Data Descriptor

A 1973–2008 Archive of Climate Surfaces for NW Maghreb

Alberto Ruiz, Maria E. Sanjuan, Juan Puigdefabregas and Gabriel del Barrio *

Estacion Experimental de Zonas Aridas (CSIC), Ctra. Sacramento sn, 04120 La Cañada, Almeria, Spain; aruiz@eeza.csic.es (A.R.); marieta@eeza.csic.es (M.E.S.); puigdefa@eeza.csic.es (J.P.)

* Correspondence: gabriel@eeza.csic.es; Tel.: +34-950-281-045

Academic Editors: Marco Helbich and Jamal Jokar Arsanjani

Received: 21 March 2016; Accepted: 22 June 2016; Published: 27 June 2016

Abstract: Climate archives are time series. They are used to assess temporal trends of a climate-dependent target variable, and to make climate atlases. A high-resolution gridded dataset with 1728 layers of monthly mean maximum, mean and mean minimum temperatures and precipitation for the NW Maghreb (28°N–37.3°N, 12°W–12°E, ~1-km resolution) from 1973 through 2008 is presented. The surfaces were spatially interpolated by ANUSPLIN, a thin-plate smoothing spline technique approved by the World Meteorological Organization (WMO), from georeferenced climate records drawn from the Global Surface Summary of the Day (GSOD) and the Global Historical Climatology Network-Monthly (GHCN-Monthly version 3) products. Absolute errors for surface temperatures are approximately 0.5 °C for mean and mean minimum temperatures, and peak up to 1.76 °C for mean maximum temperatures in summer months. For precipitation, the mean absolute error ranged from 1.2 to 2.5 mm, but very low summer precipitation caused relative errors of up to 40% in July. The archive successfully captures climate variations associated with large to medium geographic gradients. This includes the main aridity gradient which increases in the S and SE, as well as its breaking points, marked by the Atlas mountain range. It also conveys topographic effects linked to kilometric relief mesoforms.

Keywords: Maghreb; Morocco; Algeria; Tunisia; climate; temperature; precipitation; spatial interpolation; ANUSPLIN

1. Summary

Climate is a major factor in many terrestrial processes in environmental science and is crucial in explaining ecosystem structures and processes in extreme environments with low homeostasis such as drylands. Climatic effects are often assessed by interoperation of geo-spatial databases, which requires correspondence between continuous fields of a biological or geographical attribute and climate. This continuity must be both spatial and temporal, leading to the concept of archived time-series.

In this context, an archive of gridded mean maximum, mean and mean minimum surface temperatures and precipitation was computed for the NW Maghreb, with monthly resolution covering 1973 through 2008, and a spatial resolution of 0.00833 degrees (~1 km in a maximum circle). The study area covers a 1,759,330 km² area in NW Africa, with the southern limit at 28°N and the eastern limit at 12°E (Figure 1). It includes all the affected United Nations Convention to Combat Desertification (UNCCD) areas in Tunisia and Algeria, and most of those in Morocco. There are a total of 1728 layers (36 years × 12 months × 4 variables). The surfaces were spatially interpolated from georeferenced climate records drawn from the Global Surface Summary of Day product (GSOD) [1].
The increasing demand for long-term, archived time-series of spatially explicit climate data forms the background for this dataset. In contrast with climate atlases, where period summaries are made before interpolation to yield a single layer of the corresponding variable (e.g., WorldClim [2]), climate archives consist of sequences of layers throughout the period at the working temporal resolution. Climate archives are used directly for topics that include, for example, observed climate change, shifting desert boundaries [3], impact dynamics of drought events [4], or relationships between climate and vegetation productivity [5]. Indirect uses involve post-processing the monthly sequences to produce a climate atlas. The resulting layers can then be used as input data for climatic regionalization [6,7], predictive models of species distribution [8], and even urban models [9], among many others.

Global climate archives do exist, but often at coarse spatial resolutions of around 0.5° (e.g., ERA-Interim [10], BEST [11], CRUTEM4 [12]) that do not match the finer detail of the target dependent variable (e.g., vegetation Net Primary Productivity (NPP) or observed species distribution), and therefore have a limited explanatory value. Regions with fully developed data infrastructures foster the development of climate archives, such as the European ECA&D [13], to meet their research and management needs. However, emerging economies (where, by the way, critical issues of climatic vulnerability and natural resource management often coincide) provide less support to this activity, and large regions of the globe remain blank in terms of climate data coverage. This is the target gap of the NW Maghreb climate archive presented here.

The dataset was computed using the ANUSPLIN algorithm [14], a thin-plate smoothing spline technique for spatial interpolation of noisy multi-variate data approved by the World Meteorological Organization (WMO). Precedent applications have used this approach in North America [15], China [16], and Iberian Peninsula [17]. The protocol for computing and assessing NW Maghreb climate surfaces was derived from those studies.

The NW Maghreb climate archive was constructed under the EC DeSurvey: A Surveillance System for Assessing and Monitoring of Desertification Project. The 2dRUE technique developed in the project [17] estimates land condition based on the paradigm that Rain Use Efficiency (i.e., the ratio of NPP to precipitation) decreases as land degradation proceeds. At the same time, land condition trends are determined by stepwise regressions of vegetation biomass over time and inter-annual aridity. Hence, land degradation is detected after the effects of climate on vegetation have been separated. This
The technique is based on archived vegetation density time-series and corresponding climate fields. The NW Maghreb was a case study for this approach [18].

The end year (2008) of the climate archive coverage period was selected to produce a baseline application right at the start of the UNCCD Strategy (2008–2018). The starting year (1973) was selected to enable a 30-year sequence as commonly required in climate studies. However, the input data product GSOD is continuously updated, and surfaces are interpolated individually. There should therefore be no problem in extending this coverage as necessary.

Queries to the NW Maghreb climate archive are now being implemented under the EC Exploiting the European Open Data Strategy to Mobilise the Use of Environmental Data and Information (MELODIES) Project, as part of an open database service to produce desertification indicators (e.g., aridity index) for national and international agencies.

Despite those particular applications, climate data under an open license are fully multi-purpose elements for a variety of purposes. The NW Maghreb climate archive is offered as a set of four basic variables at high spatial and temporal resolutions. We foresee two main types of uses, primarily for studies requiring continuous sequences over several decades to assess temporal trends of a climate-dependent target variable. This includes compound variables that can be derived from the basic ones, such as Potential Evapotranspiration (PET) computed by the Hargreaves-Samani method [19], or the associated FAO-UNEP aridity index [20]. The results can then be matched to other time-series (e.g., vegetation functions, socio-economic maps) as necessary. A secondary use would be for studies requiring period summaries (means and deviations) of the basic variables and the bioclimatic variables computed from them by appropriate algorithms (e.g., ANUCLIM [21], dismo [22]). Period summaries can be used as such for describing the climate, or as input predictors for other applications (e.g., climate envelope niche modelling).

2. Data Description

2.1. Data

The data cover land ranging from 28°N to 37.3°N and from 12°W to 12°E. A slightly larger data window (see below) was specified to facilitate layer management. The archive starts in January 1973 and ends in December 2008. It contains one raster layer per month for each of these variables:

- Mean maximum temperature (°C): arithmetic mean of the daily maximum temperatures ($t_{\text{max}}$)
- Mean minimum temperature (°C): arithmetic mean of the daily minimum temperatures ($t_{\text{min}}$)
- Mean temperature (°C): arithmetic mean of the daily mean temperatures ($t_{\text{med}}$)
- Total precipitation (mm): sum of the daily precipitations ($\text{rain}$)

The values of these variables appear multiplied by 100 in the corresponding layers. For example, 37.82 was recorded as 3782. This was done to be able to store data in an integer format instead of the real one, which reduces both file size and download times. A correction must therefore be applied when using the raster layers.

The archive refers only to land. While whole surfaces were interpolated for the data window, the sea was masked out before releasing the current version to avoid errors associated with large areas with no supporting data.

2.2. Metadata

The archive is managed in a Geographic Information System (GIS). All the layers are in the network Common Data Form (netCDF) format. Table 1 describes the relevant metadata fields.
### Table 1. Maghreb climate archive metadata.

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>file format</td>
<td>netCDF ‘classic’</td>
<td></td>
</tr>
<tr>
<td>data type</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>file type</td>
<td>binary</td>
<td></td>
</tr>
<tr>
<td>columns</td>
<td>2882</td>
<td></td>
</tr>
<tr>
<td>rows</td>
<td>1320</td>
<td></td>
</tr>
<tr>
<td>ref. system</td>
<td>latlong</td>
<td>Geodetic coordinates in WGS84 datum (EPSG code: 4326)</td>
</tr>
<tr>
<td>ref. units</td>
<td>degrees</td>
<td>The units used to specify position in the layers</td>
</tr>
<tr>
<td>min. X</td>
<td>–11.9924701</td>
<td></td>
</tr>
<tr>
<td>max. X</td>
<td>12.0145889</td>
<td></td>
</tr>
<tr>
<td>min. Y</td>
<td>27.9983998</td>
<td></td>
</tr>
<tr>
<td>max. Y</td>
<td>38.9939993</td>
<td></td>
</tr>
<tr>
<td>resolution</td>
<td>0.0083300</td>
<td>Equivalent to 925.63 m in a maximum circle</td>
</tr>
<tr>
<td>scaling factor</td>
<td>0.01</td>
<td>Number which multiplies layer values to yield T or P</td>
</tr>
<tr>
<td>min. value</td>
<td>e.g., –1286</td>
<td>Must be rescaled: –1286 × scaling factor = –12.86</td>
</tr>
<tr>
<td>max. value</td>
<td>e.g., 1314</td>
<td>Must be rescaled: 1314 × scaling factor = 13.14</td>
</tr>
<tr>
<td>value units</td>
<td>°C or mm</td>
<td>Temperature or precipitation respectively</td>
</tr>
<tr>
<td>flag value</td>
<td>–32768</td>
<td>Used to mask sea and background areas</td>
</tr>
</tbody>
</table>

#### 2.3. Dataset

The archive is organized in four large compressed files, one per variable. File names are according to the following conventions:

- Characters 1–4: variable name (tmax, tmin, tmed or rain)
- Characters 5–8: year (1973 to 2008)
- Characters 9–10: month (01 to 12)

For example, rain197301 refers to the total precipitation in January 1973.

#### 3. Methods

Climate surfaces were interpolated using the ANUSPLIN package [23] from monthly summaries of georeferenced meteorological stations. This interpolation technique is a generalisation of a standard multi-variate linear regression, in which a smooth non parametric function is used instead of a parametric model. The complexity of the fitted surface is controlled by Generalized Cross Validation (GCV), an estimate of the surface predictive error. ANUSPLIN was selected because it compares favourably with kriging and other spatial interpolation techniques [24,25], and because it lends itself particularly well to automatic processing with changing configurations of data points.

Daily input data were downloaded from the Global Surface Summary of Day (GSOD) product [1]. GSOD provides historical records of synoptic daily climate events for a large number of stations distributed worldwide. Some quality control is performed, but in general the data are not adjusted for inhomogeneity. Monthly precipitation data were further added from the Global Historical Climatology Network-Monthly product (GHCN-Monthly version 3) [26].

Queries within a working window spanning 14°W to 14°E and 24°N to 39°N, yielded approximately 1,150,000 daily records from 140 stations located in Mauritania (1), Western Sahara (1), Morocco (29), Algeria (70), Tunisia (24), Libya (9) and Spain (6) (Figure 1).

The GSOD daily data were summarized in monthly records, which were subsequently filtered following WMO [27] and Global Climate Observing System (GCOS) [28] guidelines for the preparation of climate records. Meteorological stations are scarce in the desert zones south of the study area, which could lead to problems with interpolation. Therefore, the recommendation that ‘a monthly value should not be calculated if more than ten daily values are missing or five or more consecutive daily values are missing’, was relaxed. Monthly records were rejected if they did not contain a minimum of
daily data (5 for temperature or 15 for precipitation), and the corresponding year did not have at least 120 daily records and 12 monthly summaries. Missing data in the series were not filled in.

As a result, the number of stations used for interpolation was variable, ranging from 52 to 118 for temperature and from 40 to 106 for precipitation. In general, more missing data tended to appear in early years, especially before 1981.

Trivariate quadratic spline models were set up using longitude, latitude and altitude above sea-level as independent variables. ANUSPLIN was configured to minimize the GCV of each surface. The reporting station altitude was appropriately scaled to km, and precipitation was transformed by its square root to reduce skewness as recommended [29]. An inverse transformation was later applied to the precipitation grids to obtain final surfaces in mm.

The quality of the interpolated archive is reported on three levels [23]. The first one targets errors in the spatial structure of the interpolated surfaces, as measured by the ratio of the signal (i.e., the number of data points actually used in the interpolation, an indicator of the degrees of freedom associated with each surface) to the total number of data points. In a model with very high goodness-of-fit, it should be about 0.5, but 0.40 and 0.80 are acceptable limits. Higher values indicate either insufficient data or short range correlation in the data values [24,29]. In this archive, this ratio ranged from 0.36 to 0.88 (Figure 2), which in general indicates valid smoothing [15]. Approximately 18% of the surfaces yielded a ratio over 0.8, indicating wide spatial variability or lack of adequate data. This is compatible with low density and sparse distribution of stations in some cases.

The second estimator of quality targets the predictive error of the interpolated surfaces (Figure 3). The square root of GCV (RTGCV) is an absolute error in the original units of the variable. This was less...
than 1.5 °C for mean and mean minimum temperatures, but peaked up to 2.5 °C for mean maximum temperatures in the summer months. For precipitation surfaces, RTGCV ranged from 1.2 to 2.5 mm. However it is more meaningful to divide it by the mean, resulting in a relative error. This was below 10% in general, but very low precipitation in the summer months caused it to rise to 40% in July.

![Figure 3. Absolute errors of the interpolated surfaces Means, standard deviations and extremes shown by months across the climate archive: (a) Mean temperature; (b) Mean maximum temperature; (c) Mean minimum temperature; (d) Precipitation.](image)

Finally, the output archive also includes an error surface per interpolated surface. This represents a Bayesian estimate of the standard error of the model, and can be used to assess the spatial distribution of interpolation errors, for example in error propagation exercises. This set of surfaces is not provided with the main dataset. However, it may be requested by writing to the corresponding author.

The uncertainty associated with each layer is comprised of several different errors: in individual stations data, in spatial configuration of input data points, and in spline fitting. Individual station errors are difficult to track, although the filtering criteria described above are expected to keep quality regular in the general input dataset. Such filters caused spatial arrangements of input data points to vary from month to month, yielding generally higher densities in the more populated zones in the north of the study area. Because such zones also have a more complex topography (related to the Atlas mountain ranges), this bias at least worked in favour of connecting topography and data density. In view of the errors reported, and after visual inspection, we believe the archive successfully captures climate variations associated with large-to-medium geographic gradients. This includes the dominant aridity gradient increasing in the S and SE direction, as well as the breaking points marked by the system of Atlas mountain ranges. It also shows topographic effects linked to kilometric relief mesoforms: altitude, aspect and orientation with respect to prevailing winds. However, thermal inversion in smaller depressions, and in general any microclimate feature of hectometric size or smaller,
which lacks a meteorological station to support interpolation, should not be expected to be properly reflected on the surfaces.

4. User Notes

The monthly surfaces are compressed in four independent zip files, one per variable. The decompressed raster layers can be directly imported by several GIS packages or converted to the required GIS format using the gdal_translate utility program [30].

The following tasks depend to a great extent on whether enhanced time-series or period summaries are required. A typical database sequence involves organizing a series of layers, computing new compound serial variables, and finally deriving period summaries. Examples of some tools are given below.

For general work, all the above tasks may be performed with the R package raster [31] which provides several functions for building and managing stacks and raster-layer bricks, which suit the nature of the series of variables well.

Many studies of yearly drylands climate variations use an aridity index based on comparison of precipitation with Potential Evapotranspiration (PET). The r2dRue R package [32] has several utilities for exploiting climate surface time series in this vein. Especially relevant functions are:

- petHgsm to compute one layer of Potential Evapotranspiration (PET) from \( t_{\text{max}}, t_{\text{med}}, t_{\text{min}} \) and extraterrestrial solar radiation layers (internally computed) using the Hargreaves-Samani method [19]. The sister function batchpetHgsm does the same, but for a specified sequence of layers.

- aiObsMe to compute the FAO-UNEP aridity index as the precipitation ratio to PET for a sequence of layers.

- rgf.summary to query a sequence of layers and save the resulting summary in a new layer. This summary refers to the population of each grid cell in the specified sequence. Choices are sum, maximum, minimum, mean, count, range, standard deviation, variance and median. The length of the sequences is controlled by the argument step.

These programs are appropriate for building up a climate atlas for the whole period or a subset of it. It can be completed with compound descriptive variables to meet a variety of research goals. For example, the biovars function in the R package dismo [22] uses period summaries of the four basic variables (\( t_{\text{max}}, t_{\text{min}}, t_{\text{med}} \) and \( \text{rain} \)) to generate a set of 19 bioclimatic variables that describe most of the climate variation (mean annual temperature, maximum temperature of the warmest month, precipitation of the warmest quarter, etc.). Such an enhanced atlas can be used for descriptive purposes, or as input data for further modelling (e.g., species distribution models).

All the programs cited in this section are free open source software.

Acknowledgments: This dataset was constructed with funding from the following projects: DeSurvey IP (European Commission FP6 contract No. 003950), MesoTopos (Junta de Andalucia PE P08-RNM-04023) and MELODIES (European Commission FP7, contract No. 603525). The latter also covered the open access publishing costs. Two anonymous reviewers made constructive comments and suggestions that improved the manuscript.

Author Contributions: Alberto Ruiz assembled the input database and performed the spatial interpolations. Maria E. Sanjuan designed the experimental framework and made the geographic validation. Juan Pugdefabregas advised on the archive usability requirements, and edited the manuscript. Gabriel del Barrio assessed spatial errors and wrote the manuscript.

Conflicts of Interest: The authors declare they have no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.
Abbreviations

The following abbreviations are used in this manuscript:

- EPSG: European Petroleum Survey Group
- FAO: Food and Agriculture Organization of the United Nations
- GCV: Generalized Cross Validation error
- GCOS: Global Climate Observing System
- GHCN-Monthly: Global Historical Climatology Network-Monthly
- GSOD: Global Surface Summary of the Day
- NCEI: National Centers for Environmental Information
- netCDF: Network Common Data Form
- NPP: Net Primary Productivity
- PET: Potential Evapotranspiration
- RTGCV: Square Root of GCV
- UNCCD: United Nations Convention to Combat Desertification
- UNEP: United Nations Environment Program
- WMO: World Meteorological Organization

References


