March 2012

Evaluating Modeled Intra- to Multidecadal Climate Variability Using Running Mann–Whitney *Z* Statistics

Steven A. Mauget
U.S. Department of Agriculture, Lubbock, Texas

Eugene C. Cordero
San Jose State University, eugene.cordero@sjsu.edu

Patrick T. Brown
San Jose State University

Recommended Citation

Steven A. Mauget, Eugene C. Cordero, and Patrick T. Brown. "Evaluating Modeled Intra- to Multidecadal Climate Variability Using Running Mann–Whitney *Z* Statistics" *Journal of Climate* (2012): 1570-1586. doi:10.1175/JCLI-D-11-00211.1

Evaluating Modeled Intra- to Multidecadal Climate Variability Using Running Mann–Whitney *Z* Statistics

STEVEN A. MAUGET

Agricultural Research Service, U.S. Department of Agriculture, Lubbock, Texas

EUGENE C. CORDERO AND PATRICK T. BROWN

Department of Meteorology and Climate Science, San Jose State University, San Jose, California

(Manuscript received 20 April 2011, in final form 25 July 2011)

ABSTRACT

An analysis method previously used to detect observed intra- to multidecadal (IMD) climate regimes was adapted to compare observed and modeled IMD climate variations. Pending the availability of the more appropriate phase 5 Coupled Model Intercomparison Project (CMIP-5) simulations, the method is demonstrated using CMIP-3 model simulations. Although the CMIP-3 experimental design will almost certainly prevent these model runs from reproducing features of historical IMD climate variability, these simulations allow for the demonstration of the method and illustrate how the models and observations disagree. This method samples a time series's data rankings over moving time windows, converts those ranking sets to a Mann–Whitney *U* statistic, and then normalizes the *U* statistic into a *Z* statistic. By detecting optimally significant IMD ranking regimes of arbitrary onset and varying duration, this process generates time series of *Z* values that are an adaptively low-passed and normalized transformation of the original time series. Principal component (PC) analysis of the *Z* series derived from observed annual temperatures at 92 U.S. grid locations during 1919–2008 shows two dominant modes: a PC1 mode with cool temperatures before the late 1960s and warm temperatures after the mid-1980s, and a PC2 mode indicating a multidecadal temperature cycle over the Southeast. Using a graphic analysis of a *Z* error metric that compares modeled and observed *Z* series, the three CMIP-3 model simulations tested here are shown to reproduce the PC1 mode but not the PC2 mode. By providing a way to compare grid-level IMD climate response patterns in observed and modeled data, this method can play a useful diagnostic role in future model development and decadal climate forecasting.

1. Introduction

Although past projections of future climate change have emphasized mid- or late twenty-first-century conditions (Cubasch et al. 2001; Meehl et al. 2007a,b), attention has recently turned to the prediction of upcoming decadal periods. In addition to presenting longer-term climate projections, the most recent phase of the Coupled Model Intercomparison Project (CMIP-5) and the Intergovernmental Panel on Climate Change's (IPCC) Fifth Assessment Report will also include decadal hindcast and prediction experiments (Meehl et al. 2009; Taylor et al. 2008). The goals of the Met Office's Decadal Prediction System (Smith et al. 2007) include improving the understanding and prediction of decadal climate variability and eventually producing operational decadal forecasts. This interest in predicting decadal climate is driven in part by the importance of decadal time scales in human affairs. Meehl et al. (2009) note decision makers' concerns about climate over a 10–30-yr time horizon, given the decadal duration of persistent drought and the time scales of hurricane activity and fisheries regimes. Zwiers (2002) and Cane (2010) also note how the extended duration of typical climate projections are inconsistent with the decadal or multidecadal outlook of managers that make climate-sensitive decisions. Hurrell et al. (2010) point out how decadal climate forecasts might be useful in managing long-term investment strategies. As the social effects of climate variation are frequently local,

Corresponding author address: Steven A. Mauget, USDA/ARS, Plant Stress and Water Conservation Laboratory, 3810 4th St., Lubbock, TX 79415.
E-mail: steven.mauget@ars.usda.gov

DOI: 10.1175/JCLI-D-11-00211.1

attention has also been directed to predicting potential climate impacts over regional spatial scales (International ad hoc Detection and Attribution Group 2005). In reviewing the issues that need to be addressed before decadal forecasts can be practically useful, Vera et al. (2009) cite the need for research that translates decadal prediction information into the spatial scales needed to support decision making. Both Vera et al. (2009) and Cane (2010) summarize the needs of potential stakeholders and the current state of decadal climate prediction and conclude that while there is clear demand for such prediction, that demand is not being met. Although the physical basis for these forecasts is a relatively new area of research (Latif et al. 2006, 2010; Collins et al. 2006; Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2004; Murphy et al. 2009; Solomon et al. 2011), regional forecasts of decadal climate conditions are potentially useful to today's policy and decision makers.

However, before decision makers can confidently use decadal climate projections the associated models must be verified somehow. One of the most widely recognized measures for testing a model's ability to predict climate is the ability to simulate current or past climate (Boer 2000; Lambert and Boer 2001; Giorgi and Mearns 2002; Tebaldi et al. 2005; Räisänen 2007; Randall et al. 2007), which suggests that models should demonstrate a ''track record'' of either hindcasting or reproducing known intra- to multidecadal (IMD) climate variability. Yet the ability of models to reproduce historical IMD climate regimes has not been the focus in past attempts at model verification. Instead, the emphasis has been on comparing observed and modeled climate statistics calculated over multidecadal periods. For example, Giorgi and Mearns (2002) and Tebaldi et al. (2005) considered the differences in observed and modeled regional temperature means during 1961–90, and the difference between projected mean conditions and a model ensemble mean during 2071–2100. Pierce et al. (2009) calculated skill scores based on the mean squared spatial error between gridded maps of modeled and observed 1960–99 seasonal climate statistics over the western United States. Taking a similar approach but with a narrower geographical focus, Brekke et al. (2008) calculated the difference between modeled and observed means, variances, interdecadal variances, and the seasonal amplitude and phase for northern California precipitation and temperature during 1950–99. In tracing the evolution of CMIP-1, -2, and -3 model output, Reichler and Kim (2008) calculated a model performance index based on the difference between modeled and observed climatologies for a range of climate variables during, for the most part, 1979–99. Less emphasis has been placed on verifying if models can reproduce the observed variation over time. Although trend analysis has been used in attribution studies that try to link climate forcing influences to observed climate change (Knutson et al. 1999; Hegerl and Allen 2002; Karoly and Wu 2005; Knutson et al. 2006), trend fitting assumes more or less linear behavior in time and is not suited to detecting general climate variability (e.g., abrupt regime shifts or irregular decadal climate cycles). Graphical methods exist for summarizing the mean squared difference, the ratio of variance, and the correlation between observed and modeled time series (Lambert and Boer 2001), but correlation values by themselves give an incomplete picture of how modeled and observed data covary in time. Specifically, they give no information about a model's ability to reproduce the onset and duration of IMD climate periods in the historical record, or of leading or lagging relationships between modeled and observed climate shifts. Statistical and graphical methods that provide such information may be useful in model verification, as decadal prediction would require the ability to correctly predict the onset, and possibly the duration of IMD climate regimes.

Previous work here focused on detecting nonlinear and regimelike climate behavior via a method that calculates Mann–Whitney Z (MWZ) statistics over running time windows. This approach to time series analysis has been used to identify significant IMD periods in U.S. temperature, precipitation, and streamflow (Mauget 2003, 2004; Cordero et al. 2010) and in South American snowpack and streamflow (Masiokas et al. 2010). In the current work, this ranking-based method is adapted to compare observed and modeled decadal temperature regimes at the model's grid resolution. In this demonstration the running MWZ method is applied to annual temperature time series derived from U.S. Historical Climatology Network data (USHCN; Menne et al. 2009) and modeled temperature data derived from the CMIP-3 Climate of the Twentieth Century Project (20C3M: Folland et al. 2002; Covey et al. 2003) to produce Z series. The Z series, which are normalized and low-passed transformations of the original data series, are then used to compare the timing and significance of observed and modeled twentieth-century U.S. IMD temperature regimes. This comparison is expressed in a Z error metric that is graphically presented to show the difference in modeled versus observed IMD variability at the highest spatial resolution that the model allows.

It should be understood that the CMIP-3 ensemble average model runs that we test are used only as examples of modeled climate behavior; there is little expectation that they should reproduce the location and timing of past IMD climate regimes. The initial conditions of the 20C3M experiments were derived from the state of preindustrial control (PICNTRL) runs at various times during the nineteenth century, which, as will be seen, has likely influenced the model's evaluations. Decadal climate variation is considered a function of both initial conditions and boundary forcing over time (Meehl et al. 2009; Latif et al. 2010), and correct initialization is thought to be a requirement when reproducing natural decadal-scale ocean processes (Solomon et al. 2011). As a result, the random initialization of the 20C3M models' sea surface temperature (SST) and sea ice conditions will almost certainly lead to differences in the decadal features of modeled and observed twentieth-century climate. This initialization error is not a problem in the current work but is actually useful, as the main purpose here is to demonstrate a statistical/graphical method for detecting errors or differences in modeled IMD climate. However, although the 20C3M experiments were not initialized in a way that would, in principle, allow them to reproduce past IMD climate regimes, future CMIP-5 hindcast modeling experiments will be conducted in such a manner (Taylor et al. 2008; Meehl et al. 2009). By using this method to clearly identify where and when the modeled and observed IMD climates differed in initialized hindcasts, future developers might be better equipped to identify and correct systematic errors in the model results, test different initialization schemes, and consider the general questions surrounding decadal predictability.

The next section (section 2) will describe three models evaluated via the process described above, and the formation of a gridded dataset for annual temperature over the continental U.S. during 1919–2008. Section 3 will demonstrate the formation of a *Z* series from a gridded average of USHCN annual temperature records via the running Mann–Whitney *Z* method, and will then describe similar results calculated over 92 continental U.S. grid locations. Section 4 will describe the *Z* error metric and an ad hoc process for estimating a *Z* error threshold from the three model's PICNTRL runs. Section 5 will present and discuss the IMD temperature regimes in the three 20C3M simulations revealed by the running MWZ method, and the corresponding distributions of *Z* error in space and time. Section 6 will summarize how the running MWZ method was used here to evaluate the three models, and will briefly discuss how it might be used as a diagnostic tool in decadal climate prediction.

2. Modeled and observed temperature data

The three coupled atmosphere–ocean global climate models (AOGCMs) that were evaluated are: the Flexible Global Ocean–Atmosphere–Land System Model gridpoint version 1.0 (FGOALS; Yu et al. 2002, 2004), the Model for Interdisciplinary Research on Climate 3.2, medium-resolution version (MIROC; Hasumi and Emori 2004), and the National Center for Atmospheric Research's (NCAR) Parallel Climate Model (PCM; Washington et al. 2000; Meehl et al. 2004). The FGOALS, MIROC, and PCM 20C3M model runs begin at various times in the nineteenth century and end in either December 1999 or December 2000. As described in section 1, the initial conditions of the 20C3M runs were derived from the state of preindustrial control runs. Each model was run globally with estimated historical radiative forcings and has an identical T42 spatial grid resolution over the continental United States (Fig. 1). None of these coupled models were flux adjusted. Although all three models' 20C3M runs included the effects of well-mixed greenhouse gases, the forcing associated with known variations in anthropogenic and volcanic aerosol, land use, ozone and solar variability differed between the models. For more details see the Program for Climate Model Diagnosis and Intercomparison (PCMDI) online documentation (http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php) and the discussion of 20C3M forcing scenarios in Kunkel et al. (2006). Details of the 20C3M forcing for PCM can be found in Meehl et al. (2003, 2004). Grid-level annual temperature time series were formed from the ensemble average of annual temperatures from three runs of each model. Using an approach similar to Hegerl et al.'s (2007), the resulting 20C3M annual temperature time series were extended to 2008 by the addition of similar ensemble average series from the IPCC Special Report on Emissions Scenarios (SRES) A1B scenario simulations at each grid location. Each model's three SRES A1B simulations were initialized by the end state of the corresponding run in the 20C3M simulations.

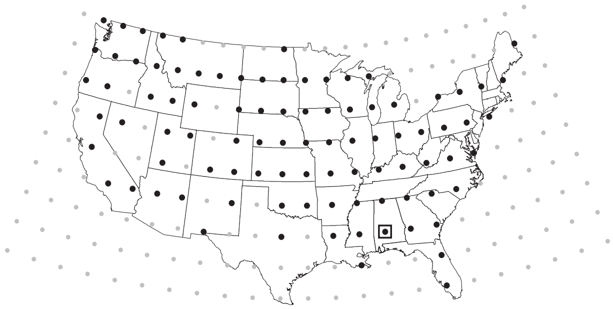

FIG. 1. North American–region T42 grid locations. Black grid locations mark continental U.S. grid areas, each having at least five USHCN stations with less than 20% estimated monthly data during 1919–2008. The black square marks the center of the southern Alabama grid location.

Time series of observed annual mean surface temperatures over the continental United States were derived from USHCN monthly temperature station data.

The USHCN station network is a subset of the U.S. cooperative network (U.S. National Weather Service 2000), which has undergone an adjustment to account for historical changes in station locations, instrumentation, and observing practices, as well as the effects of urbanization and nonstandard siting (Menne and Williams 2005, 2009). To provide as direct a comparison as possible with the three models' gridded temperature records, the annual temperatures from the station data were averaged over the semi-equal-angle areas surrounding each T42 grid location. Although missing USHCN monthly temperature values are infrequent, values estimated from nearby station data are more common, particularly in the western United States during the early twentieth century. Here, T42 grid averages were derived from station records that had less than 20% estimated monthly data during 1919–2008, and the area surrounding a T42 grid location was required to contain at least five such stations to calculate a 90-yr annual temperature series for that grid location. The five-station threshold was chosen based on an analysis similar to that of Janis et al. (2004), and the need for a reasonably representative distribution of observed gridded temperature records in the western United States. Given these data requirements, the formation of observed annual temperature time series for 1919–2008 was limited to 92 continental U.S. T42 grid locations (Fig. 1). That distribution of grid locations allows for observed versus modeled comparisons to be made over most of the United States east of the 100th meridian during that period, but includes some spatial gaps in the west.

3. Time series analysis via running Mann–Whitney *Z* statistics

At each of the 92 T42 grid locations, the running MWZ method ranks a time series' data values, samples those rankings over moving time windows of n_I years' duration, then converts each sample of rankings into a Mann–Whitney *U* statistic (Mann and Whitney 1947). A *U* statistic for a sample of rankings within an n_I-yr time window can be calculated based on the sample's size and rank sum (Mendenhall et al. 1990; Wilks 1995), but can be understood more intuitively as the total number of data values outside the sampling window that precede each sample value when all data values are arranged by rank (Hollander and Wolfe 1999). Stated otherwise, for a time window spanning an arbitrary n_I-yr period in the time series, for each ranking value within that period, count the total number of values from outside the window whose ranks are less than that value; the *U* statistic is the sum of all such counts for each ranking value in the n_I-yr time window. Thus, for a 90-yr time series divided into an n_I = 10 yr sample window and n_{II} = 80 yr outside the sample window, the highest possible *U* statistic would occur when the sample contains the 10 highest-ranked years ($U = 80 \times 10$), while the lowest value would result from a sample containing the 10 lowest-ranked years ($U = 0 \times 10$). Randomly sampled sets of 10 rankings produce *U* statistics that are normally distributed between those two extreme values and are generally proportional to the incidence of high rankings in the sample. That distribution's mean is equal to the average of the minimum and maximum *U* values; for example,

$$\mu = 0.5[(0\times 10) + (80\times 10)] = 0.5 n_I n_{II}, \qquad (1)$$

while the standard deviation can be estimated via the expression (Mendenhall et al. 1990)

$$\sigma = \{[n_I n_{II}(n_I + n_{II} + 1)]/12\}^{1/2}. \qquad (2)$$

Gaussian *U* statistics can be *Z* normalized using these null parameters, with significantly high (low) *Z* values indicating a significant incidence of high (low) annual temperature rankings relative to a null hypothesis that assumes random and independent sampling (H_0):

$$Z = \frac{U - \mu}{\sigma}. \qquad (3)$$

To demonstrate how *Z* series are formed from the Eq. (3) *Z* statistics, Figs. 2a–e show the running MWZ method applied to a grid-averaged annual temperature record calculated over the southern Alabama grid area marked in Fig. 1. Figure 2a shows the annual temperature time series and its mean, while Fig. 2b shows the *Z* statistics for temperature rankings sampled over moving 10-yr time windows. The latter figure's horizontal lines mark negative and positive significance levels relative to H_0 at two-sided 95%, 99%, and 99.9% confidence levels. Figure 2c's horizontal black lines show the 10-yr ranking regimes marked as significant at a 95% or better confidence level in Fig. 2b superimposed on the actual data. The vertical placement of those lines marks the regime's corresponding *Z* statistic, as measured by the figure's right axis.

To extend the Fig. 2c analysis to a wider range of time scales, the calculation of running *U* statistics is repeated with sampling windows between 6 and 30 yr. Those *U* statistics are then normalized into *Z* statistics using the μ and σ parameters for each of the 25 sample sizes calculated via Eqs. (1) and (2). The Eq. (3) normalization, in addition to estimating the significance of rankings for a fixed sample size, also allows for comparing the significance of *Z* statistics derived from different sample sizes. Thus, after the running *U* statistics from each analysis were normalized by the appropriate null parameters, the

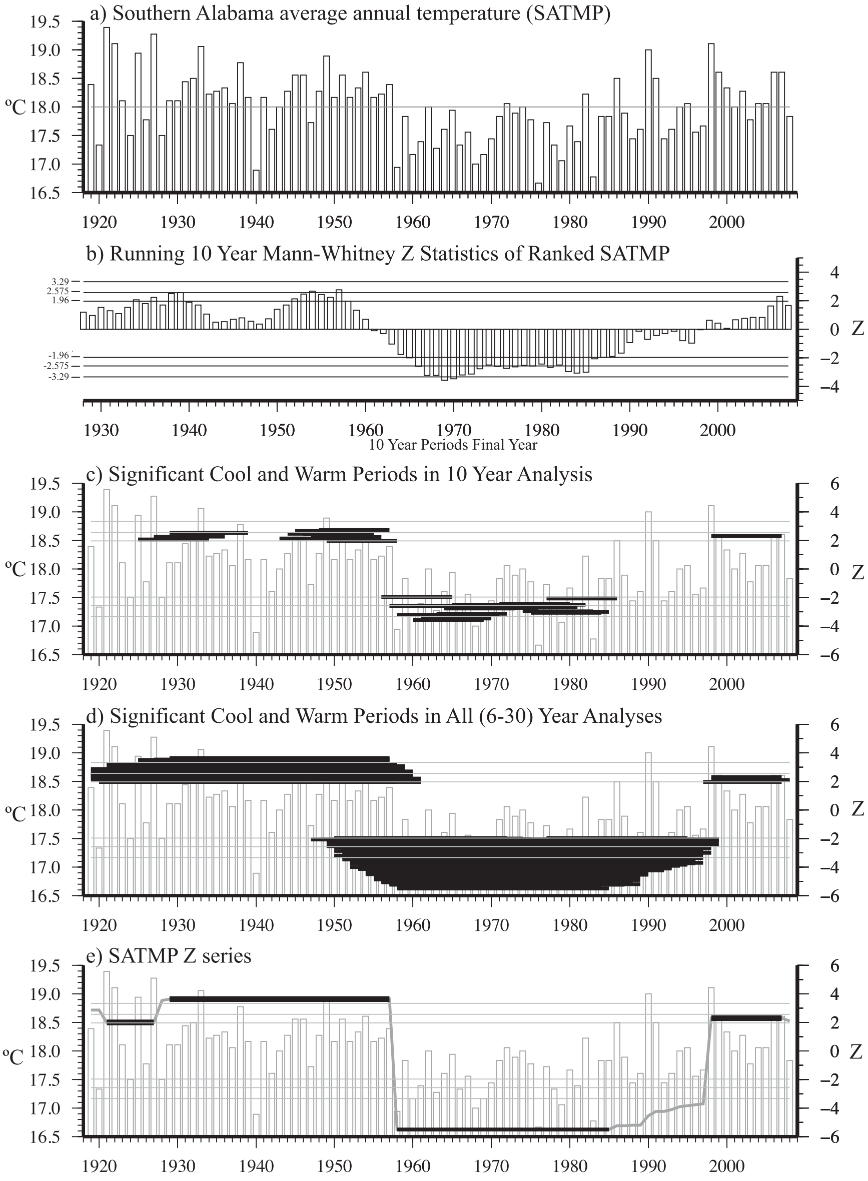

FIG. 2. (a) Time series of annual temperature averaged from southern Alabama USHCN stations during 1919–2008. Horizontal line shows the 90-yr mean. (b) MWZ statistics of ranked annual temperatures sampled over running 10-yr time windows. Horizontal lines mark two-sided 95% ($Z = \pm1.96$), 99% ($Z = \pm2.575$), and 99.9% ($Z = \pm3.29$) confidence intervals. (c) As in (a), but with the horizontal width of the black bars showing significant 10-yr cool and warm periods, as indicated in (b). Vertical placement of bars shows corresponding Z values as marked by the right axis. (d) Significant cool and warm periods indicated by running MWZ analyses with 6-, 7-, . . . , 30-yr sampling windows. (e) The Z series for southern Alabama annual temperature. Black sections show most significant cool and warm periods in (d) occurring over nonoverlapping time windows. The Z values for remaining years (gray traces) are defined by the year's maximum absolute Z value in all the analyses.

ranking regimes from all 25 tests with Z statistics that exceeded a two-sided 95% confidence threshold are combined as in Fig. 2d. Those pooled results are then screened to identify the periods that result in the greatest absolute significance over nonoverlapping time windows, which are marked by the thick black horizontal traces in Fig. 2e. This screening process begins by sorting all the regime periods by the absolute value of the period's Z statistic ($|Z|$) and recording the most significant statistic and its period. The Z series values for that period are assigned that Z value. Then, the next most significant $|Z|$ statistic with a period that does not overlap with that of the most significant statistic is recorded, and those years are assigned that value in the Z series. In the Fig. 2e Z series these two leading statistics occurred during a 1958–85 cool period ($Z = -5.486$) and a 1929–57 warm period ($Z = 3.618$). The process iteratively continues by recording the next most significant $|Z|$ statistic with a period that does not overlap with all previously recorded periods, assigns that statistic to the corresponding years in the Z series, and proceeds until all the significant ranking regimes in all 25 tests have been considered. In the Fig. 2e Z series this process identifies two other warm periods during 1998–2007 (Z = 2.31) and 1921–27 ($Z = 1.996$). The Z series values for each year between these distinct and optimally significant warm and cool periods are individually assigned according to the maximum significant $|Z|$ value for that year in all of the pooled analyses, which are shown in Fig. 2e's gray traces. The general effect is to form continuous Z series that tend to define the outer envelope of the Z statistics in the Fig. 2d pooled analyses, with sign transitions at years with overlapping positive and negative significant statistics (e.g., the 1997–98 transition in Fig. 2e).

In previous applications, the goal of the running MWZ process was to detect ranking sequences that were inconsistent with null hypotheses that assumed hypothetical stationary climate conditions. As stationary climate variability is usually required to possess the interannual persistence of observed data, those hypotheses accounted for that persistence. For example, the null hypothesis of Mauget (2004) held that 60-yr streamflow records possessed year-to-year persistence, but contained no IMD variability (H_1). The Eq. (3) μ and σ null parameters consistent with H_1 were estimated via filtering and autoregressive modeling of each annual streamflow series, as well as Monte Carlo (MC) simulations to generate U null distributions from the rankings of the resulting noise series. However, unlike Mauget (2004), the ultimate goal of the Z error test is not to test for nonstationarity in a sample of rankings, but to find dissimilarity between ranking samples in observed and modeled climate data. A key requirement in achieving that goal is that identical ranking sequences in a grid location's modeled and observed data values, which produce identical U statistics, must result in identical Z statistics. To satisfy this requirement, the same null parameters must be used in Eq. (3) to normalize modeled and observed U statistics. This condition is generally not met by H_1 null parameters estimated from modeled and observed data, as that MC protocol can produce U null distributions with varying σ values at a fixed sample size. As the emphasis here is on the consistent normalization of U statistics, and not on testing for nonstationary climate variation, the fixed null parameters in Eqs. (1) and (2) are used for each of the 25 sample sizes considered. Thus, in this comparison role the null hypothesis for the Eq. (3) Z values holds that a sequence of rankings in a time series is consistent with random sampling (H_0). This null hypothesis assumes no interannual persistence; as a result, it tends to assign higher significance levels than hypotheses that do assume stationary "red" climate variability.

The running MWZ method samples a time series over moving time windows to identify significant ranking sequences in climate data. This sampling process is repeated over a range of intra- to multidecadal time windows to identify the most significant runs of the rankings within those periods, and thus the process has the properties of an adaptive low-pass filter. Because U statistics are derived from rankings, the method is resistant to the presence of outliers and can be applied to normally or nonnormally distributed data. As stated before, the transformation of the normal U statistics into Z statistics allows for the comparison of ranking sets from different sample sizes. In addition, this ranking, U-transformation, and Z-normalization procedure allows for making observed versus modeled and model versus model comparisons with AOGCM data records that may have incorrect variance and/or biased means (Bell et al. 2000; Randall et al. 2007). However, because the method only evaluates rankings, it gives no insights into the nature or magnitude of those biases. Unlike its methodological cousin wavelet analysis (Lau and Weng 1995; Torrence and Compo 1998), the running MWZ algorithm is not based on an underlying assumption of harmonic behavior. Instead, it extends the generality of moving window methods with a more general assumption, that is, that climate variation consists of simple noncyclic ranking regimes that occur over a range of time scales and have arbitrary onset times. Given the ability to detect such regimes, the method can detect a wide range of climate variability. Abrupt climate shifts might reveal themselves as significant cool and warm periods that are immediately adjacent in time. An example of such a shift in southern Alabama temperatures in the late 1950s is apparent in the significant 10-yr ranking regimes marked in Fig. 2c and in the Fig. 2e Z series, which shows a significant warm period during 1929–57 immediately followed by a highly significant cool period during 1958–85. Similarly abrupt cyclic behavior can be revealed by Z values of alternating sign marking periods of significant high- and low-ranked data, although such alternating behavior is also found in Z series generated from smoothly harmonic variation (see Figs. A1a–d in the appendix). Thus, significant positive and negative Z statistics that are immediately adjacent in time may not always indicate

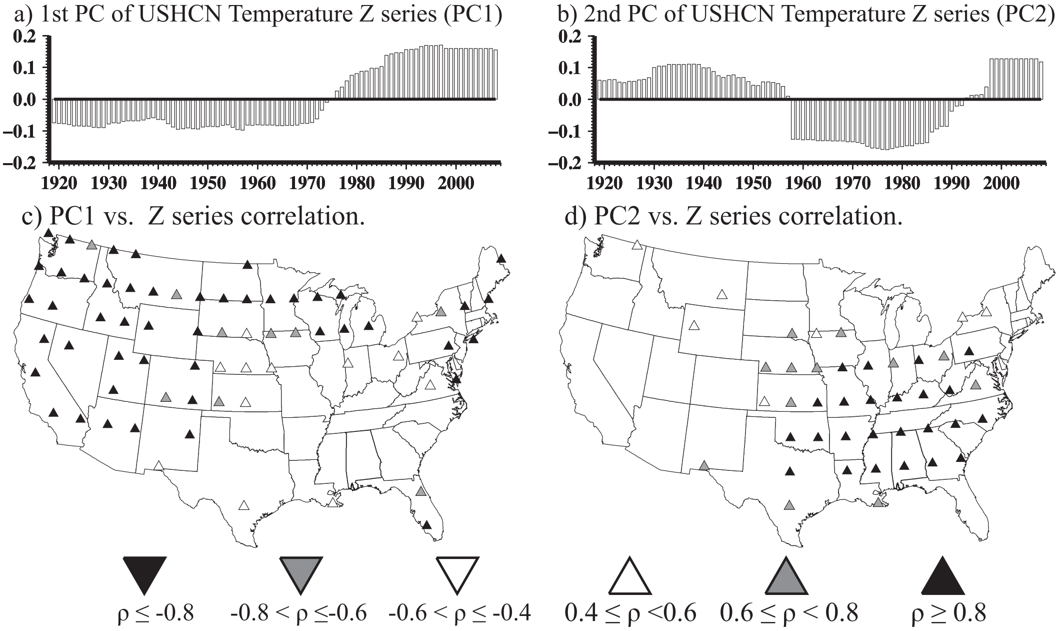

FIG. 3. (a) Normalized PC1 times series from unrotated PCA of *Z* series at Fig. 1's 92 black grid locations. (b) As in (a), but for normalized PC2 time series. (c) Correlation of PC1 with the *Z* series at each grid location. (d) Correlation of PC2 with the *Z* series at each grid location.

a regime shift in the data. A linear trend signature consists of significant *Z* statistics of opposite sign at the beginning and end of a time series, separated by periods of less significant or insignificant statistics (Fig. A1e). Overall, the running MWZ procedure could be considered as a robust, adaptive, and normalized low-passed filter.

a. Principal component analysis of USHCN temperature Z series

A comparison of unrotated and Varimax rotated (not shown) principal component (PC) analyses (PCAs) of the *Z* series from all 92 grid locations indicated that the unrotated PCA provided higher levels of correlations with the *Z* series that were more consistent and spatially coherent. The unrotated and grid-area-weighted PCA shows that U.S. temperature during 1919–2008 projects mainly onto one of two low-frequency modes. Figures 3a,b are the normalized first (PC1) and second (PC2) principal component time series of that analysis, which explain 53% and 33% of the total *Z* series variance, respectively. The correlations of those PC series with each grid's *Z* series are found in Figs. 3c,d. Figure 3c's high (>0.8) positive PC1 correlations over southern Florida and much of the United States outside of the Southeast show that that the PC's form closely represents that grid location's *Z* series. Thus, those areas experienced a fairly stable cool period until the late 1960s, followed by a transition to warmth after the mid-1980s. Over other grid locations in the southeastern and midwestern states, similarly high positive correlations in Fig. 3d indicate IMD variations that are more consistent with the PC2 time series. A warm period between 1919 and the late 1950s abruptly gave way to a cool period extending to about the mid-1980s, which was in turn followed by a more gradual shift to a second warm regime after the late 1990s. The area covered by Fig. 3d's positively correlated grid locations coincides with one of the few global land areas that Folland et al. (2001) and Trenberth et al. (2007) show as having negative annual temperature trends during the twentieth century. However, Fig. 3b's PC2 series suggests that the signs of the trends in these areas would depend on the periods over which the trends were fitted; for example, trends fitted after 1970 would more likely be positive.

Robinson et al. (2002), Pan et al. (2004), and Kunkel et al. (2006) have all noted weakened warming trends or cooling trends in the central and eastern United States in the latter half of the twentieth century. Pan et al. (2004) referred to the lack of warming in central U.S. summer temperatures as a "warming hole." Although the goal of these studies was not to identify the general location and timing of this cool regime, the cool period described by Robinson et al. (2002) and Kunkel et al. (2006) roughly coincides with the cool period of the PC2 temperature cycle in Fig. 3b. As a result, we will also use the term warming hole to refer to the period of cool temperatures in the southeastern United States between the late 1950s and the mid-1980s.

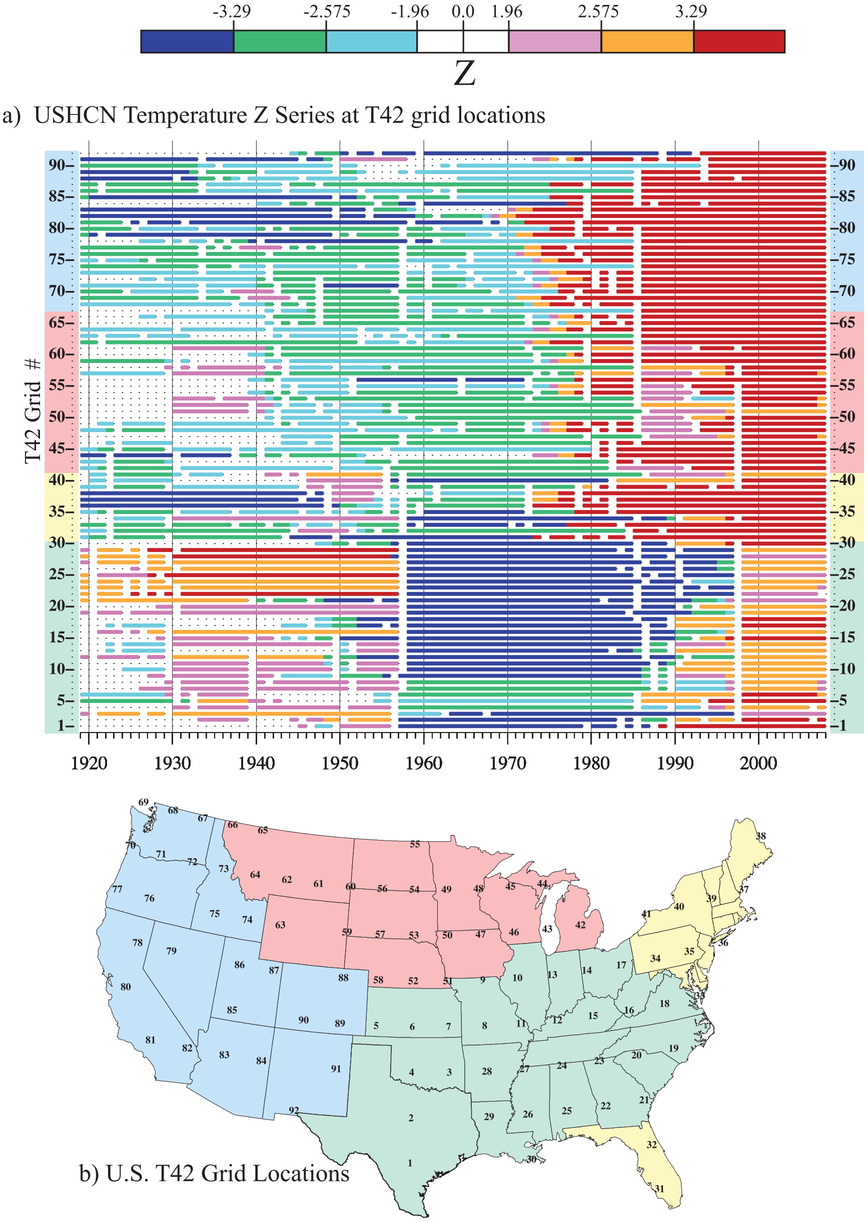

FIG. 4. (a) Positive (warm shade) and negative (cool shade) Z series magnitudes from running MWZ analyses of the annual temperature time series at each of Fig. 1's black T42 grid locations. Positive and negative significance are marked by the top legend's shading scheme. The vertical axis marks the grid location number. (b) Grid location number corresponding to the vertical axis index number in (a). Grids in the yellow-, red-, and blue-shaded regions correlate with Fig. 3a's PC1 series, while grids in the green-shaded region correlate with Fig. 3b's PC2 series.

b. *Graphic analysis of Z series*

The normalizing property of the running MWZ method makes it possible to graphically compare IMD temperature variations with different means and variances on a common scale. If a shading scheme for significance is defined, Z series similar to that in Fig. 2e can be collapsed onto one horizontal axis, and the results for multiple series can be compared. Using a cool–warm shading scheme showing negative and positive significances at 95%, 99%, and 99.9% confidence levels, Fig. 4a plots the Z series for each of the 92 black T42 grid locations in Fig. 1. As mapped in Fig. 4b, each grid location is in one of four color-coded U.S. regions, and the Fig. 4a analyses are arranged such that as the grid number increases from 1 to 92 the associated grid's collapsed *Z* series are plotted from bottom to top. The light-green region in Fig. 4b coincides with those grid locations whose *Z* series projects onto Fig. 3b's PC2 series (grids 1–30), while *Z* series in the yellow (grids 31–41), red (grids 42–66), and blue (grids 67–92) regions are generally correlated with PC1.

The *Z* series for grids 1–30 in Fig. 4a show the PC2 multidecadal temperature cycle, which is similar to that found in Fig. 2e's southern Alabama temperature record. The cycle's cool period was notably significant in some areas, with *Z* statistics <-4.0 at grids 12 and 15–29 between the late 1950s and the 1980s. When measured in terms of the incidence of low-ranked cool years during 1919–2008, the warming hole in annual temperatures is geographically centered over these southern U.S. grid locations. The 1958–85 southern Alabama (grid 25) cool period noted above produced the most significant concentration of low-ranked annual temperature values ($Z = -5.486$) in all of the analyses of the gridded USHCN data. Although prolonged warm periods are rare in the data before 1960, the southern areas at grids 22–29 all experienced a multidecadal warm regime between the early 1930s and the late 1950s. All of the 92 grid location's *Z* series end in warm regimes, but the warm periods in Fig. 4b's green region (grids 1–30) during 1998–2008 have later onset and occur at a generally lower significance level than the remaining 62 grid locations.

Similar to PC1's general form in Fig. 3a, the late century warm period over the United States outside of the Southeast is first evident in the mid- to late 1970s in Fig. 4a. Although recent warming over the southeastern United States began after the mid-1990s and is generally less significant, the southern Florida grid location (grid 31) shows highly significant warmth during 1985–2008 ($Z = 5.210$). Warm regimes with similar timing are also found at grid locations that include coastal areas of the northeast (grids 33, 35–38). Highly significant ($Z > 5.0$) continuous late-century warm periods are found at grid locations in Fig. 4b's blue-shaded western region (grids 74, 75, 77, 81–83, 85–87, and 90). The most significant warm period during 1919–2008 at any grid location ($Z = 6.147$) is found at grid 82 in southern California during 1980–2008.

4. The *Z* error metric

The *Z* error (ZE) is simply the difference between the *Z* series of modeled and observed annual temperature records at the same grid location during year *t*:

$$\mathrm{ZE}_t = Z_{\mathrm{Mod}(t)} - Z_{\mathrm{Obs}(t)}. \tag{4}$$

Figure 5a plots the time series of modeled FGOALS annual temperature at the southern Alabama grid location, with the superimposed *Z* series resulting from the running MWZ analysis. Figure 5b shows the grid's corresponding observed USHCN temperature record and *Z* series, and Fig. 5c plots the ZE series, that is, the difference between the *Z* series in Figs. 5a,b. The measure of the difference between a grid point's modeled and observed IMD temperature variations is the mean absolute *Z* error (MAZE) of the ZE series during 1919–2008; that is,

$$\mathrm{MAZE} = \frac{1}{90}\sum_{t=1}^{90}\mathrm{abs}(\mathrm{ZE}_t). \tag{5}$$

As constant *Z* normalizing parameters are used for each of the running sample sizes, time series with identical sequences of rankings result in identical *Z* traces and MAZE values of 0.0. Differences in the timing or significance of cool or warm regimes produce MAZE > 0.0.

Greater dissimilarity in modeled and observed *Z* series leads to increased positive MAZE values, but because the ZE metric and Eq. (5) are new statistical measures, there are no standard methods for estimating MAZE thresholds that indicate the significant difference between two *Z* series. To estimate such a threshold, MAZE values were calculated between the 92 U.S. temperature *Z* series in Fig. 4a and *Z* series derived from 90-yr annual temperature records from PCM, FGOALS, and MIROC preindustrial control (PICNTRL) runs at the same grid location. The PICNTRL runs, which vary in length in the three models, are forced by fixed solar variability and by preindustrial ozone and greenhouse gas concentrations. As these model runs are marked by random internal climate variability characteristic of an ocean–atmosphere system in equilibrium, the *Z* series derived from the control run temperature series are representative of random climate variation. As a result, a distribution of MAZE values calculated between the PICNTRL *Z* series and Fig. 4a's observed *Z* series can be used as a null distribution to estimate a MAZE dissimilarity threshold. Such a null distribution was formed from nine MAZE values calculated at each of the 92 grid locations. Three values were derived from years 51–140 of three 150-yr FGOALS PICNTRL runs, and the remaining six values were calculated from three nonoverlapping 90-yr segments from the single available PCM and MIROC PICNTRL runs. The resulting 828 MAZE values are approximately normally distributed, with a minimum of 1.157, a maximum of 6.447, and a median of 3.186. MAZE values calculated between the observed and modeled temperature records will be considered dissimilar when they are greater than that distribution's one-sided 95th percentile ($\varepsilon = 1.819$).

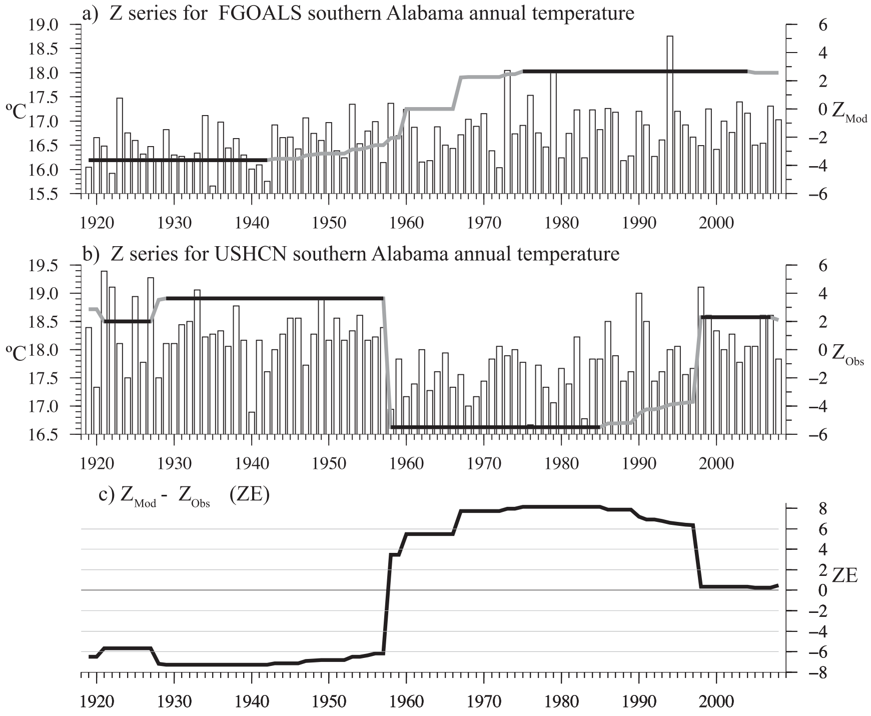

FIG. 5. (a) Time series of modeled FGOALS annual temperature at the southern Alabama grid location (grid 25) with a superimposed *Z* series resulting from the running MWZ procedure. (b) As in (a), but for the southern Alabama grid average of USHCN annual temperatures. (c) The *Z* error series formed by subtracting the USHCN *Z* value for each year in (b) from the corresponding FGOALS *Z* value in (a).

5. *Z* series, *Z* error, and MAZE patterns of the FGOALS, MIROC, and PCM models

Figures 6a–c show the results of the running MWZ analyses for the FGOALS, MIROC, and PCM ensemble mean temperature records, respectively. In each figure, the *Z* series are arranged by grid location as mapped in Fig. 4b, and are plotted according to Fig. 4a's cool–warm shade coloring scheme. As a result, there is direct correspondence between the 1919–2008 *Z* series of modeled temperature and Fig. 4a's observed *Z* series at each grid location. Figures 6d–f show the three models' ZE series at each grid location, after they are collapsed on a horizontal axis using the blue–red-shade coloring scheme shown above (Fig. 6d). Given the form of Eq. (4), negative (positive) ZE values are marked by blue (red) shades and show when and where the model was generating lower- (higher-) ranked annual mean temperatures relative to observed rankings. The horizontal bars of Figs. 6g–i plot the MAZE values for the corresponding ZE trace to the left in Figs. 6d–f, and those figures' green vertical lines mark the ε MAZE dissimilarity threshold.

Figure 6a–c's modeled *Z* series are generally consistent with the first PC of observed *Z* series in Fig. 3a, that is, with relatively cool temperatures before the early to mid-1970s, followed by warm periods of varying significance that extend to the end of the data record. An exception to that pattern of behavior is seen in the FGOALS *Z* series (Fig. 6a), which shows highly significant ($Z < -3.29$) cool conditions over almost every U.S. grid area between 1919 and the mid-1940s. At grids 1–40 the model shows a shift from early cool conditions to insignificant *Z* values during the 1960s, and then to significant positive values after the early 1970s. This gradual shift is more consistent with a noisy linear trend in temperature over these eastern grid areas during 1919–2008 (see the discussion of Fig. A1e in the appendix). In contrast, the MIROC and PCM model's *Z* series (Figs. 6b,c) show fairly uniform evidence of a shift from cool regimes to warm conditions over almost all of the 92 grid areas in the mid- to late 1970s. In Fig. 6b the MIROC model's preceding cool regimes at grid locations east of the Rockies (grids 1–65) during the mid-1940s to mid-1970s are generally more significant than the corresponding PCM *Z* values in Fig. 6c. However, while the 'PC1 like' variation in observed temperatures is limited to grid locations in the yellow, red, and blue shaded areas of the United States in Fig. 4b, the trend or regimelike behavior in modeled temperatures is found at almost all 92 U.S. grid locations in

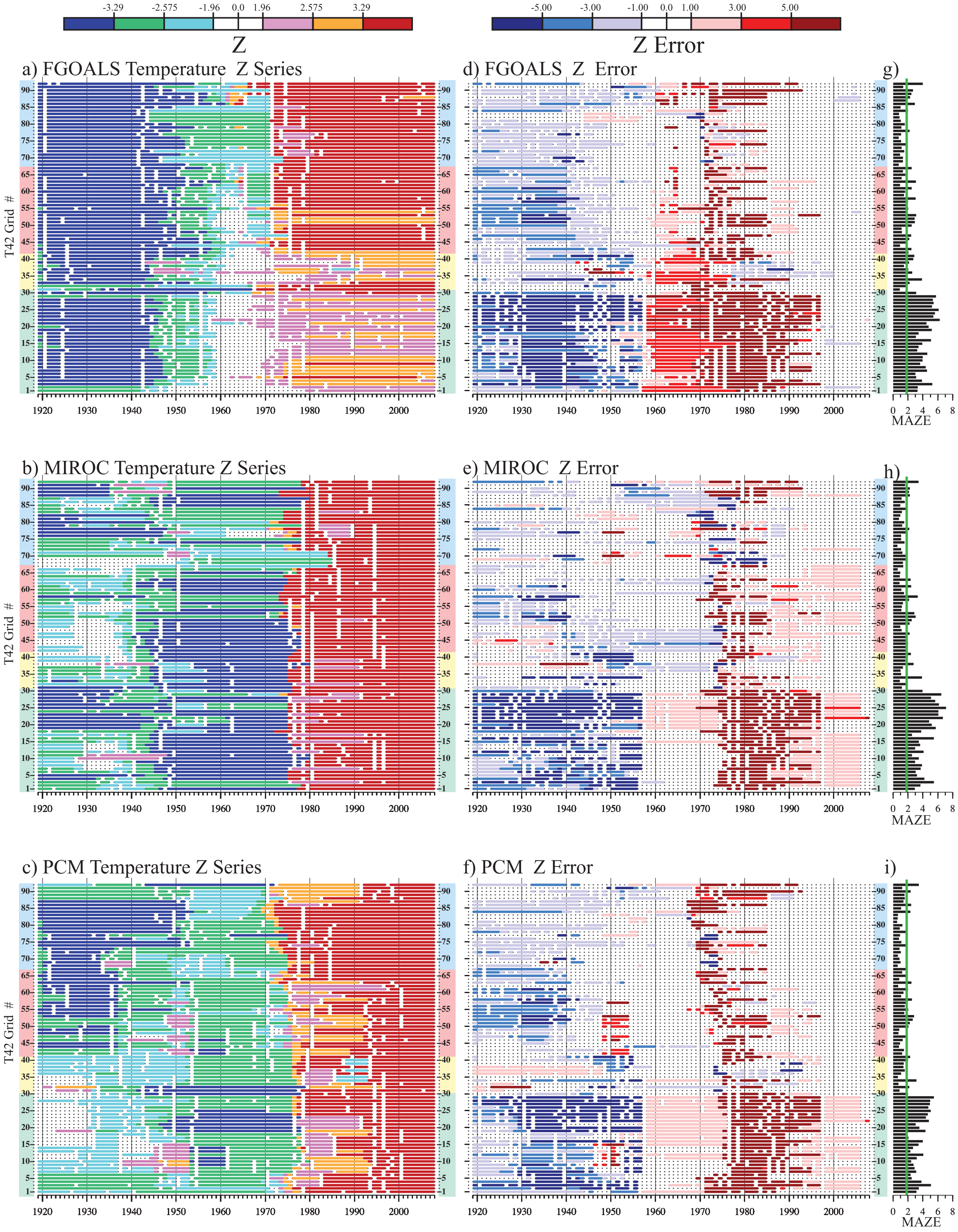

FIG. 6. (a) As in Fig. 4a, but for *Z* series derived from FGOALS 1919–2008 annual temperature. (b) As in Fig. 4a, but for MIROC annual temperature *Z* series. (c) As in Fig. 4a, but for PCM annual temperature *Z* series. Shading schemes for (a)–(c) are indicated in the legend above (a). (d) FGOALS ZE series calculated by subtracting USHCN-derived *Z* series in Fig. 4a from the corresponding FGOALS *Z* series in (a). (e) MIROC ZE series calculated by subtracting USHCN-derived *Z* series in Fig. 4a from the corresponding MIROC *Z* series in (b). (f) PCM ZE series calculated by subtracting USHCN-derived *Z* series in Fig. 4a from the corresponding PCM *Z* series in (c). Shading schemes for (d)–(f) are indicated in the legend above (d). (g) FGOALS MAZE at each T42 grid location. The green vertical line marks a MAZE dissimilarity threshold (1.819). (h) As in (g), but for MIROC MAZE. (i) As in (g), but for PCM MAZE.

Figs. 6a–c. The observed low-frequency temperature cycle apparent at grids 1–30 in Fig. 4a (i.e., the "PC2 like" variation of Fig. 3b) is not found in the corresponding modeled *Z* traces. As a result, the grid locations where the observed Z traces correlate with PC2 (i.e., the southeastern U.S. areas marked by the light-green shade in Figs. 4a,b) show clear patterns of ZE variation in Figs. 6d–f.

In Fig. 6a, the FGOALS *Z* traces show highly significant cool periods at grids 1–30 before the late 1950s, while the observed data for the same time period shows significant warmth at grids 21–29. The model's tendency to produce low-ranked temperatures where high-ranked temperatures were observed is evident in the dark-blue-shaded negative ZE values in Fig. 6d at those grid locations before 1958. Although FGOALS temperatures were trending upward in the southeastern United States after 1958, that year marked the beginning of the warm hole period in observed temperatures. The drop in observed temperature rankings combined with rising modeled rankings leads to red-shaded positive ZE values in Fig. 6d, which increase in magnitude until the end of the cool period in the late 1980s. The warming in the observed temperatures during the 1990s and after 2000 brought the observed and modeled *Z* series values into relative agreement ($|\mathrm{ZE}| < 1.0$). The MIROC and PCM ZE values at grids 1–30 in Figs. 6e,f are similar in form to the FGOALS values, but because those models were producing significantly cool conditions in the Southeast during the late 1950s to mid-1970s, the modeled and observed *Z* trace values are closer in value ($0.0 < \mathrm{ZE} < 3.0$). Although the cool period in the Southeast lasted until the mid- to late 1980s in the observed data in Fig. 4b, after the mid-1970s all three models made a transition to significant warmth in the Southeast. This transition resulted in high positive ZE values ($\mathrm{ZE} > 5.0$) at grids 1–29 in Figs. 6d–f, which inversely mirror the latter half of the southeastern warm hole period in Fig. 4a. The significance of the late southeastern warming in the FGOALS model in Fig. 6a is generally consistent with that found in the observed *Z* values in Fig. 4a ($1.96 < Z < 3.29$). In contrast, the MIROC and PCM results reflect a greater incidence of high-ranked annual temperatures ($Z > 3.29$) at southeastern grid locations, which produce weakly positive ZE values ($1.0 < \mathrm{ZE} < 3.0$) during the 1990s and after 2000 in Figs. 6e,f.

Over U.S. grid areas where the *Z* series correlate with Fig. 3a's first PC (i.e., grid locations in the yellow, red, and blue areas in Fig. 4b), the *Z* traces of modeled and observed temperatures are generally in closer agreement. In Fig. 6d the FGOALS ZE values tend to be weakly negative in these areas before 1950, which shows generally lower temperature rankings relative to the observed rankings. In Figs. 6d–f brief runs of high positive ZE values are evident during the 1970s and 1980s, which shows that the onset of late-century warming in all three models preceded the observed warming. However, unlike the late period modeled *Z* values in the southeastern PC2 region, the *Z* values in the PC1 areas are closer to the observed *Z* values. This agreement is most evident in the absence of shaded ZE values in the FGOALS and PCM models after 1990 in Figs. 6d,f.

The relative agreement of observed and modeled *Z* series in the PC1-correlated areas can be seen in the corresponding MAZE values in Figs. 6g–i. At those grid locations (grids 31–92) values less than ε (1.819) are fairly common in all three models. In contrast, in the PC2-correlated region (grids 1–30) MAZE values consistent with significant dissimilarity ($\mathrm{MAZE} > \varepsilon$) are more the rule in all the models. Over the southeastern grids where the PC2 low-frequency temperature cycle is most clear in Fig. 4a's *Z* series (grids 21–29), the MAZE values for all three models clearly exceed the ε threshold.

The ability to reproduce all of the observed U.S. *Z* series is measured by an overall figure of merit (FM) equal to the grid-area-weighted spatial average of each model's 92 MAZE values:

$$\mathrm{FM} = \frac{\sum_{\mathrm{igrd}=1}^{92} \mathrm{area}(\mathrm{igrd}) \times \mathrm{MAZE}(\mathrm{igrd})}{\sum_{\mathrm{igrd}=1}^{92} \mathrm{area}(\mathrm{igrd})}. \tag{6}$$

Like the MAZE values, lower FM values show greater agreement between the observed and modeled IMD climates. When calculated for the FGOALS, MIROC, and PCM models, those values for U.S. surface temperature during 1919–2008 are 2.759, 2.496, and 2.054, respectively.

6. Summary and discussion

A time series analysis method used previously to identify intra- to multidecadal climate variations in historical data records (Mauget 2003, 2004; Cordero et al. 2010; Masiokas et al. 2010) was adapted to test three AOGCMs' abilities to reproduce observed U.S. temperature regimes. The approach samples annual temperature rankings over moving time windows, converts those samples to Mann–Whitney *U* statistics, and then normalizes the *U* statistics into *Z* statistics (Figs. 2a–c). The process is repeated using moving windows of varying duration to identify the most significant ranking regimes in a data record (Fig. 2d). In the present application, *Z* statistics for years between these regimes are assigned according to the year's maximum absolute *Z* value in all the running analyses to form continuous *Z* series (Fig. 2e).

A principal component analysis of the Z series at 92 continental U.S. grid locations shows that observed annual temperature during 1919–2008 generally projected onto one of two low-frequency modes: a PC1 mode with cool conditions before the late 1960s followed by a shift to warmth after the mid-1980s (Figs. 3a,c), and a southeastern PC2 mode with warm conditions before the late 1950s and after the late 1990s, separated by a cool regime between those times (Figs. 3b,d).

Because it is based on the analysis of rankings, the process that produces the Z series is resistant and robust. The method's moving window approach allows it to identify ranking regimes of varying length and arbitrary timing, and to expose a wide range of IMD climate variation (see the appendix). The reexpression of ranking sequences into Z statistics transforms the data variation onto a common scale and allows for a simplified graphic analysis of the observed (Fig. 4a) and modeled Z series (Figs. 6a–c) at a model's grid locations. This normalization also allows for the calculation of a Z error (ZE) metric that provides a year-to-year comparison of a grid location's observed and modeled Z series (Fig. 5). Graphic analysis of the resulting ZE series shows when and where the model's IMD temperature regimes depart from observations (Figs. 6d–f). Mean absolute Z error (MAZE) figures calculated from the ZE series provide a figure of merit at each grid location (Figs. 6g–i), which can be spatially averaged to assign an overall figure of merit (FM) for each model. Because the running MWZ method is a sensitive and robust way of detecting IMD climate regimes, the ZE series provide a correspondingly sensitive way of contrasting modeled and observed climate patterns of behavior. However, because this process tests rankings, it gives no information about magnitude-related errors (e.g., biases in the means or variances of modeled climate data relative to observations). Such biases would have to be identified through a complementary statistical analysis [e.g., the method of Lambert and Boer (2001) or Taylor (2001)].

Although the NCAR Parallel Climate Model (PCM) produces the lowest FM value, the three model's IMD temperature patterns in Figs. 6a–c are generally similar during 1919–2008. That is, although the models either reproduce the PC1 mode of U.S. temperature shown in Figs. 3a,c, or show trendlike behavior, none reproduces the cyclic PC2 mode of Figs. 3b,d. Based on modeling simulations with forcing scenarios that included observed SST, Robinson et al. (2002) proposed that eastern U.S. cooling during 1951–97 was the consequence of warming tropical Pacific SST and the associated increased cloud cover. Although this might explain the PC2 mode's cool regime, their simulations did not consider the multidecadal temperature cycle of which that cool period was a part. Even so, as the FGOALS, MIROC, and PCM 20C3M coupled model runs were not initialized with observed SST, their results suggest that the model's failure to generate the PC2 temperature cycle might be due to an inability to reproduce the correct oceanic boundary forcing. Of course, as noted in section 1, it is unsurprising that 20C3M model runs initialized by model-generated preindustrial boundary conditions might fail to generate the observed Pacific SST in the latter half of the twentieth century. However, some of the proposed CMIP-5 30-yr hindcast experiments will be initialized with observed ocean state and sea ice conditions in 1960 and 1980 (Taylor et al. 2008; Meehl et al. 2009). With the addition of natural and anthropogenic radiative boundary forcings, those experiments might be expected to reproduce observed IMD regimes in SST and the associated climate effects during those 30-yr time windows. With some modifications (e.g., limiting the sampling windows to 6–15-yr duration), the graphic analysis method and MAZE and FM metrics described here could be used to test the CMIP-5 model's ability to reproduce those effects.

Developing models that reproduce the decadal features of historical climate may be an important first step in producing useful projections of future decadal climate. Such work may depend, in turn, on sensitive diagnostic methods that can help to identify and troubleshoot the shortcomings of current models and initialization schemes. Given its ability to compare observed and modeled IMD climate regimes at a model's grid resolution, the running MWZ method may be used as in Fig. 6, that is, as a diagnostic tool to show a model's space–time response to specified initial conditions and full forcing and compare that response with observations. This demonstration evaluated U.S. surface temperatures, but the method could be used with any climate variable that provides reliable observed values at a model's grid resolution. For example, Latif et al. (2004, 2010) cite the North Atlantic as an area of strong and potentially predictable decadal climate variation. Thus, modeled SST anomaly values might be compared with gridded reanalysis values (e.g., Kaplan et al. 1998) over that ocean basin. The method might also be adapted to current detection and attribution techniques (e.g., Zwiers and Zhang, 2003; Santer et al. 2003; Zhang et al. 2006; Lee et al. 2006; Hegerl et al. 2007) to calculate grid resolution response patterns for various climate forcings, and then solve for the combination of those responses that yield the lowest FM value. Patterns of anthropogenically forced and natural decadal variabilities might be estimated by comparing response patterns in initialized and uninitialized hindcast experiments (Solomon et al. 2011). As climate change caused by increasing CO_2 levels in the coming decades may be

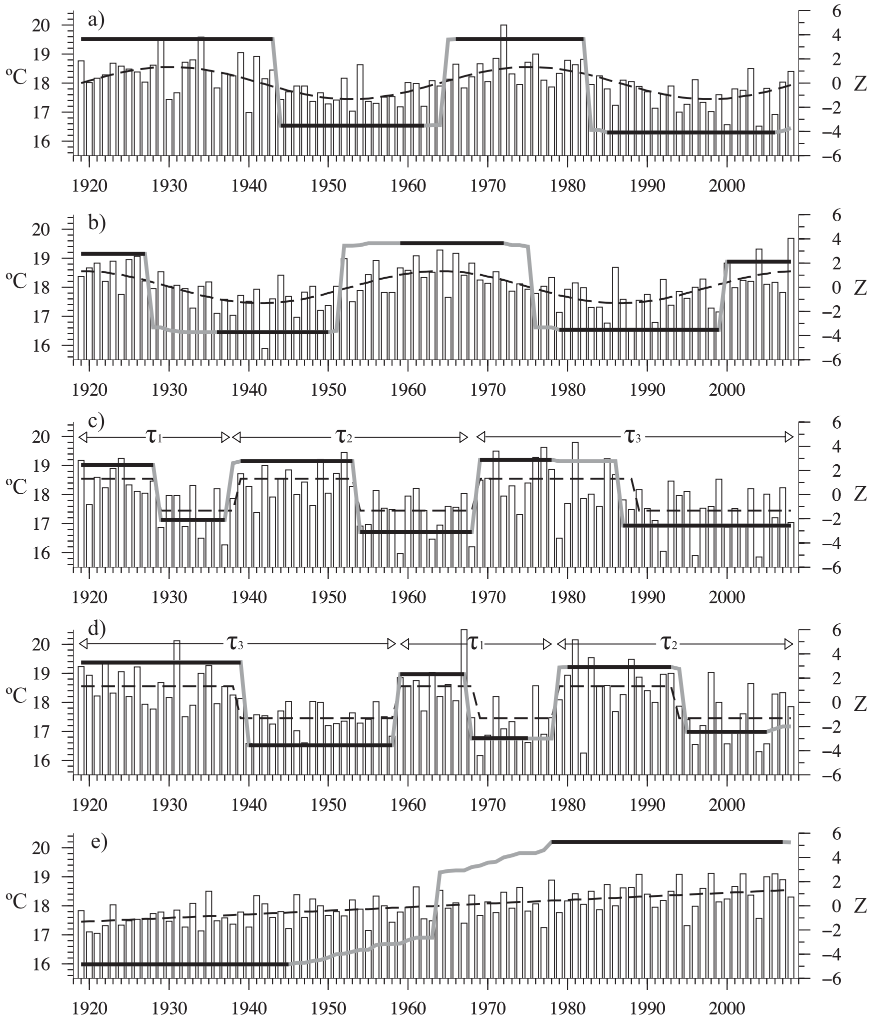

FIG. A1. MWZ series for five 90-yr artificial temperature records modeled after the Fig. 2a SATMP temperature series. Black and gray traces show *Z* series as in Fig. 2e, dashed traces show idealized climate signals, and the bar values show the sum of the those climate signals and AR(3) noise processes. (a) Regular harmonic cycle with a 45-yr period. (b) As in (a), but with phase retarded by $\pi/2$. (c) Alternating cyclic regimes with increasing periods of $\tau_1 = 20$ yr, $\tau_2 = 30$ yr, and $\tau_3 = 40$ yr. (d) As in (c), but with the sequence of the cyclic regimes reordered as τ_3–τ_1–τ_2. (e) A linearly increasing temperature signal during 1919–2008.

modulated by natural decadal climate variability (Smith et al. 2007; Keenlyside et al. 2008; Ting et al. 2009; Hurrell et al. 2010; Semenov et al. 2010), such analysis might aid in developing models that respond correctly to anthropogenic and natural forcing, and combinations of those forcings.

Acknowledgments. We thank the organizers of the Program for Climate Model Diagnosis and Intercomparison and the developers of FGOALS g1.0, the MIROC3.2 medium-resolution model, and NCAR's Parallel Climate Model. Without their work, this analysis would not have been possible. The Parallel Climate Model project (information online at www.cgd.ucar.edu/pcm) is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy and the Directorate for Geosciences of the National Science Foundation. Thanks also to Clare Cordero for editing and proofreading and to preliminary reviewers Mariano Masiokas and Kevin Clifford for their helpful comments. All figures were produced using Generic Mapping Tools (Wessel and Smith, 1995). This work was supported in part by NSF's Faculty Early Career Development (CAREER) Program, Grant ATM-0449996 (E. Cordero and P. Brown). The USDA is an equal opportunity provider and employer.

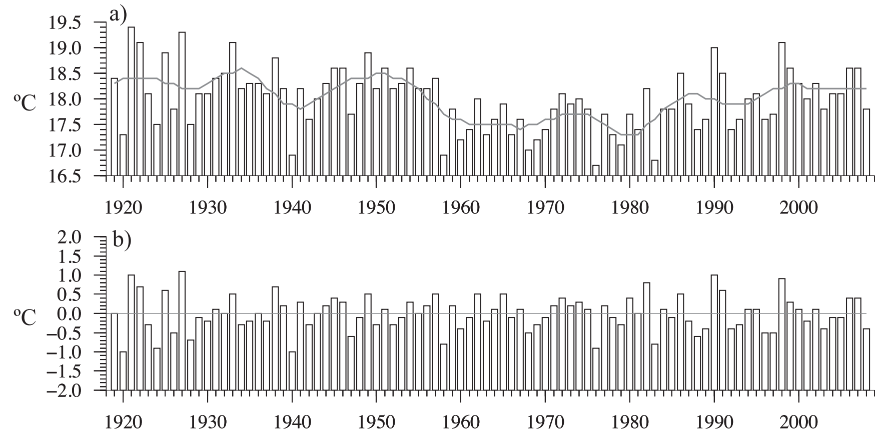

FIG. A2. (a) SATMP series of annual temperature as in Fig. 2a. Gray trace shows the series' low-frequency component ($\nu < 10^{-1}$ yr) as defined by a low-pass Lanczos filter. (b) High-pass residual series derived from subtracting the SATMP low-frequency component from the annual SATMP values.

APPENDIX

Artificial Temperature Time Series Tests

To demonstrate the running MWZ method's general ability to resolve varying climate signals, Figs. A1a–e shows the Z series that result when the method is applied to time series with known signal and noise characteristics. These series consist of autoregressive (AR) noise processes that have been added to idealized low-frequency climate signals to form five artificial 90-yr temperature records. The Fig. A1 series are modeled after the Fig. 2 southern Alabama temperature (SATMP) time series, which is reproduced in Fig. A2a. To define a ratio of low-frequency signal variance to AR noise variance in the artificial series similar to that of the SATMP series, the SATMP series was first subjected to a low-pass Lanczos filter (Duchon 1979). As the shortest window used in the running MWZ sampling is 6 yr, with a corresponding cyclic period of ~12 yr, this filter was assigned a half-power cutoff frequency of $\nu = 10^{-1}$ yr. The variance of the SATMP low-frequency component, shown in gray in Fig. A2a, is 0.124°C^2, while the variance of the high-pass residual series (Fig. A2b) is 0.207°C^2. The ratio of the variance of the artificial low-frequency signals (the dashed traces in Figs. A1a–e) to the variance of the series' AR noise component was thus set to 0.600 in each of the examples by adjusting the AR processes' variance. This signal-to-noise (s–n) ratio is an important factor in determining whether the method can ''find'' the signals in these artificial series; higher ratios ensure success, while higher AR noise variance and lower ratios obscure the signal and ensure failure. The s-n ratios for the annual temperature series at the 92 grid locations in Fig. 4b range between 0.282 and 1.04, with a median value of 0.502. The 0.600 s–n ratio used in the Fig. A1 tests is thus above the median, but not unrepresentative. The AR noise processes added to the artificial signals were formed using an AR(3) model, as that model yielded the minimum Akaike information criteria score (Akaike 1974) in modeling the Fig. A2b high-pass residual series as AR(1), AR(2), and AR(3) processes.

Figure A1a shows the Z trace for a 45-yr harmonic temperature signal with a mean equal to that of the SATMP series (18.0°C), while Fig. A1b shows the Z trace for the same signal retarded in phase by $\pi/2$. The general similarity of those figures' cyclic Z traces to their corresponding idealized climate signals, and the retarded phase of the Fig. A1b Z series, shows the method's ability to detect the existence and phase of smoothly varying cyclic variation. Figure A1c shows the Z trace resulting from a signal consisting of three abrupt cyclic regimes with increasing periods of $\tau_1 = 20$ yr, $\tau_2 = 30$ yr, and $\tau_3 = 40$ yr. Figure A1d repeats the test of Fig. A1c, but with the ordering of the three temperature cycles changed to a τ_3–τ_1–τ_2 sequence. Although random number generator initialization and/or low s-n ratios can prevent the detection of the temperature signal in these tests, the method is generally successful in identifying the Fig. A1c regime signals and the rearranged signals in Fig. A1d. This demonstrates an ability to detect abrupt climate regimes of varying duration and arbitrary onset. However, the similarity in the abrupt shifts between significant positive and negative Z series values in Figs. A1a–d shows the approach is not effective at detecting the difference between smoothly harmonic and abruptly transitional cyclic patterns of behavior. Figure A1e shows the Z series that results from a linear increase in the temperature signal with the same magnitude as the Figs. A1c,d regime changes. Thus, the linear trend Z signature under the 0.600 s–n conditions consists of the most significant Z statistics occurring with opposite sign at the beginning and end of a time series, separated by periods of gradually decreasing significance. As the s–n ratio is decreased in these trend tests, the significance of the cool and warm periods at the series' beginning and end decreases, and midperiod Z values become insignificant.

REFERENCES

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Automat. Contr.,* **AC-19,** 716–723.

Bell, J., P. B. Duffy, C. Covey, and L. Sloan, 2000: Comparison of temperature variability in observations and sixteen climate model simulations. *Geophys. Res. Lett.,* **27,** 261–264.

Boer, G. J., 2000: Analysis and verification of model climate. *Numerical Modeling of the Global Atmosphere in the Climate System,* P. Mote and A. O'Neill, Eds., Kluwer Academic, 59–104.

Brekke, L. D., M. D. Dettinger, E. P. Maurer, and M. Anderson, 2008: Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Climatic Change,* **89,** 371–394.

Cane, M., 2010: Decadal predictions in demand. *Nat. Geosci.,* **3,** 231–232, doi:10.1038/ngeo823.

Collins, M., and Coauthors, 2006: Interannual to decadal climate predictability in the North Atlantic: A multimodel ensemble study. *J. Climate,* **19,** 1195–1203.

Cordero, E. C., W. Kessomkiat, J. Abatzoglou, and S. Mauget, 2010: The identification of distinct patterns of California temperature trends. *Climatic Change,* **108,** 357–382.

Covey, C., K. M. AchutaRao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor, 2003: An overview of results from the Coupled Model Intercomparison Project (CMIP). *Global Planet. Change,* **37,** 103–133.

Cubasch, U., and Coauthors, 2001: Projections of future climate change. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 525–582.

Duchon, C. E., 1979: Lanczos filtering in one and two dimensions. *J. Appl. Meteor.,* **18,** 1016–1022.

Folland, C. K., and Coauthors, 2001: Observed climate variability and change. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 99–181.

——, J. Shukla, J. L. Kinter, and M. Rodwell, 2002: The Climate of the Twentieth Century Project. *CLIVAR Exchanges,* Vol. 7, International CLIVAR Project Office, Southampton, United Kingdom, 37–39.

Giorgi, F., and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. *J. Climate,* **15,** 1141–1158.

Hasumi, H., and S. Emori, Eds., 2004: K-1 coupled model (MIROC) description. K-1 Tech. Rep. 1, Center for Climate System Research, University of Tokyo, Tokyo, Japan, 34 pp. [Available online at www.ccsr.u-tokyo.ac.jp/kyosei/hasumi/MIROC/tech-repo.pdf.]

Hegerl, G. C., and M. R. Allen, 2002: Origins of model–data discrepancies in optimal fingerprinting. *J. Climate,* **15,** 1348–1356.

——, and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 663–745.

Hollander, M., and D. A. Wolfe, 1999: *Nonparametric Statistical Methods*. Wiley and Sons, 787 pp.

Hurrell, J. W., and Coauthors, 2010: Decadal climate prediction: Opportunities and challenges. *Proc. OceanObs'09: Sustained Ocean Observations and Information for Society,* Vol. 2, Venice, Italy, European Space Agency Publ. WPP-306. [Available online at http://www.oceanobs09.net/proceedings/cwp/cwp45/index.php.

International ad hoc Detection and Attribution Group, 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *J. Climate,* **18,** 1291–1314.

Janis, M. J., K. G. Hubbard, and K. T. Redmond, 2004: Station density strategy for monitoring long-term climatic change in the contiguous United States. *J. Climate,* **17,** 151–162.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.,* **103,** 18 567–18 589.

Karoly, D. J., and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Climate,* **18,** 4337–4343.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature,* **453,** 84–88.

Knutson, T. R., T. L. Delworth, K. Dixon, and R. J. Stouffer, 1999: Model assessment of regional surface temperature trends (1949–97). *J. Geophys. Res.,* **104,** 30 981–30 996.

——, and Coauthors, 2006: Assessment of twentieth-century regional surface temperature trends using the GFDL–CM2 coupled models. *J. Climate,* **19,** 1624–1651.

Kunkel, K. E., X. Liang, J. Zhu, and Y. Lin, 2006: Can CGCMs simulate the twentieth-century "warming hole" in the central United States? *J. Climate,* **19,** 4137–4153.

Lambert, S. J., and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dyn.,* **17,** 83–106.

Latif, M., and Coauthors, 2004: Reconstructing, monitoring, and predicting multidecadal-scale changes in the North Atlantic thermohaline circulation with sea surface temperature. *J. Climate,* **17,** 1605–1614.

——, M. Collins, H. Pohlmann, and N. Keenlyside, 2006: A review of predictability studies of the Atlantic sector climate on decadal time scales. *J. Climate,* **19,** 5971–5987.

——, and Coauthors, 2010: Dynamics of decadal climate variability and implications for its prediction. *Proc. OceanObs'09: Sustained Ocean Observations and Information for Society,* Vol. 2, European Space Agency Publ. WPP-306. [Available online at http://www.oceanobs09.net/proceedings/cwp/cwp53/index.php.]

Lau, K. M., and H. Weng, 1995: Climate signal detection using wavelet transform: How to make a time series sing. *Bull. Amer. Meteor. Soc.,* **76,** 2391–2402.

Lee, T. C. K., F. W. Zwiers, X. Zhang, and M. Tsao, 2006: Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. *J. Climate,* **19,** 5305–5318.

Mann, H. B., and D. R. Whitney, 1947: On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.,* **18,** 50–60.

Masiokas, M. H., R. Villalba, B. H. Luckman, and S. Mauget, 2010: Intra- to multidecadal variations of snowpack and streamflow records in the Andes of Chile and Argentina between 30° and 37°S. *J. Hydrometeor.,* **11,** 822–831.

Mauget, S., 2003: Multidecadal regime shifts in U.S. streamflow, precipitation, and temperature at the end of the twentieth century. *J. Climate,* **16,** 3905–3916.

——, 2004: Low frequency streamflow regimes over the central United States: 1939–1998. *Climatic Change,* **63,** 121–144.

Meehl, G. A., W. M. Washington, T. M. L. Wigley, J. M. Arblaster, and A. Dai, 2003: Solar and greenhouse gas forcing and climate response in the twentieth century. *J. Climate,* **16,** 426–444.

——, ——, C. Ammann, J. M. Arblaster, T. M. L. Wigley, and C. Tebaldi, 2004: Combinations of natural and anthropogenic forcings and twentieth-century climate. *J. Climate,* **17,** 3721–3727.

——, and Coauthors, 2007a: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–845.

——, C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, and R. J. Stouffer, 2007b: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

——, and Coauthors, 2009: Decadal prediction. *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Mendenhall, W., D. D. Wackerly, and R. L. Sheaffer, 1990: *Mathematical Statistics with Applications*. PWS-Kent, 818 pp.

Menne, M. J., and C. N. Williams Jr., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate,* **18,** 4271–4286.

——, and ——, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate,* **22,** 1700–1717.

——, ——, and R. S. Vose, 2009: The U.S. Historical Climatology Network Monthly Temperature Data, version 2. *Bull. Amer. Meteor. Soc.,* **90,** 993–1007.

Murphy, J., and Coauthors, 2009: Towards prediction of decadal climate variability and change. *Proc. World Climate Conference-3 (WCC-3),* Geneva, Switzerland, World Meteorological Organization, 24 pp. [Available at http://www.clivar.org/organization/decadal/references/WCC3_Decadal_WhitePaper.pdf.]

Pan, Z., R. W. Arritt, E. S. Takle, W. J. Gutowski, C. J. Anderson, and M. Segal, 2004: Altered hydrologic feedback in a warming climate introduces a ''warming hole.'' *Geophys. Res. Lett.,* **31,** L17109, doi:10.1029/2004GL020528.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci. USA,* **106,** 8441–8446.

Pohlmann, H., M. Botzet, M. Latif, A. Roesch, M. Wild, and P. Tschuck, 2004: Estimating the decadal predictability of a coupled AOGCM. *J. Climate,* **17,** 4463–4472.

Räisänen, J., 2007: How reliable are climate models? *Tellus,* **59A,** 2–29.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 589–662.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311.

Robinson, W. A., R. Reudy, and J. E. Hansen, 2002: General circulation model simulations of recent cooling in the east-central United States. *J. Geophys. Res.,* **107,** 4748, doi:10.1029/2001JD001577.

Santer, B. D., and Coauthors, 2003: Contributions of anthropogenic and natural forcing to recent tropopause height changes. *Science,* **301,** 479–483, doi:10.1126/science.108412.

Semenov, V. A., M. Latif, D. Dommenget, N. Keenlyside, A. Strehz, T. Martin, and W. Park, 2010: The impact of North Atlantic–Arctic multidecadal variability on Northern Hemisphere surface air temperature. *J. Climate,* **23,** 5668–5677.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science,* **317,** 796–799, doi:10.1126/science.1139540.

Solomon, A., and Coauthors, 2011: Distinguishing the roles of natural and anthropogenically forced decadal climate variability. *Bull. Amer. Meteor. Soc.,* **92,** 141–156.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192.

——, R. J. Stouffer, and G. A. Meehl, cited 2008: A summary of the CMIP5. Experimental design. [Available online at http://www.clivar.org/organization/wgcm/references/Taylor_CMIP5.pdf.]

Tebaldi, C., R. W. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *J. Climate,* **22,** 1469–1481.

Torrence, C., and G. P. Compo, 1998: A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.,* **79,** 61–78.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 235–336.

U.S. National Weather Service, cited 2000: Cooperative Observer Program (COOP). U.S. National Weather Service. [Available online at http://www.nws.noaa.gov/os/coop/Publications/coop.PDF.]

Vera, C., and Coauthors, 2009: Needs assessment for climate information on decadal time scales and longer. *Proc. World Climate Conference-3 (WCC-3),* Geneva, Switzerland, World Meteorological Organization. [Available online at http://www.wcc3.org/sessions.php?session_list=WS-9#doc.]

Washington, W. M., and Coauthors, 2000: Parallel Climate Model (PCM) control and transient simulations. *Climate Dyn.,* **16,** 755–774.

Wessel, P., and W. H. F. Smith, 1995: New version of the Generic Mapping Tools released. *Eos, Trans. Amer. Geophys. Union,* **76,** 329.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 464 pp.

Yu, Y., R. Yu, X. Zhang, and H. Liu, 2002: A flexible global coupled climate model. *Adv. Atmos. Sci.,* **19,** 169–190.

——, X. Zhang, and Y. Guo, 2004: Global coupled ocean–atmosphere general circulation models in LASG/IAP. *Adv. Atmos. Sci.,* **21,** 444–455.

Zhang, X., F. W. Zwiers, and P. A. Stott, 2006: Multimodel multisignal climate change detection at regional scale. *J. Climate,* **19,** 4294–4307.

Zwiers, F. W., 2002: The 20-year forecast. *Nature,* **416,** 690–691.

——, and X. Zhang, 2003: Towards regional scale climate change detection. *J. Climate,* **16,** 793–797.