be overfitted to some degree. Ultimately, only prediction errors from cross-validation should be considered for judging model performance because of this issue. The single tuned hyperparameter is known as the learning rate, which restricts the reduction in error contributed by individual decision trees and is important for controlling the degree of fitting of the training data. For tuning, the learning rate was varied across orders of magnitude from 0.0001 to 0.1 with 13 intervals, resulting in models with widely varying degrees of both overfitting and underfitting. The learning rate has a strong interaction with the total number of decision trees (termed weak learners in the context of the ensemble). This hyperparameter controls the complexity of the GB model and was held constant at $n = 5,000$ for all GB models. The suggested defaults of the Sci-kit Learn GB implementation were used for remaining GB hyperparameters. The prediction error for each candidate model was considered as the average RMSE of a K-folds cross-validation using 5 folds, in which each datapoint in the training dataset was reserved to assess prediction error exactly once. The resulting candidate models were ordered with respect to their prediction error, and models were selected with the lowest prediction errors with an underfitted training error within 50% of the prediction error. In total, this model selection process
resulted in 5 models × 13 learning rate values × 5 K-fold cross-validations = 325 model fittings.

Summary information for the accepted GB models is shown in Table 3. The standard deviation of RMSE across folds is included as a measure of internal heterogeneity within the data. Model overfitting occurs with varying degrees for each of the five GB models, as does overall prediction error given by the K-folds average RMSE. Plots of the fitted GB models are shown in Figure 5 that indicate the slopes of linear regressions of fitted estimates against observed values. Each slope suggests underestimation bias, which is confirmed by slightly negative PBIAS of prediction errors for all accepted models except MX.5P (discussed later). Slope coefficients less than one result from underestimation due to increased variance towards the greater ranges of variable values that correspondingly tend to be underestimated. This estimation bias could be eliminated by increasing the degree of model overfitting and model complexity, or by introducing further predictors that can explain the non-linearity present towards greater values; however, there are limitations of computational requirements and the availability of data that could contribute further explanatory power.

The fitting outcomes of MEAN P and MX.5P are especially crucial for an accurate representation of sub-daily rainfall intensity. These two parameters are reasonably well fitted, with slopes within 3% of the 1:1 slope. Adjusted r² values are also close to 1 for these parameters. There is a poorer performance for SKEW P and RATIO. Values of RATIO tend to be particularly underestimated in ranges towards greater values. It can also be seen that some observed RATIO values repeatedly occurred due to the fact that CLIGEN parameter values have precision to only the hundredth decimal; e.g. in the case of a value of 50, P(W|D) had the minimum possible non-zero value of 0.01, and P(W|W) had a value of 0.50. It is important to note that errors in RATIO only impact the sequencing of dry/wet days and do not influence accumulation totals, the total number of wet days, or sub-daily dynamics.
3. Validation of estimated CLIGEN precipitation parameters

Forms of validation were done to assess error in CLIGEN precipitation input parameters that include cross-validations and comparison to additional ground data. The K-folds cross-validation with $K = 5$ folds used in the GB model selection process uses 80% of the data for training, and therefore, the number of samples used for training of individual models in K-folds is different from the accepted models which use 100% of the data for training. Leave-one-out cross-validation (LCV) was also applied, which in this case considered precipitation parameters from every ground record except for a single station of interest. For this reason, LCV may provide the best representation of model performance in terms of estimation of parameter values for the 5 GB models, though LCV is used to derive prediction errors for relatively few stations compared to K-folds. In this case, 30 stations from both Africa and South America were selected (60 total) by their distribution across wide spatial extents and climates, as much as was possible, resulting in a total of 60 fitted models for LCV analysis.

3.1. K-folds cross-validation

Since each ground station was used to assess prediction error exactly once in K-folds, the distributions in Figure 6 represent error for the entirety of respective ground networks and show both absolute error and relative error distributions. The skewness of each distribution towards large errors resulted in RMSE values that were considerably higher than median errors. For MEAN $P$ and MX.5P, the two most crucial parameters, the median values of the error distribution show that 50% of errors are below 0.95 mm and 1.96 mm/hr, respectively. Corresponding relative errors are shown in the bottom row of plots in Figure 6 and indicate that most relative errors are below 50%. It is important to reiterate that the MX.5P ground network was primarily distributed within the contiguous U.S., and therefore, no stations were within the target regions. For this reason, errors in MX.5P may be expected to be greater in application.

These same error distributions were used to assess the importance of the density of ground network spatial distribution on estimates of MEAN $P$. Spatial densities were
calculated for each station in the international CLIGEN network to a 1 arc deg. resolution using a 3 arc deg. radius search window from the centre points of the 1 arc deg. grid. These spatial densities were correlated against the average relative MEAN $P$ error for each station (plot not shown). No correlation was found, with the slope coefficient and adj. $r$-squared values being close to zero. This suggests that ground stations in areas with low station density have essentially the same degree of error as stations with high density. A similar correlation was done to assess error in MEAN $P$ along trends in accumulation by regressing average absolute MEAN $P$ error for each station against the annual accumulation of each station. Again, no correlation was found suggesting that error is essentially constant over precipitation gradients and different ranges of precipitation totals. However, for the ground station density analysis, a disproportionate number of stations had greater station density because of clustering of stations, and the correlation may therefore be skewed by these stations. Considering this, the average relative error of MEAN $P$ was also assessed for stations grouped into seven station density bins within the range 0–35 stations-degree$^{-2}$. All average relative errors for these bins were within 16–18% and showed no trend with decreasing station density.

### 3.2. Leave-one-out cross-validation of precipitation parameters

To represent the spatial extents and range of climates of the international CLIGEN ground network within the target regions, we performed the leave-one-out cross-validation (LCV) for a distribution of 60 selected stations in Africa and South America taken from those shown in Figure 1. In LCV, one GB model is fitted for each station being cross-validated using all training set datapoints except for those from the selected station. Unlike with the K-folds analysis, resulting error distributions were used to assess performance on the level of individual stations in addition to pooling the data to compare errors between African and South American continents. Prediction errors of the pooled data from LCV are shown in Table 4. The RMSE values are consistent with those obtained from the K-folds analysis. However, prediction errors tend to be higher in Africa than in South America. In the case of the 30 selected station in Africa, there is apparent underestimation bias for some parameters, judging by percent bias and slope. In other cases, percent bias values may be negative (underestimation) despite a slope coefficient close to 1:1, which may be explained by large relative errors towards greater value ranges.

**Table 4. Summary of leave-one-out cross-validation for 30 stations in each continent.**

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>PBIAS</th>
<th>Adj. $r^2$</th>
<th>Slope coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MEAN $P$ estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>3.00 mm</td>
<td>-19.32%</td>
<td>0.914</td>
<td>1.06</td>
</tr>
<tr>
<td>South America</td>
<td>2.03 mm</td>
<td>-5.78%</td>
<td>0.966</td>
<td>1.02</td>
</tr>
<tr>
<td><strong>SDEV $P$ estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>3.89 mm</td>
<td>-10.15%</td>
<td>0.886</td>
<td>0.97</td>
</tr>
<tr>
<td>South America</td>
<td>2.61 mm</td>
<td>-6.55%</td>
<td>0.963</td>
<td>1.03</td>
</tr>
<tr>
<td><strong>SKEW $P$ estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>23.13 mm</td>
<td>3.44%</td>
<td>0.853</td>
<td>0.85</td>
</tr>
<tr>
<td>South America</td>
<td>19.97 mm</td>
<td>-3.12%</td>
<td>0.895</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>RATIO estimation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>5.41 mm</td>
<td>-19.78%</td>
<td>0.431</td>
<td>0.51</td>
</tr>
<tr>
<td>South America</td>
<td>1.58 mm</td>
<td>-8.20%</td>
<td>0.631</td>
<td>0.64</td>
</tr>
</tbody>
</table>
The MEAN P results for a selection of individual stations are shown in Figure 7, represented by 25 stations from the 60 stations used for LCV. From these plots, it is evident that seasonal trends are well represented for the variety of climates and geographical locations. Values of MEAN P can be seen to range from 0 to 30 mm within both African and South American stations. Evidence of systematic bias occurring spatially is possibly shown by the fact that stations grouped nearby show a similar error. For example, three stations are shown in Venezuela (bottom right of Figure 7) that have similar overestimation bias despite seasonal trends and relative differences being well represented. This is not necessarily due to low station density because in Pietersburg, South Africa, for example, where stations have high spatial density, similar estimation bias occurs. In contrast, other stations such as Cocobeach, Gabon and Itaqui, Brazil show full ranges of MEAN P values with no apparent estimation bias.

The estimated values obtained in the LCV were used to parameterize the generation of 20-year CLIGEN simulations for the same 60 stations as before, and the resulting time-series were used to determine daily accumulation frequency distributions and annual rainfall erosivity. For this parameterization, the required MX.5P input values were unknown, and therefore, estimated MX.5P values were used to parameterize both the observation-derived and estimation-derived CLIGEN timeseries. As such, the combined errors in MEAN P, SDEV P, SKEW P, and the transition probabilities are reflected in the comparisons. Overall, the daily accumulation frequency distributions shown in Figure 8 compare well and are not statistically different according to K-S significance testing. Visually, the normalized frequency distributions are similar except for minor deviations.
3.3. Validation of estimated MX.5P using high temporal resolution data

In-situ, high temporal resolution (5-minute or finer) precipitation data from four ground stations in Brazil and South Africa enabled validation of MX.5P for locations within the target area, which was not possible using the training data for MX.5P that came from the ASOS network. However, the difficult data requirement of 20-year, continuous ground records could not be met, and consequently, shorter ground records were used for the comparison. The higher average RMSE of 6.04 mm/hr compared to that obtained from the K-folds analysis (4.695 mm/hr) is expected for two reasons: (1) shorter ground record
lengths and (2) for all four stations, data from the nearest 0.25 arc deg. grid point was used to represent the station locations, and in each case, grid points were far enough away that the stations were outside of the GPM 0.1 arc deg. grid cell covering the grid point, which was not a source of error in the K-folds analysis. In other words, the possibility is tested of a grid point being used to represent a nearby point of interest, and therefore, spatial variability within the grid resolution is an additional source of error. Despite larger error, seasonal trends and value ranges for \( MX.5P \) were reasonably represented. Comparison of the 20-year \( MX.5P \) estimates to ground-derived \( MX.5P \) measurements is shown in Figure 10. Two stations are in eastern Brazil and two stations are in the mountainous Great Escarpment region of South Africa. Each station had variability in reference periods.

The Guaraíra Experimental Basin (GEB) is located in the State of Paraíba, Brazil, 10 km from the Atlantic coast, in a tropical wet climate (Aw in the Köppen climate classification). The reference period for the 1-min GEB ground record was from 2004 through 2019. The 5-min record for the Aiuaba Experimental Basin (AEB), located inland in the State of Ceará, Brazil with a hot semi-arid climate (BSh) was from 2003 through 2018 with some temporal gaps. Precipitation dynamics have been studied at GEB and AEB in such works as Barbosa et al. (2018). The Sani Pass station is located on the border of South Africa and Lesotho in the Mkhomazi Wilderness Area with a subtropical highland climate (Cwb). The 5-min Sani Pass record has a reference period from 2001 to 2006 with some gaps in temporal coverage. The Sentinel Peak station in South Africa is located at high elevation (2,900 m), also with a subtropical highland climate (Cwb). The reference period for the 5-min Sentinel Peak record is the shortest period of the four records, from 2002 through 2005. To account for the coarser temporal resolution of the three 5-min records, a temporal downscaling correction factor was applied to \( MX.5P \) based on Köppen classification (Fullhart et al., 2020).

![Annual Erosivity Validation for LCV Stations](image)

**Figure 9.** Annual rainfall erosivity comparison for LCV stations with values derived from 20-year CLIGEN simulations.
The validation of these four stations is shown in Figure 9, including the raw GPM-derived \textit{MX.5P} values so that the importance of the raw grid-scale \textit{MX.5P} values on estimated \textit{MX.5P} can be judged. The two stations in Brazil show the highest intensity, with the range of intensity values being better represented at AEB than GEB. In particular, the greatest intensity at AEB is well estimated. For both Brazilian stations, the raw \textit{MX.5P} values were substantially underestimated. For the higher elevation stations in South Africa, all series are generally in agreement at Sani Pass, while \textit{MX.5P} values tend to be overestimated at Sentinel Peak. The Sentinel Peak comparison is subject to error from large topographic gradients of 1,000 m within the grid resolution. The shortness of records, particularly in South Africa, may explain the large excursions between successive monthly ground-based \textit{MX.5P} values such as the August value at AEB and the December value at Sani Pass. These values may reflect the impact of events that would otherwise have less weight when longer ground records are averaged.

4. Discussion

Applications for this CLIGEN parameterization dataset are varied and should consider the existing sources of error. Long-term accumulations are reasonably well represented given the fact that monthly accumulations according to TerraClimate are maintained by the parameter adjustment method used to balance Equation 6. Daily-scale outputs such as daily accumulation and the frequency and sequencing of wet days are also reasonably well represented, with \textit{MEAN P}, \textit{SDEV P}, \textit{SKEW P}, and the wet/dry transitional probabilities...
being major determiners of daily-scale precipitation. Moreover, Figure 11 shows the predicted annual N days with precipitation, which as a variable, has an important role in ensuring accurate daily intensities. However, sub-daily dynamics like the intensity-time profile of a storm event, event duration, and the timing and magnitude of peak intensity, are influenced by essentially all precipitation input parameters and are therefore subject to the combination of error from all parameters. This accounts for the high variance in the annual rainfall erosivity comparison in Figure 9. Furthermore, in Figure 11, areas are indicated where sub-hourly intensities are suspected of being too low for extreme events. This problem was identified by extreme value analysis for various sub-hourly durations, which showed that output data from grid points in this area had a poor fit to the generalized extreme value distribution. The issue is evident for sub-hourly rainfall accumulations during events with greater than two-year return period. The issue relates to mechanisms in CLIGEN that dampen 30-minute intensities in locations with large storm accumulations and number of monthly wet days. Applications requiring accurate sub-daily precipitation timeseries in this region will be affected.

Typically, with stochastic timeseries, modelling is done for risk-assessment or impact studies rather than for deterministic modelling driven by observed climate data. Within this scope, possible applications for this dataset include large-scale modelling in
ungauged areas, as may be done with models like Soil and Water Assessment Tool (SWAT, Arnold et al., 1998) and Kinematic Runoff and Erosion (KINEROS, Woolhiser et al., 1990), allowing researchers to avoid the need for downscaling climate data. SWAT can be operated using daily timesteps, and in this case, would not be subject to error in sub-daily precipitation. In fact, SWAT has the capability of using a stochastic weather generator that has essentially the same input parameters as CLIGEN, which enables the possibility of either CLIGEN input parameterizations or output timeseries to be used. Several plot-scale and hillslope-scale, physically based water erosion models require sub-daily inputs, such as the Rangeland Hydrology and Erosion Model (RHEM), which produces runoff and erosion estimates using long-term CLIGEN timeseries. The annual erosivity validation suggests that long-term erosion estimates will be reasonably estimated, though variance will be considerably greater in locations where greater intensity and accumulation of rainfall occurs. Other erosion models that are more empirically based, such as USLE and its revised version, directly use rainfall erosivity as climate input instead of a timeseries, and therefore, this dataset may be applied with the degree of error shown in Figure 9.

Several potential sources of error are inherent in the data used but were not quantified in this analysis. For example, extreme rainfall is not well represented by the ground networks used and is often affected by bias in global climate products. Several ground stations in the tropics showed annual precipitation exceeding 2,500 mm but these were a small proportion of stations. This issue is further impacted by the fact that climate models often underestimate extreme precipitation (Sun and Liang, 2020). Data limitations are also an issue for analysing the accuracy of estimated precipitation parameters in mountainous regions with strong, topographically controlled, climate gradients. The Sentinel Peak and Sani Pass ground stations in Section 3.4 represent a mountainous region that may not have as strong of climate gradients as is possible in other places in the target region. Moreover, ERA5 and GPM are known to have higher uncertainty in complex terrains (Derin et al., 2019; Hu and Yuan, 2021). Another source of error that results from data limitations in the ground data relates to the ranges of values for predictor variables in the GB training phase. Ideally, the target region in which the GB models are applied in a predictive fashion should have predictor variables with ranges of values encompassed by those used for training. This condition was not always met; e.g. latitude (as a predictor variable in the MX.5P GB model) had value ranges corresponding to the ASOS ground network, while the target region had a wider range of values. Similarly, the GHCN-daily network did not represent latitudes as far south as the southern extents of Argentina and Chile. Because most of the gridded climate products have spatial units in arc degrees, the result is that variability in the surface area of square arc degrees along meridian lines is not accounted for at some latitudes. Furthermore, values outside of training data ranges are an issue for climatological predictor variables.

The analysis of generated 20-year CLIGEN output in Section 3.2 showed that 3% more total accumulation occurred in the subset of stations used for LCV than the corresponding observation-derived CLIGEN outputs. This seems to contradict the relationship shown in Figure 3 showing an apparent slope bias of 7% and percent bias metric of 2% towards underestimation by TerraClimate monthly accumulation. This may be partially explained by the difference in time scales being compared. The differences in the spatial distribution of LCV stations compared to all GHCN-Daily stations used in the study is also likely to be
affected by isolated errors in the global climate products. The relationship shown in Figure 3 may also be skewed by outliers that influence the regression slope. Further possibilities exist for comparing aspects of precipitation in the gridded dataset of this paper to spatial climate data such as the TerraClimate dataset. Even with these limitations, this effort has provided estimates of CLIGEN parameters for data sparse regions that enable more confident application of a number of models in cases where locally observed high temporal resolution climate data are unavailable.

5. Conclusion

This paper discusses a gridded parameterization for a stochastic weather generator, CLIGEN, of South America and Africa with 0.25 arc deg. resolution. The various required CLIGEN input parameters were derived from selected climate datasets and included downscaled precipitation parameters that allow downscaled stochastic timeseries to be generated at any grid point. The gridded CLIGEN parameterization product of this paper may be used for several hydrological and erosional modelling applications, in addition to others, and may facilitate research in ungauged areas where point-scale precipitation timeseries are a data limitation. Aspects of long-term and daily-scale precipitation events were demonstrated to be realistic, but more uncertainty exists for generated sub-daily outputs. Therefore, applications utilizing this dataset should consider these potential sources of error.

Acknowledgements

Thanks to the Agricultural Research Service and the U.S. Department of Agriculture for supporting this work. This research was a contribution from the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture. Sincere appreciation is expressed for the many people who are dedicated to collecting and maintaining long-term climatic data and for making the data available to climate researchers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Andrew Fullhart, Ph.D., is currently a post-doctoral researcher at the USDA-ARS Southwest Watershed Research Center in Tucson, Arizona, U.S. Andrew earned a B.Sc. in Geoscience from the University of Iowa, a M.Sc. in Environmental Science from Bemidji State University, and a Ph.D in Hydrologic Science from the University of Wyoming. Andrew’s primary research interests are in hydrological modelling and hydrometeorology. Recently, Andrew has been involved in international soil erosion research projects and creating large-scale geospatial climate datasets.

ORCID

Andrew Fullhart  http://orcid.org/0000-0003-1223-6451
Guillermo E. Ponce-Campos  http://orcid.org/0000-0003-4332-338X
Menberu B. Meles  http://orcid.org/0000-0002-1980-9243
Data availability statement

The data that support the findings of this study are openly available in Ag Data Commons at https://doi.org/10.15482/USDA.ADC/1524754.

Open Scholarship

This article has earned the Center for Open Science badge for Open Data. The data are openly accessible at https://doi.org/10.15482/USDA.ADC/1524754.

References


