SCIENTIFIC DATA

OPEN

Data Descriptor: MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling

Received: 12 July 2016
Accepted: 28 April 2017
Published: 20 June 2017

Greta C. Vega[1], Luis R. Pertierra[1] & Miguel Ángel Olalla-Tárraga[1]

Species Distribution Models (SDMs) combine information on the geographic occurrence of species with environmental layers to estimate distributional ranges and have been extensively implemented to answer a wide array of applied ecological questions. Unfortunately, most global datasets available to parameterize SDMs consist of spatially interpolated climate surfaces obtained from ground weather station data and have omitted the Antarctic continent, a landmass covering c. 20% of the Southern Hemisphere and increasingly showing biological effects of global change. Here we introduce MERRAclim, a global set of satellite-based bioclimatic variables including Antarctica for the first time. MERRAclim consists of three datasets of 19 bioclimatic variables that have been built for each of the last three decades (1980s, 1990s and 2000s) using hourly data of 2 m temperature and specific humidity. We provide MERRAclim at three spatial resolutions (10 arc-minutes, 5 arc-minutes and 2.5 arc-minutes). These reanalysed data are comparable to widely used datasets based on ground station interpolations, but allow extending their geographical reach and SDM building in previously uncovered regions of the globe.

Design Type(s)	observation design • time series design
Measurement Type(s)	temperature of air • atmospheric water vapour
Technology Type(s)	computational modeling technique
Factor Type(s)	temporal_interval • spatial resolution • climatic variable
Sample Characteristic(s)	Earth

[1]Department of Biology and Geology, Physics and Inorganic Chemistry, Rey Juan Carlos University, Calle Tulipán s/n, Móstoles (Madrid) 28933, Spain. Correspondence and requests for materials should be addressed to G.C.V. (email: greta.vega@urjc.es).

Background & Summary

The application of species distribution modelling (SDM) has boomed during the past ten years in the fields of biogeography, macro-ecology and conservation biology[1]. SDMs combine information on species occurrence with environmental characteristics to estimate the suitable distributional area[2]. The theory behind this relationship has been developed since the beginning of the 20th century[3]. From a macro-ecological perspective, climate-richness models based on water-energy dynamics[4] have also displayed solid predictive ability to forecast responses to climate change (e.g., woody plants[5]). These models are built with environmental variables such as temperature and specific humidity, which are also physiologically meaningful[6–12] in different parts of the globe[13–15]. The advent of GIS and the increased availability of global environmental data in recent years have favoured the proliferation of diverse kinds of SDMs intended to answer a wide range of applied ecological questions[2] (e.g., discovering biodiversity, conservation planning, health security, invasion ecology).

In the current macro-ecological research scene, WorldClim[16] has become a most valuable and widely used source to retrieve high-resolution GIS climatic layers to build SDMs. These layers consist of spatially interpolated climate surfaces for global land areas obtained from weather station data using splines. WorldClim provides among other datasets 19 bioclimatic variables derived from precipitation and temperature records for the period 1950 to 2000. This set of bioclimatic variables describes temperature and water related annual tendencies, seasonality and extreme climatic conditions, including a combination of both environmental factors.

Despite the extensive application of WorldClim data in SDM approaches, some limitations have been recently identified as inherent to the usage of climatic datasets based on ground station interpolations[17]. While a high number of weather stations are spatially scattered to intensively survey the climatic conditions of highly urbanised countries, some large areas of the globe are not covered by a dense number of weather stations. For instance, some geographical areas at high latitudes and altitudes (such as Greenland), which are forecasted to undergo a dramatic temperature increase under current climate change scenarios[18,19], lack direct climatic information from weather stations. Furthermore, a complete continental landmass as the Antarctic is omitted in WorldClim. Investigating the climate-driven redistribution of biodiversity in a warming planet would benefit from a detailed climatic description of these zones.

In parallel to the development and wide circulation of WorldClim, global-level satellite data collections have also become increasingly available and reanalyses of this information have served to deliver a set of physical and chemical variables to characterize the climatic conditions of the Earth's surface[1]. These reanalyses combine a background forecast model and data assimilation routines. Then, the data assimilation fuses the available observations with the forecasts to produce uniform gridded data. Therefore, those areas accumulating more observation tools (grounded and remote) have higher accuracy levels, while those with low sampling effort are estimated using the forecast model. Remotely sensed information has improved the performance of SDMs[17], including models aimed to assess the establishment of non-indigenous species in Antarctica[20,21]. In this context, the Modern Era Retrospective-analysis for Research and Applications (MERRA) is a NASA atmospheric data reanalysis of satellite information containing 28 data products with several variables each[22].

Here, we have reproduced the computation and interpolation methods of WorldClim[23] to generate MERRAclim, a global set of satellite-based bioclimatic variables. MERRAclim consists of three datasets of

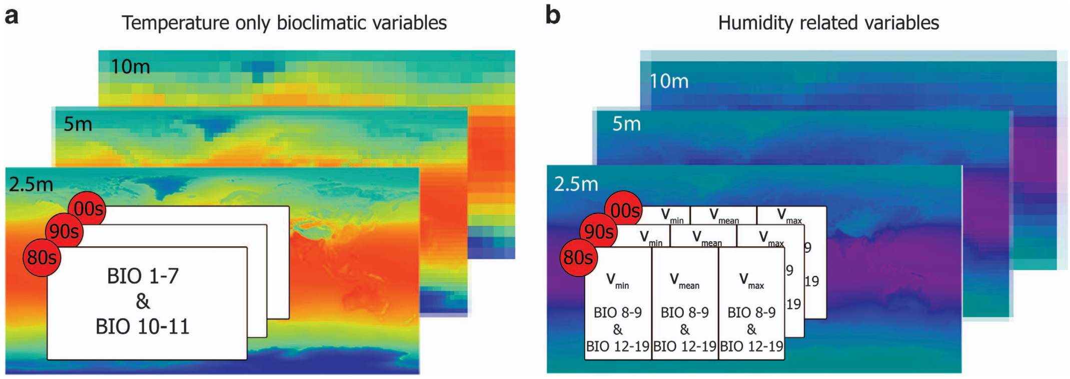

Figure 1. Structure of the MERRAclim dataset. (**a**) the temperature-only bioclimatic variables (BIO1-BIO7 & BIO10–11) are provided at 3 resolutions (coloured maps; 2.5 arc-minutes, 5 arc-minutes, 10 arc-minutes). For each resolution, a single dataset is available per decade (white boxes with red label; 1980s, 1990s, 2000s); (**b**) The humidity-related bioclimatic variables (BIO8–9 & BIO12–19) are provided at 3 resolutions (coloured maps). For each resolution three alternative versions are available (V_{max}, V_{mean}, V_{min}) per decade (white boxes with red label).

19 bioclimatic variables that have been built for each of the last three decades using hourly data from the 1st of January 1981 to the 31st of December 2010. MERRAclim bioclimatic variables are computed from geographically homogeneous temperature and specific humidity gridded data and, hence, benefit from the same assimilation technique across the globe, including Antarctica. MERRAclim (Data Citation 1) datasets are derived from MERRA data, which has been extensively validated in the literature. We also provide a quantitative comparison of MERRA data and Antarctic Meteorological stations. The resolution of the gridded data has been done using a spline method to provide MERRAclim bioclimatic variables at three different resolutions (Fig. 1). We provide a comparison with WorldClim[16] to facilitate the interpretation of future MERRAclim-based results with past research based on WorldClim.

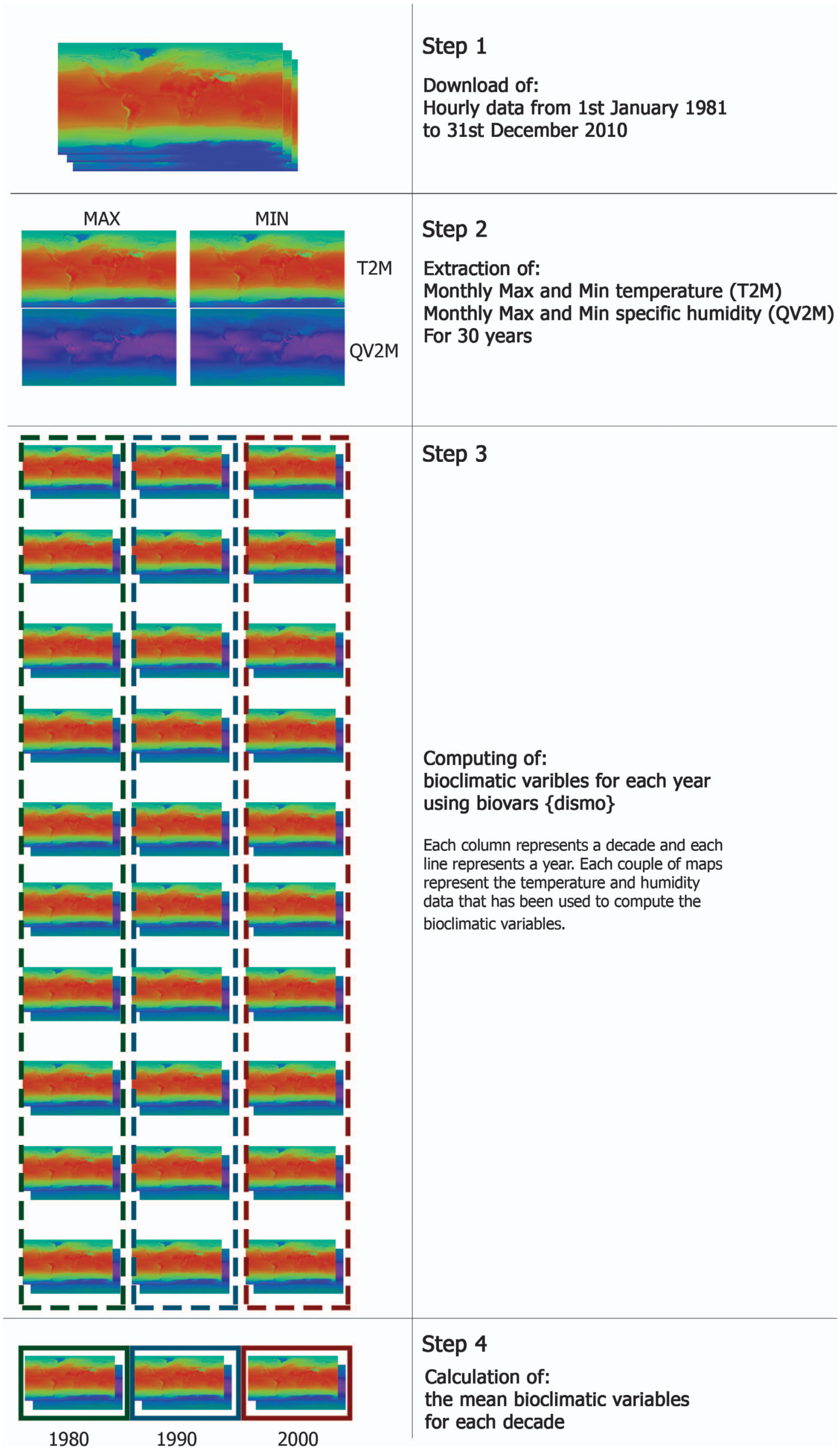
Step 1
Download of:
Hourly data from 1st January 1981
to 31st December 2010
MAX
MIN
T2M
QV2M
Step 2
Extraction of:
Monthly Max and Min temperature (T2M)
Monthly Max and Min specific humidity (QV2M)
For 30 years
Step 3
Computing of:
bioclimatic varibles for each year
using biovars {dismo}
Each column represents a decade and each
line represents a year. Each couple of maps
represent the temperature and humidity
data that has been used to compute the
bioclimatic variables.
1980
1990
2000
Step 4
Calculation of:
the mean bioclimatic variables
for each decade

Figure 2. MERRAclim processing step by step. Computational steps followed to create the bioclimatic variables for each decade.

Methods

Step 1: We used 2 m air temperature (Kelvin degrees) and 2 m specific humidity (kg of water/kg of air) hourly data from the Modern Era Retrospective Analysis for Research and Applications Reanalysis[22] (MERRA) 2D Incremental Analysis Update atmospheric single-level diagnostics product (short name: MAT1NXSLV) provided by the NASA Global Modelling and Assimilation Office from the 1st of January 1981 to the 31st of December 2010 (Fig. 2). Specific humidity is an absolute measure of humidity which indicates the real amount of water present in the atmosphere that, contrarily to relative humidity, is not affected by changes in pressure or temperature[24].

Step 2: After opening the downloaded NetCDF files using the R package *RNetCDF*[25], for each month of the 30 year series, minimum and maximum temperature and specific humidity were extracted.

Step 3: For each year, three sets of bioclimatic variables were generated using the 'biovars' function of the R package *dismo*[26]. This function uses monthly minimum and maximum temperature and precipitation (mm) of the 12 months of a year following WorldClim protocols. Bioclimatic variables in WorldClim are: BIO1: Annual Mean Temperature, BIO2: Mean Diurnal Range, BIO3: Isothermality, BIO4: Temperature Seasonality, BIO5: Max Temperature of Warmest Month, BIO6: Min Temperature of Coldest Month, BIO7: Temperature Annual Range, BIO8: Mean Temperature of Wettest Quarter, BIO9: Mean Temperature of Driest Quarter, BIO10: Mean Temperature of Warmest Quarter, BIO11: Mean Temperature of Coldest Quarter, BIO12: Annual Precipitation, BIO13: Precipitation of Wettest Month, BIO14: Precipitation of Driest Month, BIO15: Precipitation Seasonality, BIO16: Precipitation of Wettest Quarter, BIO17: Precipitation of Driest Quarter, BIO18: Precipitation of Warmest Quarter and BIO19: Precipitation of Coldest Quarter.

For MERRAclim we used specific humidity (kg of water/kg of air) instead of precipitation (mm) values. To allow users' choice of the most appropriate data for their ecological work we produced three versions of bioclimatic variables which depend on the specific humidity value used to produce them: a first one using monthly maximum specific humidity (V_{max}), a second one using monthly mean specific humidity (V_{mean}) and a third one using monthly minimum specific humidity (V_{min}).

Step 4: Once the 30 datasets of bioclimatic variables (one for each year) and their respective three versions (V_{max}, V_{mean}, V_{min}) were created, we merged them by calculating the mean for each decade (1980s, 1990s, 2000s) thus obtaining the following datasets: 80s(V_{max}), 80s(V_{mean}), 80s(V_{min}); 90s(V_{max}), 90s(V_{mean}), 90s(V_{min}); 00s(V_{max}), 00s(V_{mean}), 00s(V_{min}). Spatial resolution of these datasets corresponds to the one of original MERRA raw data: 40 min of latitude and 30 min of longitude.

Step 5 (Fig. 1): Each dataset has been interpolated using the Spline geoprocess of type regularised, which yields a smooth surface and smooth first derivatives in ArcMap[27], to obtain the datasets at the same three coarsest resolutions available in WorldClim (10 arc-minutes, 5 arc-minutes and 2.5 arc-minutes). Since its initial release to the public in 2005, WorldClim has been cited by 6,060 scientific papers, of which almost one fifth had a focus on SDMs (ISI Web of Science literature survey based on the search-string: TOPIC = 'SDM' OR 'Species Distribution Model*' OR 'ENM' OR 'Environmental Niche Model*'; 19th of December 2016).

Spline is a deterministic interpolation method that has been shown to deliver similar results and sometimes slightly underperform when compared to Kriging[28–31] (a stochastic method). Nevertheless, it has been commonly considered as appropriate for interpolation of densely sampled environmental variables[32], for instance to produce WorldClim[16], as it does not assume the process is normal nor stationary. Instead, the spline approach is based on the assumptions that the interpolation function passes through the data points and at the same time is as smooth as possible. This assumption is important as it implies that the data between two points that might be very different because of their physical characteristics will differ more depending on the interpolation technique used. Indeed, the absolute difference between the values obtained via Kriging and Spline of MERRAclim show that the littoral and high elevation areas have the larger bias that might reach 10 °C for BIO1 and 0.004 kg/kg for BIO12 (Supplementary Fig. 1).

Step 6: The final values have been multiplied by 10 for the temperature related variables (BIO1-BIO11) and by 100,000 the humidity related variables (BIO12-BIO19) to store the information as integers and therefore using rasters with a smaller depth of pixel allowing a faster download and easier manipulation in GIS software.

Step 7: As the biovars function was designed to be used with precipitation, not specific humidity, some of the resulting bioclimatic variables needed to be divided to have ecological meaning. Accordingly, the resulting BIO12 has been divided by 12 to obtain the final MERRAclim BIO12, which describes the annual mean of specific humidity instead of cumulative annual rainfall. The resulting BIO16, BIO17, BIO18 and BIO19 have all been divided by 3 so that the corresponding final MERRAclim variables inform on quarterly means instead of cumulative quarterly precipitation.

Code availability

Code is available in Supplementary File 1.

Data Records

The MERRAclim dataset (Data Citation 1,Fig. 1) is provided for three decades (1980s, 1990s and 2000s) in three versions (V_{min}, V_{mean} and V_{max}) and at three spatial resolutions (10 arc-minutes, 5 arc-minutes

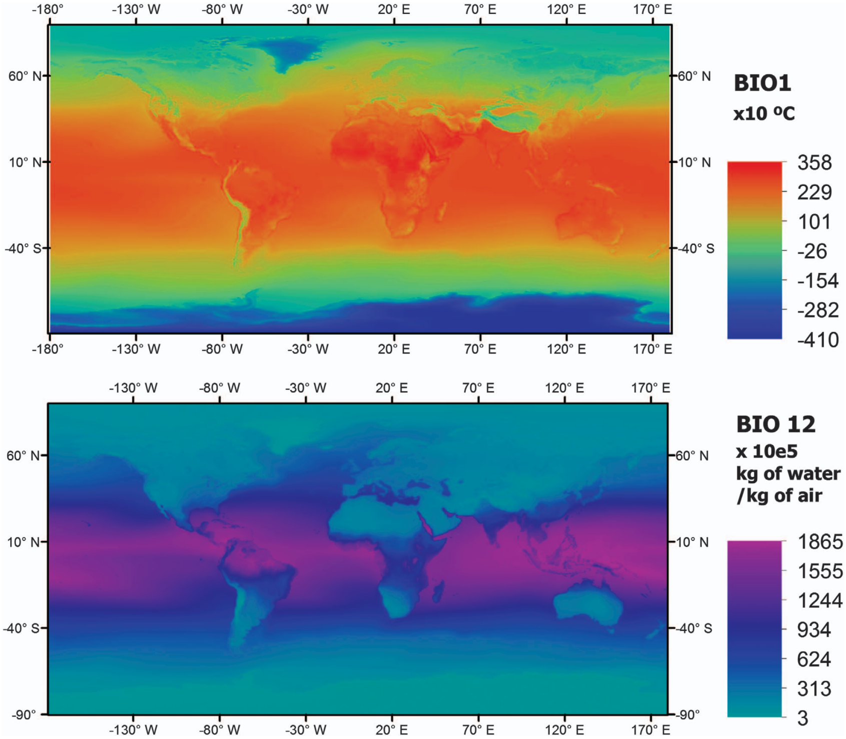

Figure 3. BIO1 (top; Annual mean temperature) and BIO2 (bottom; Annual mean humidity) from the 2000s decade V_{min} at 10 arc-minutes resolution.

and 2.5 arc-minutes). We provide users with the three versions so they can choose the one that best meets their research needs. Example layers for BIO1 and BIO12 from the 2000s decade V_{min} at 10 arc-minutes resolution are depicted in Fig. 3. The datasets are zipped folders and are named following the convention: resolution_version_decade. Each folder contains the 19 bioclimatic variables (Table 1) as georeferenced GEOtiff files and are titled with the standard combination: resolution_version_decade_bioclimatic.tif. BIO1- BIO11 represent temperature (in degree Celsius multiplied by 10) and BIO12-BIO19 are specific humidity (kg of water/kg of air multiplied by 100,000). Each of these zip folders can be downloaded individually.

Temperature-related bioclimatic variables (BIO1-BIO7 and BIO10-BIO11) are identical in the three versions of the dataset because they do not rely on specific humidity data which is the variable that is inputted in three different versions (see Methods). The remaining bioclimatic variables show very little variation among the three different versions (see Usage Notes).

MERRAclim is derived from MERRA, a global reanalysis that assimilates available ground and satellite observations with a background model forecast. Thus, its uncertainty, as a reanalysis, is related to the location of in situ and remote observations. Consequently, developed nations in the Northern Hemisphere have smaller uncertainty than isolated areas[33]. MERRA has been evaluated and compared to other reanalyses since its release, we refer to the literature in the Technical Validation to justify its suitability to derive bioclimatic variables. In addition to this, we have performed a quantitative comparison between MERRA and Antarctic ground stations which shows a strong correlation although the values from MERRA are colder.

We provide a comparison between MERRAclim and WorldClim (see Usage Notes) to assist with the choice of version. Water-related variables (BIO12-BIO17) and the combined bioclimatic variables (BIO8, BIO18 and BIO19) from version V_{min} are the ones that correlate the most with their corresponding bioclimatic variables from WorldClim, whereas BIO9 correlates more strongly with its WorldClim counterpart from V_{max} or V_{mean}. Overall, MERRAclim varies the most compared to WorldClim for those bioclimatic variables sensitive to extremes.

Technical Validation

The Modern-Era Retrospective Analysis for Research and Applications (MERRA) was made public in 2011 aiming to improve upon the hydrologic cycle described in earlier reanalyses[22]. We chose MERRA to produce global bioclimatic variables as, in several evaluations englobing several decades, it showed high reliability for water[34] and energy variables[35] at different scales and in different regions. To reinforce these validations, we have carried on a quantitative comparison between MERRA and Antarctic ground stations.

Variable shortname	Variable description	Units	Variable type	Resolution	Version	Decade	Naming convention	Example
BIO1	Annual Mean Temperature	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO1.tif	10m_min_00s_BIO1.tif (BIO1 from 00 s decade at 10 m resolution from version Vmin)
BIO2	Mean Diurnal Range Temperature	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO2.tif	2_5m_mean_90s_BIO2.tif (BIO2 from 90 s decade at 2.5 m resolution from version Vmean)
BIO3	Isothermality (BIO2/BIO7) (* 100)	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO3.tif	10m_max_80s_BIO3.tif (BIO3 from 80 s decade at 10 m resolution from version Vmax)
BIO4	Temperature Seasonality (standard deviation *100)	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO4.tif	5m_min_00s_BIO4.tif (BIO4 from 00 s decade at 5 m resolution from version Vmin)
BIO5	Max Temperature of Warmest Month	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO5.tif	2_5_mean_90s_BIO5.tif (BIO5 from 90 s decade at 2.5 m resolution from version Vmean)
BIO6	Min Temperature of Coldest Month	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO6.tif	10m_min_80s_BIO6.tif (BIO6 from 80 s decade at 10 m resolution from version Vmin)
BIO7	Temperature Annual Range (BIO5-BIO6)	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO7.tif	5m_max_00s_BIO7.tif (BIO7 from 00 s decade at 5 m resolution from version Vmax)
BIO8	Mean temperature of most humid quarter	*10 °C	combined	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO8.tif	10m_mean_90s_BIO8.tif (BIO8 from 90 s decade at 10 m resolution from version Vmean)
BIO9	Mean temperature of least humid quarter	*10 °C	combined	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO9.tif	2_5m_min_80s_BIO9.tif (BIO9 from 80 s decade at 2.5 m resolution from version Vmin)
BIO10	Mean Temperature of Warmest Quarter	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO10.tif	5m_max_00s_BIO10.tif (BIO10 from 00 s decade at 5 m resolution from version Vmax)
BIO11	Mean Temperature of Coldest Quarter	*10 °C	temperature-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO11.tif	2_5m_mean_80s_BIO11.tif (BIO11 from 80 s decade at 2.5 m resolution from version Vmean)
BIO12	Annual Mean Specific Humidity	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO12.tif	10m_min_90s_BIO12.tif (BIO12 from 90 s decade at 10 m resolution from version Vmin)
BIO13	Specific Humidity of most humid Month	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO13.tif	5m_max_80s_BIO13.tif (BIO13 from 80 s decade at 5 m resolution from version Vmax)
BIO14	Specific Humidity of least humid Month	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO14.tif	5m_mean_90s_BIO14.tif (BIO14 from 90 s decade at 5 m resolution from version Vmean)
BIO15	Specific Humidity seasonality (Coefficient of variation)	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO15.tif	2_5m_min_00s_BIO15.tif (BIO15 from 00 s decade at 2.5 m resolution from version Vmin)
BIO16	Specific Humidity Mean of most humid quarter	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO16.tif	5m_mean_90s_BIO16.tif (BIO16 from 90 s decade at 5 m resolution from version Vmean)
BIO17	Specific Humidity Mean of least humid quarter	100000 * kg of water/kg of air	water-related	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO17.tif	10m_max_80s_BIO17.tif (BIO17 from 80 s decade at 10 m resolution from version Vmax)
BIO18	Specific Humidity Mean of warmest quarter	100000 * kg of water/kg of air	combined	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO18.tif	5m_mean_00s_BIO18.tif (BIO18 from 00 s decade at 5m resolution from version Vmean)
BIO19	Specific Humidity Mean of coldest quarter	100000 * kg of water/kg of air	combined	10 m, 5 m, 2.5 m	min, mean, max	80 s, 90 s, 00 s	resolution_version_decade_BIO19.tif	2_5m_max_00s_BIO19.tif (BIO19 from 00 s decade at 2.5 m resolution from version Vmax)

Table 1. Summary of MERRAclim's bioclimatic variables. There are 11 temperature-related variables of which two are combined and 7 water-related variables of which two are combined.

Reference	Covered Dates	Variables	Location
Ashouri *et al.*[33]	1979–2010	Water	USA
Bracegirdle & Marshall[43]	1979–2008	Energy	Antarctica
Essou *et al.*[34]	1979–2003	Water and Energy	USA
Lader *et al.*[45]	1979–2009	Water and Energy	Alaska (USA)
Bosilovich *et al.*[37]	1979–2012	Water	Central USA
Roberts *et al.*[35]	2000–2010	Water	West Africa
Lindsay *et al.*[42]	1981–2010	Water and Energy	Arctic
Lorenz *et al.*[36]	1989–2010	Water	Global
Bosilovich *et al.*[39]	1979–2009	Water and Energy	Global
Trenberth *et al.*[38]	1979–2005	Water	Global
Cullather & Bosilovich[40]	1979–2005	Water	Polar Regions
Cullather & Bosilovich[44]	1979–2005	Energy	Polar Regions
Serreze *et al.*[41]	1979–2010	Water and Energy	Arctic

Table 2. Summary of the references evaluating and comparing the water and energy variables of the MERRA dataset.

Location	Region	Time span	Number of records	Time step	Pearson's correlation	Slope	Intercept	Explained variance	Bias
Palmer	East Antarctica	2007–10	31066	1 h	0.82	0.86	− 1.57 °C	0.68	± 2.4 °C
McMurdo	West Antarctica	1990s	10865	6 h	0.87	0.73	− 8.9 °C	0.76	± 5.4 °C
Amundsen South Pole	Interior Antarctica	1990s	87618	1 h	0.92	0.69	− 11 °C	0.86	± 4.5 °C

Table 3. Description of the time series from the United States Antarctic Program meteorological stations used to fit linear regressions and to calculate Pearson's correlation.

At large scales, the comparisons with other reanalyses showed that MERRA performed similarly[36,37] or outperformed some[38,39] (Table 2). Although it presents some weaknesses these have also been found in past reanalyses[40]. Regionally, MERRA has been compared for Polar Regions where it has shown to be one of the most consistent for both water[41,42] and energy variables[43,44]. Although it has been demonstrated that it contains some errors for the energy budgets, these are not directly related with temperature[45].

At more regional scales, the comparisons between reanalyses show that each modern reanalyses perform better from one area to another, for instance, in Alaska, MERRA is the best reanalysis for interior areas, while other reanalyses are more suitable in North and the South-eastern Alaska[46].

We have compiled hourly data from the University of Wisconsin-Madison Automatic Weather Station Program (http://amrc.ssec.wisc.edu) to estimate the regional variability, the correlation and the bias of MERRA data in Antarctica. We used temperature time series for three United States Antarctic Program (USAP) bases each located in a different Antarctic region: Palmer (West Antarctica, 2007–2010, data not available from June to November 2010), McMurdo (East Antarctica, 1990s decade) and Amundsen-Scott South Pole (Interior Antarctica, 1990s decade) (Table 3). For Palmer and Amundsen time series the data is available hourly, whereas for McMurdo the time series step is every 6 h. Pearson's correlation coefficients for the three regions show a high correlation (over 0.8) between MERRA and USAP ground stations, being stronger for higher latitudes (Table 3). The same relationship trend is shown by the linear regressions (with a slope between 0.69 and 0.86) that explained over 68% of the variance and showed that MERRA records were colder, with the largest difference in Amundsen (11 °C) (Fig. 4). However, the residuals of the linear regressions show that bias over 5 °C are rare (Fig. 5). Indeed, Inter Quartile Range of the residuals are between 2.4 and 5.4 °C (Table 3) and are larger for the coldest months when the low temperature does not allow ecological activity (Fig. 5). Furthermore, we have compared the resulting bioclimatic variables using both datasets (Table 4), this comparison leads to the same conclusion as the hourly comparison: MERRA records are colder than USAP records but, as they are summarised values, the difference is smaller than when comparing hourly.

Everything considered, MERRA data is one of the best reanalyses available with a global extent as the evaluations at different scales in different regions have shown. In the Antarctic region in particular, the temperatures recorded by MERRA are colder than the ground stations, but this bias is small during the summer months, when the biological activity takes place and it is more visible in extremely high latitudes, i.e., the South Pole, where ecologically viable temperatures are never reached.

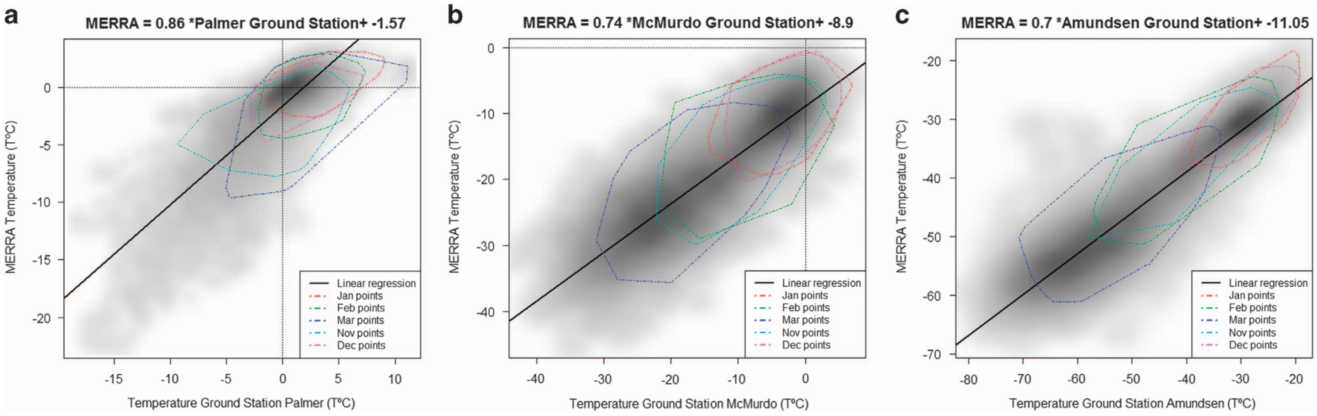

Figure 4. Density scatterplot of temperature time series for MERRA and United States Antarctic Program meteorological stations in (a) Palmer (2007–10), (b) McMurdo and (c) Amundsen-Scott South Pole. Darker grey represents a higher density of points. The dashed polygons represent the distribution of the points for the warmest months: January (red), February (green), March (dark blue), November (light blue) and December (pink). The estimated parameters of the fitted linear relationship are at the top.

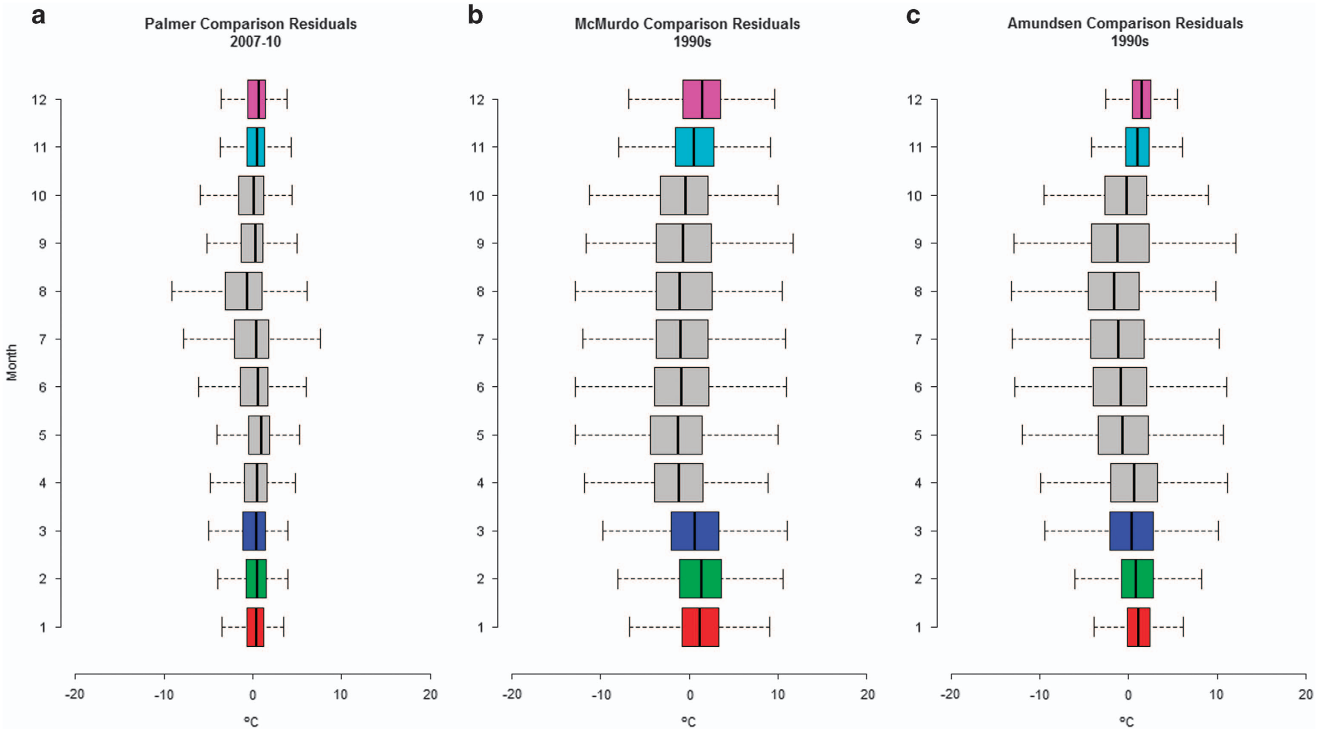

Figure 5. Distribution of the residuals of the linear models for MERRA and United States Antarctic Program meteorological stations in (a) Palmer (2007–10), (b) McMurdo and (c) Amundsen-Scott South Pole by month.

Usage Notes

Comparison between MERRAclim (80s and 90s decades) and WorldClim

MERRAclim datasets were created using temperature and specific humidity and following the methods described for WorldClim[23], to derive 19 bioclimatic variables that can be used in ecology. This section provides a comparison of MERRAclim and WorldClim to find possible patterns of spatial congruence or discordance. We calculated Pearson's correlation coefficients and fitted linear regressions to assess the relationship between both datasets[47].

The comparison is geographically limited to those areas where WorldClim data are not interpolated, i.e., around weather stations that were used to compile information. As WorldClim is temporally limited

Location	Dataset	bio1	bio2	bio3	bio4	bio5	bio6	bio7	bio10	bio11
Palmer	MERRA	− 4.3	11.3	47.7	346.0	3.2	− 20.6	23.8	0.0	− 8.5
	USAP	− 1.3	12.1	48.3	345.4	8.8	− 16.1	25.0	2.9	− 5.6
McMurdo	MERRA	− 21.3	18.9	47.7	721.4	− 1.7	− 41.2	39.6	− 11.4	− 28.2
	USAP	− 16.0	20.9	48.1	808.8	4.8	− 38.7	43.5	− 4.9	− 23.5
Amundsen South Pole	MERRA	− 45.3	19.3	45.4	907.8	− 23.1	− 65.9	42.8	− 32.1	− 53.2
	USAP	− 48.5	27.7	49.9	1130.2	− 21.0	− 76.6	55.6	− 32.2	− 58.0

Table 4. **Temperature-only bioclimatic variables computed with MERRA data and United States Antarctic Program (USAP) meteorological stations.** The unit for all the values is °C. NOTE: For Palmer station only records from 2007 to 2009 have been used.

to climatic records ranging from 1950 to 2000, the MERRAclim datasets could only be compared for the 80s and 90s decades.

Framing the geographical extent for the comparison WorldClim vs. MERRAclim

MERRA raw data is composed of a grid made of 540 columns and 360 rows. Each grid cell covers 2/3 degrees of latitude and < degrees of longitude, i.e., each cell covers an area of 1/3 square degrees. To make both datasets comparable we geographically limited the WorldClim dataset by creating an area of influence around each weather station as a buffer zone that covers 1/3 square degrees (equivalent to an area of ≈ 4,000 km^2 near the Equator and ≈250,000 km^2 for the farthest north weather station included in WorldClim at a latitude of 82°). Two comparisons were conducted: a first one for weather stations with available temperature observations (25,576 stations covering 73,109,913 km^2, roughly 22% of the WorldClim coverage) and a second one for weather stations with available precipitation observations (47,675 stations covering 67,893,375 km^2, roughly 20% of the WorldClim coverage).

Validation methodologies

We tested the relationship between MERRAclim and WorldClim bioclimatic variables calculating Pearson's correlation coefficients and fitting linear regressions. For all versions of the 19 bioclimatic variables we found linear correlations between both datasets that explained most of the variance. Linear regressions for the 1980s and the 1990s revealed that both decades have a similar relationship with WorldClim (Supplementary File 2). The absence of WorldClim data for the 2000s prevented a comparison for this decade.

Temperature-related bioclimatic variables (BIO1-BIO7, BIO10-BIO11)

Pearson coefficients testing the correlation between temperature-related bioclimatic variables (BIO1-BIO7, BIO10-BIO11) from WorldClim and MERRAclim were very high (>0.8) in all cases, except for BIO2 ($r = 0.6$). BIO2, a variable representing diurnal range and thus highly sensitive to temperature extremes, showed the highest discrepancy between datasets. Overall, MERRAclim yielded higher temperature values than WorldClim with a positive and close to unity slope (Supplementary File 1). Mean temperatures of the most extreme months (BIO5 and BIO6) show the largest differences between datasets: the warmest month in MERRAclim is ~10 °C higher than in WorldClim, whereas the coldest month is around 7 °C lower. Due to this important difference between datasets we fitted linear regressions using subsets depending on the absolute difference whose geographical distribution is depicted in Supplementary Fig. 2. Firstly, we used the points that showed an absolute difference smaller than 10 °C and we obtained for BIO5 the same trend but only 7 °C higher, for BIO6 the difference was of 3 °C. Secondly, we used a subset of those points with an absolute difference smaller than 5 °C, this time the intercept for BIO5 was 5 °C and for BIO6 there were no differences with the previous subset.

Water-related bioclimatic variables (BIO12-BIO17)

Comparisons of water-related MERRAclim bioclimatic variables for each decade with WorldClim are consistent, but important differences between versions were detected (Supplementary File 1). Bioclimatic variables from the V_{min} version show the strongest correlation with the bioclimatic variables in WorldClim and also the highest proportion of explained variance. In general, water-related bioclimatic variables were less strongly correlated with WorldClim than temperature-related ones. Bioclimatic variables describing the extreme lack of environmental water availability, both monthly (BIO14) and quarterly (BIO17), had the lowest congruence with WorldClim (Pearson correlation coefficient ~0.37 and ~0.4, respectively). Water seasonality (as described by BIO15) greatly varies in its correlation with WorldClim depending on the version, being the correlation with V_{min} four times stronger than with V_{max}.

Combined bioclimatic variables (BIO8-BIO9, BIO18-BIO19)

Combined bioclimatic variables depend on temperature and humidity information (Table 1) to describe the most extreme quarters. Although linear associations remained similar between MERRAclim versions (V_{min}, V_{mean} and V_{max}) and WorldClim, the strongest correlation coefficient was found for V_{min}. Among temperature-dependent combined variables, BIO8 showed the greatest difference between datasets (~15 °C higher in MERRAclim). BIO9 is ~2 °C warmer in WorldClim. Water-dependent combined variables (BIO18 and BIO19) followed the same trends as other water-related variables and, again, the V_{min} version showed the highest variance explained (Supplementary File 1).

Geographic location of the differences between WorldClim and MERRAclim

We located those geographical areas where MERRAclim (V_{min} version) and WorldClim vary the most using the outliers of residuals from linear regressions for each bioclimatic variable (Supplementary Figs 3 and 4). We defined outliers using the IQR (InterQuartile Range) of the residuals, for which we calculated the first and third quartiles (Q1 and Q3) and estimated the values outside the range Q1—(1.5*IQR) and Q3+(1.5*IQR) as outliers. Both datasets showed an outstanding spatial congruence and the average area of outliers for each bioclimatic variable covers less than 5% of the compared geographical space (only BIO9 has a larger extent of outliers, summing up to 7%, Supplementary File 1). To get a more detailed information of these variations we have also drawn the bias from the fitted linear regressions (Supplementary Fig. 5).

The differences between the datasets, as identified by the outliers and the bias, are geographically clustered (Supplementary Figs 3 and 4); which can probably be explained by the fact that WorldClim was built from heterogeneous regional networks of weather stations some of which are also compiled from several datasets[48] (e.g., Latin America, The Caribbean, the Altiplano in Peru and Bolivia, European Nordic Countries, the United States of America, Australia, New Zealand and Madagascar) that depend on different sources of information and techniques[16].

References

1. Franklin. *Mapping Species Distributions: Spatial Inference And Prediction* (Cambridge University Press, 2010).
2. Peterson, A. T. *et al. Ecological Niches And Geographic Distributions* (Princeton University Press, 2011).
3. Grinnell, J. The niche-relationships of the California thrasher. *Auk* **34,** 427–433 (1917).
4. O'Brien, E. M. Water-energy dynamics, climate, and prediction of woody plant species richness: an interim general model. *J. Biogeogr.* **25,** 379–398 (1998).
5. O'Brien, E. M. Climatic gradients in woody plant species richness: towards an explanation based on an analysis in southern Africa's woody flora. *J. Biogeogr.* **20,** 181 (1993).
6. Getz, L. L. Influence of water balance and microclimate on the local distribution of the redback vole and white-footed mouse. *Ecology* **49,** 276–286 (1968).
7. Williams, J. B., Ostrowski, S., Bedin, E. & Ismail, K. Seasonal variation in energy expenditure, water flux and food consumption of Arabian oryx Oryx leucoryx. *J. Exp. Biol.* **204,** 2301–2311 (2001).
8. Webster, M. D. & King, J. R. Temperature and humidity dynamics of cutaneous and respiratory evaporation in pigeons, Columba livia. *J. Comp. Physiol. B Biochem. Syst. Environ. Physiol* **157,** 253–260 (1987).
9. Bambach, N., Meza, F. J., Gilabert, H. & Miranda, M. Impacts of climate change on the distribution of species and communities in the Chilean Mediterranean ecosystem. *Reg. Environ. Chang.* **13,** 1245–1257 (2013).
10. Shaman, J. & Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc. Natl. Acad. Sci. USA* **106,** 3243–3248 (2009).
11. Marsden, B. J., Lieffers, V. J. & Zwiazek, J. J. The effect of humidity on photosynthesis and water relations of white spruce seedlings during the early establishment phase. *Can. J. For. Res* **26,** 1015–1021 (1996).
12. Seefeldt, M. W., Hopson, T. M. & Warner, T. T. A Characterization of the variation in relative humidity across West Africa during the dry season. *J. Appl. Meteorol. Climatol.* **51,** 2077–2089 (2012).
13. Colesie, C. *et al.* Terrestrial biodiversity along the Ross Sea coastline, Antarctica: lack of a latitudinal gradient and potential limits of bioclimatic modeling. *Polar Biol.* **37,** 1197–1208 (2014).
14. Martius, C. Rainfall and air humidity: non-linear relationships with termite swarming in Amazonia. *Amazoniana* **17,** 387–397 (2002).
15. Song, S. *et al.* Impacts of environmental heterogeneity on moss diversity and distribution of Didymodon (Pottiaceae) in Tibet, China. *PLoS ONE* **10,** e0132346 (2015).
16. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
17. Deblauwe, V. *et al.* Remotely sensed temperature and precipitation data improve species distribution modelling in the tropics. *Glob. Ecol. Biogeogr* **25,** 443–454 (2016).
18. Bentley, M. J. in *Exploring The Last Continent* (eds Liggett, D., Storey, B., Cook, Y. & Meduna, V.) Ch. 25 (Springer International Publishing, 2015).
19. Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (IPCC, 2014).
20. Chown, S. L. *et al.* Continent-wide risk assessment for the establishment of nonindigenous species in Antarctica. *Proc. Natl. Acad. Sci. USA* **109,** 4938–4943 (2012).
21. Pertierra, L. R. *et al.* Global thermal niche models of two European grasses show high invasion risks in Antarctica. *Glob. Chang. Biol.* **23,** (2017).
22. Rienecker, M. M. *et al.* MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Clim* **24,** 3624–3648 (2011).
23. Xu, T. & Hutchinson, M. ANUCLIM Version 6.1 User Guide (2011).
24. Cassano, J. J. in *Antarctica: Global Science From A Frozen Continent* (ed. Walton, D. W. H.) Ch. 4 (Cambridge University Press, 2013).
25. Michna, P. & Woods, M. RNetCDF: Interface to NetCDF Datasets (2016).
26. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. dismo: Species Distribution Modeling (2016).
27. ESRI. ArcGIS 10.2.2 for Desktop (2014).
28. Laslett, G. M. Kriging and splines: an empirical comparison of their predictive performance in some applications. *J. Am. Stat. Assoc.* **89,** 391–409 (1994).
29. Dubrule, O. Comparing splines and kriging. *Comput. Geosci.* **10,** 327–338 (1984).

30. Hutchinson, M. F. & Gessler, P. E. Splines—more than just a smooth interpolator. *Geoderma* **62,** 45–67 (1994).
31. Laslett, G. M., McBratney, A. B., Pahl, P. J. & Hutchinson, M. F. Comparison of several spatial prediction methods for soil pH. *J. Soil Sci* **38,** 325–341 (1987).
32. Hutchinson, M. F. Interpolating mean rainfall using thin plate smoothing splines. *Int. J. Geogr. Inf. Syst* **9,** 385–403 (1995).
33. Langland, R. H., Maue, R. N. & Bishop, C. H. Uncertainty in atmospheric temperature analyses. *Tellus, Ser. A Dyn. Meteorol. Oceanogr.* **60 A,** 598–603 (2008).
34. Ashouri, H. *et al.* Evaluation of NASA's MERRA precipitation product in reproducing the observed trend and distribution of extreme precipitation events in the United States. *J. Hydrometeorol.* **17,** 693–711 (2016).
35. Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F. & Poulin, A. Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling? *J. Hydrometeorol* **17,** 1929–1950 (2016).
36. Roberts, A. J., Marsham, J. H. & Knippertz, P. Disagreements in low level moisture between (re)analyses over summertime West Africa. *Mon. Weather Rev.* **143,** 1193–1211 (2015).
37. Lorenz, C. & Kunstmann, H. The hydrological cycle in three state-of-the-art reanalyses: Intercomparison and performance analysis. *J. Hydrometeorol.* **13,** 1397–1420 (2012).
38. Bosilovich, M. G., Chern, J. D., Mocko, D., Robertson, F. R. & Da Silva, A. M. Evaluating observation influence on regional water budgets in reanalyses. *J. Clim.* **28,** 3631–3649 (2015).
39. Trenberth, K. E., Fasullo, J. T. & Mackaro, J. Atmospheric moisture transports from ocean to land and global energy flows in reanalyses. *J. Clim.* **24,** 4907–4924 (2011).
40. Bosilovich, M. G., Robertson, F. R. & Chen, J. Global energy and water budgets in MERRA. *J. Clim.* **24,** 5721–5739 (2011).
41. Cullather, R. I. & Bosilovich, M. G. The moisture budget of the polar atmosphere in MERRA. *J. Clim.* **24,** 2861–2879 (2011).
42. Serreze, M. C., Barrett, A. P. & Stroeve, J. Recent changes in tropospheric water vapor over the Arctic as assessed from radiosondes and atmospheric reanalyses. *J. Geophys. Res. Atmos* **117,** 1–21 (2012).
43. Lindsay, R., Wensnahan, M., Schweiger, A. & Zhang, J. Evaluation of seven different atmospheric reanalysis products in the arctic. *J. Clim.* **27,** 2588–2606 (2014).
44. Bracegirdle, T. J. & Marshall, G. J. The reliability of Antarctic tropospheric pressure and temperature in the latest global reanalyses. *J. Clim.* **25,** 7138–7146 (2012).
45. Cullather, R. I. & Bosilovich, M. G. The energy budget of the polar atmosphere in MERRA. *J. Clim.* **25,** 5–24 (2012).
46. Lader, R., Bhatt, U. S., Walsh, J. E., Rupp, T. S. & Bieniek, P. A. 2-m Temperature and Precipitation from Atmospheric Reanalysis Evaluated for Alaska. *J. Appl. Meteorol. Climatol.*, **55,** 901–922 (2016).
47. Varela, S., Lima-Ribeiro, M. S. & Terribile, L. C. A short guide to the climatic variables of the last glacial maximum for biogeographers. *PLoS ONE* **10,** e0129037 (2015).
48. Peterson, T. C. & Vose, R. S. An overview of the Global Historical Climatology Network temperature database. *Bull. Am. Meteorol. Soc.* **78,** 2837–2849 (1997).

Data Citation

1. C. Vega, G., Pertierra, L. R. & Olalla-Tárraga, M. Á. *Dryad Digital Repository* http://dx.doi.org/10.5061/dryad.s2v81 (2016).

Acknowledgements

Research was funded by Spanish MINECO grant (CTM2013-47381-P) to MAO-T. LRP was awarded a SCAR-COMNAP fellowship 2013-14. We acknowledge the Global Modeling and Assimilation Office (GMAO) and the GES DISC for the dissemination of MERRA. We appreciate the support of the University of Wisconsin-Madison Automatic Weather Station Program for the data set, data display, and information, NSF grant number ANT-1543305. We are grateful to Joaquin Hortal for his constructive comments.

Author Contributions

G.C.V. conceived and designed the study, processed and validated the dataset, and drafted a first version of the manuscript. L.R.P. conceived and designed the study. M.A.O.-T. conceived and designed the study and helped draft the manuscript. All authors revised the manuscript and gave final approval for publication.

Additional Information

Supplementary Information accompanies this paper at http://www.nature.com/sdata

Competing interests: The authors declare no competing financial interests.

How to cite this article: C. Vega, G. *et al.* MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling. *Sci. Data* 4:170078 doi: 10.1038/sdata.2017.78 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.