SCIENTIFIC DATA

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Palaeoclimate
» Palaeoecology
» Climate-change ecology
» Biogeography

Data Descriptor: Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD

David J. Lorenz[1], Diego Nieto-Lugilde[2], Jessica L. Blois[3], Matthew C. Fitzpatrick[2] & John W. Williams[1,4]

Increasingly, ecological modellers are integrating paleodata with future projections to understand climate-driven biodiversity dynamics from the past through the current century. Climate simulations from earth system models are necessary to this effort, but must be debiased and downscaled before they can be used by ecological models. Downscaling methods and observational baselines vary among researchers, which produces confounding biases among downscaled climate simulations. We present unified datasets of debiased and downscaled climate simulations for North America from 21 ka BP to 2100AD, at 0.5° spatial resolution. Temporal resolution is decadal averages of monthly data until 1950AD, average climates for 1950–2005 AD, and monthly data from 2010 to 2100AD, with decadal averages also provided. This downscaling includes two transient paleoclimatic simulations and 12 climate models for the IPCC AR5 (CMIP5) historical (1850–2005), RCP4.5, and RCP8.5 21st-century scenarios. Climate variables include primary variables and derived bioclimatic variables. These datasets provide a common set of climate simulations suitable for seamlessly modelling the effects of past and future climate change on species distributions and diversity.

Design Type(s)	data integration objective
Measurement Type(s)	climate change
Technology Type(s)	computational modeling technique
Factor Type(s)	
Sample Characteristic(s)	North America

[1]Center for Climatic Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [2]University of Maryland Center for Environmental Science, Appalachian Lab, Frostburg, Maryland 21532, USA. [3]School of Natural Sciences, University of California, Merced, California 95343, USA. [4]Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. Correspondence and requests for materials should be addressed to J.W.W. (email: jww@geography.wisc.edu).

Background & Summary

A productive new synthesis is being forged among ecologists, conservation biologists, biogeographers, and paleoecologists[1–7], motivated by a common goal of understanding how climate change has affected and will affect the diversity and distribution of species. Species extinction rates currently are 100 times background rates[8] and climate change is expected to further stress threatened populations[9–12]. Metaanalyses suggest that 8% of species will go extinct if global temperatures rise above 3 °C (ref. 11). Given these projections, there is an urgent need to understand how environmental conditions govern the distribution and abundance of organisms, and to apply this knowledge to help species adapt to climate change.

Within this synthesis, the geohistorical record—with many well-documented instances of past climate changes and biological responses—is essential[5,9,10,13,14]. Most notably, the glacial-interglacial cycles of the last million years provide the closest equivalent to the rates and magnitudes of climate change expected for this century[13,15]. Ecological research opportunities created by these periods of transition include understanding (1) the stability of ecological niches during periods of environmental change[16–18], (2) rates of biological response to abrupt climate change[19–23], (3) the linkage between no-analogue communities and no-analogue climates[13,24–26]; (4) the predictive ability of ecological forecasting models and their underlying assumptions[27,28]; (5) the influence of past climate changes and human activity on current genetic diversity[29,30]; (6) glacial refugia, areas of climate stability, and other mechanisms by which species persisted during past adverse climates[31–33]; and (7) the causes and consequences of the late-Quaternary extinctions of large-bodied vertebrates[34,35].

However, despite calls for the full synthesis of contemporary observations, geohistorical data, experiments, and 21st-century projections to build an integrated science of climate change and biodiversity assessment[6,36,37], progress has been limited by the scarcity of climate simulations that seamlessly extend from the past into the future. As a result, future- and paleo-oriented studies of biodiversity dynamics generally remain divided, with few directly integrating the fossil record with future projections[38].

Because Earth system models (ESMs) are computationally expensive, most simulations are run for limited time windows (decades to centuries) according to prescribed experimental protocols[39,40]. Common modelling targets include the middle Pliocene (3.3 to 3.0 million years ago), last interglacial (125,000 years ago), last glacial maximum (21,000 years ago), mid-Holocene (6,000 years ago), and last millennium[39]. Some ESM's are run for longer periods to assess the evolving response of the earth system to transient forcings, e.g., to shifting orbital variations, rising greenhouse gases, and meltwater pulses during the last 21,000 years[41,42], or to variations in solar luminosity, volcanic eruptions, orbital variations, greenhouse gases, and land cover over the last millennium[43,44].

ESM simulations usually have systematic biases compared to observations and their native spatial resolutions often are too coarse for biological modelling. Simply using their simulations without a common debiasing and downscaling approach will lead to systematic biases among the past, present, and future climate simulations, deeply confounding any ecological inferences based on these simulations. Moreover, ensembles comprising multiple ESMs generally fit better to observations than individual ESMs[45]. Hence, an ecologist seeking to model the effects of past, present, and future climate change on biodiversity is forced to stitch together climate simulations from many models, a labour-intensive process that often results in unnecessary duplication of efforts among research teams.

Here we present debiased and downscaled climate simulations for North America, from 21,000 years ago to 2100AD (Data Citation 1), using a common set of observational datasets and methods. The paleoclimatic datasets are based on transient simulations from two ESMs[41,46,47], while the 20th- and 21st-century simulations are based on 12 ESMs using the CMIP5 historical, RCP4.5, and RCP8.5 scenarios[40]. These downscaled datasets include standard indices of monthly temperature and precipitation, other climate variables useful to modelling surface energy and moisture budgets, and derived variables such as growing degree days and actual and potential evapotranspiration. All variables are available as decadal averages of monthly values for the last 21,000 years, and as monthly values from 1950AD to 2100AD. From the paleoclimatic simulations, we extract century-scale averages and other statistics, spaced 500 years apart, while from the 21st-century simulations we extract decadal-scale averages, spaced a decade apart. Note that this debiasing and downscaling does not guarantee that the ESM simulations are accurate predictors of past and future climates, merely that they have been standardized to be consistent with each other and with contemporary observational data. As always, ESMs and their simulations should be used carefully and critically, and paleoclimatic simulations should be checked for consistency with paleoclimatic proxy data, which, of course, carry their own uncertainties.

Methods

Overview

Statistical downscaling and debiasing followed a multi-step approach (Fig. 1). First, most of the primary climate variables were debiased and downscaled using the standard change-factor approach[48] applied separately to each month, or to each season if monthly data are unavailable. Climate variables archived only as seasonal values then were interpolated to monthly values. Primary climate variables are mean daily maximum temperature, mean daily minimum temperature, total monthly precipitation, water vapour pressure, downward and upward shortwave radiation at the surface, net longwave radiation at the
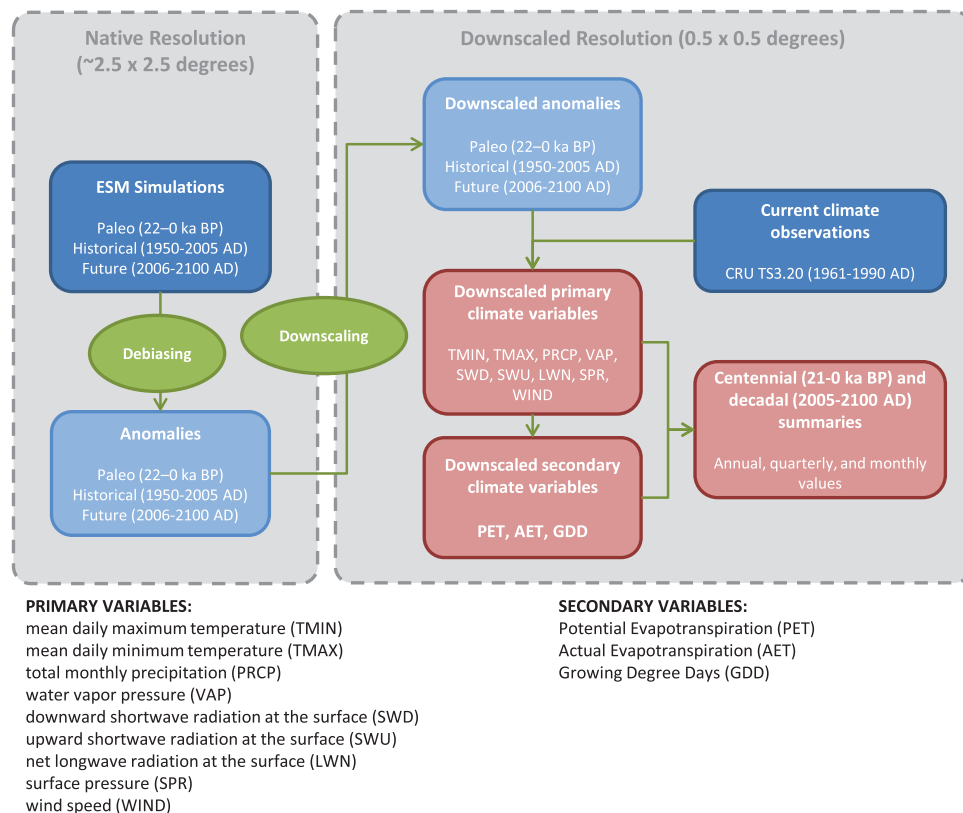
PRIMARY VARIABLES:
mean daily maximum temperature (TMIN)
mean daily minimum temperature (TMAX)
total monthly precipitation (PRCP)
water vapor pressure (VAP)
downward shortwave radiation at the surface (SWD)
upward shortwave radiation at the surface (SWU)
net longwave radiation at the surface (LWN)
surface pressure (SPR)
wind speed (WIND)

SECONDARY VARIABLES:
Potential Evapotranspiration (PET)
Actual Evapotranspiration (AET)
Growing Degree Days (GDD)

Figure 1. Workflow diagram summarizing the major steps used to generate the debiased and downscaled paleoclimate and 21st-century datasets described here. The primary climate variables are first debiased by differencing (or similar calculation, see Section 3) each paleoclimate or future climate simulation from a climate simulation representing present climates. These anomalies (also known as factors) are then downscaled via bilinear interpolation to a higher-resolution grid corresponding to the resolution of the modern observational dataset. The anomalies are then added to (or multiplied by, see Section 3) the observational data to produce the downscaled primary variables. From these downscaled primary variables, secondary bioclimatic variables (GDD, AET, PET) are calculated. Finally, statistical summaries of all variables are calculated at centennial (paleoclimatic simulations) and decadal resolution (21st-century projections).

surface, and wind speed (Fig. 1, Table 1 (available online only)). From these primary variables, a secondary set of climate variables were calculated for each month: potential and actual evapotranspiration (PET and AET) and growing degree days (base 0° and 5 °C; GDD0 and GDD5). For the paleoclimatic simulations, some grid cells were land in the past but are under water today due to ice sheet melting and sea level rise during the last deglaciation. For these grid cells, paleoclimates were inferred by spatial extrapolation.

For the paleoclimate simulations, two climate models were downscaled: CCSM3 (refs 41,46) and ECBilt-CLIO[47]. For the 21st-century climate simulations, 12 models from the CMIP5 model archive were downscaled (Table 2), using the historical, RCP4.5, and RCP8.5 scenarios.

The spatial domain is North America (48° to 173°W, 10° to 80°N) and the spatial resolution of the downscaled climate simulations is 0.5×0.5 degrees. The temporal domain for the paleoclimate simulations is the last glacial maximum to the 20th century (CCSM3: 22,000 years before 1950AD (ka BP) to 1990AD, ECBilt: 21 ka BP to 1950AD). The temporal domain for the 21st century climate simulations is 1950AD to 2100AD. The temporal resolution of the downscaled paleoclimate simulations is decadal means for each calendar month, while the temporal resolution of the 21st-century climate simulations is monthly data for each year. This temporal resolution is determined by the archived climatic simulations. The native temporal resolution of atmospheric circulation models typically is subhourly, but not all time steps are publicly archived, so part of the downscaling of the paleoclimatic simulations involved inferring monthly values from archived seasonal means.

After initial downscaling, paleoclimate datasets were prepared for biogeographic hindcasting by extracting century-scale bins of the data spaced 500 years apart (each bin was 200 years thick and centered on the 500 year time step), from 21 ka BP to the 20th century. For each century-scale bin, annual and seasonal averages and other statistical summaries of climate variables were calculated from the

Modeling Center or Group	Model Name
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	ACCESS1.3
Canadian Centre for Climate Modelling and Analysis	CanESM2
Community Earth System Model Contributors	CESM1(CAM5)
Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	CNRM-CM5
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-Mk3.6.0
NOAA Geophysical Fluid Dynamics Laboratory	GFDL-CM3
NASA Goddard Institute for Space Studies	GISS-E2-R
Met Office Hadley Centre (additional HadGEM2-ES realizations contributed by Instituto Nacional de Pesquisas Espaciais)	HadGEM2-ES
Institute for Numerical Mathematics	INM-CM4
Institut Pierre-Simon Laplace	IPSL-CM5A-MR
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC5
Meteorological Research Institute	MRI-CGCM3

Table 2. CMIP5 climate models used for 21st-century projections (RCP4.5 and RCP8.5).

Variable	Annual		Quarterly		Monthly	
	Summaries	Variability	Min	Max	Min	Max
Tmin	an-avg-TMIN	an-sd-TMIN	qt-lwr-TMIN		mo-lwr-TMIN	
Tmax	an-avg-TMAX	an-sd-TMAX		qt-hgr-TMAX		mo-hgr-TMAX
Prec	an-sum-PRCP	an-cv-PRCP	qt-lwr-PRCP	qt-hgr-PRCP	mo-lwr-PRCP	mo-hgr-PRCP
GDD	an-sum-GDD	an-sd-GDD	qt-lwr-GDD	qt-hgr-GDD	mo-lwr-GDD	mo-hgr-GDD
AET	an-sum-AET	an-cv-AET	qt-lwr-AET	qt-hgr-AET	mo-lwr-AET	mo-hgr-AET
PET	an-sum-PET	an-cv-PET	qt-lwr-PET	qt-hgr-PET	mo-lwr-PET	mo-hgr-PET
ETR (AET/PET)	an-avg-ETR	an-cv-ETR	qt-lwr-ETR	qt-hgr-ETR	mo-lwr-ETR	mo-hgr-ETR
WDI (PET—Prec)	an-sum-WDI	an-cv-WDI	qt-lwr-WDI	qt-hgr-WDI	mo-lwr-WDI	mo-hgr-WDI

Table 3. Naming convention for variable names in raster files generated for ecological modeling with centennial and decadal summaries.

decadal mean monthly data in the 100 years before and after the time step, and values were averaged across the 20 decades (Table 3). For the 21st-century climate simulations, 20-year monthly averages and other statistical summaries were extracted at 10-year time intervals (for 2011–2030, 2021–2040, etc.). All steps are described in more detail below. The centennial and decadal summaries are saved as geotiffs (Table 1 (available online only)).

Data sources
Observational climate datasets for the 20th and 21st centuries. The change factor approach requires contemporary observational datasets against which the ESM simulations are debiased and downscaled[48]. Monthly precipitation, maximum and minimum daily temperature, and vapour pressure (1901–2011) are from the Climate Research Unit (CRU) TS v.3.20 dataset[49]. Wind speed is from a 12 month climatological CRU dataset for the time period 1961–1990 (ref. 50). The spatial resolution of all CRU data is 0.5×0.5 degrees. The radiation data are from the NASA/GEWEX Surface Radiation Budget (SRB) Release-3.0 dataset[51,52]. The GEWEX radiation data are interpolated from a 1×1 degree grid to the CRU grid using bilinear interpolation. For the GEWEX shortwave and longwave radiation, the monthly anomalies from the climatology are considered significantly less reliable than the average climatology itself, and therefore only the climatology is used, for 1983–2007. The top-of-atmosphere solar radiation from 21 ka BP to present is calculated using the algorithm of Berger[53]. Surface pressure in the present is estimated from elevation data using guidelines recommended by Allen et al.[54]. The elevation data are from the Global Land One-km Base Elevation Project (GLOBE)[55]. The elevation is averaged spatially over the CRU grid boxes to create a lower resolution elevation dataset consistent with the other data. We estimate past surface pressure from the dataset of land and ice elevation of Peltier[56].

To maximize uniformity among downscaled variables and models, we sought to use a common time period for the observational time period. However, perfect uniformity was impossible because of differences among model simulations in their 20th century start/end dates and differences among observational datasets in their temporal extent. For temperature, precipitation, and vapour pressure, the time period used for the baseline climate was 1901 to 2011 for the paleoclimatic simulations and 1950 to

2005 for the CMIP5 simulations. For wind speed, the baseline time period is 1961–1990 for all downscaling analyses. For shortwave and longwave radiation, the baseline time period is 1983–2007.

Climate models. The paleoclimate simulations consist of a 22,000 year transient simulation from the Community Climate System Model (CCSM3)[41,46] and a 21,000 year transient simulation using ECBilt-CLIO[47]. Decadal means of paleoclimatic simulations were obtained from CCSM3 (Feng He, pers. comm.) and EC-BILT (http://apdrc.soest.hawaii.edu/datadoc/sim2bl.php). CCSM3 was forced by prescribed trends in orbital parameters, ice sheet extent and height, sea level, greenhouse gases, and meltwater pulses to the North Atlantic, while ECBilt-CLIO was forced by prescribed trends in orbital parameters, ice sheet extent and height, and greenhouse gases. Often, when using output from climate models, all simulations are given equal weight in derived ensembles. However, the ECBilt-CLIO simulation carries known simplifications in its model structure relative to the CCSM3 simulation, including: (1) a quasi-geostrophic atmospheric model that constrains vertical static stability; (2) 3 vertical levels versus 26 for CCSM3; (3) a radiation code linearized about present day conditions, and (4) prescribed seasonally and spatially varying cloud cover climatology. Therefore, we recommend that any use of these simulations place greater weight on the CCSM3 simulations. Note too that the ECBilt-CLIO archive did not include vapour pressure and wind speed, so these variables are not downscaled for this model.

Output from twelve climate models were downloaded from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website (http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html), for the historical, RCP4.5, and RCP8.5 scenarios, as reported in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5). One model was chosen from each modelling centre and models were selected only if they archived all climate variables used for the analyses here. Among IPCC models, the quality of the simulations vary and, depending on the variable, some models are more skilful than others, with no clear signal of some models consistently outperforming others. Therefore, it is generally understood that the projections of all models should be considered with equal weight. Note, however, that two of the models (ACCESS1.3 and MIROC5) do not conserve atmospheric water mass[57]. Analyses relying on these individual models should be aware of the in-built hydrological imbalances and the associated biases in modelled radiative forcing. Nevertheless, skill tests show that mean climates from model ensembles consistently outperform all individual models in nearly every respect[58].

Debiasing and downscaling of primary variables from paleoclimatic simulations

Overview. To debias the variables, we calculated the difference between the modelled paleoclimate and the modelled present climate (Fig. 1). The resulting anomaly was then downscaled through bilinear interpolation to the higher resolution of the observational climatology to produce the debiased and downscaled paleoclimate at higher resolution (Fig. 2). This differencing removes any systematic difference between modeled and simulated climates as long as that bias is constant through time[59]. For example, for temperature and net longwave radiation at the surface, we calculated the difference between the modelled variable and the modelled present state of the climate variable (defined for this step as the last 110 years of the paleoclimatic model run) at each grid cell. This anomaly was then bilinearly interpolated to the 0.5° grid of the CRU TS 3.20 dataset. For climate variables bounded by zero, such as wind speed and water vapour pressure, we employed the factor approach, i.e., the ratio between the modelled paleoclimate and the modelled present climate was calculated, interpolated, and then multiplied by the observed climatology. Some variables required special handling during debiasing and downscaling, which we describe further below.

Shortwave radiation. A significant portion of the change in surface shortwave radiation over the past 21,000 years is not related to the internal physics of the climate models but is instead externally forced by changes in the earth's orbit. Therefore when calculating shortwave anomalies, we normalize both the upward and downward solar radiation at the surface by the maximum top of atmosphere downward solar radiation: $\hat{S} = S/B$, where S is either the raw downward or raw upward solar radiation at the surface, B is the top-of-atmosphere solar radiation which is a function of latitude and the earth's orbital parameters at time t^{53}, and $\hat{S}$ is the normalized downward or upward solar radiation. $\hat{S}$ must fall between 0 and 1. To automatically preserve this 0 to 1 range, we express the value of $\hat{S}$ in an alternate climate $(\hat{S}_{alt})$ as the present climate $(\hat{S}_0)$ raised to a positive power, γ:

$$\hat{S}_{alt} = \hat{S}_0^{\gamma} \tag{1}$$

To use equation 1 for downscaling, we solve for γ:

$$\gamma = \log(\hat{S}_{alt})/\log(\hat{S}_0) \tag{2}$$

where both $\hat{S}$'s are the low-resolution climate model values. The low-resolution γ is then interpolated to the 0.5° grid of the observed climate data. Finally the high resolution γ is used to transform the present climate $\hat{S}_0$ using equation 1. This approach is analogous to the traditional change-factor approach. The only difference is that the transformation is modified in order to rescale modeled solar radiation to the range of values allowable given past orbital configurations.
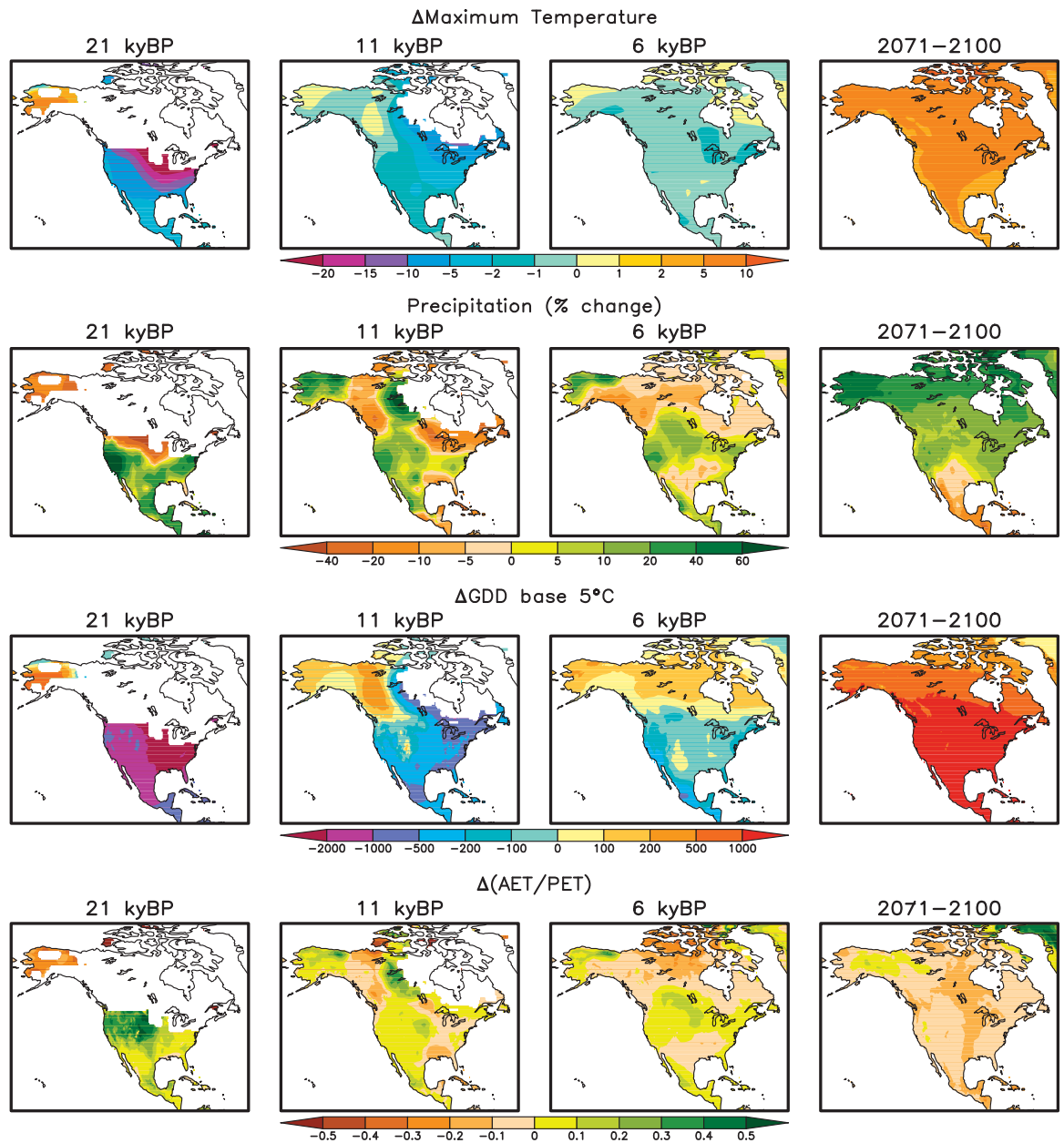
Figure 2. Anomaly maps showing examples of the downscaled differences between simulated past and future climates to the 20th century baselines. Top row: difference maps for mean maximum daily temperature, for 21 ka BP, 11 ka BP, 6 ka BP and 2071–2100AD. The paleosimulations are for the downscaled SynTrace CCSM3 simulation, with locations under the ice sheet masked out. The upper-right plot is the difference between 2071–2100 and 1950–2005, averaged over 12 climate models and for the RCP8.5 scenario. Second row: as top row, but for the percent change in precipitation compared to present. Third row: as top row, but for growing degree days (GDD) base 5C. Fourth row: as top row, but for the ratio of mean annual AET to mean annual PET.

Precipitation. For precipitation, which is bounded by zero, the factor method described above is a commonly used downscaling approach. Unfortunately, locations that are very dry tend to have very large fractional changes in precipitation and vice versa. This creates a problem for regions that are very dry in the climate model but very wet in observations. In a dry region, the climate model might predict a factor-of-10 increase in precipitation. While this is reasonable in a desert, it almost certainly artificial in a wet region. Therefore it is unrealistic to assume that the factor to multiply the observed climatological precipitation can be taken directly from the climate model.

To avoid the above issue, we use a method based on quantile mapping[60]. Quantile mapping uses the fact that the precipitation variability in dry regions is also large compared to wet regions. The historical

relationship between the variability in the climate model and the observational dataset is used to map the simulated change in precipitation to a corresponding change in observations. The details are as follows: consider a 100-year record of July precipitation in observations and in a climate model, where the observations are first averaged in space to the low-resolution climate model grid. Both the observations and the climate model time series are then independently sorted from the smallest July precipitation to the largest July precipitation. Let the sorted observations and climate model precipitation be denoted as p_j and q_j, respectively. The empirical, monotonic function, $p_j = f(q_j)$, describes a mapping from modelled precipitation to observations that transforms the modelled probability density function (PDF) to the observed PDF. To use the function, f, for past or future modelled precipitation (q_{alt}), we use linear interpolation to estimate f at points between the q_j's and at points beyond the range of 'observed' climate model precipitation (i.e., less than q_1 or greater than q_{100}). The resulting (low resolution) observed precipitation estimate, $p_{alt} = f(q_{alt})$, is then normalized by the low-resolution climatology and, like the factor approach, the resulting factors are then interpolated to the high-resolution observational grid. The high-resolution observed climatology is then multiplied by the high-resolution factors to make the final downscaled precipitation.

The above methodology is essentially the standard quantile mapping approach of Wood $et\ al.$[60] and is also used to downscale the precipitation data for the future climate simulations. We modify the method for the paleoclimatic downscaling because we are limited to decadal data, which means we have only 11 decades of observations over the 20th and early 21st centuries. Because the limited sample can lead to noise in the function, f, we instead use the 11 observations of p_j and q_j to fit a linear least-squares line ($p = aq + b$) and we use the line as the function giving p_j as a function of q_j. As in the standard case, potential issues arise when one has values of q_{alt} that are beyond the range of 'observed' climate model precipitation. In our case, we use the above linear fit when $b \geq 0$ or when $q_{alt} \geq q_1$. When $b < 0$ and $q_{alt} < q_1$, on the other hand, $p = q\,(aq_1 + b)/q_1$. This solution is continuous with $p = aq + b$ at $q = q_1$ and it automatically satisfies the constraint that $p > 0$ when $q > 0$.

Interpolating seasonal to monthly variables

At the time of analysis, many of the variables for CCSM3 and all of the variables for ECBilt-CLIO were available only as seasonal means. We interpolated these variables to monthly values, both for consistency and for the calculation of potential and actual evapotranspiration, which benefit from having monthly resolved variables. However, linearly interpolating seasonal data to monthly values strongly dampens the annual cycle and moreover produces monthly data that is usually not consistent with the seasonal data from which it is derived (e.g., summer temperature is not equal to the average of the June, July and August temperature in the derived monthly data, see examples in technical validation section). Here we develop a method for creating consistent monthly data from seasonal data.

Temperature and other anomaly variables. The simplest case is for (maximum and minimum) temperature where, for all practical purposes, there is no lower or upper bound on allowable values. Let $S1$, $S2$, $S3$ and $S4$ be the seasonal mean temperature for winter, spring, summer and fall, respectively. Let T_j ($j = 1,12$) be the monthly mean temperature for the 12 calendar months, which we want to estimate from S_j. The T_j must satisfy the following constraints:

$$
\begin{aligned}
T_1 + T_2 + T_{12} &= 3S_1 \\
T_3 + T_4 + T_5 &= 3S_2 \\
T_6 + T_7 + T_8 &= 3S_3 \\
T_9 + T_{10} + T_{11} &= 3S_4.
\end{aligned}
\tag{3}
$$

The specification of T_j is obviously under-constrained. One solution is to let monthly temperature be equal to the corresponding seasonal temperature (i.e., $T_1 = T_2 = T_{12} = S_1$, etc.). This solution is highly unlikely because it produces an annual cycle that is not very smooth. We hypothesize that the most likely solution is the 'smoothest' solution that also satisfies equation 3. To quantify smoothness we use the simple second order approximation of the second derivative: $T_{j-1} - 2T_j + T_{j+1}$. When the second derivative is small the 'smoothness' is high, so we want to minimize:

$$
min\left(\frac{1}{2} \sum_{j=1}^{12} \left(T_{j-1} - 2T_j + T_{j+1} \right)^2 \right)
\tag{4}
$$

where T_0 is December temperature and T_{13} is January temperature. We solve equations 3 and 4 using Lagrange multipliers, which leads to a linear 16×16 matrix equation. For the downscaling, this procedure is applied to the debiased low-resolution anomalies.

In summary, the steps are: 1) calculate the difference between the modelled seasonal temperature and the modelled current seasonal temperature, 2) convert the seasonal anomalies to monthly anomalies using the above procedure, 3) interpolate the monthly anomalies to the high resolution grid, and 4) add the resulting high resolution monthly anomalies to the observed climatology. The same procedure is used for the net longwave radiation at the surface. See the Technical Validation section for an assessment of this method for converting seasonal variables to monthly variables.

Precipitation and other factor variables. For precipitation the method is more complicated because one must ensure that all monthly precipitation amounts are at least zero. We solve this problem by minimizing the second derivative of the logarithm of the precipitation:

$$min\left(\frac{1}{2}\sum_{j=1}^{12}\left(log\left(P_{j-1}\right)-2log\left(P_j\right)+log\left(P_{j+1}\right)\right)^2\right) \tag{5}$$

where P_j is the estimated monthly precipitation and, as before, P_0 is December precipitation and P_{13} is January precipitation. The constraints are exactly analogous to equation 3 for temperature. Using Lagrange multipliers, the solution can be represented by a system of 16 equations, some of which are nonlinear. We solve the system of equations using MINPACK[61], which solves a system of nonlinear equations using a modification of the Powell hybrid method. See the Technical Validation section for an assessment of this method for converting seasonal variables to monthly variables.

For downscaling past precipitation, we represent past precipitation as a fraction (f) of the present precipitation (i.e., $P_{\text{past}} = f(P_{\text{present}})$). Therefore, the above method for the mean precipitation needs to be modified. Let P_j be the observed 12 month climatology of precipitation in the present, f_j be the monthly fraction of present precipitation that we want to calculate, and F_j be the seasonal fraction of present precipitation from the climate model (where $j=1,4$ corresponds to winter, spring, summer and fall, respectively). The four constraints are:

$$
\begin{aligned}
f_1P_1 + f_2P_2 + f_{12}P_{12} &= F_1(P_1 + P_2 + P_{12})\\
f_3P_3 + f_4P_4 + f_5P_5 &= F_2(P_3 + P_4 + P_5)\\
f_6P_6 + f_7P_7 + f_8P_8 &= F_3(P_6 + P_7 + P_8)\\
f_9P_9 + f_{10}P_{10} + f_{11}P_{11} &= F_4(P_9 + P_{10} + P_{11}).
\end{aligned}
\tag{6}
$$

The function we want to minimize subject to the above constraints is:

$$min\left(\frac{1}{2}\sum_{j=1}^{12}\left(log\left(f_{j-1}\right)-2log\left(f_j\right)+log\left(f_{j+1}\right)\right)^2\right) \tag{7}$$

The same method is used for water vapour pressure and wind speed.

Shortwave radiation. For shortwave radiation, we simply apply the temperature equations (3 and 4) to the power (γ) except that we first take the logarithm of the seasonal power, to help satisfy the constraint of a positive power. Once the monthly $log(\gamma)$ is found we then transform back to the raw powers, γ.

Secondary variables

After all primary variables are debiased, downscaled, and transformed to monthly resolution, they are used to derive the secondary variables as described below.

Potential evapotranspiration. We estimate the potential (reference) evapotranspiration using the Penman-Monteith method of Dobrowski et al.[62] Their method is very similar to the standard defined in Allen et al.[54], but they incorporated several modifications to create a reference that behaves more realistically under cold and snowy conditions. Our only modifications from Dobrowski et al.[62] are (1) the use of downscaled albedo (specifically, upward shortwave radiation at the surface) instead of one or two predetermined values depending on the presence of snow cover, and (2) we did not estimate the effect of fine-scale local topography on surface radiation.

Actual evapotranspiration. Estimating actual evapotranspiration requires an estimate of soil moisture. We used the 'single-bucket' methodology of Lutz et al.[63] to model actual evapotranspiration (AET) but with modifications to the snowmelt model, as described elsewhere[62]. We use a constant soil water holding capacity of 150 mm. Estimates of actual evapotranspiration are not particularly sensitive to reasonable alternative water holding capacities. For example, re-running the model using observed soil water holding capacity, Dunne and Willmott[64] demonstrated that constant water capacity leads to a domain-averaged error in the annual average AET of only 3.8%. Furthermore, given that soil water holding capacity is not constant over millennia, we thought it was best to simply assume a constant capacity rather than assume the present water-holding capacity applies to the past.

Growing degree days. With daily temperature data, the calculation of Growing Degree Days (GDD) is straightforward. Unfortunately, we only have monthly temperature ($\bar{T}$). If we make some assumptions about the shape of the probability density function (PDF) for daily temperature variability, however, we can calculate GDD given daily temperature variance. For each day, the GDD base T_0 is defined as:

$$GDD = max(T - T_0, 0) \tag{8}$$

where T is the 'average' daily temperature ($T = (T_{\max} - T_{\min})/2$). If we know the form of the daily temperature PDF (here referred to as $f(T)$), then the climatological GDD is:

$$GDD = \int_{-\infty}^{\infty} f(T) \ max(T - T_0, 0) \ dT = \int_{T_0}^{\infty} f(T)(T - T_0)dT \tag{9}$$

If $f(T)$ is a normal distribution with mean $\bar{T}$ and standard deviation σ, then equation 9 can be solved to give a relationship between GDD, $\bar{T}$ and σ, given T_0:

$$GDD = \frac{\sigma}{\sqrt{2\pi}} \ exp\left(-\frac{1}{2}\left(\frac{\bar{T} - T_0}{\sigma}\right)^2\right) + \frac{\bar{T} - T_0}{2} \ erfc\left(\frac{T_0 - \bar{T}}{2\sigma}\right) \tag{10}$$

where *erfc* is the complementary error function. We test the assumptions used to derive equation 10 by analyzing all daily NOAA COOP weather stations with fewer than 20 missing days from 1950 to 2009 ($= 108$ stations). For each station we calculate the actual average monthly GDD (for base 0 °C and 5 °C) as well as the monthly mean temperature and the monthly mean of daily temperature standard deviation. From the monthly mean and the monthly mean daily standard deviation, we estimate GDD using equation 10. We also apply equation (8) directly to the monthly mean temperature, which is the usual way GDD is calculated from monthly data. See the Technical Validation section for an assessment of the GDD estimator.

To use equation 10, we need the daily standard deviation σ^2_{daily}. For this project we estimate the daily standard deviation from the monthly standard deviation using the same daily COOP stations. If the daily temperature is uncorrelated in time (i.e., white noise) then the relationship between the daily and monthly precipitation is:

$$\sigma^2 = n\sigma^2_{monthly} \tag{11}$$

where n is the number of days in the month. In reality, temperature autocorrelation is >0 so that the actual monthly σ is greater than that given by equation 11. Some of the extra auto-correlation or 'memory' comes from internal 'weather' variability and some comes from the annual cycle (e.g., after removing the monthly mean in April, there is still an upward trend in temperature (on average) that manifests itself as a positive auto-correlation). The memory due to the annual cycle is calculated by assuming the daily annual cycle is a piecewise linear function based on the climatological monthly-mean annual cycle, where the daily-mean cycle is precisely equal to the monthly mean at the center of each month. Next, after removing the monthly mean, we calculate the standard deviation of the daily annual cycle over each individual month. The formula for the daily standard deviation due to the annual cycle at month j ($= 1$, 12) is:

$$\sigma_{annual} = n_j\sqrt{\frac{5a_+^2 + 6a_+a_- + 5a_m^2}{192}} \tag{12}$$

where n_j is the number of days in month j, $a_+ = 2(T_{j+1} - T_j)/(n_{j+1}+n_j)$ and $a_- = 2(T_j - T_{j-1})/(n_j+n_{j-1})$. Here we assume that the extra memory from internal variability is the same for all stations, so that its effect can be estimated empirically by:

$$\sigma^2_{daily} = an\sigma^2_{monthly} + \sigma^2_{annual} \tag{13}$$

where a is the empirical constant. Equation 13 is fit using least squares (with no intercept), and the value of the constant a is 0.178.

Using equation 13, we calculate the daily standard deviation for each grid point in the CRU temperature data. We chose not to estimate the standard deviation for the past simulated temperatures, because we only have decadal means for temperature and the time-scales are very different from the daily observations available today. Instead, we apply the daily standard deviation from the modern data to the paleosimulations and assume that the daily standard deviation is constant across time. The mean temperatures for equation 10 are taken directly from the downscaled temperatures.

Treatment of former land grid cells and ice-covered surfaces

Sea level rise has submerged portions of coastal North America that were above sea level during the last glacial period. To identify grid cells that were land in the past but are marine now, we used the paleoshoreline maps for North America developed by Patrick Bartlein at the University of Oregon, which are based on digitizing the topographic anomalies from Peltier[56] and interpolating them onto a contemporary digital elevation model[65]. These paleoshoreline reconstructions are available at 1,000 year intervals. The digitized shapefiles were clipped and rasterized to the same spatial domain and resolution as the climatic data. Each decade was assigned to the temporally closest paleoshoreline (e.g., 1 ka BP paleoshorelines were assigned to decadal data from 0.5 ka BP to 1.5 ka BP). To estimate the past climates for these now-submerged grid cells, we extrapolate from nearby grid points that are land in the current climate. One approach would be to use the closest land grid point but this is more subject to local noise. Alternatively, one could average all land grid points within a certain radius of the target grid point. However, the optimal averaging radius depends on the distance to the closest land grid point. Instead we

incrementally increase the 'search radius' and use the smallest radius that includes land grid points. In addition, because we expect climates to vary less in longitude than in latitude we consider the grid points in an ellipse (rather than a circle) with a major axis trending east-west. This modified ellipse-based distance is

$$a = \sqrt{(1-w)d^2 + w(R|\Delta\phi|)^2} \qquad (14)$$

where d is the distance, R is the radius of the earth, $\Delta\phi$ is the latitudinal difference between the grid points and w is a weighting that determines the deviation from circular ($w = 0.75$). Let ε be the grid spacing in latitude. We incrementally increase the 'search a' ($= a_{max}$) by ε starting from 1.5ε. When there are land grid points within a_{max}, we average these grid points together with an inverse-distance weighting $1/a$ to get the extrapolated terrestrial paleoclimates for the now-submerged target grid cell. Note that these extrapolated estimates of paleoclimates for now-submerged grid cells are likely to be more prone to systematic bias and higher uncertainty, so caution is warranted when making ecological or other inferences for now-submerged grid cells.

No special consideration was given to ice-covered surfaces. However, the Lutz et al.[63] scheme has evaporation (AET) only when there is liquid water (rain or snow melt), so AET is zero over most ice-covered grid cells.

Extracting century-scale averages from the decadal data

These downscaled datasets were originally developed for use with late-Quaternary fossil pollen records. Because of uncertainties in radiocarbon dates and other age controls, and because dating quality varies widely over paleoecological records that have been collected over the last 50 years, mapped syntheses of late-Quaternary paleoecological data typically have a maximum temporal resolution of ca. 500 years[66], although the dating for individual records can be much more precise (10^1 to 10^2 years). Hence, for use with the synthesized fossil data, we extracted century-scale climatic averages from the paleoclimatic simulations.

Century-scale averages were calculated at 500 year intervals by averaging climate variables for all decades within 100 years of the target date. For example, for 0.5 ka BP, we averaged climatic values for 0.4 to 0.6 ka BP). The one exception is the 0 ka BP time window (0 ka BP is set to 1950 AD) where this interval is truncated depending on when the paleoclimatic simulation ended, producing an average spanning 1850 to 1990 AD for CCSM3 and 1850 to 1950 AD for ECBilt.

For each century-scale time window, we calculated several summary annual statistics. Statistics were first computed for individual decades, then averaged across all decades in the 200-year window. The statistics computed were annual sum for GDD, AET, PET, and precipitation and annual average for minimum and maximum daily temperature. We also computed estimates of variability, using the coefficient of variation for AET, PET, and precipitation. Because the coefficient of variation can be meaningless for data in an interval scale, we used the standard deviation for temperature related variables (GDD, minimum daily temperature, and maximum daily temperature). We also derived two indices of water availability: Evapotranspiration Ratio (ETR; AET/PET) and Water Deficit Index (WDI; PET—precipitation)[67,68]. Finally, we computed the highest and lowest monthly and seasonal (quarterly) values of each variable.

Downscaling the IPCC AR5 simulations

The downscaling method for the IPCC AR5 simulations for the late 20th and 21st century closely followed the paleoclimatic simulations. However, several of the customized methods developed to work with the paleoclimatic simulations were unnecessary, because all IPCC AR5 variables were available as monthly values for each year, rather than decadal averages of monthly or seasonal values. Data were obtained from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website (http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html).

Simulations for 12 models (Table 2) were downscaled with three simulations per model: Historical, RCP4.5, RCP8.5. The historical scenario covers 1950–2005 AD. The two future scenarios (RCP4.5 and RCP8.5) are from 2006 to 2100. All downscaled climate variables are available as monthly values for every year. Summary statistics were calculated as for the century-scale averages, but were calculated for 1950–2005 average in the historical dataset and for 20-year averages at 10-year intervals (2011–2030, 2021–2040, …, 2081–2100 AD) in the two climate scenarios (i.e., RCP4.5 and RCP8.5), instead of 200-year windows centred on 500-year intervals. This use of 20-year averages spaced 10 years apart creates interdependencies in the decadal averages (to avoid this, use decadal averages from every other decade), but enables calculation of climate means and other statistics that were not overly influenced by low sample size and interannual climate variability. Users wishing to construct alternate averages and summary statistics can make use of the monthly data available in the NetCDF files (Table 1 (available online only)).

Code availability

All scripts used to develop the raster files with the centennial-scale and decadal-scale summaries are available on GitHub at the following repository: https://github.com/fitzLab-AL/climateSims21kto2100.

Data Records

All datasets have been deposited with the Dryad Digital Repository (Data Citation 1). See Table 1 (available online only) for a full list of all datasets, associated metadata, and DOIs.

Technical Validation

Seasonal to monthly interpolation

Temperature. We assess the performance of the method used to predict the monthly annual cycle from the seasonal annual cycle by comparing the observed seasonally averaged cycle, the monthly annual cycle estimated from the linearly interpolated seasonal cycle, and the results from the new method described above. Both the maximum and minimum temperature are shown for a point near Madison, WI ((a) and (b)) and a point with a very different maximum temperature annual cycle ((c) and (d), northwest of Oaxaca, Mexico) (Fig. 3). For Madison the method works very well. For the Mexico location, the method also works reasonably well although the method does not pick out the higher-frequency variations in the annual cycle (in this case, some of the high-frequency variations may be sampling noise).

Precipitation. We also assess the performance of these methods for deriving the monthly precipitation from the seasonal precipitation data (Fig. 4). The new method is certainly better than standard methods but sometimes there is simply not enough information in the seasonal averages to reproduce the high frequency components of the monthly annual cycle. In some cases, the new method does surprisingly well despite the fact there are strong changes in adjacent months. For example, the method captures much of the sudden onset of the North American monsoon in July (Fig. 4c). For a very similar situation a bit further north where winter precipitation is more pronounced, however, the method does not reproduce
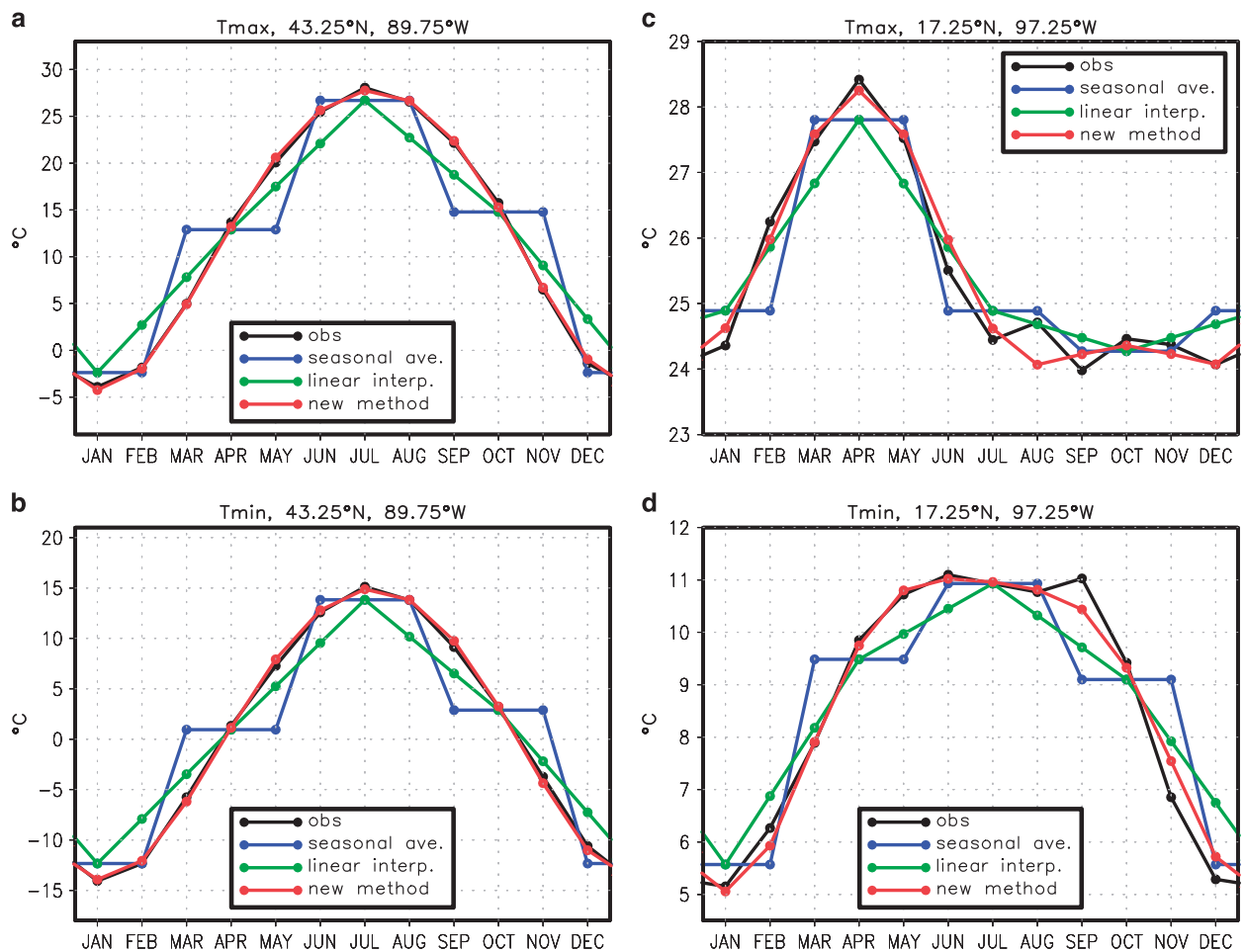

Figure 3. Assessment of the process used to predict the monthly annual cycle from the seasonal annual cycle for temperature. This plot shows maximum or minimum temperature annual cycle for Madison, WI ((a) and (b)) and a point northwest of Oaxaca, Mexico ((c) and (d)). The observed monthly annual cycle is in black. The observed monthly annual cycle averaged together into seasonal means is in blue. The monthly annual cycle estimated from the seasonal means by simple linear interpolation is in green. The monthly annual cycle estimated from the seasonal means by the new method described in the text is in red.
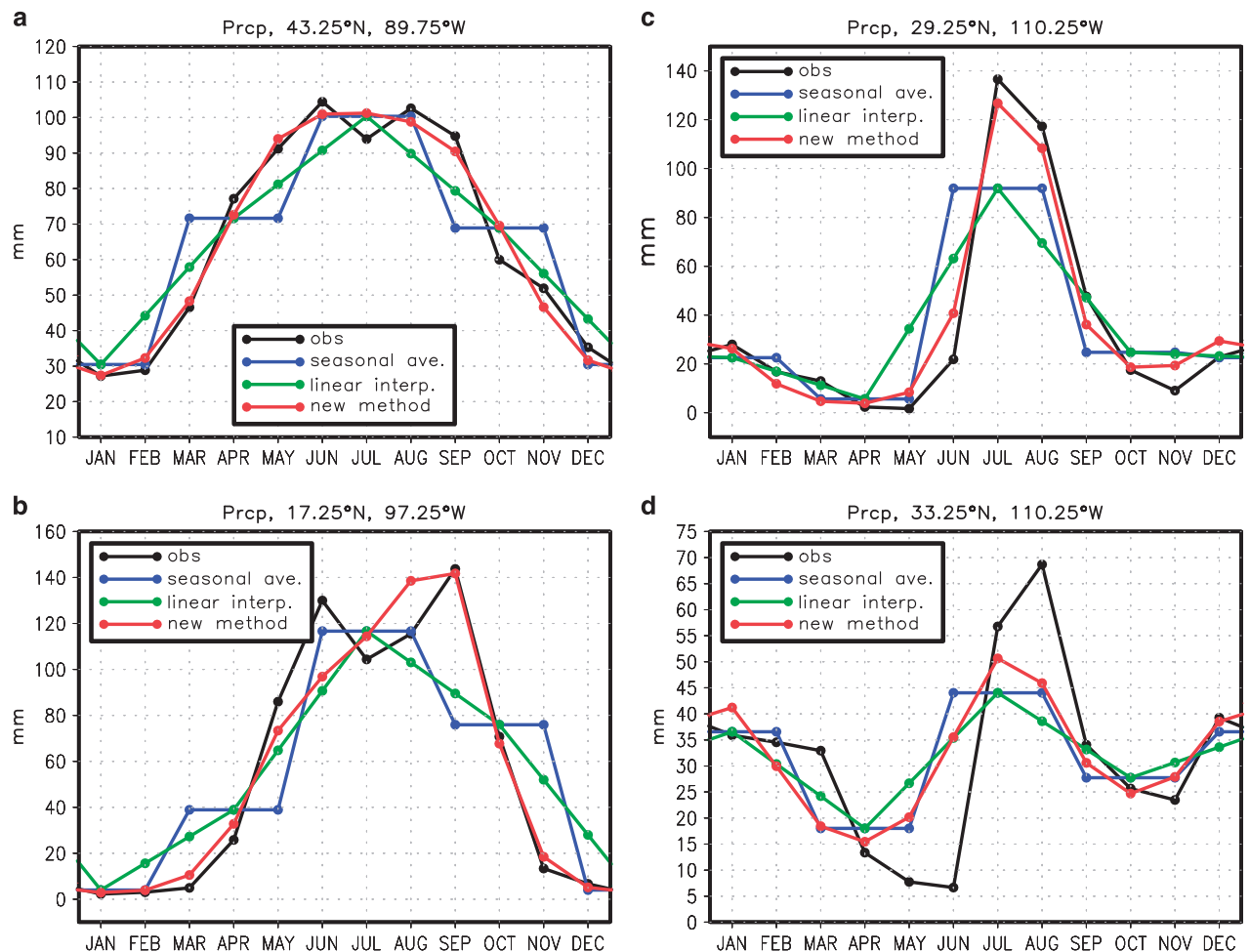
Figure 4. Estimating the monthly annual cycle from the seasonal annual cycle for precipitation. Shown is the precipitation annual cycle for (**a**) Madison, WI, (**b**) a point northwest of Oaxaca, Mexico, (**c**) a point in northwest Mexico, and (**d**) a point 4° north in Arizona. The observed monthly annual cycle is in black. The observed monthly annual cycle averaged together into seasonal means is in blue. The monthly annual cycle estimated from the seasonal means by linear interpolation is in green. The monthly annual cycle estimated from the seasonal means by the new method described in the text is in red.

the abrupt onset of the monsoon. Hence, we recommend some caution and critical judgment when using the monthly precipitation data; these data are not well suited for capturing precipitation extremes and abrupt intra-annual shifts in rainfall. To avoid these limitations, we recommend that public archives of climate simulations store data at daily or monthly resolution, rather than seasonal resolution.

Growing degree days: effect of using estimators of daily data

We assess the effectiveness of using estimators of daily standard deviations in temperature to calculate GDD (Fig. 5). During the warm months both the standard approach (simply subtracting monthly temperatures from a base temperature (T_0), equation 8) and the inclusion of estimates of daily standard deviations with the monthly data (equation 10) closely approximate the actual GDD. In the cool parts of the year, however, equation 10 does a much better job of estimating actual GDD.

Cross-correlation structure in downscaled variables

Downscaling will debias each individual variable, but it does not necessarily preserve realistic correlations among climate variables. A biased correlation structure among downscaled climate variables can bias the calculation of derived variables such as PET and AET, or bias ecological simulations based on these climate datasets. Here, we compare the correlation structure in observations from the observed CRU dataset versus the correlation structure in monthly anomalies in the downscaled CMIP5 historical simulations. Because some variables, such as wind speed, are only available as climatologies, this correlation analysis is restricted to precipitation, mean maximum and minimum daily temperature, and vapour pressure.
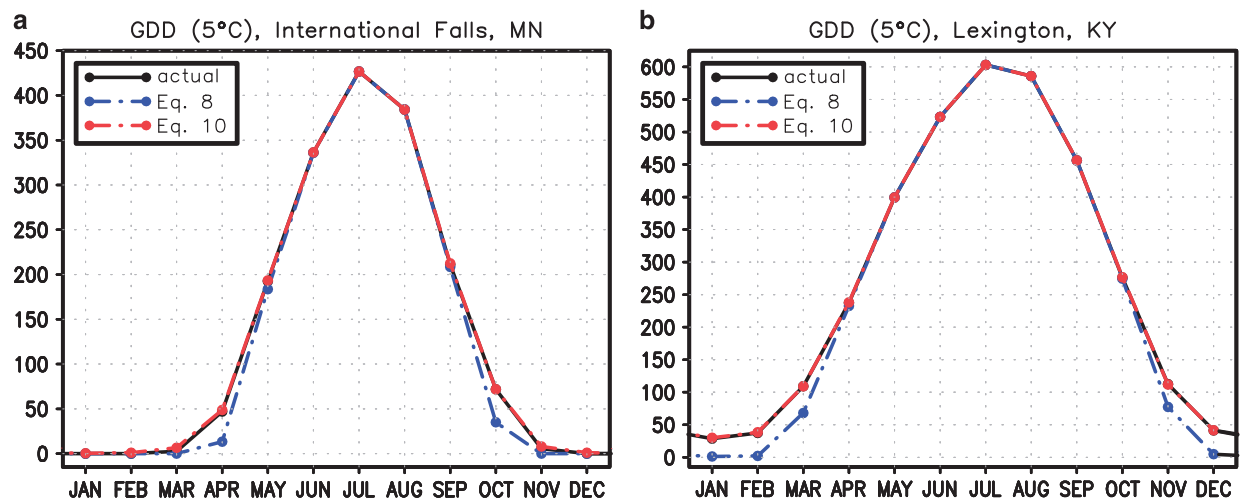
Figure 5. Estimating the monthly annual cycle of GDD base 5° using the monthly mean temperature only (blue) and the monthly mean temperature and the daily standard deviation (red). The actual GDD is in black. (**a**) International Falls, MN, (**b**) Lexington, KY.

The correlation between precipitation and temperature variables tends to match well between observations and the downscaled historical simulations. For example, the downscaling generally captures the correlation between maximum temperature and precipitation (Fig. 6a,b). The other correlations among precipitation and minimum and maximum temperature are similarly well captured. Correlation structure, however, is not well preserved between maximum temperature and vapour pressure (Fig. 6c,d) and precipitation and vapour pressure (Fig. 6e,f). The main discrepancies are in Mexico and the southwestern U.S and these discrepancies are also present in the raw climate model data. Diagnosing the reasons for these biases is beyond the scope of this research, but for maximum temperature and vapour pressure, we hypothesize that air temperature in the CMIP5 models is coupled too tightly to the land model, such that dry soil anomalies (which are correlated with vapour pressure) have too strong an effect on maximum temperature, via changes in sensible and latent heating. In actuality, temperature and vapour pressure likely are more controlled by atmospheric advection from remote locations. For precipitation and vapour pressure, perhaps the convection schemes in the CMIP5 models are too sensitive to boundary layer humidity as opposed to humidity in the atmosphere above the boundary layer, e.g. ref. 69.

We also checked to see whether correlation biases involving vapour pressure affect derived variables such as PET and AET (Figs 7 and 8). Both mean PET and AET agree well between observations and the downscaled simulations (Figs 7 and 8) and do not appear to be affected by these correlation biases. The good fit likely emerges because the total variation in AET and PET is a combination of the mean annual cycle and the monthly anomalies from the mean annual cycle. Apparently the mean annual cycle dominates and therefore the mean PET is well simulated. However, the correlation biases do impact the standard deviation of the PET anomalies from the mean annual cycle (Fig. 7c,d). The variance in downscaled PET is too large because the correlation between maximum temperature (T_{max}) and vapour pressure (e_s) is too small or even negative. Therefore the difference between the saturation vapour pressure (e_{sat}) at temperature T_{max} (i.e., $e_{sat}(T_{max})$) and e_s varies more than if e_{sat} (T_{max}) and e_s were highly correlated. Because one of the two terms in PET is proportional to e_{sat} (T_{max}) $- e_s$, PET also has too much variance.

Like PET, mean AET is also very well reproduced in the downscaling (Fig. 8). Unlike PET, however, the standard deviation for AET is also quite good in the downscaling and, moreover, the bias tends to have the opposite sign. This behaviour is due to the competing effects of the maximum temperature/vapour pressure bias and the precipitation/vapour pressure bias. For example, in the southwest U.S., increased vapour pressure leads to decreased PET (and likewise AET) via the former bias, but leads to increased precipitation, wetter soil and increased AET via the latter bias. Because the main biases are all related to vapour pressure, for future downscaling efforts, we recommend parameterizing vapour pressure (or dew point temperature) in terms of temperature and/or precipitation instead of downscaling it directly.

Usage Notes
General considerations
In both paleoclimatic simulations, the timescale is expressed as time before present, where 'present' follows radiocarbon dating conventions and is defined as 0 ka BP (1950). Decade 0 is defined as from 1st
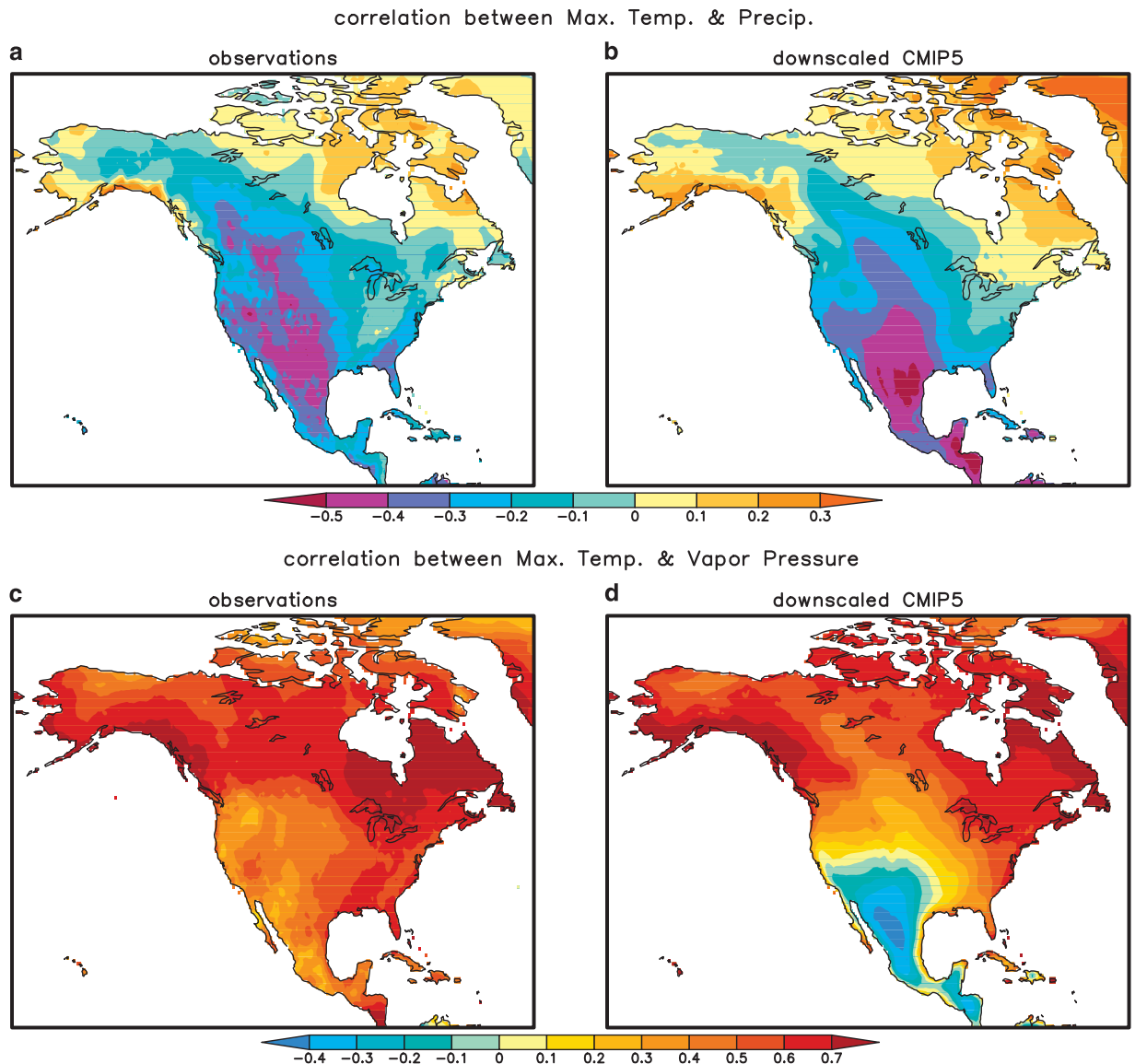
Figure 6. Maps of the temporal correlation between monthly anomalies for selected climate variables, to check whether the downscaled dataset has preserved the correlational structure in observational data. (**a**) Observed correlation between maximum temperature and precipitation (1950–2005). (**b**) As in (**a**) but for the downscaled CMIP5 data. This is the mean correlation averaged over 12 climate models. (**c,d**) As (**a,b**) but for maximum temperature and vapour pressure. The cross-correlation structure between temperature and precipitation is generally well-preserved, while the cross-correlation structure is less well preserved for vapour pressure.

January 1951 to 31st December 1960. CCSM3 data extends from decade −2200 to +3, which means it ends on December 31st, 1990, whereas ECBilt-CLIO goes from −2100 to −1, which means this simulation ends on December 31st, 1950.

Because debiasing and downscaling is imperfect, whenever possible, comparative analyses among climatic simulations should rely upon simulations from the same model or rely upon ensembles based on the same sets of models (e.g., the ACCESS3-1 simulations for the historical and the RCP4.5 scenarios). Similarly, future and paleoclimatic simulations should be compared to the downscaled climate simulations from the historical IPCC AR5 simulation or the available decades for the last centennial from the paleoclimatic simulations (1850–1990 for CCSM3 and 1850–1950 for ECBilt-CLIO), rather than to observational datasets for the 20th and 21st century, including to the original CRU dataset. Some model switching is unavoidable when going from the paleoclimatic to future climate simulations. However, by using standard observational baselines and methods we have minimized the effect of this switching.

Note that these paleoclimatic simulations contain no estimates of uncertainty, nor do they assess whether there is systematic bias between the paleoclimatic simulations and inferences based on
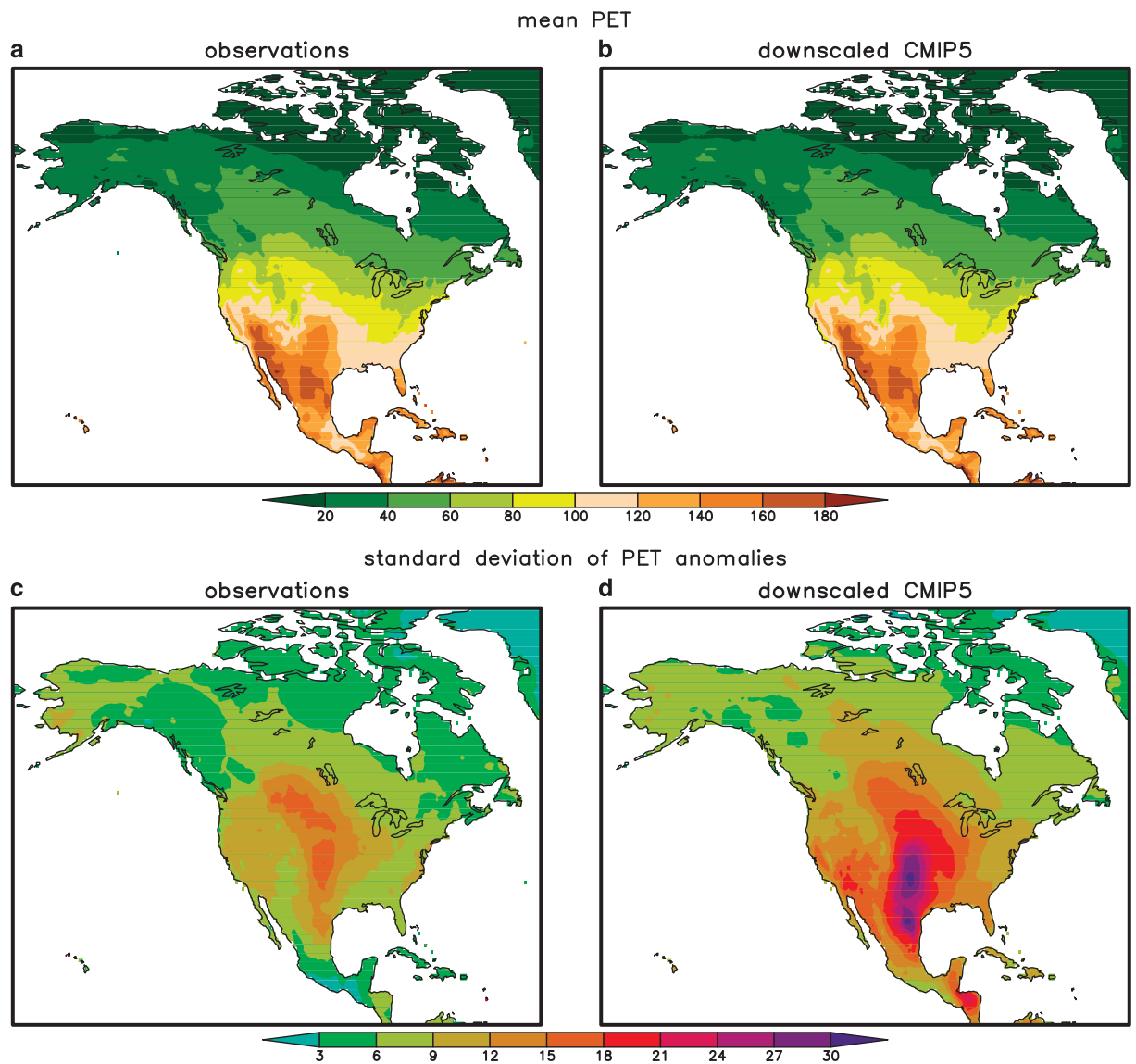
Figure 7. Comparison of observed and downscaled simulated potential evapotranspiration (PET), represented as annual averages of monthly data. (**a**) PET derived from observed variables such as temperature and vapour pressure. (**b**) As in (**a**) but for an average of 12 downscaled CMIP5 models. (**c**) Standard deviation of PET anomalies calculated from observational data. (**d**) As in (**c**) but for the downscaled CMIP5 models.

paleoclimatic proxies. Paleoclimatic model simulations carry inherent uncertainty both in model structure (how processes are represented in models) and parameterization (the values of parameters used within the model). There is a rich literature in paleoclimatic data-model comparisons and syntheses, particularly for the Paleoclimatic Intercomparison Modeling Project, e.g. (refs 39,70–72) and for the SynTrace simulations that are downscaled here[41,42,46,73,74]. Users are encouraged to refer to this literature to check for known data-model discrepancies before beginning work with the paleoclimatic simulations. A critical and on-going need is to understand how uncertainties in model choice, structure, and parameterization propagate to uncertainties in paleoclimatic simulations.

Primary and secondary variables (NetCDF)
All primary and secondary variables are stored as NetCDF files (see Table 1 (available online only) for file names and contents). For the paleoclimatic simulations (CCSM3 and EC-BILT), the individual netCDFs for each variable are archived separately (Table 1 (available online only)). For the 21st-century simulations, the individual netCDFs are organized into a directory structure as follows: cmip5\[emissions scenario]\[ESM model name]\[variable name.nc]. For example, cmip5\rcp45\ACCESS1-3\ET.nc holds the evapotranspiration simulations produced by the ACCESS1-3 ESM (Table 2) for emissions scenario RCP4.5.
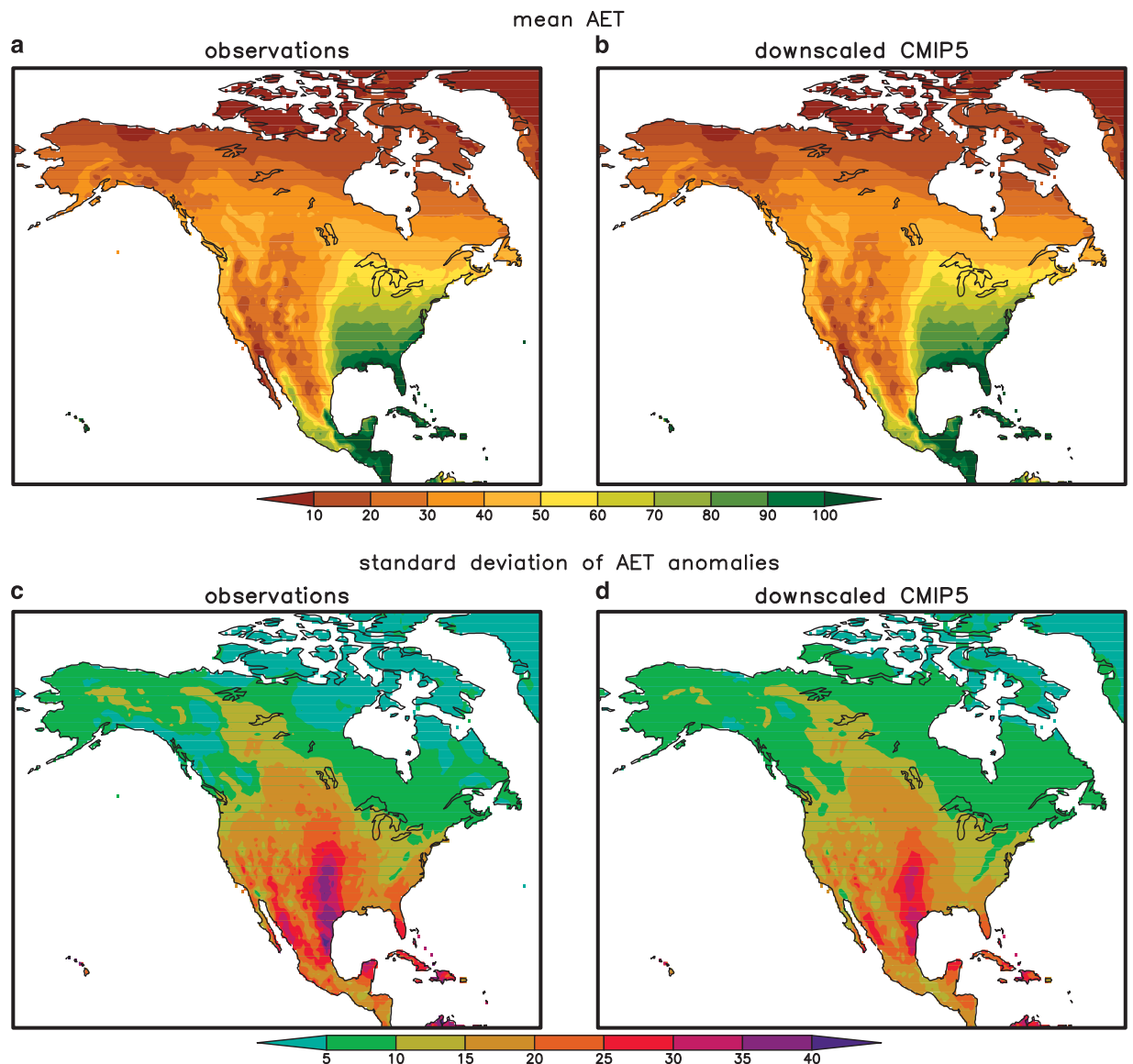
Figure 8. As Fig. 7, but for actual evapotranspiration (AET). (**a**) AET derived from observed variables such as temperature and vapour pressure. (**b**) As in (**a**) but for an average of 12 downscaled CMIP5 models. (**c**) Standard deviation of AET anomalies calculated from observational data. (**d**) As in (**c**) but for the downscaled CMIP5 models.

In the NetCDF files, the data have been packed into short integers (each requiring 2 bytes) instead of real numbers (requiring 4 bytes) to save space. One must unpack that data to get the correct floating point representation of the data. Each netCDF variable that has been packed has an add_offset and scale_factor attribute associated with it. The formula to unpack the data is:

unpacked value = add_offset + ((packed value) * scale_factor)

For more information see: http://www.unidata.ucar.edu/software/netcdf/docs/attribute_conventions.html.

The 'missing_value' attribute in the NetCDF data is set to −32768. Only grid points outside the downscaling domain are given the missing data value. The ncdf4 package in R (https://cran.r-project.org/web/packages/ncdf4/index.html) allows unpacking of these files with automatic correction for the offset, scale factor or the missing values.

Centennial and decadal variables (geotiffs)

The centennial- and decadal-scale summaries produced for input into ecological models are stored as geotiffs (TIFF format, .tif). Each raster holds a unique combination of climate variable, ESM, and time period, producing a large number of individual rasters. The rasters are zipped into compressed files organized by paleoclimatic simulation and CMIP5 emissions scenario (Table 1 (available online only)). Rasters are organized into a directory structure that differs slightly for the paleoclimatic and CMIP5

simulations. For the paleoclimatic rasters, the directory structure is [model name]\[time period]\[variable name].tif. For the 21st-century rasters, the directory structure is [CMIP5 scenario name]\[ESM model name]\[time period]\[variable name].tif. For example, RCP4.5\ACCESS1-3\2100\mo-lwr-TMIN.tif stores the lower monthly value of minimum daily temperatures, expected for year 2100 for the ACCESS1-3 ESM simulation under the CMIP5 RCP4.5 emissions scenario (Table 3). The naming conventions for variables and raster files are described in Table 3.

References

1. Moritz, C. & Agudo, R. The future of species under climate change: Resilience or decline? *Science* **341**, 504–508 (2013).
2. Fritz, S. A. *et al.* Diversity in time and space: wanted dead and alive. *Trends Ecol. Evol.* **28**, 509–516 (2013).
3. Blois, J. L., Zarnetske, P. L., Fitzpatrick, M. C. & Finnegan, S. Climate change and the past, present, and future of biotic interactions. *Science* **341**, 499–504 (2013).
4. Jackson, S. T. & Blois, J. L. Community ecology in a changing environment. *Proceedings of the National Academy of Sciences* **112**, 4915–4921 (2015).
5. Kidwell, S. M. Biology in the Anthropocene: Challenges and insights from young fossil records. *Proceedings of the National Academy of Sciences* **12**, 4922–4929 (2015).
6. Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C. & Mace, G. M. Beyond predictions: Biodiversity conservation in a changing climate. *Science* **332**, 53–58 (2011).
7. Fordham, D. A., Brook, B. W., Moritz, C. & Nogués-Bravo, D. Better forecasts of range dynamics using genetic data. *Trends Ecol. Evol.* **29**, 436–443 (2014).
8. Pimm, S. L. *et al.* The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752 (2014).
9. Botkin, D. B. *et al.* Forecasting the effects of global warming on biodiversity. *Bioscience* **57**, 227–236 (2007).
10. Willis, K. J. & MacDonald, G. M. Long-term ecological records and their relevance to climate change predictions for a warmer world. *Annual Review of Ecology, Evolution, and Systematics* **42**, 267–287 (2011).
11. Urban, M. C. Accelerating extinction risk from climate change. *Science* **348**, 571–573 (2015).
12. Thomas, C. D. *et al.* Extinction risk from climate change. *Nature* **427**, 145–148 (2004).
13. Williams, J. W. *et al.* Model systems for a no-analog future: Species associations and climates during the last deglaciation. *Annals of the New York Academy of Sciences* **1297**, 29–43 (2013).
14. Willis, K. J. & Birks, H. J. B. What is natural? The need for a long-term perspective in biodiversity conservation. *Science* **314**, 1261–1265 (2006).
15. Overpeck, J. T., Whitlock, C., Huntley, B. in *Paleoclimate, global change and the future* (eds Bradley R. S., Pedersen T. F., Alverson K. D. & Bergmann K. F.) 81–103 (Springer-Verlag, 2003).
16. Pearman, P. B., Guisan, A., Broennimann, O. & Randin, C. F. Niche dynamics in space and time. *Trends Ecol. Evol.* **23**, 149–158 (2008).
17. Martínez-Meyer, E., Peterson, A. T. & Hargrove, W. W. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Glob. Ecol. Biogeogr* **13**, 305–314 (2004).
18. Veloz, S. *et al.* No-analog climates and shifting realized niches during the late Quaternary: Implications for 21st-century predictions by species distribution models. *Global Change Biology* **18**, 1698–1713 (2012).
19. Ordonez, A. & Williams, J. W. Climatic and biotic velocities for woody taxa distributions over the last 16 000 years in eastern North America. *Ecology Letters* **16**, 773–781 (2013).
20. Birks, H. H. South to north: Contrasting late-glacial and early-Holocene climate changes and vegetation responses between south and north Norway. *Holocene* **25**, 37–52 (2015).
21. Ammann, B., von Grafenstein, U. & van Raden, U. J. Biotic responses to rapid warming about 14,685 yr BP: Introduction to a case study at Gerzensee (Switzerland). *Palaeogeography, Palaeoclimatology, Palaeoecology* **391**, 3–12 (2013).
22. Williams, J. W., Blois, J. L. & Shuman, B. N. Extrinsic and intrinsic forcing of abrupt ecological change: Case studies from the late Quaternary. *J. Ecol.* **99**, 664–677 (2011).
23. Svenning, J.-C. & Sandel, B. Disequilibrium vegetation dynamics under future climate change. *American Journal of Botany* **100**, 1266–1286 (2013).
24. Roberts, D. R. & Hamann, A. Predicting potential climate change impacts with bioclimate envelope models: a palaeoecological perspective. *Global Ecology & Biogeography* **21**, 121–133 (2012).
25. Jackson, S. T. & Overpeck, J. T. Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology* **26**(Supplement): 194–220 (2000).
26. Williams, J. W., Jackson, S. T. & Kutzbach, J. E. Projected distributions of novel and disappearing climates by 2100AD. *Proceedings of the National Academy of Sciences* **104**, 5738–5742 (2007).
27. Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T. & Ferrier, S. Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences* **110**, 9374–9379 (2013).
28. Williams, J. W. *et al.* The ice age ecologist: Testing methods for reserve prioritization during the last global warming. *Global Ecology & Biogeography* **22**, 289–301 (2013).
29. Sandel, B. *et al.* The influence of late Quaternary climate-change velocity on species endemism. *Science* **334**, 660–664 (2011).
30. Svenning, J.-C. & Skov, F. The relative roles of environment and history as controls of tree species composition and richness in Europe. *Journal of Biogeography* **32**, 1019–1033 (2005).
31. Hampe, A. & Jump, A. S. Climate relics: Past, present, future. *Annual Review of Ecology and Evolutionary Systematics* **42**, 313–333 (2011).
32. Gavin, D. G. *et al.* Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist* **204**, 37–54 (2014).
33. Carnaval, A. C., Hickerson, M. J., Haddad, C. F. B., Rodrigues, M. T. & Moritz, C. Stability Predicts Genetic Diversity in the Brazilian Atlantic Forest Hotspot. *Science* **323**, 785–789 (2009).
34. Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* **479**, 359–364 (2011).
35. Doughty, C. E. Preindustrial Human Impacts on Global and Regional Environment. *Annual Review of Environment and Resources* **38**, 503–527 (2013).
36. Nogués-Bravo, D. Predicting the past distribution of species climatic niches. *Global Ecology & Biogeography* **18**, 521–531 (2009).
37. Dietl, G. P. & Flessa, K. W. Conservation paleoecology: Putting the dead to work. *Trends Ecol. Evol.* **26**, 30–37 (2011).
38. Maiorano, L. *et al.* Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe. *Glob. Ecol. Biogeogr* **22**, 302–317 (2013).
39. Braconnot, P. *et al.* Evaluation of climate models using palaeoclimatic data. *Nature Clim. Change* **2**, 417–424 (2012).

40. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. A summary of the CMIP5 Experiment Design (2009).

41. Liu, Z. *et al.* Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming. *Science* **325,** 310–314 (2009).

42. He, F. *Simulating transient climate evolution of the last deglaciation with CCSM3*, PhD thesis (University of Wisconsin-Madison, 2010).

43. Schmidt, G. A. *et al.* Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). *Geoscientific Model Development* **4,** 33–45 (2011).

44. Bothe, O., Jungclaus, J. H., Zanchettin, D. & Zorita, E. Climate of the last millennium: ensemble consistency of simulations and reconstructions. *Clim. Past* **9,** 1089–1110 (2013).

45. Randall, D. A. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon S. *et al.*) 589–662 (Cambridge University Press, 2007).

46. He, F. *et al.* Northern Hemisphere forcing of Southern Hemisphere climate during the last deglaciation. *Nature* **494,** 81–85 (2013).

47. Timm, O. & Timmermann, A. Simulation of the last 21 000 years using accelerated transient boundary conditions. *J. Clim* **20,** 4377–4401 (2007).

48. Wilby, R. L. *et al. Guidelines for use of climate scenarios developed from statistical downscaling methods* (IPCC Task Group on Data and Scenario Support for Impacts and Climate Analysis, 2004).

49. Harris, I., Jones, P., Osborn, T. & Lister, D. Updated high-resolution grids of monthly climatic observations–the CRU TS3.10 dataset. *Int. J. Climatol.* **34,** 623–642 (2014).

50. New, M. G., Hulme, M. & Jones, P. D. Representing twentieth century space-time climate variability. Part I: Development of a 1961–1990 mean monthly terrestrial climatology. *J. Clim.* **12,** 829–856 (1999).

51. Zhang, T. *et al.* The validation of the GEWEX SRB surface shortwave flux data products using BSRN measurements: A systematic quality control, production and application approach. *Journal of Quantitative Spectroscopy and Radiative Transfer* **122,** 127–140 (2013).

52. Gupta, S. K. *et al.* Improvement of surface longwave flux algorithms used in CERES processing. *Journal of Applied Meteorology and Climatology* **49,** 1579–1589 (2010).

53. Berger, A. Long-term variations of daily insolation and Quaternary climatic changes. *Journal of the Atmospheric Sciences* **35,** 2362–2367 (1978).

54. Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. Crop evapotranspiration-Guidelines for computing crop water requirements. *FAO irrigation and drainage paper* **300,** 6541 (1998).

55. Hastings, D. A. & Dunbar, P. K. *Global Land One-kilometer Base Elevation. Report No. 34, 147* (National Oceanic and Atmospheric Administration, 1999).

56. Peltier, W. R. Ice age paleotopography. *Science* **265,** 195–201 (1994).

57. Liepert, B. G. & Lo, F. CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environmental Research Letters* **8,** 029401 (2013).

58. Gleckler, P. J., Taylor, K. E. & Doutriaux, C. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres* **113,** D6 (2008).

59. Tabor, K. & Williams, J. W. Globally downscaled climate projections for assessing the conservation impacts of climate change. *Ecol. Appl.* **20,** 554–565 (2010).

60. Wood, A. W., Maurer, E. P., Kumar, A. & Lettenmaier, D. P. Long range experimental hydrologic forecasting for the eastern US. *Journal of Geophysical Research* **107,** 4429 (2002).

61. Moré, J. J., Garbow, B. S. & Hillstrom, K. E. *User guide for MINPACK-1* 48 (Argonne National Laboratory, 1980).

62. Dobrowski, S. Z. *et al.* The climate velocity of the contiguous United States during the 20th century. *Global Change Biology* **19,** 241–251 (2013).

63. Lutz, J. A., van Wagtendonk, J. W. & Franklin, J. F. Climatic water deficit, tree species ranges, and climate change in Yosemite National Park. *Journal of Biogeography* **37,** 936–950 (2010).

64. Dunne, K. A. & Wilmott, C. J. *Global distribution of plant extractable water capacity of soil* (Oak Ridge National Laboratory Distributed Active Archive Center, 2000).

65. Williams, J. W., Shuman, B. N., Webb, T. III, Bartlein, P. J. & Leduc, P. L. Late Quaternary vegetation dynamics in North America: Scaling from taxa to biomes. *Ecological Monographs* **74,** 309–334 (2004).

66. Blois, J. L., Williams, J. W., Grimm, E. C., Jackson, S. T. & Graham, R. W. A methodological framework for improved paleovegetation mapping from late-Quaternary pollen records. *Quaternary Science Reviews* **30,** 1926–1939 (2011).

67. Stephenson, N. L. Actual evapotranspiration and deficit: biologically meaningful correlates of vegetation distribution across spatial scales. *Journal of Biogeography* **25,** 855–870 (1998).

68. Woodward, F. I., Kelly, C. K. in *Plant Functional Types: Their Relevance to Ecosystem Properties and Global Change* (eds Smith T. M., Shugart H. H. & Woodward F. I.) 47–65 (Cambridge University Press, 1997).

69. Derbyshire, S. H. *et al.* Sensitivity of moist convection to environmental humidity. *Quarterly Journal of the Royal Meteorological Society* **130,** 3055–3079 (2004).

70. Hargreaves, J. C., Annan, J. D., Yoshimori, M. & Abe-Ouchi, A. Can the Last Glacial Maximum constrain climate sensitivity? *Geophysical Research Letters* **39,** L24702 (2012).

71. Schmidt, G. A. *et al.* Using palaeo-climate comparisons to constrain future projections in CMIP5. *Climates of the Past* **10,** 221–250 (2014).

72. Harrison, S. P. *et al.* Climate model benchmarking with glacial and mid-Holocene climates. *Clim. Dyn.* **43,** 671–688 (2013).

73. Clark, P. U. *et al.* Global climate evolution during the last deglaciation. *Proceedings of the National Academy of Sciences* **109,** E1134–E1142 (2012).

74. Liu, Z. *et al.* The Holocene temperature conundrum. *Proceedings of the National Academy of Sciences* **111,** E3501–E3505 (2014).

Data Citation

1. Lorenz, D. J., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C. & Williams, J. W. *Dryad Digital Repository*. http://dx.doi.org/10.5061/dryad.1597g (2016).

Author Contributions

D.J.L. conducted all statistical downscaling analyses and co-led manuscript writing. D.N.L. processed paleoshorelines and calculated centennial and decadal summaries of climate variables from the downscaled monthly simulations, converted the NetCDF files to raster formats suitable for input to ecological models, and contributed to manuscript writing. J.L.B. co-designed the research study and contributed to manuscript writing. M.C.F. co-designed the research study and contributed to manuscript writing. J.W.W. co-designed the research study and co-led manuscript writing.

Additional Information

Table 1 is only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lorenz, D. J. *et al.* Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD. *Sci. Data* 3:160048 doi: 10.1038/sdata.2016.48 (2016).