

An evaluation of the statistical homogeneity of the Twentieth Century Reanalysis

Craig R. Ferguson · Gabriele Villarini

Received: 17 January 2013 / Accepted: 4 November 2013 / Published online: 27 November 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract The Twentieth Century Reanalysis (20CR) holds the distinction of having the longest record length (140-year; 1871–2010) of any existing global atmospheric reanalysis. If the record can be shown to be homogenous, then it would be the first reanalysis suitable for long-term trend assessments, including those of the regional hydrologic cycle. On the other hand, if discontinuities exist, then their detection and attribution—either to artificial observational shocks or climate change—is critical to their proper treatment. Previous research suggested that the quintupling of 20CR’s assimilated observation counts over the central United States was the primary cause of inhomogeneities for that region. The same work also revealed that, depending on the season, the complete record could be considered homogenous. In this study, we apply the Bai-Perron structural change point test to extend these analyses globally. A rigorous evaluation of 20CR’s (in)homogeneity is performed, composed of detailed quantitative analyses on regional, seasonal, inter-variable,

and intra-ensemble bases. The 20CR record is shown to be homogenous (natural) for 69 (89) years at 50 % of land grids, based on analysis of the July 2 m air temperature. On average 54 % (41 %) of the grids between 60°S and 60°N are free from artificial inhomogeneities in their February (July) time series. Of the more than 853,376 abrupt shifts detected in 26 variable fields over two monthly time series, approximately 72 % are non-climate in origin; 25 % exceed 1.8 standard deviations of the preceding time series. The knock-on effect of inhomogeneities in 20CR’s boundary forcing and surface pressure data inputs to its surface analysis fields is implicated. In the future, reassessing these inhomogeneities will be imperative to achieving a more definitive attribution of 20CR’s abrupt shifts.

Keywords Twentieth Century Reanalysis · Change point detection · Climate trend analysis · Observational shocks · Sparse data assimilation

Electronic supplementary material The online version of this article (doi:10.1007/s00382-013-1996-1) contains supplementary material, which is available to authorized users.

C. R. Ferguson (✉)
Department of Hydrology and Water Resources Engineering,
Institute of Industrial Science, The University of Tokyo,
Tokyo, Japan
e-mail: crferguson@albany.edu

C. R. Ferguson
Department of Environmental Resources Engineering, The State
University of New York College of Environmental Science
and Forestry, Syracuse, NY, USA

G. Villarini
IIHR, Hydroscience & Engineering, The University of Iowa,
Iowa City, IA, USA

1 Introduction

Understanding how the global hydrologic cycle has responded to climate change (natural or anthropogenic) in the past and will likely respond to climate change in the future is imperative to ensuring the efficacy of adaptive planning measures that aim to minimize the adverse socio-economic and environmental impacts of climate change. Increases in the frequency and severity of floods and droughts (e.g., Sheffield and Wood 2008; Huntington 2006), heatwaves (e.g., Schar et al. 2004), wildfires (e.g., Westerling et al. 2006; Moritz et al. 2012), and strains on water and food security (e.g., Lobell et al. 2008), have all been linked to climate change. Without advanced warning

or sufficient resources to mitigate their effects, even modern societies risk destabilization as a consequence of climate extremes (e.g., Hsiang et al. 2011).

In theory, observation-based global reanalyses such as NASA's Modern Era Retrospective-analysis for Research and Applications (MERRA; Rienecker et al. 2011; Rienecker et al. 2008) provide a robust means for detecting long-term trends and/or abrupt shifts in the hydrologic cycle, and more importantly, enable attribution to their root mechanisms. In practice, inhomogeneities caused by forecast model biases and/or the observations they assimilate limits their applicability for long-term trend assessment (Thorne and Vose 2010; Dee et al. 2011b). Detecting and correcting for observational biases is complicated by the diversity of the observations (e.g., in source, coverage, and record length). The challenge becomes how to distinguish between real climate shifts and artificial observational "shocks." From an operational perspective, diagnosing the latter (e.g., sensor-related breakpoints) is critical to the success of variational bias adjustment methods (Dee et al. 2011a).

So-called "climate quality" reanalyses of which the Twentieth Century Reanalysis (20CR; Compo et al. 2011) is the first, seek to ameliorate the issue of unphysical time-varying biases through the assimilation of only those data streams that are stable over long periods of time. For example, 20CR assimilates only synoptic surface and sea-level pressure observations that span a period of 140-years (1871–2010). If shown to be homogenous, it would provide the first comprehensive (i.e., multivariate, multi-level) and consistent long-term climate record suitable for trend assessment, including assessment of whether the hydrologic cycle is intensifying (Huntington 2006).

In a previous study (Ferguson and Villarini 2012), we suggested that the fivefold increase of 20CR's assimilated observation counts in the 1940's over the central U.S. caused inhomogeneities during the same period. We also showed that, depending on the season, the complete (140-year) record could be considered homogenous. The purpose of this paper is to provide a comprehensive global follow-up to our previous work using a similar methodology. Specifically, we address the following questions:

- (1) Is the finding that inhomogeneities in 20CR are linked to underlying observational density unique to the central U.S., or globally-representative? If other artificial (non-climate) inhomogeneities are detected, what is their frequency relative to those that are naturally occurring (as a product of climate variability)?
- (2) For what fraction of the globe is 20CR homogeneous over the period of record? And, for surface air temperature and precipitation, how does this compare with available global gridded in situ datasets?

- (3) What is the size distribution of the discontinuities? And how are they distributed in time?
- and (4) How varied are inhomogeneity characteristics among 20CR's variable fields?

Our approach is to assess homogeneity on a global grid point basis and summarize results regionally for a large subset of variables. The overarching motivation for this study, which is to use 20CR to identify the key processes, feedback mechanisms, and hydrometeorological variables that drive long-term changes in the hydrologic cycle at regional scales (e.g., Troy et al. 2012), dictates that the homogeneity assessment be conducted at seasonal time step, but this is not always practical. In our case, we choose to focus on the minimum and maximum months in the seasonal cycle of global homogeneity.

The paper is organized as follows. Section 2 describes the 20CR, comparison datasets, and the full methodology, including the statistical test that we apply. Results for each of the experiments are presented in Sect. 3. Section 4 includes a brief summary and conclusions.

2 Data and Methods

2.1 20CR

The 20CR is a global atmospheric reanalysis spanning the 140-year period from 1871 to 2010 at 2.0° spatial resolution and 6-hourly temporal resolution with 24 atmospheric levels (Compo et al. 2011). It is remarkable not only because it more than doubles the pre-existing reanalysis record length but because only two surface observations are used. Namely, six-hourly surface- and sea-level pressure observations from the International Surface Pressure Databank (ISPD v2.2.4) and monthly sea surface temperature (SST) and sea-ice concentration fields from the Hadley Centre Sea Ice and SST dataset (HadISST v1.1; Rayner et al. 2003). The ISPD v2.2.4 contains millions of observations from International Comprehensive Ocean–Atmosphere DataSet (ICOADS) v2.2 (Worley et al. 2005) as well as newly digitized data from land stations that have never before been used. HadISST v1.1 (described in Sect. 2.2.3 below) incorporates many types of observations, in situ as well as from satellites.

The 20CR employs a deterministic Ensemble Kalman Filter (EKF) based on the ensemble square root filter algorithm of Whitaker and Hamill (2002). Background first guess fields are obtained from a short-term forecast ensemble run in parallel, consisting of 56 9-hour integrations of the April 2008 experimental version of the U.S. National Centers for Environmental Prediction

Table 1 The 26 variable subset of 20CR analyzed in this study and their abbreviated names

Variable name	Definition
T_a	2 m temperature
TMIN	Minimum 2 m temperature
TMAX	Maximum 2 m temperature
Q	Total runoff
P	Total precipitation
CP	Convective precipitation
PWV	Total column precipitable water vapor (*)
C	Total column moisture convergence (*^)
E	Atmospheric-inferred evapotranspiration (*^)
LH	Latent heat flux
SH	Sensible heat flux
G	Ground heat flux
SW↓	Downward shortwave radiation flux
LW↓	Downward longwave radiation flux
SW↑	Upward shortwave radiation flux
LW↑	Upward longwave radiation flux
WSPD	10 m wind speed (^)
HPBL	Planetary boundary layer height
CAPE	Convective available potential energy (*)
CTP	Convective triggering potential (*^)
CIN	Convective inhibition (*)
HI	Low-level humidity index (*^)
LCL	Lifting condensation level (^)
P_{surf}	Surface pressure (*)
SPFH	2 m specific humidity
PWVf	Total column precipitable water vapor

All variables were taken from the forecast fields, unless noted by an *, in which case the variables were taken from the analysis. Derived quantities are denoted by an ^

(NCEP) Global Forecast System (GFS; Kanamitsu et al. 1991; Moorthi et al. 2001; Saha et al. 2006)—each initiated using the previous 6-hour analysis. The GFS is coupled with the four-layer NOAH v2.7 land surface model (Ek et al. 2003) and run at a horizontal resolution of T62 (192 × 94 Gaussian longitude/latitude) and 3-hourly time step with 28 hybrid sigma-pressure levels. For each time-iteration of the assimilation (6-hourly) and forecast (3-hourly) systems, the ensemble mean and uncertainty estimate (i.e., ensemble spread) are recorded for all variable fields. ISPD surface pressure and sea level pressure observations are independently quality controlled during the assimilation cycle (i.e., ISPD quality controls were not used) through a multi-step procedure that includes a basic check for meteorological plausibility, comparisons with the first guess ensemble and neighboring observations, and for the station (land

pressure observations only, an adaptive time-varying platform-by-platform bias correction scheme. The scheme corrects for statistically significant differences between the first guess and observational means over 60-day increments of the assimilation, which has the effect of smoothing out any sudden shift in observations over a period of a few months (Compo et al. 2011; their Appendix 2). No bias correction was performed on the marine and tropical cyclone ‘best track’ pressure observations and reports. For maximum computational efficiency, 5-year production streams (with 14-month spin-up) were used, with the exception of a single 6-year stream for the period 1946–1951, and the most modern period 2001–2010.

We analyze the monthly or annual time-average of 26 variables in total: four analysis fields, 17 forecast first guess fields, and 5 derived quantities (Table 1). We focus on the ensemble mean fields, except in Appendix 1, where we analyze the every-member ($n = 56$) data available at two locations: Geneva, Switzerland (46.20°N, 6.15°E) and Rondonia, Brazil (24.0°S, 51.0°W). The ensemble spread fields are analyzed jointly in order to diagnose the nature of the breakpoint (i.e., real or artificial). Note that the ensemble spread is an estimate of the uncertainty in the 6-hourly analyses and the 3-hourly forecast values, not the estimate of uncertainty in the monthly mean values themselves. Also important to note is that the uncertainty estimates do not account for uncertainties in the SST, which are considerable (Rayner et al. 2003; Kennedy et al. 2011a, 2011b).

The derived quantities were selected on the basis of their relevance to studies of land–atmosphere interaction and the global water and energy cycles (e.g., Ferguson and Wood 2011; Betts 2009; McVicar et al. 2008; Ferguson et al. 2012). They are: total column moisture convergence (C), atmospheric-inferred evapotranspiration (E), 10 m wind speed, convective triggering potential (CTP), low-level humidity index (HI), and lifting condensation level (LCL). To be clear, E is computed by:

$$E = P - C + \frac{dw}{dt}, \quad (1)$$

where P is obtained from the forecast, C is calculated using the analysis surface pressure and multi-level humidity and wind fields, and dw/dt , the change in total column moisture, is calculated by taking the difference between first of the month (e.g., February 1 minus January 1) total column precipitable water vapor/ice (PWV) analysis fields. The CTP is a measure of departure from the moist adiabatic temperature lapse rate from 100 to 300 hPa above ground level (AGL). The HI is defined by the 50–150 hPa AGL dew point depression. The LCL is computed from a parcel originating at 2 m and lifted along a dry adiabat.

2.2 Comparison data

2.2.1 CRU

The Climate Research Unit (CRU) time series dataset version 3.1 (TS3.1) is a 0.5° gridded record of monthly land surface climate (precipitation, mean temperature, diurnal temperature range, and other secondary variables) for the period 1901–2009, derived entirely from daily surface meteorological observations (Mitchell and Jones 2005; Mitchell et al. 2004; New et al. 2000). For this study, we use the 2 m air temperature and corresponding station count data only. TS3.1 fields are the product of an angular-distance-weighted interpolation of monthly climate anomalies relative to the 1961–1990 mean, subsequently recombined with an equivalent grid of normals for the same baseline period (New et al. 1999). In estimating each grid point, TS3.1 uses the eight nearest station records, regardless of direction, within an empirically derived correlation decay distance (CDD) of 1,200 km for temperature (New et al. 2000). If a grid point lies beyond the CDD of any stations, the grid is ‘relaxed’ to the 1961–1990 mean. We found that an entire year (i.e., 12 consecutive months) or longer is relaxed at some point (generally, earlier) in the record at 45 % of the grids. Major sources of error in the TS3.1 include instrumental measurement error, insufficient station density, and interpolation errors (New et al. 2000).

The temperature database on which CRU TS is based was assembled in the late 1990’s; only updates from the Monthly Climatic Data for the World (MCDW), monthly climatological data (CLIMAT), and various Bureau of Meteorology (BOM) reports are routinely incorporated. Unlike in previous versions of the TS (TS2.1; Mitchell and Jones 2005), neither homogeneity assessment nor homogenization of the ingested data streams is performed in the production of TS3.1. Nevertheless, a large number of its data sources, including the Global Historical Climatology Network (GHCN; Lawrimore et al. 2011), are received by CRU in homogenized state.

2.2.2 GPCC

The Global Precipitation Climatology Centre (GPCC; Becker et al. 2012, Schneider et al. 2013) produces a monthly gauge-based precipitation reanalysis at 0.5° spatial resolution that spans the twentieth century (1901–2010). It is commonly referred to as the Full Data Reanalysis. In this study, we use the latest release, version 6 (December 2011). It is based on the world’s largest and most comprehensive collection of precipitation gauge data. This includes: data from over 190 national weather service networks; daily surface synoptic observations (SYNOP) and monthly CLIMAT messages transmitted via the World

Meteorological Organization (WMO) global telecommunication system (GTS); global precipitation data collections from CRU, GHCN, and the Food and Agriculture Organisation (FAO); in addition to numerous other regional datasets. Only stations with ten years of data or more are included. After the gauge data (and/or metadata) are received, they are subjected to rigorous comparative analysis (screening) against different sources of data relevant for the same or neighboring stations, as well as a gridded background anomaly. Once screened and (if necessary) corrected, GPCC applies a modified version of the SPHEREMAP method (Willmott et al. 1985) to spatially interpolate station anomalies to grid anomalies, drawing from the data of 16 nearby stations. In the present version (v6), the normal fields are the product of observations from approximately 67,200 stations. A count of contributing gauges is provided for each estimate of P .

The Full Data Reanalysis was not designed to achieve temporal homogeneity and is therefore not recommended for climate trend analysis. An alternative GPCC analysis, VASCLIMO, which includes only those stations with data coverage for 90 % (45 years) of its record length, was intended for this purpose. It covers, however, only a fraction (1951–2000) of our period of interest and for that reason it is not used in this study.

2.2.3 HadISST v1.1

The Met Office Hadley Centre’s globally complete monthly sea ice concentration and SST dataset version 1.1 (HadISST v1.1; Rayner et al. 2003) covers the period from 1870–2010 at 1.0° spatial resolution. It uses gridded, quality-controlled in situ observations for 1871–1981, merged with night-time bias-adjusted National Oceanic and Atmospheric Administration (NOAA) satellite-borne Advanced Very High Resolution (AVHRR) observations from January 1982 onwards. The gridded data for 1871–1941 were bias-adjusted to account for uncertainty in sampling methods following Folland and Parker (1995). A two-stage (global and inter-annual) reduced space optimal interpolation (RSOI; Kaplan et al. 1997) procedure was applied to reconstruct the complete (spatial and temporal) SST fields. Quality-improved (homogenized for variance) gridded data is blended with the reconstructed fields to restore local (~ 500 km) variance attributes. HadISST v1.1 is of particular relevance to our work because it supplies the boundary conditions for the 20CR (see Sect. 2.1).

2.2.4 HadSLP2

The Met Office Hadley Centre’s globally complete monthly mean sea level pressure (SLP) dataset version 2 (HadSLP2; Allan and Ansell 2006) covers the period from

1850 to 2004 at 5° spatial resolution. It is an RSOI reconstruction (like HadISST v1.1) based on a blending of monthly mean SLP observations from 2,228 land stations with gridded marine SLP observations from ICOADS v2.2. The marine component of the ISPD v2.2.4 used in the 20CR is extracted from ICOADS versions 2.4 and 2.5 for the periods of 1952–2010 and 1871–1951, respectively. Consequently, there is a high degree of overlap in the observational content.

The HadSLP2 terrestrial data are subject to a large number of quality control procedures including temporal and spatial consistency checks and a Kolmogorov-Smirnov (K–S) test (Press et al. 1992) for inhomogeneities in the seasonal mean. In the case of the marine data, the Hadley Centre Marine Data System (MDS) version 2 was applied. MDS includes climatology and near-neighbor spatial consistency checks. Time series quality-cleared by MDS were then subjected to further correction according to procedures described in Ansell et al. (2006).

Along with SLP, the observation counts and uncertainty estimates are also provided. Of course, uncertainty estimates are only calculable for the month and grid point for which data has been assimilated. For HadSLP2's period of overlap with 20CR (1871–2004), only 2.2 % ($n = 60$) of grid points offer a continuous record of uncertainty.

HadSLP2 was extended from 2005 to present using NCEP-National Center for Atmospheric Research (NCAR) reanalysis (denoted as “R1”; Kalnay et al. 1996; Kistler et al. 2001). However, this more modern record, named HadSLP2r, is not homogeneous with the earlier time series (see <http://www.metoffice.gov.uk/hadobs/hadslp2/>). For the above reasons, we use only HadSLP2 and its corresponding observational count record.

2.2.5 COBE SST

The Japan Meteorological Agency's globally complete monthly sea ice concentration and SST dataset (COBE SST; Ishii et al. 2005) covers the period from 1891 to present at 1.0° spatial resolution. It uses quality-controlled SSTs from ICOADS v2.0, the Japanese Kobe Collection, the Canadian Marine Environmental Data Service (MEDS) buoy dataset, as well as ship reports. As in HadISST v1.1, biases in bucket observations before 1941 are removed using the method of Folland and Parker (1995). The objective analyses are based on optimum interpolation and a monthly reconstruction with empirical orthogonal functions intended to homogenize the data.

2.2.6 ERSST v3b

The NOAA Extended Reconstruction Sea Surface Temperature version 3b (ERSST v3b; Smith and Reynolds

2003) is a globally complete monthly sea ice concentration and SST dataset covering the period from 1854 to present at 2° spatial resolution. It is based upon statistical interpolation of quality-controlled ICOADS v 2.4 data and does not include satellite data due to a cold bias in the satellite-derived SSTs that proved difficult to correct. The spatial variance ratio of the SST is measurably less than that of HadISST v1.1 because filtering of modes is applied to reduce small-scale noise.

2.3 Methods

In this study, we rely exclusively on the results of the Bai-Perron structural change point test (Bai and Perron 2003). This test is well suited to our purpose for two reasons: it is objective and it has the capability to detect multiple breakpoints. Multiple breaks are the norm rather than the exception in 20CR's 140-year record (Ferguson and Villarini 2012). Seventy-three percent of all grid locations have more than one break in annual mean T_a (not shown). The test must be objective because it needs to be automatable for bulk application on a global grid basis ($n = 16,200 @ 2.0^\circ$). Finally, we found the Bai-Perron test to be of comparable skill to the Pettitt test (Pettitt 1979) in zero and single break cases; the Pettitt test corroborates 90 % of homogenous series and 73 % of single break dates (not shown).

2.3.1 Bai-Perron test

The Bai-Perron test (Bai and Perron 2003) enables the simultaneous estimation of multiple change points of unknown timing. The data are assumed to come from a distribution belonging to the exponential family (e.g., Gaussian, exponential, Poisson). The test represents an extension of F statistical tests against a single-shift (e.g., Andrews 1993) for multiple break applications. It is based on a standard linear regression model for which the null hypothesis of structural stability is tested against the alternative that at least one coefficient varies with time. We use a constant as the regressor for our model. The minimum permissible segment length (i.e., trimming parameter) is set by the user. We chose to set this parameter, h , to 0.15 (default value in the package we used), which equates to allowing up to five breaks in the 140-year (1871–2010) 20CR. Note this parameter also dictates the earliest and latest possible break date. First, the number of breaks is selected using BIC (Bayesian Information Criterion; Schwarz 1978). Then, dating of the change points is accomplished via a dynamic programming approach that minimizes the objective function (RSS) (Bai and Perron 2003). The statistical confidence intervals corresponding to each change point are computed using the distribution

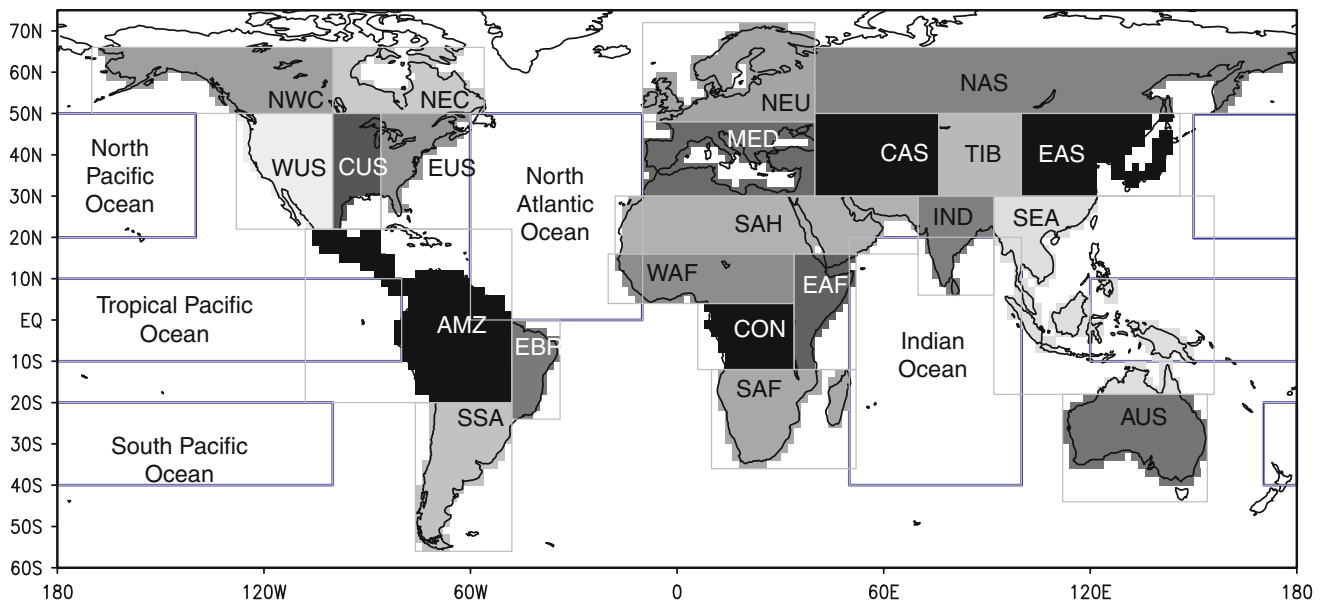


Fig. 1 Regions over which the analysis was conducted ($n = 27$; see Table 2). The delineation over land is based on that of Giorgi and Francisco (2000), but modified to better reflect the land–atmosphere

coupling and climatological wetness regimes shown by Ferguson et al. (2012). Ocean domains are defined according to standard Japan Meteorological Agency conventions

function of Bai (1997), although in a limited number of cases (<1.5 % in this study), errors preclude their estimation (i.e., singular gradient). When calculable, we include the 95 % confidence interval. Modifying the confidence level affects only the length of the confidence interval(s), not the number of change points detected. The Bai-Perron test is non-optimal for cases in which the record length is short, the break sizes are small, and/or the breaks are clustered (Bai and Perron 2003). In addition to abrupt shifts in the mean, changes in series variability or the presence of gradual trends can also lead to breakpoint detection.

Our results were obtained using R 2.14.2 (R Development CoreTeam 2008) with the packages *strucchange* v1.4-6 (Zeileis and Kleiber 2005; Zeileis et al. 2003), *sandwich* v2.2-9, and *zoo* v1.7-7.

2.3.2 Experimental design

We apply the Bai-Perron test on a 2° grid-by-grid basis to the ensemble mean and uncertainty estimate fields of 26 variables (see Sect. 2.1, Table 1). We summarize the results for the area of 60°S – 60°N as well as 27 constituent climatic regions of which 22 are over land and the rest are over ocean (Fig. 1; Table 2). Whether a time series is inhomogeneous is case dependent. From a statistical perspective a time series may be considered inhomogeneous if it has any breakpoints while in a climate sense inhomogeneous means affected by changes that are not of climate origin. In this study, we will focus on inhomogeneities in the second sense, although information on climate-related breakpoints will be presented in figures.

We focus on T_a and P because of their prominent role in global climate, but also because (along with surface pressure) they are the most widely (and accurately) monitored meteorological quantities. Due to high confidence in their measurements (especially T_a), they commonly serve as benchmarks in model performance. One of our objectives is to inform users of their homogeneity characteristics so that the fields are not applied inappropriately in some form of climate model evaluation.

We focus on 20CR's mean fields (i.e., official 20CR product) because they are the most widely applied. However, we acknowledge that homogeneity will vary among 20CR's 56 ensemble members and between ensemble members and the ensemble mean. In Appendix 1, we present results from our every-member analyses of T_a and P for Geneva, Switzerland and Rondonia, Brazil. We found that coincident discontinuities in as few as five ensemble members could lead to a detectable shift in the ensemble mean.

We define a non-climate (i.e., unphysical or artificial) break as a breakpoint in the time series of the mean variable field whose 95 % confidence interval overlaps (for any number of years) with the 95 % confidence interval of a breakpoint in the time series of the corresponding ensemble spread (e.g., P and P spread). Substitute variable spread fields are used for derived variables that have no associated spread field. The meridional wind (v_{grd}) ensemble spread is used to detect non-climate breaks in 10 m windspeed (WSPD); the convective available potential energy (CAPE) ensemble spread is used to detect non-climate breaks in CTP; the 2 m minimum air temperature (TMIN) ensemble

Table 2 Description of the regions used in this study and the number, n , of 2° grid cells they comprise

Region	Acronym	Latitude(°)	Longitude(°)	n
<i>North America</i>				
Northwestern Canada	NWC	50°N–66°N	170°W–100°W	203
Northeastern Canada	NEC	50°N–66°N	100°W–56°W	131
Western United States	WUS	22°N–50°N	128°W–100°W	143
Central United States	CUS	22°N–50°N	100°W–86°W	78
Eastern United States	EUS	22°N–50°N	86°W–60°W	85
<i>South America</i>				
Amazon	AMZ	20°S–22°N	108°W–48°W	273
Eastern Brazil	EBR	24°S–EQ	48°W–34°W	60
Southern South America	SSA	56°S–20°S	76°W–48°W	152
<i>Africa</i>				
Sahara	SAH	16°N–30°N	18°W–70°E	277
Western Africa	WAF	4°N–16°N	20°W–34°E	148
Congo	CON	12°S–4°N	6°E–34°E	92
Eastern Africa	EAF	12°S–16°N	34°E–52°E	85
Southern Africa	SAF	36°S–12°S	10°E–52°E	149
<i>Europe</i>				
Northern Europe	NEU	48°N–72°N	10°W–40°E	213
Mediterranean	MED	30°N–48°N	10°W–40°E	179
<i>Asia</i>				
Northern Asia	NAS	50°N–66°N	40°E–180°E	488
Central Asia	CAS	30°N–50°N	40°E–76°E	180
Tibetan Plateau	TIB	30°N–50°N	76°E–100°E	120
Eastern Asia	EAS	30°N–50°N	100°E–146°E	176
India	IND	6°N–30°N	70°E–92°E	83
Southeast Asia	SEA	18°S–30°N	92°E–156°E	230
<u>Australia</u>	AUS	44°S–18°S	112°E–154°E	183
<i>Ocean domains</i>				
North Atlantic Ocean	NAO	EQ–50°N	60°W–10°W	560
Indian Ocean	IO	40°S–20°N	50°E–100°E	688
South Pacific Ocean	SPO	40°S–20°S	170°E–100°W	450
North Pacific Ocean	NPO	20°N–50°N	150°E–140°W	525
Tropical Pacific Ocean	TPO	10°S–10°N	120°E–80°W	751

As per the shading in Fig. 1, only land grids are considered over the continents; only ocean grids are considered over ocean domains

spread is used to detect non-climate breaks in: C, E, HI, and LCL. Our definition of non-climate breaks may be sensitive to the size of confidence intervals in cases in where the confidence intervals for both fields are relatively

wide. In the case of monthly T_a , confidence intervals range from 3 to 84 years in length, with a median length of 18 years. For monthly spread in T_a the range is similar, although the median length is substantially shorter (11 years).

The underlying rationale for our non-climate definition derives from the fact that the ensemble spread typically varies as an inverse function of the assimilated observation count (Ferguson and Villarini 2012; their Fig. 1). If break dates are coincident between variable mean and spread fields, then the logic follows that observational network changes are likely the discontinuities source. However, over ocean we found that the inverse relationship between T_a spread and assimilated observation count is not always upheld (i.e., Figs. 5b, h, i, 16). While one plausible explanation is that the ensemble is tightly constrained to a time invariant constant by the specified HadISST v1.1 field, it does not explain how there can still be variability in the TMIN spread (Fig. 16). Because its time series appears more realistic, we use TMIN spread in place of T_a spread.

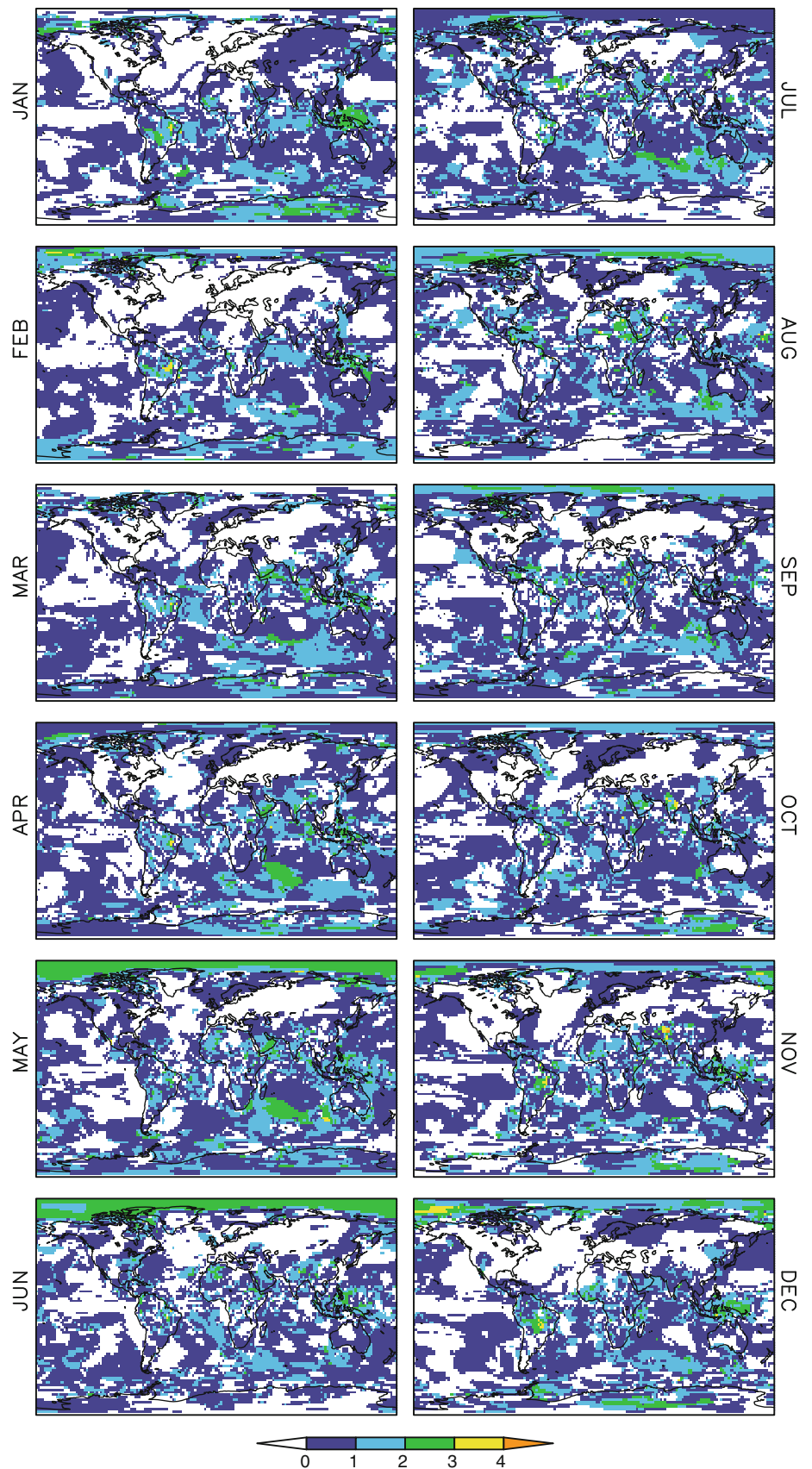
3 Results

3.1 Homogenous fraction and seasonality

Since we first reported evidence of observational shocks in 20CR's record over the central U.S. (Ferguson and Villarini 2012), an open question has been: how pervasive are such effects globally and how do they vary seasonally? In Fig. 2, we present global maps of non-climate breakpoint counts for each monthly time series of T_a . Consistent with earlier work, we find a sizeable seasonal component to 20CR's inhomogeneities (and their detectability), especially in the northern extratropics (Fig. 3). For the period 1871–2010, the global fraction of statistically homogenous grids (grids with natural breaks-only) can be seen to range from 10 % in July and August (28 % in May) to 21 % in February (39 % in January; Fig. 3). Both climate and non-climate inhomogeneities in 20CR's other variable fields track the same general seasonality (i.e., the homogenous fraction peaks during northern hemisphere winter and dips during the northern hemisphere summer; not shown). Because February and July typically constitute the months of maximum and minimum homogeneity, respectively, we chose to focus the remainder of our analysis on them. Relatively greater homogeneity observed in the northern hemisphere (Figs. 2, 3, and S1) might have been anticipated from the hemispheric local anomaly correlation results presented in Compo et al. (2006; their Figs. 7 and 10).

In Fig. 4, the February and July inter-variable and inter-regional differences in the areal extent of non-climate

Fig. 2 Global monthly non-climate breakpoint count in 20CR T_a . For the total breakpoint counts, including both physical and non-physical breaks, see Fig. S1



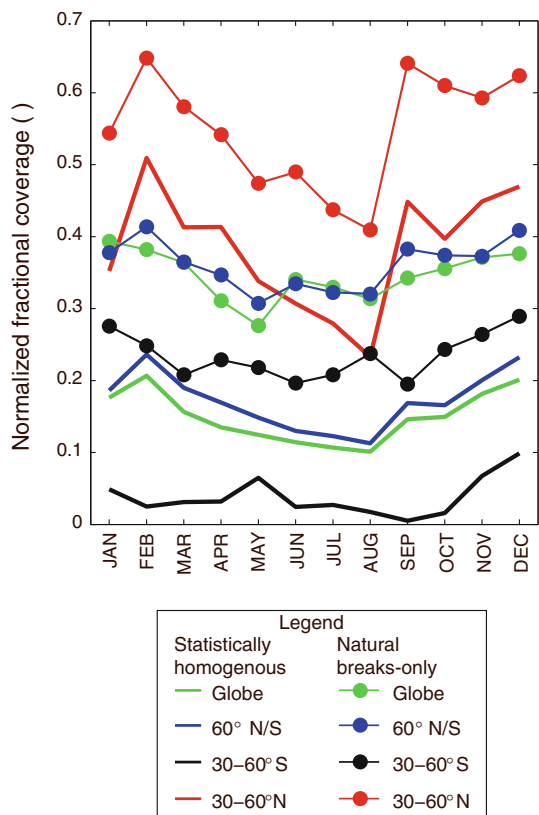


Fig. 3 For the globe, 60°S–60°N, and northern and southern extratropics, the normalized fractional coverage in 20CR T_a that is unaffected by inhomogeneities (i.e., statistically homogenous; solid line) or affected only by changes of climate origin (line with filled circular marker). Note that the difference of the sum of these terms from unity is approximately (because, over the 140-year record, both climate and non-climate breaks can occur at a single point) equal to the fractional area contaminated with non-climate changes

(unphysical) changes are summarized for a 26 variable subset (Table 1) of 20CR over 60°S–60°N in addition to 27 smaller regions (see Table 2). It shows that on-average 46–59 % (February–July) of grids between 60°S and 60°N are affected by non-climate changes, which is less than that for T_a (February: 0.59; July: 0.68) but more than that for P (February: 0.33; July: 0.53). The upward longwave radiation flux ($LW\uparrow$) and 10 m wind speed (WSPD) are the least and most contaminated with artificial shocks, respectively. Their 60°S–60°N affected coverage range from 16 % ($LW\uparrow$) and 72 % (WSPD) in February to 22 % ($LW\uparrow$) and 79 % (WSPD) in July. Overall, Fig. 4b, c can serve as a valuable reference for users looking to isolate regions where long-term trend assessment is currently feasible (or not). Conversely, Fig. S2, which shows the areal extent of climate-related changes is typically between 20 and 23 % of grid points, is valuable for further analysis of climate variability.

According to the multivariate median, Northern Europe (NEU) is the least affected by non-climate breaks, both in

February (<1 %) and July (31 %). In February, north-western Canada (NWC), northeastern Canada (NEC), western U.S. (WUS), central U.S. (CUS), eastern U.S. (EUS), and Mediterranean (MED) are also relatively unaffected. Eastern Africa (EAF; 0.78) and Sahara (SAH; 0.82) are the most affected domains in February and July, respectively. February–July differences in the multivariate median affected area fraction (Fig. 4b, c, bold black line) average 21 %, but range between less than 1 % (Amazon: AMZ; Indian Ocean: IO; and southern Africa: SAF) and as much as 52 % (EUS). The February–July difference is less than 10 % for the following regions: EBR, southern South America (SSA), Congo (CON), EAF, SAF, southeast Asia (SEA), and IO.

In general, these findings hold qualitatively for climate-related changes as well (see Fig. S2). The areal extent of climate-related breakpoints is highest in AMZ, EBR, and Africa (except SAF). Remarkably little area (2–4 %) in Australia (AUS) is affected by changes of climate origin (Fig. S2b, c).

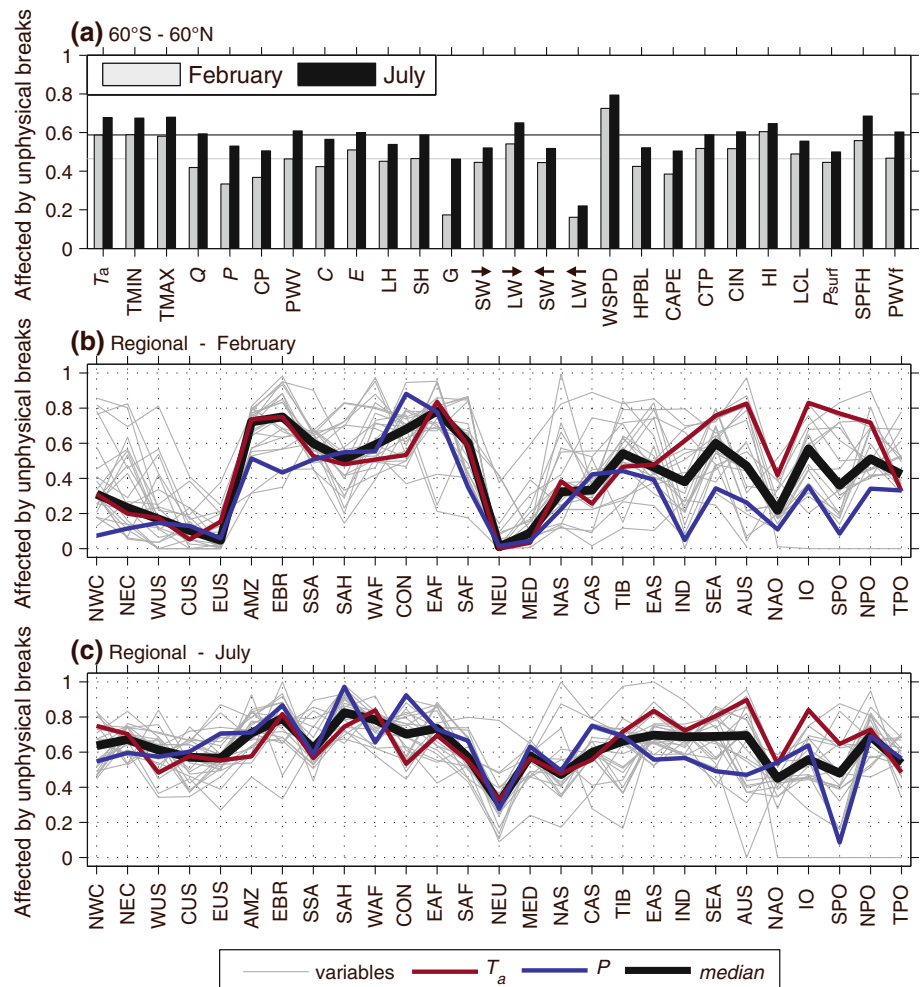
3.2 Breakpoint size distribution and detectability

Considering that 20CR is statistically inhomogenous at the majority of grids (Fig. S1), a key question is: what is the typical jump size associated with these breaks? In Fig. 5, we provide a sampling of eight inhomogenous grid records each for T_a and P from around the world. They are representative of the array of detectable inhomogeneity, ranging from instantaneous (e.g., Fig. 5a) to gradual (e.g., Fig. 5c), and with varying jump sizes. For added reference, the spread time series and breakpoint record in comparison datasets, CRU T_a and GPCC P , are included. Although no breakpoints were detected in GPCC P . Breakpoint summaries for the full 26-variable subset of 20CR (Table 1) at these same grid points are provided in Fig. S3. While the breaks in T_a and P highlighted in Fig. 5 do pervade through multiple (if not most) modeled variables, Fig. S3 shows this is not the rule.

In several instances the abrupt shifts in variable mean correspond with those in the variable ensemble spread (Figs. 5 and S3). Such coincidences are strongly suggestive of unphysical inhomogeneities (Ferguson and Villarini 2012). The fact that only one breakpoint in one location is corroborated by a break (95 % confidence interval; not shown) in CRU T_a (Fig. 5d) further supports this conclusion. Finally, the ocean grids for which the T_a spread is time invariant (Fig. 5b, h, i) are examples of why TMIN spread is used instead for diagnosing non-climate breakpoints (see Sect. 2.3.2).

In Fig. 6, the full distribution of detected breakpoints ($n = 853,376$) for 26 variables and 2 months (February and July) is summarized according to jump size. The jump

Fig. 4 Bai-Perron test results for February and July time series of 26 20CR variables (see Table 1 for variable name definitions). **a** For 60°S–60°N, the fraction of grid cells (land and ocean combined, except for Q and G , which are defined over land only) that are affected by non-climate breaks at some point over the period of availability (1871–2010). **b** and **c** same as in (a) but on a regional basis (see Fig. 1, Table 2). In (a), the multivariate median values for February (gray) and July (black) are marked by horizontal lines. Results for T_a (red), P (blue), and the multivariate median (black) are highlighted in (b) and (c). See Fig. S2 for the complimentary figure (i.e., the fraction of grid cells affected by climate-related breaks)



sizes are normalized by the standard deviation of the preceding series segment to enable inter-variable comparison (Fig. 6a). The variable frequency polygons are generally in close agreement with regards to mean, spread, and skew. There is remarkably strong consensus that the distribution is positively skewed; 72 % of breaks exceed one standard deviation in magnitude; 50 % of breaks exceed 1.3 standard deviations in magnitude; and 25 % of breaks exceed 1.8 standard deviations of the preceding time series. Figure 6b, c shows the absolute jump size distributions for T_a and P , respectively. One-half of breaks are shown to exceed 0.7 °C and 11.9 mm month⁻¹, respectively.

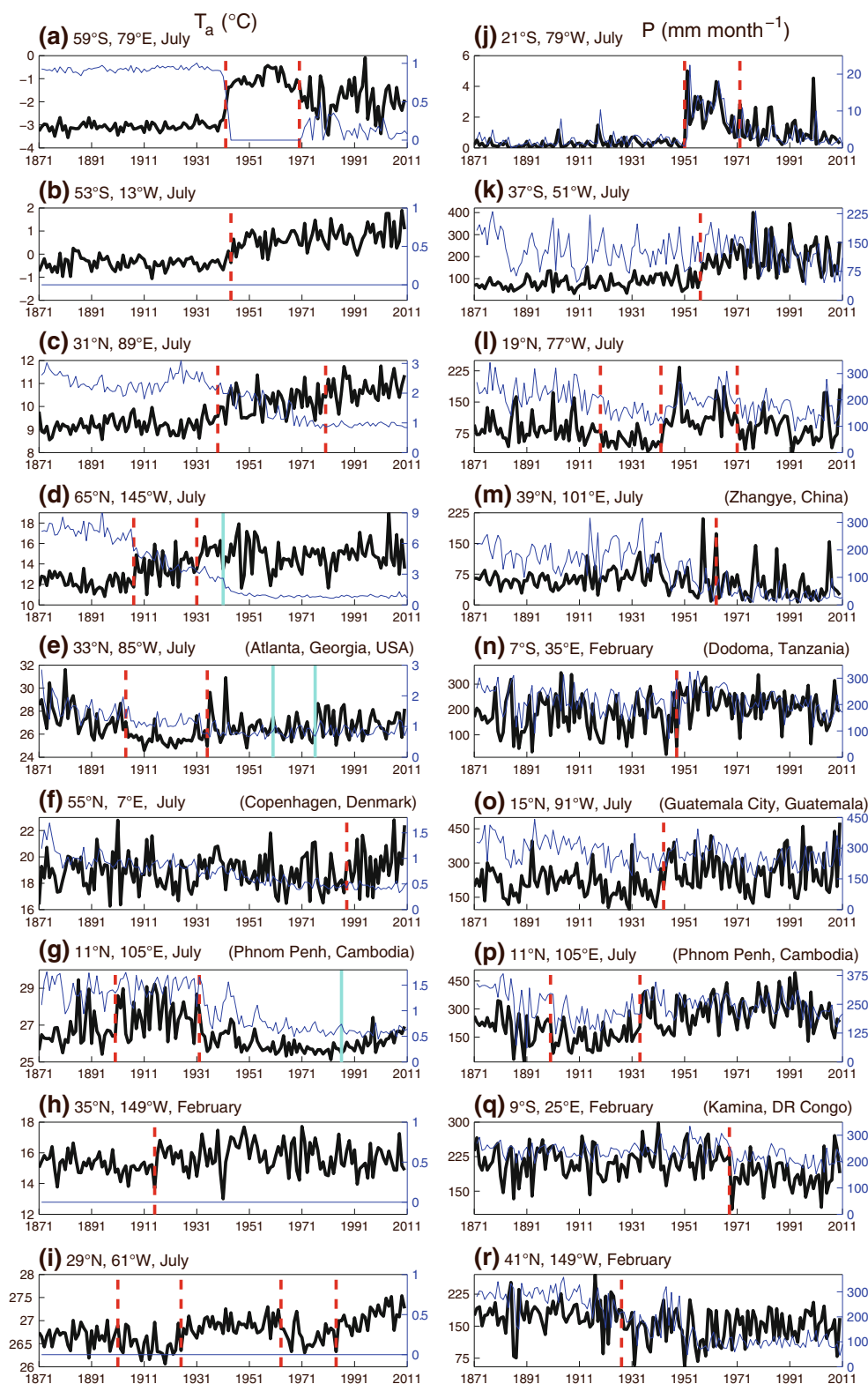
In many applications, especially in an operational setting, knowing the detectability limits is desirable. For the Bai-Perron test, we recommend assuming a detectability limit of 0.7 standard deviations computed from the time series prior to the change point, equivalent to the fifth percentile of multivariate detected jump size (0.3 °C and 1.4 mm month⁻¹ for T_a and P), globally (Fig. 6). It could be the case that test sensitivity exceeds that which is required or meaningful for the application at hand. For example, Figs. 7 and 8 illustrate the monthly global distribution of minimum

detected jump sizes in T_a and P , respectively. The smallest shifts over the Tropics and coastal areas for T_a and deserts of Africa for P might be inconsequential. Notably, we found no substantial difference in the mean jump sizes of non-climate breaks related to observational network changes (next section) and natural breaks related to climate variability.

3.3 Non-climate breakpoints

Non-climate (unphysical) inhomogeneities are diagnosed using the joint confidence intervals of breaks detected in the variable and spread fields (Sect. 2.3.2). Their fraction of the total February and July breakpoint counts is summarized in Fig. 9. As before, a 26-variable subset of 20CR is considered over 60°S–60°N (Fig. 9a) and 27 constituent climatic regions (Fig. 9b, c). The multivariate mean non-climate fraction for 60°S–60°N found to be approximately 0.72 (for both February and July), which slightly exceeds that of T_a (February: 0.70; July: 0.64) but not P (February: 0.80; July: 0.82). Non-climate breaks constitute the least proportion (0.20) of breaks in surface upward longwave radiation (LW \uparrow). Regionally, the largest non-climate fractions (0.95)

Fig. 5 For selected grid points, the time series of 20CR **a–i** T_a and **j–r** P (in black) and their respective ensemble spread (in blue). Vertical dashed red lines denote detected breaks in the variable mean fields (confidence intervals not shown). CRU T_a and GPCC P records, available over land points (**c–e, g, and m–q**), were also evaluated for breaks. Breaks detected in the CRU T_a are denoted by vertical cyan lines (confidence intervals not shown); no breaks were detected in GPCC P over the selected grid points. The nearest major city is noted, within reason



are reported for CAS (February) and Australia (AUS; February and July). The smallest non-climate fraction (0.60) is found for tropical Pacific Ocean (TPO; February) and North Atlantic Ocean (NAO; July). Neither the 60°S–60°N results

nor the domain results exhibit seasonality in their multivariate median non-climate fraction (Fig. 9b, c; except in the case of CUS and EUS, for which the February breakpoint population size was insufficient).

Figure 10 details the spatio-temporal distribution of non-climate breaks. For 60°S–60°N, the inter-variable range in median non-climate break dates is 1924–1959 for February and 1936–1950 for July; the multivariate mean of median non-climate break dates is 1947 and 1944 for February and July, respectively (Fig. 10a). Ninety percent of all non-climate breaks between 60°S and 60°N (both February and July) are detected prior to 1979, when modern satellite-era atmospheric reanalyses such as MERRA and NCEP’s Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) begin. Figure 10b gives the February and July multivariate median: 10th percentile, mean and 90th percentile non-climate break date for each of the 27 climatic regions considered. The means of these values, taken over all domains, are: 1904, 1934/1939 (February/July), and 1967, respectively. In Fig. 10c, the area-normalized multivariate median non-climate breakpoint count is plotted. By multiplying this value by the number of 2° grids in the domain (see Table 2), the actual median count can be computed. Because the 60°S–60°N non-climate fraction is insensitive to seasonality (Fig. 9a), we expect the non-climate break count to scale linearly with the count of all breaks, which it does. On average, 0.21 domain areas more non-climate breaks are detected in July (mean = 0.82) as compared to February (mean = 0.61) (Fig. 10c).

The box plots and bulk statistics of Figs. 9 and 10 can only go so far towards isolating the inaccuracies of 20CR. Using Fig. 11 it is possible to visually pin-down time windows for each region that deserve greater scrutiny. It illustrates the full 140-year time series of physical and non-climate breaks in T_a for February or July- whichever is least homogenous. Red shading indicates the times when

breaks are mostly of non-climate origin. Hollow black bars, on the other hand, denote times when natural breaks dominate. A great example of the robustness of our approach is the 1976–1977 climate shift of the North Pacific basin (e.g., Meehl et al. 2009; Powell and Xu 2011), which is properly diagnosed as real (Fig. 11aa).

The results for CUS contrast-but do not completely contradict- our previous assertion that the mid 1940’s breakpoint is unphysical (Ferguson and Villarini, 2012). In this study, a total of 21 breaks are detected in the 1940’s (of which all occur in 1949) and only three are diagnosed as non-climate (Fig. S4e). Accordingly, the mid-1940’s break appears to have competing observational network and climate explanations. In more general terms, this case highlights sensitivity to the choice of statistical test. Recall that previously we applied a hands-on segmented test to the ensemble spread field while we are previously presently applying an automated Bai-Perron test.

It is important to point out that the ratio of non-climate breaks to total breaks in Fig. 11, as well as the absolute count, is inconsistent with previous results (Fig. 10). That is because a different accounting convention was applied. In Fig. 11, breakpoints contribute to the tally in every year of their 95 % confidence interval. For example, if the year lies within the joint confidence interval of breakpoints in both the T_a mean and TMIN spread, then the non-climate breakpoint count is increased by one. Alternatively, if the year lies within the confidence interval of a breakpoint in T_a mean, but not for TMIN spread, then the physical breakpoint count is increased by one. While it is true that the breakpoint confidence intervals can be very wide (see Sect. 2.3.2), accounting for their lengths is the only way to achieve the true uncertainties inherent to the breakpoint

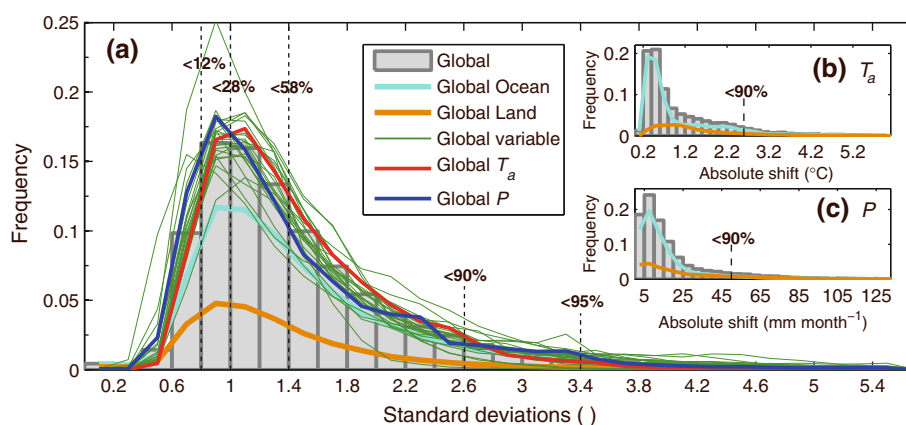


Fig. 6 For the globe (90°S–90°N), the normalized frequency polygons of breakpoint jump size in (a) units of standard deviations, and for b T_a and c P , as an absolute quantity. The gray bars in (a) characterize the set of all detected breakpoints (i.e., physical and non-climate breaks for February and July) in 26 variable mean fields ($n = 853,376$). Each separate variable polygon (green, blue, and red

lines) is normalized by its own respective count total; the global ocean and global land polygons are normalized by the total global count. Annotated percentiles correspond with the multivariate global (gray bars) polygon only. Polygons have a bin size of 0.2 in (a–b) and 5.0 in (c). The polygon for surface pressure is the most peaked with 47 % of breaks less than one standard deviation in size

Fig. 7 For 20CR T_a , the global monthly distribution of minimum detected jump sizes (1°C)

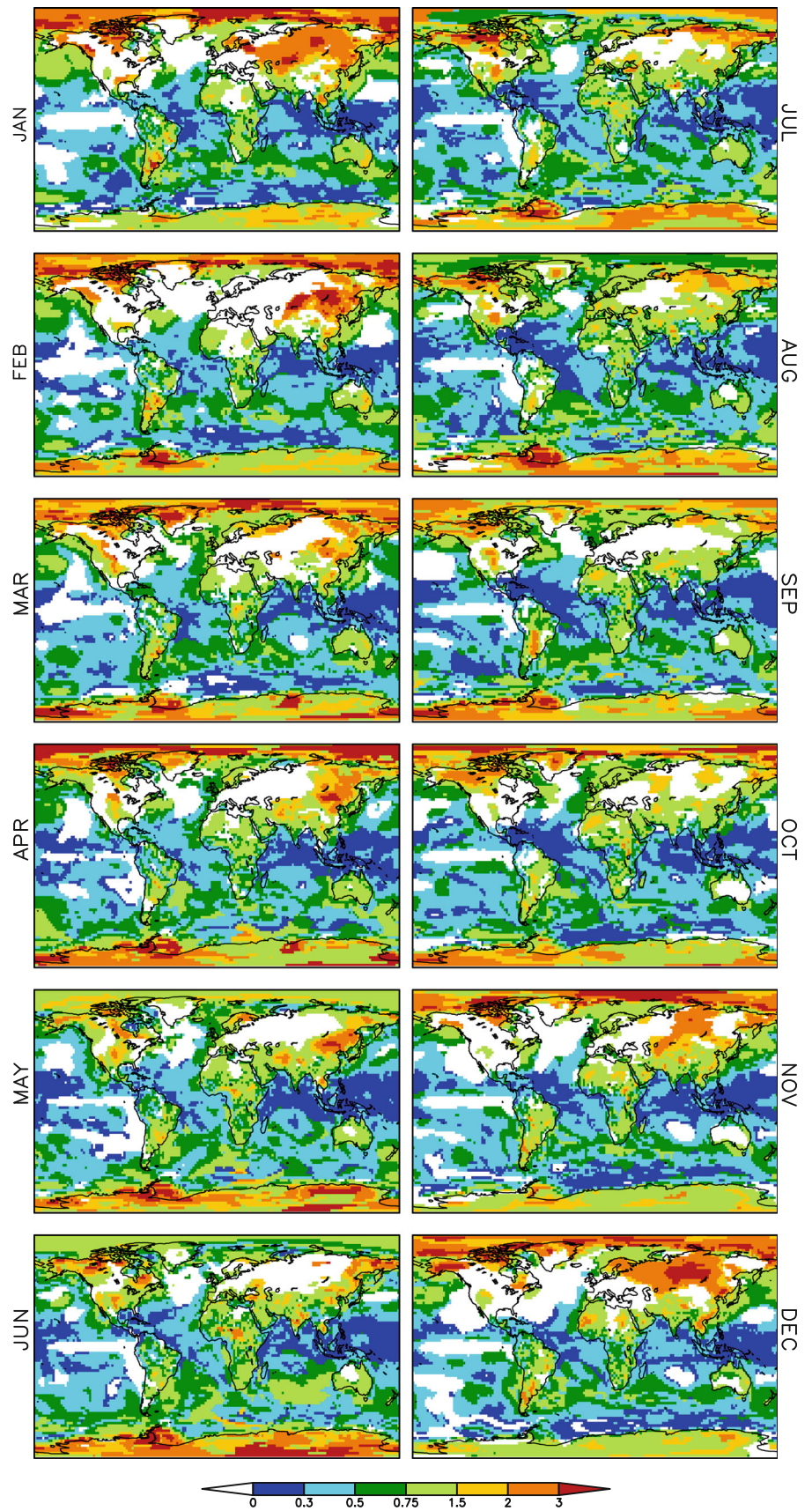


Fig. 8 As in Fig. 6, for 20CR
 P (lmm month^{-1})

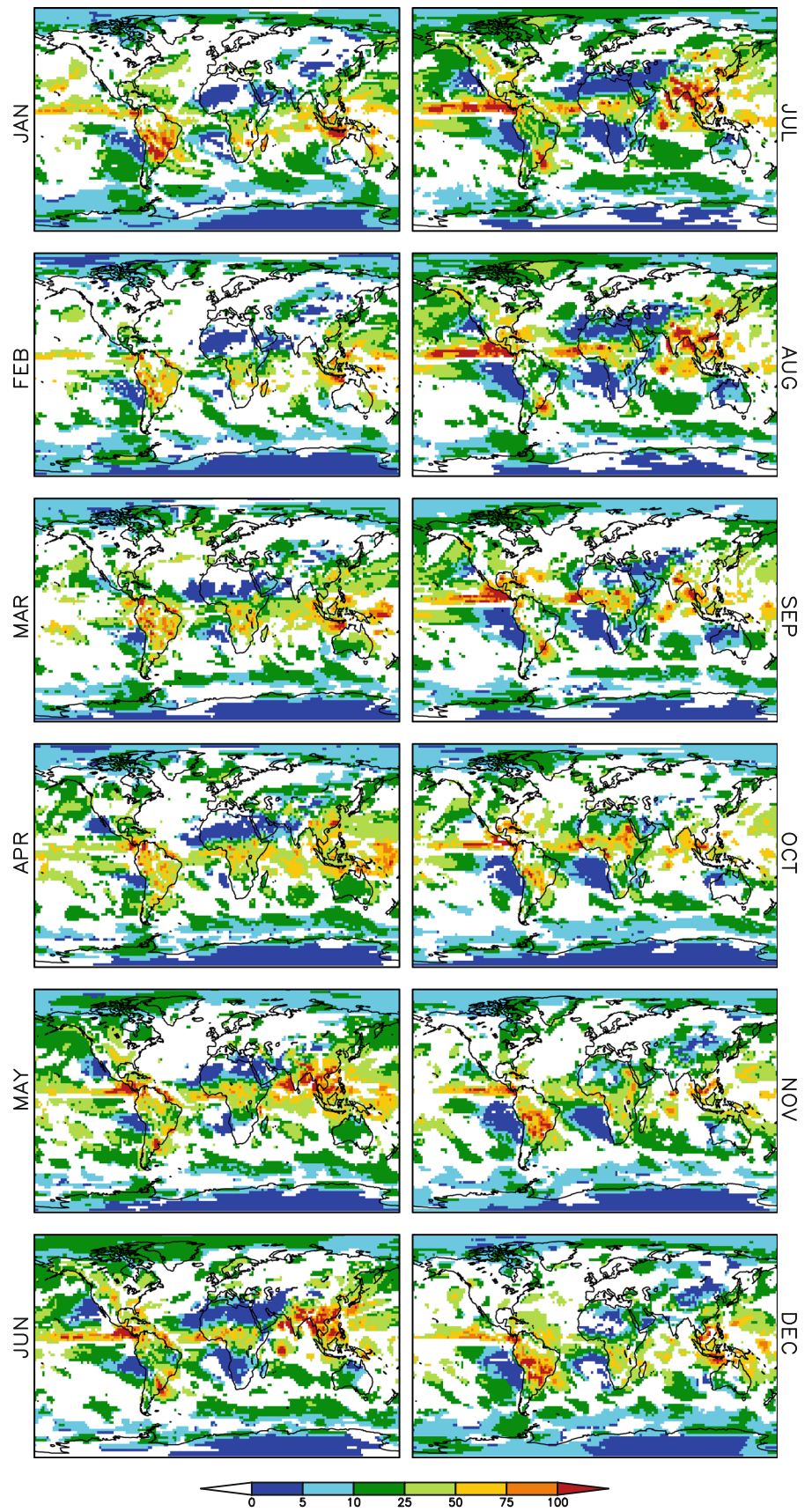
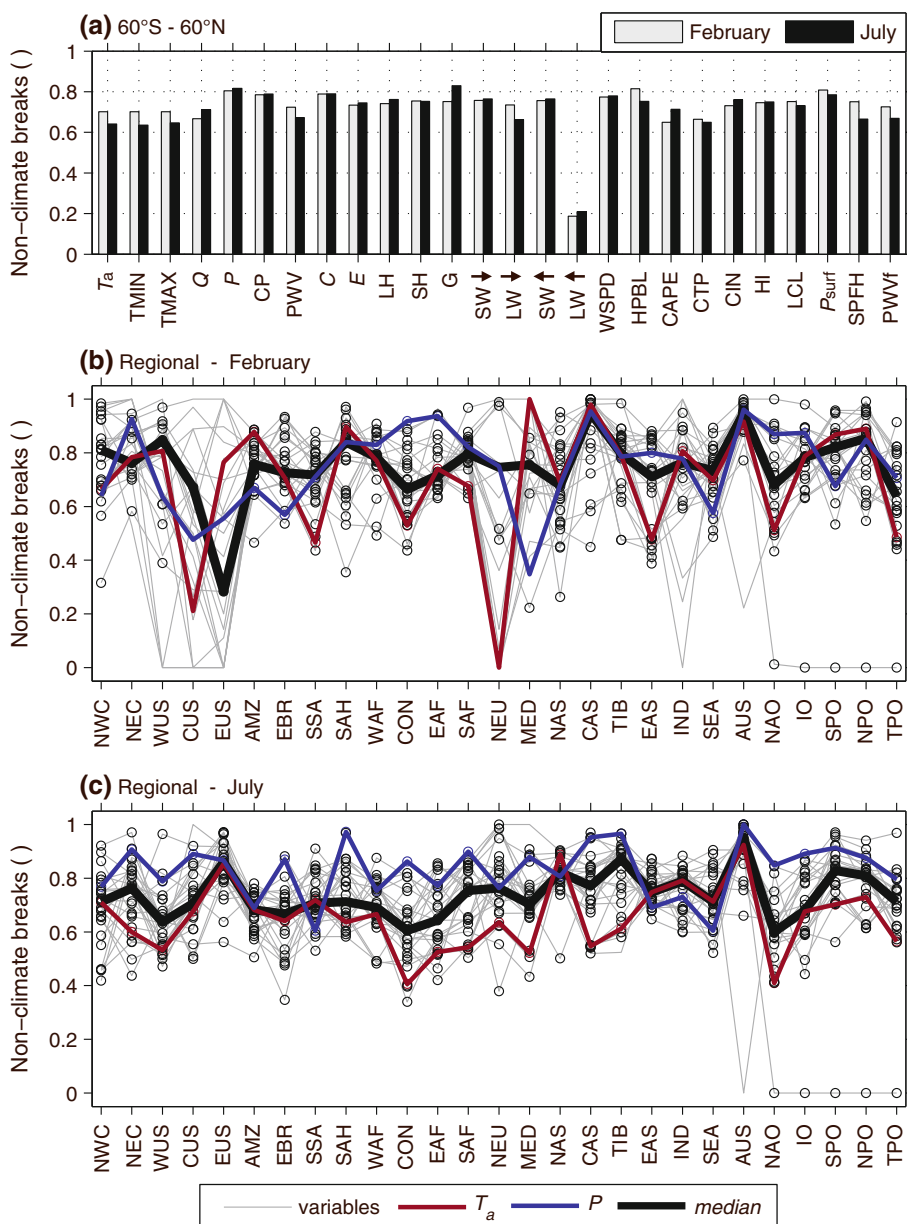


Fig. 9 Fraction of all detected breaks in the February and July time series that can be attributed to non-climate (i.e., observational network) sources. **a** For 60°S–60°N (land and ocean combined, except for *Q* and *G*, which are defined over land only), and for **b, c** each of 27 land- and ocean-only climatic regions. The multivariate regional medians [bolded black line in (b) and (c)] are computed from the set of all single-variable values with underlying sample sizes of 50 or more non-climate breaks in the specific region (denoted by a circle)



detection. The convention of this study has been and remains to be that of constraining all accounting to the year of the central break date (see Fig. S4). The merit of Fig. 11 is that it informs the precise era of overlap between the confidence intervals of the mean and ensemble spread fields (see Fig. S5 for further details).

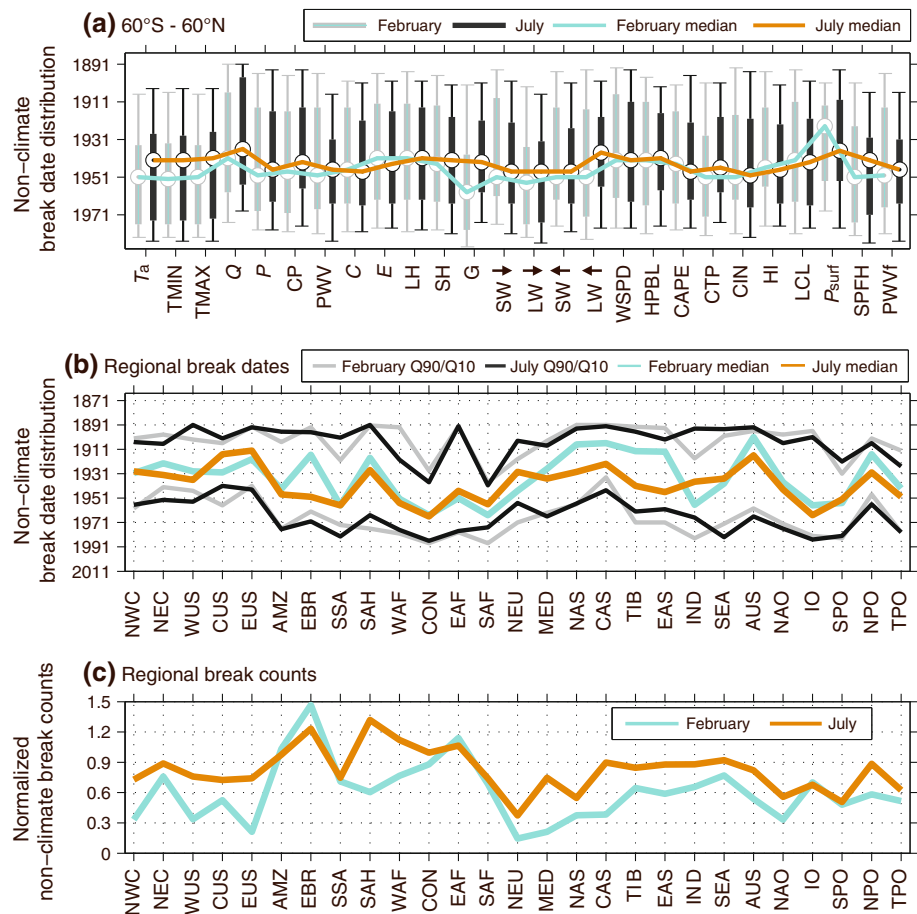
Break test results from comparison datasets are also included in Fig. 11. Over land, physical breakpoint count results from CRU T_a are plotted. Because there is no equivalent spread field for CRU, breaks in its CDD contributing station time series are used to diagnose non-climate effects. Over ocean, counts of triple coincidence among breakpoint confidence intervals in HadISST v1.1, ERSST, and COBE SST, are plotted as best estimate

physical breakpoint counts. Assuming 20CR is skillful, we would expect its physical breaks (hollow black bars) to correspond closely with those of the comparison series. However, tight correspondence only really occurs for MED (Fig. 11p) and NAO (Fig. 11x). The reality, as we shall discuss in the next section, is that these datasets come with their own uncertainties and artificial inhomogeneities, as well.

3.4 Attributing 20CR’s breakpoints

The non-climate shifts identified in this study are more than likely the lowest-hanging fruit per se. We believe that the number of natural (non-climate) breaks is much lower

Fig. 10 Boxplot summary of temporal patterns in non-climate breakpoints for **a** 60°S–60°N (land and ocean combined, except for Q and G , which are defined over land only) and **b** for each of 27 land- and ocean-only regions. **c** The area-normalized (i.e., by the contributing grid area, n , provided in Table 2) multivariate median non-climate breakpoint count. In **(a)** the *boxplot* whiskers bracket the 10th and 90th percentiles; the *circled dot* denotes the median. In **(b)**, the multivariate regional median (*cyan, orange*) and 10th and 90th percentiles (*gray, black*) are all medians of the set of like (i.e., median, Q10, and Q90) single-variable statistics supported by underlying sample sizes of 50 or more non-climate breaks in the respective region. Thus, each inclusive variable is given equal weight



(higher) in reality due to remote (in time and/or space) network effects that are not well captured by the local variable spread upon which our diagnosis is entirely dependent. The teleconnections that translate these signals also carry important implications for homogenization. Namely, that correcting for a single break can have reverberating (and perhaps, unintended) consequences (i.e., either eliminating or giving rise to secondary breaks).

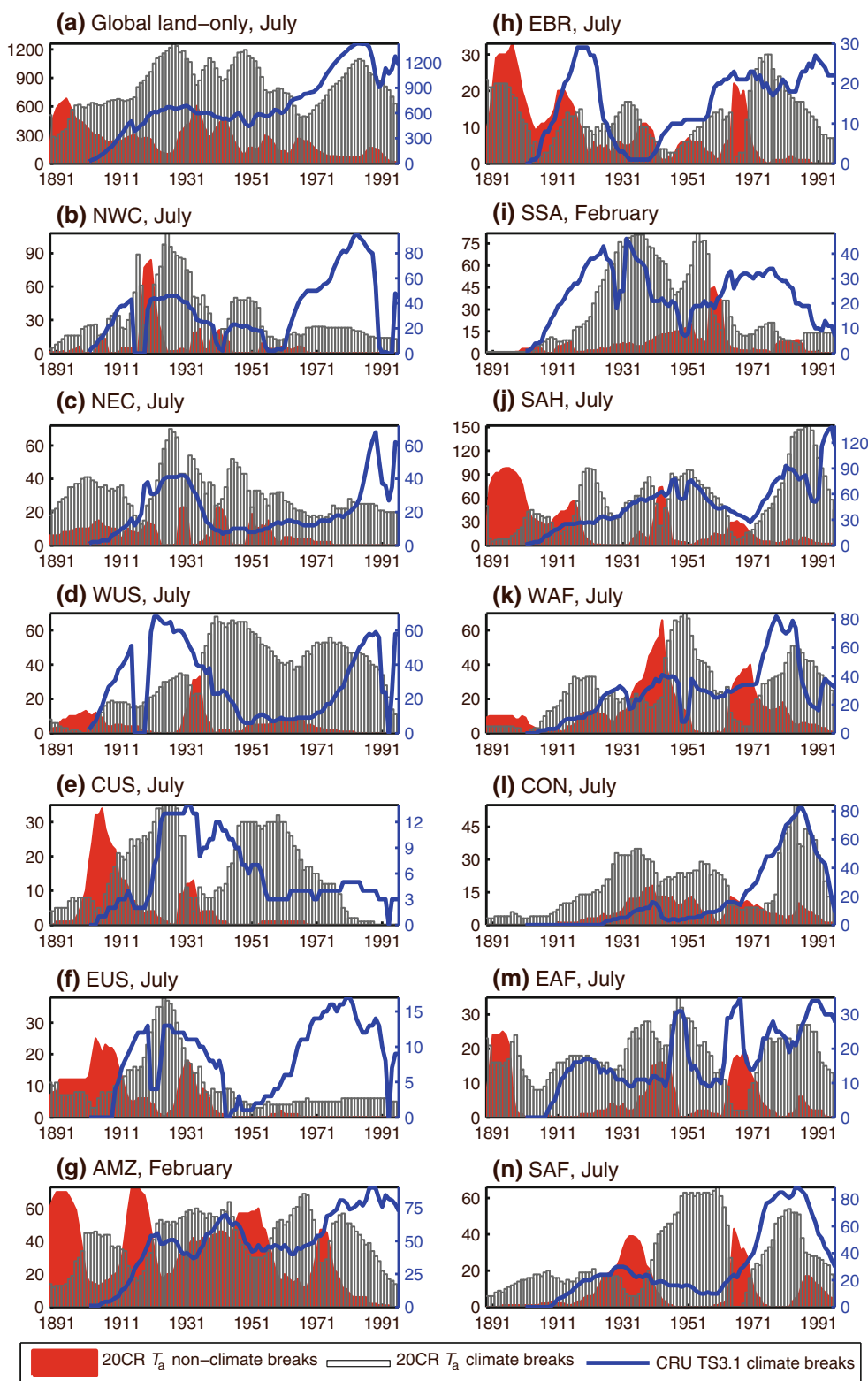
Potential sources of discontinuity include: climate events (e.g., El Niño, severe and extended drought, volcanic eruptions, loss of permanent land/sea ice) and climate change (natural or anthropogenic), model or observational bias, discontinuities in other model data (e.g., specification of greenhouse gas concentrations, volcanic aerosols, ozone, land surface conditions), reanalysis production in multiple streams, and technical mistakes in production. In the case of 20CR, stream production (described in Sect. 2.1) has been shown to affect the continuity of only slow-varying parameters (i.e., integrated column soil water content) at high latitudes (personal comm., Justin Sheffield 2012) that are not the focus of this work.

A more definitive attribution of 20CR's inhomogeneities than we have provided here will require an immense effort moving forward. But it is a necessary hurdle in the

development path towards a climate-quality reanalysis. Specifically, the homogeneity of the observational base for 20CR, which includes HadISST v1.1 and synoptic sea level pressure observations, will need to be reassessed. Long-term independent in situ datasets, of which CRU T_a and GPCC P are primary examples, can assist in verifying realism. The difficulty is that each of the data products has their own problems and uncertainties, which are not very well understood. Ultimately, the objective, as depicted in Fig. 12, is to maintain climate variability (Fig. 12a, c, e, g) in the process of correcting for non-climate breaks (Fig. 12b, d, f, h).

Figure 13 frames the inhomogeneity of HadISST v1.1, HadSLP2, CRU T_a , and GPCC P into perspective with that of 20CR's (dataset descriptions in Sect. 2). It shows the global breakpoint count map for each of the datasets at their native spatial resolution, computed from their annual mean series. For the purpose of inter-comparison, the total break counts between 60°S and 60°N from the similar analysis conducted at 2° resolution are denoted on each subplot. An abundance of breakpoints are detected in CRU T_a , HadSLP2, and HadISST v1.1, whereas GPCC P is found to be relatively homogenous. GPCC P has a homogenous fraction of 0.51 between 60°S and 60°N

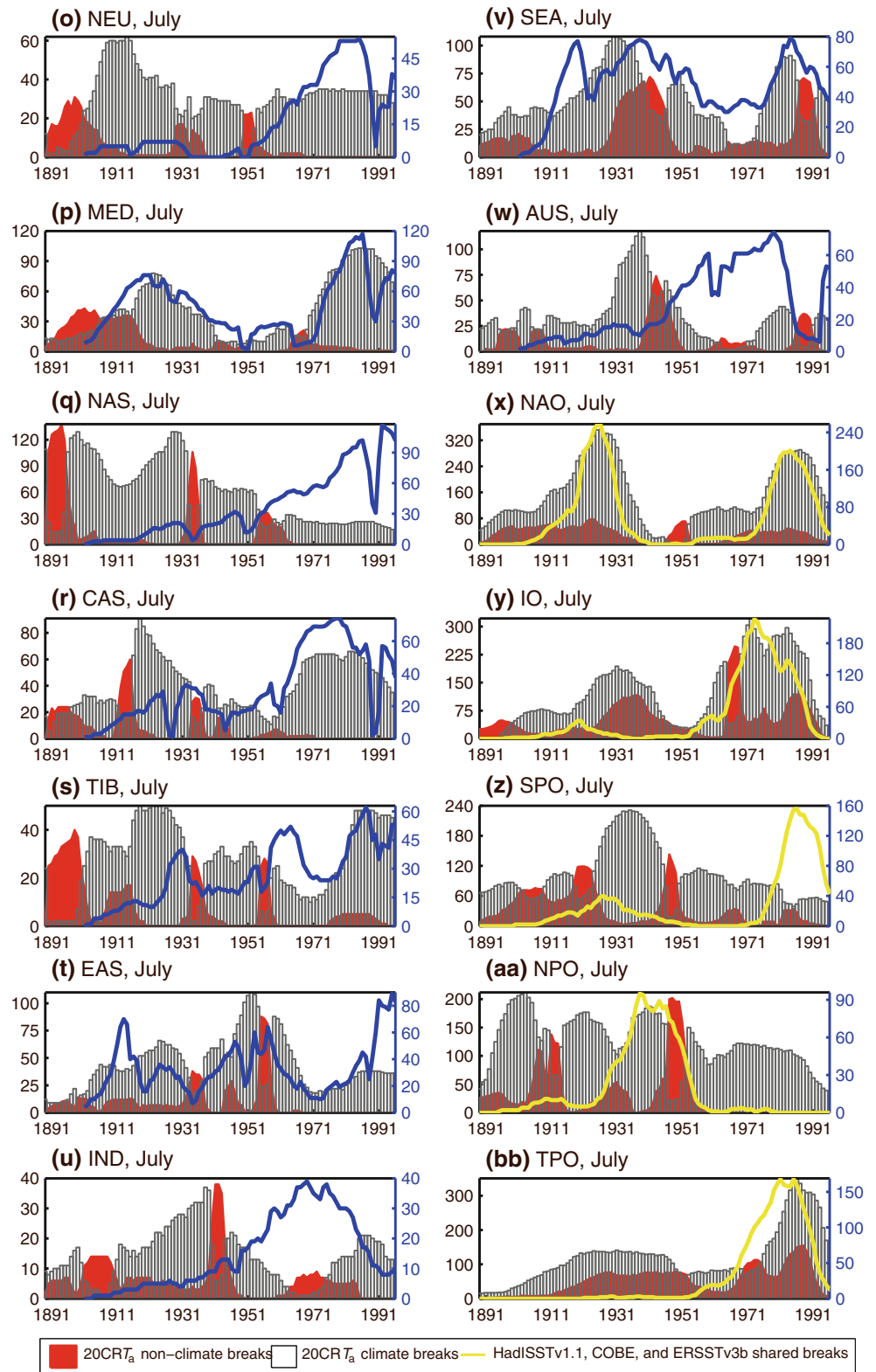
Fig. 11 For (a) global land (excluding Greenland and Antarctica) and (b-bb) each of the 27 study regions (Table 2), the (hollow black bars) grid count within the 95 % confidence interval (CI) of a breakpoint in T_a , the (red shaded area) grid count within the 95 % CI's of breakpoints in both T_a and TMIN ensemble spread, and in (x-bb), the (secondary yellow y-axis) grid count within the triple 95 % CI of the following three SST datasets: HadISST v1.1, ERSST v3b, and COBE SST. In (a-w), the number of grids within the 95 % CI of a physical break in CRU T_a is plotted on the secondary (blue) y-axis



(land-only), compared to only 0.19 for 20CR P . Moreover, only 15 % of its coverage between 60°S and 60°N is affected by multiple breakpoints. The same statistic is 46 % for 20CR P .

The fact that so many inhomogeneities persist without definitive attribution, especially in the HadISST v1.1 and HadSLP2, is reason for concern. These datasets constitute critical elements of climate modeling. All model-based

Fig. 11 continued



reanalyses as well as climate model integrations rely on similar SST/sea ice estimates for boundary forcing. Surface (sea-level) pressure observations inform time-variations in the total mass of the atmospheric column, which the 20CR

has demonstrated is sufficient information for producing a skillful reanalysis; they also comprise the earliest meteorological records. The knock-on effect of inhomogeneities in HadISST v1.1 and HadSLP2 to 20CR is apparent in

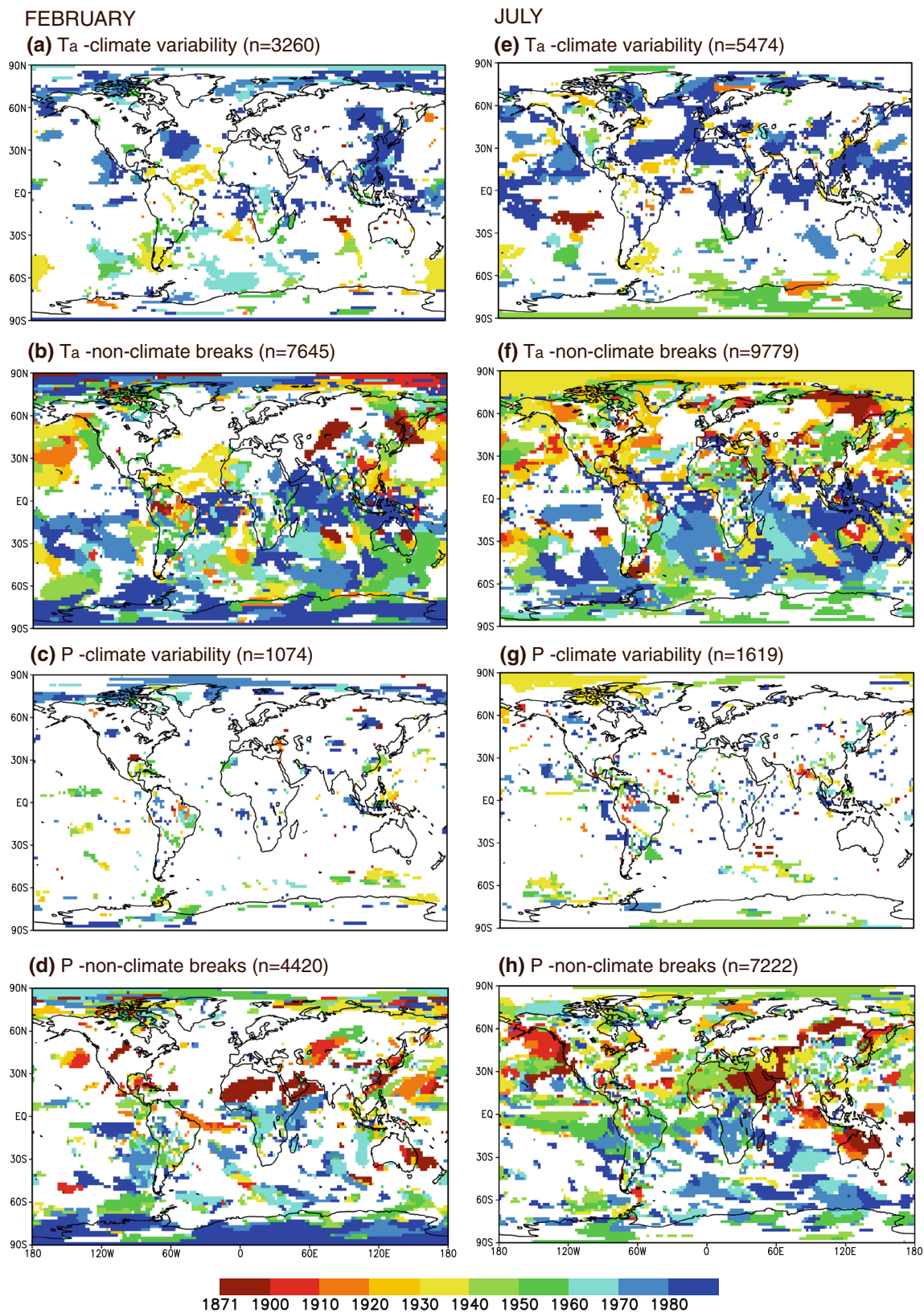


Fig. 12 Date of the most modern physical [(a, c, e), and (g)] and non-climate [(b, d, f), and (h)] breaks detected in the February and July time series of (a–d) T_a and (e–h) P . For each subplot, the total count

of all breaks (not only the most recent) detected over the region 60°S–60°N (land and ocean) is annotated

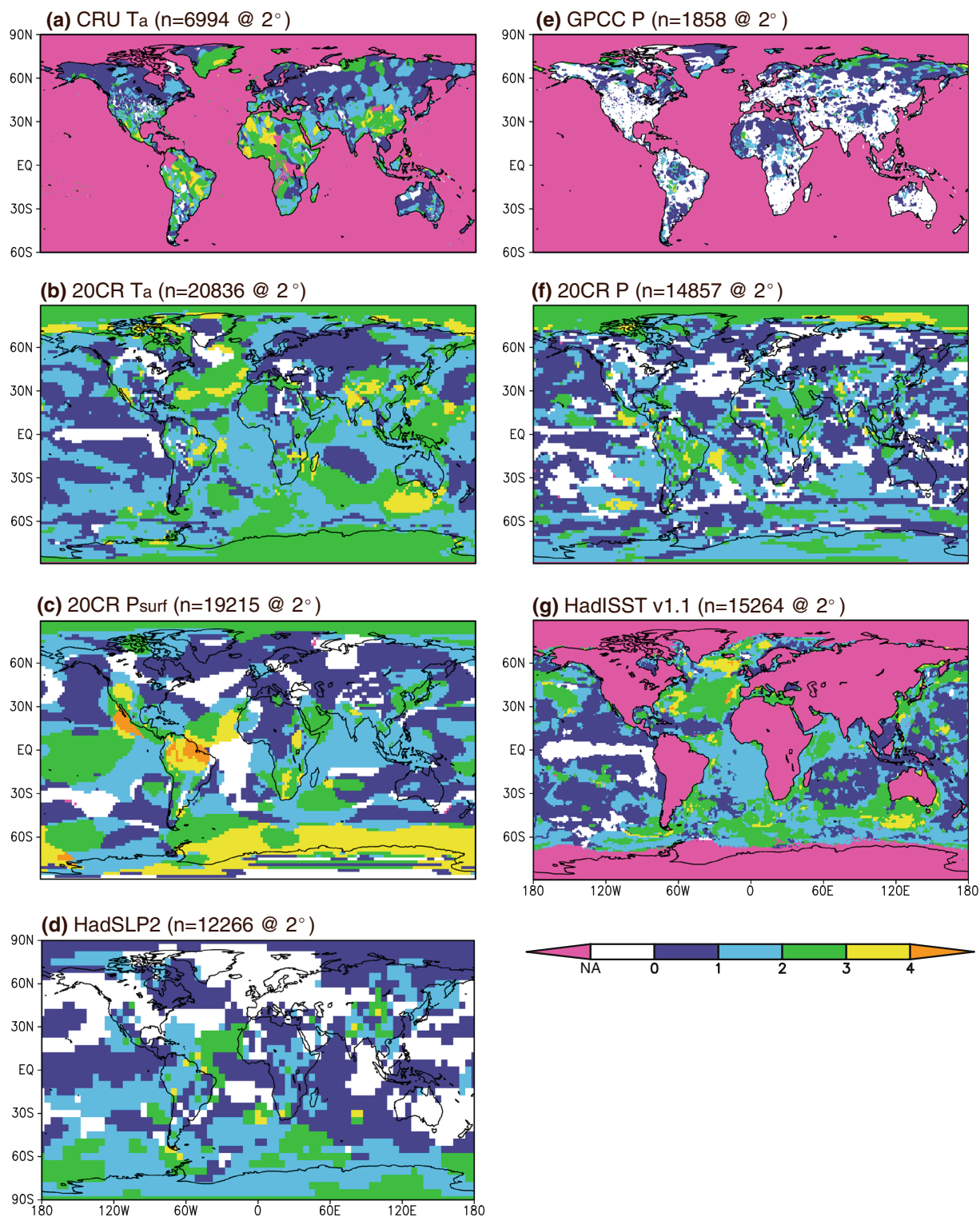


Fig. 13 Global breakpoint count maps for the yearly mean **a** CRU T_a , **b** 20CR T_a , **c** 20CR P_{surf} , **d** HadSLP2, **e** GPCC P , **f** 20CR P , and **g** HadISST v1.1. The analyses were conducted at the native product resolution for either 20CR's period of availability (1871–2010) or the datasets period of availability, whichever is shorter (i.e., CRU: 1901–2009; GPCC: 1901–2010; HadSLP2: 1871–2004; HadISST v1.1: 1871–2010). Grids in white are homogenous (at the 5 % significance level) for the period of record. Grids shaded in pink

generally denotes a coverage gap, however, it can also indicate permanent sea ice cover in (g), and in (a), the fact that climatology was imposed over a substantial portion of the record. The breakpoint counts provided in each subplot title are the results of a similar analysis (not shown) conducted over 60°S–60°N (land and ocean grids) at a standard spatial resolution of 2° and thus can be directly compared

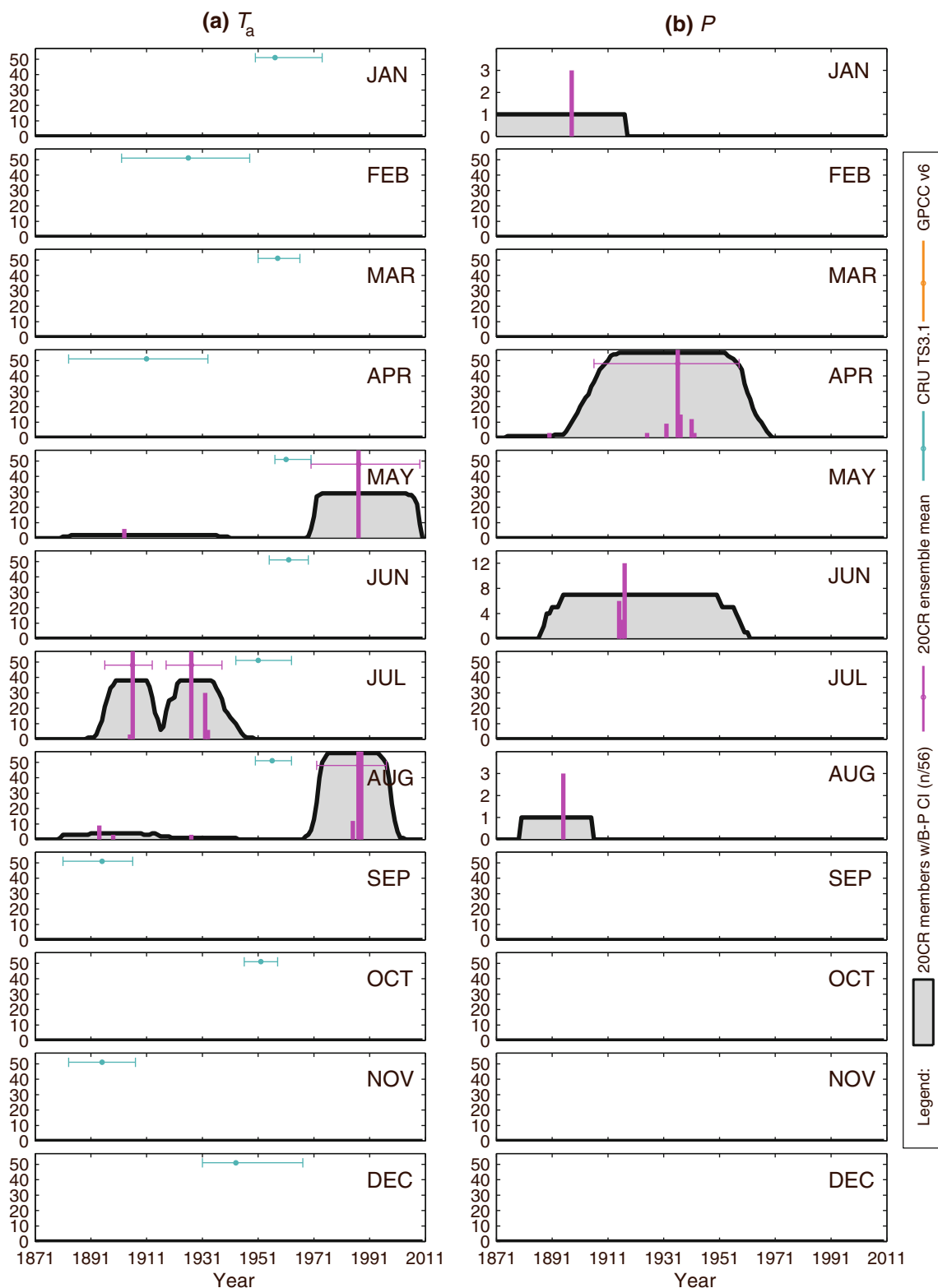


Fig. 14 Summary of the Bai-Perron test results for 20CR **a** T_a and **b** P at Geneva, Switzerland. The *horizontal magenta lines* at the top of each subpanel (where detected) are the change points and associated 95 % confidence intervals computed from the 20CR ensemble mean field. The *gray-filled area* is the histogram of cumulative confidence intervals summed over 20CR’s 56 ensemble members. *Vertical*

magenta bars show the number of ensemble members for which a change point is detected, multiplied by three for emphasis. Results from CRU T_a (*blue*) and GPCC P (*orange*) are also included (where detected). Notice that no change points are detected for the latter. Also, note that the vertical scale in (b) varies

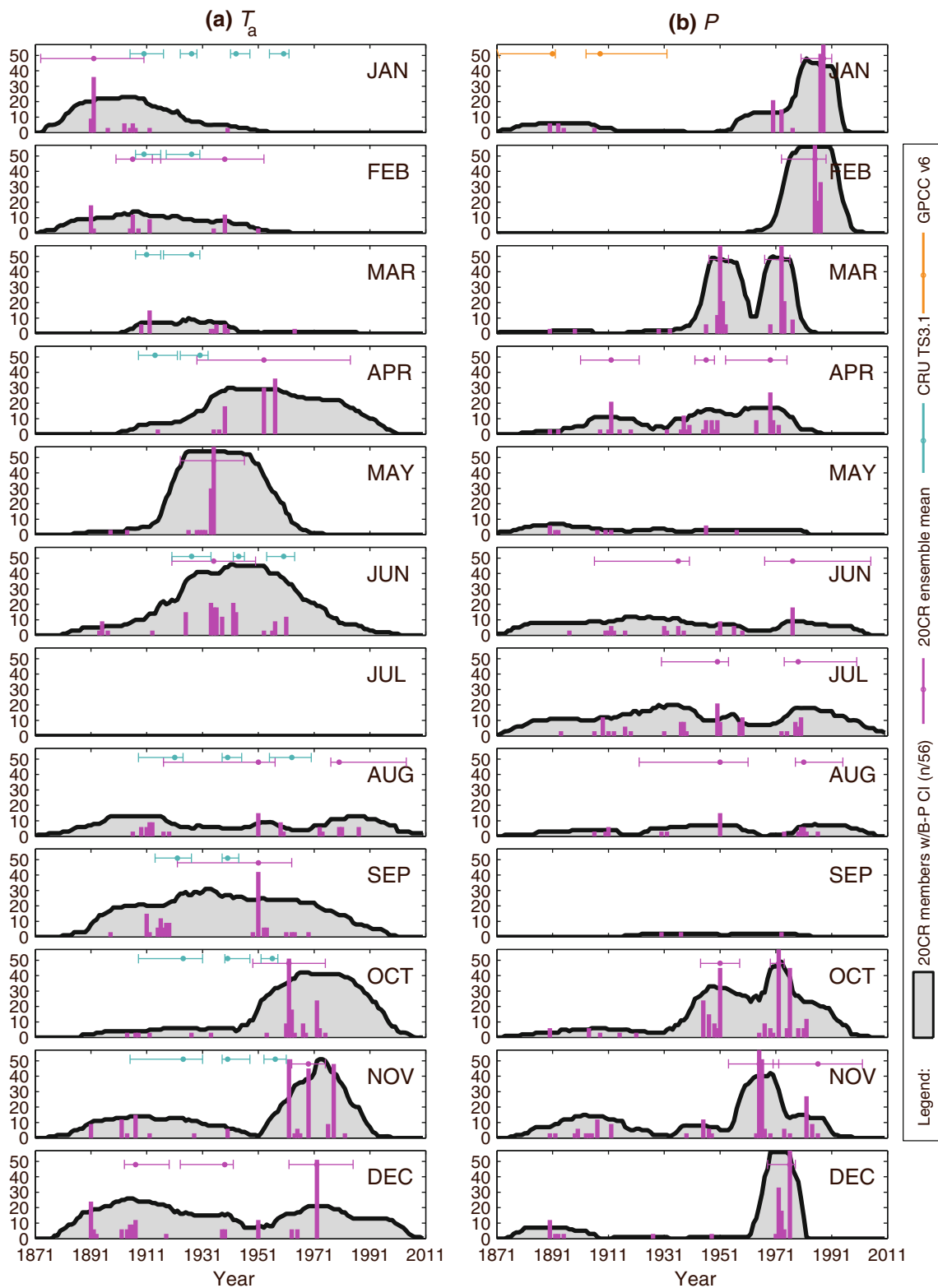


Fig. 15 As in Fig. 14 but for Rondonia, Brazil. The tripling of cumulative change point counts for visual effect places the representative *magenta bars* outside plotting extents in May 1936 (38/56 ensemble members) for T_a and the following cases for P : January 1988 (26/56 ensemble members); February 1985 (38/56 ensemble

members); March 1951 (33/56 ensemble members) and 1973 (37/56 ensemble members); October 1972 (24/56 ensemble members); November 1965 (19/56 ensemble members); December 1976 (36/56 ensemble members)

Fig. 13. The weighted pattern (Pearson) correlation coefficient between breakpoint count maps of 20CR T_a (P_{surf}) and HadISST v1.1 (HadSLP2) is 0.51 (0.36). The bottom line is that attributing inhomogeneities in these input datasets is an essential prerequisite to attributing inhomogeneities in 20CR itself.

4 Summary and conclusions

Using the Bai-Perron test for structural breakpoints we have shown that the 20CR is affected by artificial “shocks” to an extent that varies on a regional, seasonal, and parameter basis. On a grid point basis, the occurrence of multiple abrupt shifts over the 140-year (1871–2010) record is common. Collectively, the scale order of these abrupt shifts can appear overwhelming; 20,836 change points were detected in 20CR’s yearly 2 m air temperature record between 60°S and 60°N (Fig. 13).

The most important task is differentiating between breaks due to natural climate variability and breaks that are caused by observational network changes. We use the joint confidence intervals of breaks detected in the variable and spread fields to diagnose non-climate (unphysical) shifts, which we demonstrate account for approximately 72 % of all breaks (Sect. 3.3). In reality, the proportion could be even greater due to remote network effects (i.e., via teleconnections). The absolute jump size of the breaks can be sizeable. Seventy-two percent of breakpoints exceed one standard deviation of the preceding series segment, 50 % of breaks exceed 1.3 standard deviations of the preceding series segment, and 25 % of breaks exceed 1.8 standard deviations of the preceding series segment (Sect. 3.2).

On a positive note, a significant fraction of points do exist for which the full record is homogenous (Figs. 2, 3, 4, S2). And of the inhomogeneous records, often the last (most modern) homogenous segment extends back to a very early date. For example, July 2 m air temperature is statistically homogenous (free from non-climate breaks) back to 1942 (1922) for half of all land grids (excluding Greenland and Antarctica; not shown). In some cases, it is possible for the record to be considered natural (or even homogenous) over even longer intervals. This can occur when the jump size of the detected shift is deemed inconsequential to the application at hand (i.e., the change point is disregarded). Accordingly, geographically- and temporally-selective applications of the 20CR for long-term trend analysis are feasible.

The longest pre-existing global atmospheric reanalysis, NCEP-R1 (Kalnay et al. 1996; Kistler et al. 2001), does not even begin until 1948. Hence, and this should be appreciated, the 20CR constitutes a major improvement in series

continuity (spatially and temporally) relative to the status quo. Relative to long-term in situ gridded datasets over land that span the twentieth century, 20CR’s degree of homogeneity is comparable to that of CRU 2 m air temperature, but less than that of GPCC precipitation (Fig. 13).

The manifestation of 20CR’s inhomogeneities in time and space, among variables and atmospheric levels, is shown to be highly complex—much like the reanalysis system from which they were generated (e.g., Fig. S3). This makes attributing inhomogeneities to their sources a challenge. Inhomogeneities in 20CR’s boundary forcing (HadISST v1.1) and input data stream (represented by HadSLP2), as well as the comparison datasets (CRU T_a and GPCC P), make the process even more difficult. Specifically, we found that abrupt shifts in 2 m air temperature and surface pressure are related to coincident shifts in HadISST v1.1 and HadSLP2, respectively.

Currently, the presence of inhomogeneities confounds the detection and attribution of possible regional climate change signals in 20CR. Our hope is that the results of this work will serve as a valuable resource to 20CR’s broad user group as well as the developers of climate-quality reanalyses such as the planned Sparse Input Reanalysis for Climate Applications (SIRCA; Compo et al. 2012) to span 1850–2014. With sufficiently detailed metadata it is possible to track (to an extent) the propagation of artificial (observational) shocks in the record using the Bai-Perron test (as demonstrated herein) and subsequently correct the affected time series. However, attributing abrupt shifts away from their immediate source (i.e., teleconnections from sea to land) remains a challenge and will require a different approach than we have taken. Another difficulty is diagnosing shifts for which there are competing observational network and climate explanations. The many challenges to attribution suggest that an automated procedure is likely to be insufficient, while the sheer number of inhomogeneities all but necessitates one.

Acknowledgments We would like to thank Gilbert Compo and Prashant Sardeshmukh for valuable discussions on this topic. The lead author was supported by Japan Society for the Promotion of Science Postdoctoral Fellowship for Foreign Researchers P10379: Climate change and the potential acceleration of the hydrological cycle. The second author received financial support from the Iowa Flood Center, IIHR-Hydroscience & Engineering. Support for the Twentieth Century Reanalysis (20CR) Project dataset is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office. The 20CR version 2.0 and HadISST v1.1 data were obtained from the Research Data Archive (RDA; <http://dss.ucar.edu>), which is maintained by the Computational and Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) and sponsored by the National Science Foundation (NSF). 20CR every-member data

was obtained from the National Energy Research Scientific Computing Center (NERSC; http://portal.nersc.gov/project/20C_Reanalysis/). Chesley McColl provided the 20CR assimilated observation count dataset. HadSLP2 was obtained from the Met Office Hadley Centre for Climate Change (www.metoffice.gov.uk/hadobs/hadslp2/). The CRU TS3.1 dataset was obtained in May 2011 from the British Atmospheric Data Centre (BADC; <http://badc.nerc.ac.uk>). The GPCC v6 Full Data Reanalysis was obtained from the Deutscher Wetterdienst (DWD; gpcp.dwd.de), operated under the auspices of the World Meteorological Organization (WMO). The COBE SST dataset was obtained from the Japan Meteorological Agency Tokyo Climate Center (http://ds.data.jma.go.jp/tcc/tcc/products/el_nino/cobesst/cobe-sst.html). ERSST v3b was obtained from the NOAA National Climate Data Center (<ftp://ftp.ncdc.noaa.gov/pub/data/cmb/ersst/v3b/netcdf>).

Appendix 1: Every member analyses for Geneva and Rondonia

In this section, we evaluate variability among the 56 ensemble members from which we assess only the ensemble mean (i.e., official 20CR product) in our study. We focus on two grid points: Geneva, Switzerland (46.20°N, 6.15°E) and Rondonia, Brazil (24.0°S, 51.0°W). The significance of these locations is the dichotomy of their supporting observational record. The record at Geneva benefits from a steady stream of approx. 300 observations per month over the period of availability. In contrast, not a single observation was ever assimilated at Rondonia.

Figures 14 and 15 present results of the Bai-Perron test performed on the 56-member monthly T_a and P data at Geneva and Rondonia, respectively. In general, but

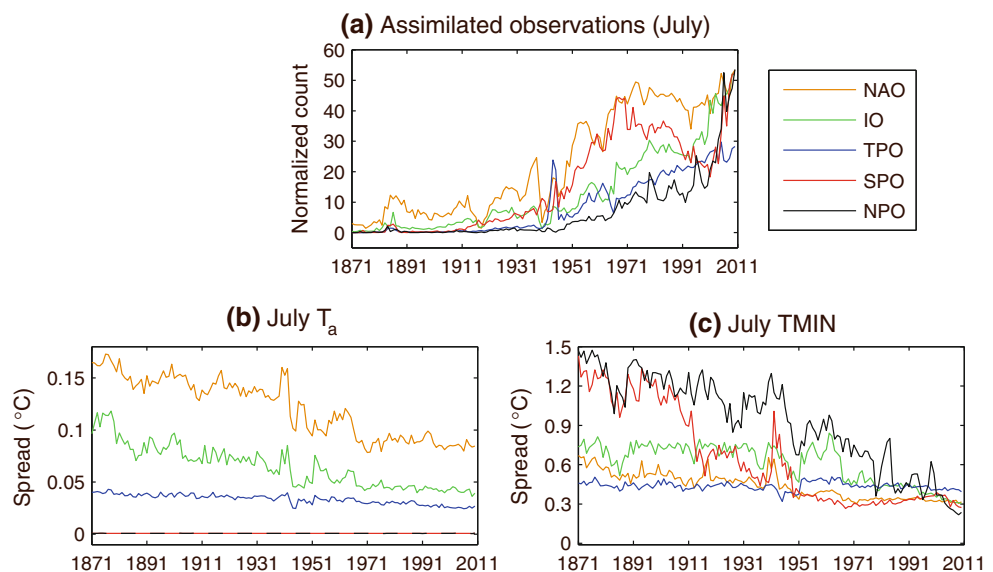
particularly for Geneva, the confidence interval for the ensemble mean change point encompasses much of the cumulative 56-member confidence interval range. The figures show that a change point in the ensemble mean could be triggered by abrupt shifts in as few as five ensemble members. On the other hand, there are cases in which clusters of 20 ensemble member change points over 9 years (Fig. 15a, September) do not cause an equally significant shift in the ensemble mean. At Geneva, most of the variable-months are homogenous. This suggests, to us, that if a climate shift did occur, it was weak. At Rondonia, on the other hand, change points manifest in all variable-months (except July for T_a), which we interpret as evidence of a strong shift.

Figure 15 exposes limitations of the Bai-Perron detection algorithm. Specifically, evidence of competing models (in a Bai-Perron sense) and detection limits can be seen [e.g., Fig. 15a: April, October, and November; Fig. 15b: October and December]. For example, at Rondonia, 17, 15, and 16 members of the T_a ensemble had change points in November 1962, 1969, and 1978, respectively. In other words, three potential shifts were identified within a 17-year time period. Because of the imposed 21-year minimum segment length (Sect. 2.3.2), only one change point was resolved.

Appendix 2: Temperature spread fields over ocean

See Fig. 16.

Fig. 16 For ocean domains, the area-averaged (at 2°) time series of 20CR July **a** assimilated observations, **b** T_a ensemble spread, and **c** TMIN ensemble spread. Notice the y-axis scales of **(b)** and **(c)** vary substantially



References

- Allan R, Ansell T (2006) A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *J Climate* 19(22):5816–5842. doi:[10.1175/Jcli3937.1](https://doi.org/10.1175/Jcli3937.1)
- Andrews DWK (1993) Tests for parameter instability and structural-change with unknown change-point. *Econometrica* 61(4): 821–856
- Ansell TJ, Jones PD, Allan RJ, Lister D, Parker DE, Brunet M, Moberg A, Jacobeit J, Brohan P, Rayner NA, Aguilar E, Alexandersson H, Barriendos M, Brandsma T, Cox NJ, Della-Marta PM, Drebs A, Founda D, Gerstengarbe F, Hickey K, Jonsson T, Luterbacher J, Nordli O, Oesterle H, Petrakis M, Philipp A, Rodwell MJ, Saladie O, Sigro J, Slonosky V, Srnec L, Swail V, Garcia-Suarez AM, Tuomenvirta H, Wang X, Wanner H, Werner P, Wheeler D, Xoplaki E (2006) Daily mean sea level pressure reconstructions for the European-North Atlantic region for the period 1850–2003. *J Climate* 19(12):2717–2742. doi:[10.1175/Jcli3775.1](https://doi.org/10.1175/Jcli3775.1)
- Bai J (1997) Estimation of a change point in multiple regression models. *Rev Econ Stat* 79(4):551–563
- Bai J, Perron P (2003) Computation and analysis of multiple structural change models. *J Appl Econom* 18(1):1–22. doi:[10.1002/Jae.659](https://doi.org/10.1002/Jae.659)
- Becker A, Finger P, Meyer-Christoffer A, Rudolf B, Schamm K, Schneider U, Ziese M (2013) A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth Syst Sci Data Discuss* 5:921–998. doi:[10.5194/essd-5-71-2013](https://doi.org/10.5194/essd-5-71-2013)
- Betts AK (2009) Land-surface-atmosphere coupling in observations and models. *J Adv Model Earth Syst* 1(4):18. doi:[10.3894/JAMES.2009.1.4](https://doi.org/10.3894/JAMES.2009.1.4)
- Compo GP, Whitaker JS, Sardeshmukh PD (2006) Feasibility of a 100-year reanalysis using only surface pressure data. *Bull Am Meteorol Soc* 87(2):175. doi:[10.1175/Bams-87-2-175](https://doi.org/10.1175/Bams-87-2-175)
- Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE, Vose RS, Rutledge G, Bessemoulin P, Bronnimann S, Brunet M, Crouthamel RI, Grant AN, Groisman PY, Jones PD, Kruk MC, Kruger AC, Marshall GJ, Maugeri M, Mok HY, Nordli O, Ross TF, Trigo RM, Wang XL, Woodruff SD, Worley SJ (2011) The twentieth century reanalysis project. *Q J Roy Meteorol Soc* 137(654):1–28. doi:[10.1002/qj.776](https://doi.org/10.1002/qj.776)
- Compo GP, Whitaker JS, Sardeshmukh PD, Giese B (2012) Developing the Sparse Input Reanalysis for Climate Applications (SIRCA) 1850–2014. Paper presented at the 4th World Climate Research Programme International Conference on Reanalyses, Silver Spring, Maryland, USA, 7–11 May 2012
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Kohler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thepaut JN, Vitart F (2011a) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J Roy Meteorol Soc* 137(656):553–597. doi:[10.1002/qj.828](https://doi.org/10.1002/qj.828)
- Dee DP, Kallen E, Simmons AJ, Haimberger L (2011b) Comments on “Reanalyses suitable for characterizing long-term trends”. *Bull Am Meteorol Soc* 92(1):65–70
- Ek MB, Mitchell KE, Lin Y, Rogers E, Grunmann P, Koren V, Gayno G, Tarpley JD (2003) Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J Geophys Res Atmos* 108 (D22)
- Ferguson CR, Villarini G (2012) Detecting inhomogeneities in the twentieth century reanalysis over the central United States. *J Geophys Res-Atmos* 117:D05123. doi:[10.1029/2011jd016988](https://doi.org/10.1029/2011jd016988)
- Ferguson CR, Wood EF (2011) Observed land-atmosphere coupling from satellite remote sensing and re-analysis. *J Hydrometeorol* 12(6):1221–1254. doi:[10.1175/2011JHM1380.1](https://doi.org/10.1175/2011JHM1380.1)
- Ferguson CR, Wood EF, Vinukollu RV (2012) A global inter-comparison of modeled and observed land-atmosphere coupling. *J Hydrometeorol* early-online. doi:[10.1175/JHM-D-11-0119.1](https://doi.org/10.1175/JHM-D-11-0119.1)
- Folland CK, Parker DE (1995) Correction of instrumental biases in historical sea-surface temperature data. *Q J Roy Meteorol Soc* 121(522):319–367
- Giorgi F, Francisco R (2000) Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Clim Dyn* 16(2–3):169–182
- Hsiang SM, Meng KC, Cane MA (2011) Civil conflicts are associated with the global climate. *Nature* 476(7361):438–441. doi:[10.1038/Nature10311](https://doi.org/10.1038/Nature10311)
- Huntington TG (2006) Evidence for intensification of the global water cycle: review and synthesis. *J Hydrol* 319(1–4):83–95
- Ishii M, Shouji A, Sugimoto S, Matsumoto T (2005) Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection. *Int J Climatol* 25(7):865–879
- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc* 77(3):437–471
- Kanamitsu M, Alpert JC, Campana KA, Caplan PM, Deaven DG, Iredell M, Katz B, Pan HL, Sela J, White GH (1991) Recent changes implemented into the global forecast system at NMC. *Weather Forecast* 6(3):425–435
- Kaplan A, Kushnir Y, Cane MA, Blumenthal MB (1997) Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J Geophys Res-Oceans* 102(C13):27835–27860
- Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M (2011a) Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J Geophys Res-Atmos* 116:D14103. doi:[10.1029/2010jd015218](https://doi.org/10.1029/2010jd015218)
- Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M (2011b) Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J Geophys Res-Atmos* 116:D14104. doi:[10.1029/2010jd015220](https://doi.org/10.1029/2010jd015220)
- Kistler R, Kalnay E, Collins W, Saha S, White G, Woollen J, Chelliah M, Ebisuzaki W, Kanamitsu M, Kousky V, van den Dool H, Jenne R, Fiorino M (2001) The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bull Am Meteorol Soc* 82(2):247–267
- Lawrimore JH, Menne MJ, Gleason BE, Williams CN, Wuertz DB, Vose RS, Rennie J (2011) An overview of the global historical climatology network monthly mean temperature data set, version 3. *J Geophys Res-Atmos* 116:D19121. doi:[10.1029/2011jd016187](https://doi.org/10.1029/2011jd016187)
- Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP, Naylor RL (2008) Prioritizing climate change adaptation needs for food security in 2030. *Science* 319(5863):607–610. doi:[10.1126/science.1152339](https://doi.org/10.1126/science.1152339)

- McVicar TR, Van Niel TG, Li LT, Roderick ML, Rayner DP, Ricciardulli L, Donohue RJ (2008) Wind speed climatology and trends for Australia, 1975–2006: capturing the stilling phenomenon and comparison with near-surface reanalysis output. *Geophys Res Lett* 35 (20)
- Meehl GA, Hu AX, Santer BD (2009) The Mid-1970s climate shift in the Pacific and the relative roles of forced versus inherent decadal variability. *J Climate* 22(3):780–792. doi:10.1175/2008jcli2552.1
- Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int J Climatol* 25(6):693–712
- Mitchell TD, Carter TR, Jones PD, Hulme M, New M (2004) A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901–2000) and 16 scenarios (2001–2100). Tyndall Centre Working Papers No 55 (Available at <http://www.tyndall.ac.uk/biblio/working-papers/>): 30
- Moorthi S, Pan H-L, Caplan P (2001) Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. Technical Procedures Bulletin 484, NOAA, NWS: Silver Spring, MD Available from <http://www.nws.noaa.gov/om/tpb/484.htm>
- Moritz MA, Parisien M-A, Battlori E, Krawchuk MA, Dorn JV, Ganz DJ, Hayhoe K (2012) Climate change and disruptions to global fire activity. *Ecosphere* 3(6). doi:10.1890/ES11-00345.1
- New M, Hulme M, Jones P (1999) Representing twentieth-century space-time climate variability. Part I: development of a 1961–90 mean monthly terrestrial climatology. *J Climate* 12(3):829–856
- New M, Hulme M, Jones P (2000) Representing twentieth-century space-time climate variability. Part II: development of 1901–96 monthly grids of terrestrial surface climate. *J Climate* 13(13):2217–2238
- Pettitt AN (1979) A non-parametric approach to the change-point problem. *Appl Stat* 28:126–135. doi:10.2307/2346729
- Powell AM, Xu JJ (2011) A new assessment of the mid-1970s abrupt atmospheric temperature change in the NCEP/NCAR reanalysis and associated solar forcing implications. *Theor Appl Climatol* 104(3–4):443–458. doi:10.1007/s00704-010-0344-1
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (eds) (1992) *Numerical recipes in C: the art of scientific computing*, 2 edn. Cambridge University Press, 994 pp
- Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J Geophys Res-Atmos* 108(D14):4407. doi:10.1029/2002jd002670
- Rienecker MM, Suarez MJ, Todling R, Bacmeister J, Takacs L, Liu H-C, Gu W, Sienkiewicz M, Koster RD, Gelaro R, Stajner I, Nielson E (2008) The GEOS-5 data assimilation system—documentation of versions 5.0.1 and 5.1.0 NASA GSFC technical report series on global modeling and data assimilation. NASA/TM-2007-104606 27:92
- Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, Bosilovich M, Schubert SD, Takacs L, Kim G-K, Bloom S, Chen J, Collins D, Conaty A, da Silva AM, Gu W, Joiner J, Koster RD, Lucchesi R, Molod A, Owens T, Pawson S, Pegion P, Redder CR, Reichle R, Robertson FR, Ruddick AG, Sienkiewicz M, Woollen J (2011) MERRA-NASA’s modern-era retrospective analysis for research applications. *J Clim*. doi:10.1175/JCLI-D-11-00015.1
- Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, Van den Dool HM, Pan HL, Moorthi S, Behringer D, Stokes D, Pena M, Lord S, White G, Ebisuzaki W, Peng P, Xie P (2006) The NCEP climate forecast system. *J Climate* 19(15):3483–3517
- Saha S, Moorthi S, Pan HL, Wu XR, Wang JD, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, Liu HX, Stokes D, Grumbine R, Gayno G, Wang J, Hou YT, Chuang HY, Juang HMM, Sela J, Iredell M, Treadon R, Kleist D, Van Delst P, Keyser D, Derber J, Ek M, Meng J, Wei HL, Yang RQ, Lord S, Van den Dool H, Kumar A, Wang WQ, Long C, Chelliah M, Xue Y, Huang BY, Schemm JK, Ebisuzaki W, Lin R, Xie PP, Chen MY, Zhou ST, Higgins W, Zou CZ, Liu QH, Chen Y, Han Y, Cucurull L, Reynolds RW, Rutledge G, Goldberg M (2010) The NCEP climate forecast system reanalysis. *Bull Am Meteorol Soc* 91(8):1015–1057. doi:10.1175/2010BAMS3001.1
- Schar C, Vidale PL, Luthi D, Frei C, Haberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. *Nature* 427 (6972):332–336. doi:10.1038/Nature02300
- Schneider U, Becker A, Finger F, Meyer-Christoffer A, Ziese M, Rudolf B (2013) GPCC’s new land-surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor Appl Climatol*. doi:10.1007/s00704-013-0860-x
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Sheffield J, Wood EF (2008) Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. *Clim Dynam* 31(1):79–105. doi:10.1007/s00382-007-0340-z
- Smith TM, Reynolds RW (2003) Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J Climate* 16(10):1495–1510
- Team RDC (2008) *R: a language and environment for statistical computing*. R Found. For Stat. Comput., Vienna
- Thorne PW, Vose RS (2010) Reanalyses suitable for characterizing long-term trends are they really achievable? *Bull Am Meteorol Soc* 91(3):353. doi:10.1175/2009BAMS2858.1
- Troy TJ, Sheffield J, Wood EF (2012) The role of winter precipitation and temperature on northern Eurasian streamflow trends. *J Geophys Res-Atmos* 117:D05131. doi:10.1029/2011jd016208
- Westerling AL, Hidalgo HG, Cayan DR, Swetnam TW (2006) Warming and earlier spring increase western US forest wildfire activity. *Science* 313(5789):940–943. doi:10.1126/science.1128834
- Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. *Mon Weather Rev* 130(7):1913–1924
- Willmott CJ, Rowe CM, Philpot WD (1985) Small-scale climate maps: a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *Am Cartographer* 12(1):5–16
- Worley SJ, Woodruff SD, Reynolds RW, Lubker SJ, Lott N (2005) ICOADS release 2.1 data and products. *Int J Climatol* 25(7):823–842. doi:10.1002/joc.1166
- Zeileis A, Kleiber C (2005) Validating multiple structural change models: a case study. *J Appl Econom* 20(5):685–690
- Zeileis A, Kleiber C, Kraemer W, Hornik K (2003) Testing and dating of structural changes in practice. *Comput Stat Data Anal* 44:109–123