June 2012

# Reproduction of Twentieth Century Intradecadal to Multidecadal Surface Temperature Variability in Radiatively Forced Coupled Climate Models

Patrick T. Brown
*San Jose State University*

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

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

# Reproduction of twentieth century intradecadal to multidecadal surface temperature variability in radiatively forced coupled climate models

Patrick T. Brown,[1] Eugene C. Cordero,[1] and Steven A. Mauget[2]

Received 14 September 2011; revised 19 January 2012; accepted 20 January 2012; published 14 June 2012.

[1] Coupled Model Intercomparison Project 3 simulations that included time-varying radiative forcings were ranked according to their ability to consistently reproduce twentieth century intradecadal to multidecadal (IMD) surface temperature variability at the 5° by 5° spatial scale. IMD variability was identified using the running Mann-Whitney Z method. Model rankings were given context by comparing the IMD variability in preindustrial control runs to observations and by contrasting the IMD variability among the ensemble members within each model. These experiments confirmed that the inclusion of time-varying external forcings brought simulations into closer agreement with observations. Additionally, they illustrated that the magnitude of unforced variability differed between models. This led to a supplementary metric that assessed model ability to reproduce observations while accounting for each model's own degree of unforced variability. These two metrics revealed that discernable differences in skill exist between models and that none of the models reproduced observations at their theoretical optimum level. Overall, these results demonstrate a methodology for assessing coupled models relative to each other within a multimodel framework.

**Citation:** Brown, P. T., E. C. Cordero, and S. A. Mauget (2012), Reproduction of twentieth century intradecadal to multidecadal surface temperature variability in radiatively forced coupled climate models, *J. Geophys. Res.*, *117*, D11116, doi:10.1029/2011JD016864.

## 1. Introduction

[2] Coupled general circulation models (CGCMs) have become the primary tools for making projections of future surface temperature changes. In order for these projections to instill confidence, models should have a history of accurately reproducing spatial and temporal characteristics of past climate variation. Despite this, much of the work regarding CGCM validation has focused on time mean state climate statistics that infer little about climate variation through the historical record [e.g., *Giorgi and Mearns*, 2002; *Knutson et al.*, 2006; *Pierce et al.*, 2009; *Reichler and Kim*, 2008; *Tebaldi et al.*, 2005]. While it is important that a model be able to produce the correct climatic mean and variance of a given variable, ideally models should also be able to reproduce historical variations of such variables in time.

[3] One reason for restricting climate model evaluation to time mean state statistics is that free-running CGCMs cannot be expected to reproduce unforced variability throughout the historical record (variability that emerges simply from the internal dynamics of the coupled climate system). For instance, the El Niño–Southern Oscillation (ENSO) influences global temperature change on the interannual time scale [*Neelin et al.*, 1998]. Free-running CGCMs have their own ENSO variability, and the phasing of this cycle is not expected to match historical observations. Even on the multidecadal scale, continental temperature is heavily constrained by oceanic temperature variation [*Compo and Sardeshmukh*, 2009]. As a result, it is likely that unforced oceanic circulations, such as those associated with Pacific Decadal Variability (PDV) and the Atlantic Meridional Overturning Circulation (AMOC), have heavily influenced global surface temperature evolution on the multidecadal scale [*DelSole et al.*, 2011; *Kravtsov and Spannagle*, 2008; *Swanson et al.*, 2009; *Wu et al.*, 2011; *Zhang et al.*, 2007]. A portion of this oceanic variability, however, may be constrained by external radiative forcings [*Goosse and Renssen*, 2004; *Ottera et al.*, 2010]. In this case, retrospective CGCM simulations may be expected to correctly reproduce some portion of the multidecadal scale temperature variability over the twentieth century. Indeed, the influence of external radiative forcings on the global surface temperature has been shown to exceed that of internal variability as the time scale of interest increases from the interannual to the multidecadal level [*Solomon et al.*, 2011; *Hegerl et al.*, 2007].

[4] The present study evaluates intradecadal to multidecadal (IMD) temperature variability in retrospective CGCM simulations from the World Climate Research Program's

[1]Department of Meteorology and Climate Science, San Jose State University, San Jose, California, USA.
[2]Plant Stress and Water Conservation Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Lubbock, Texas, USA.

Corresponding author: P. T. Brown, Department of Meteorology and Climate Science, San Jose State University, San Jose, CA 95192-0104, USA. (patrick.brown@sjsu.edu)

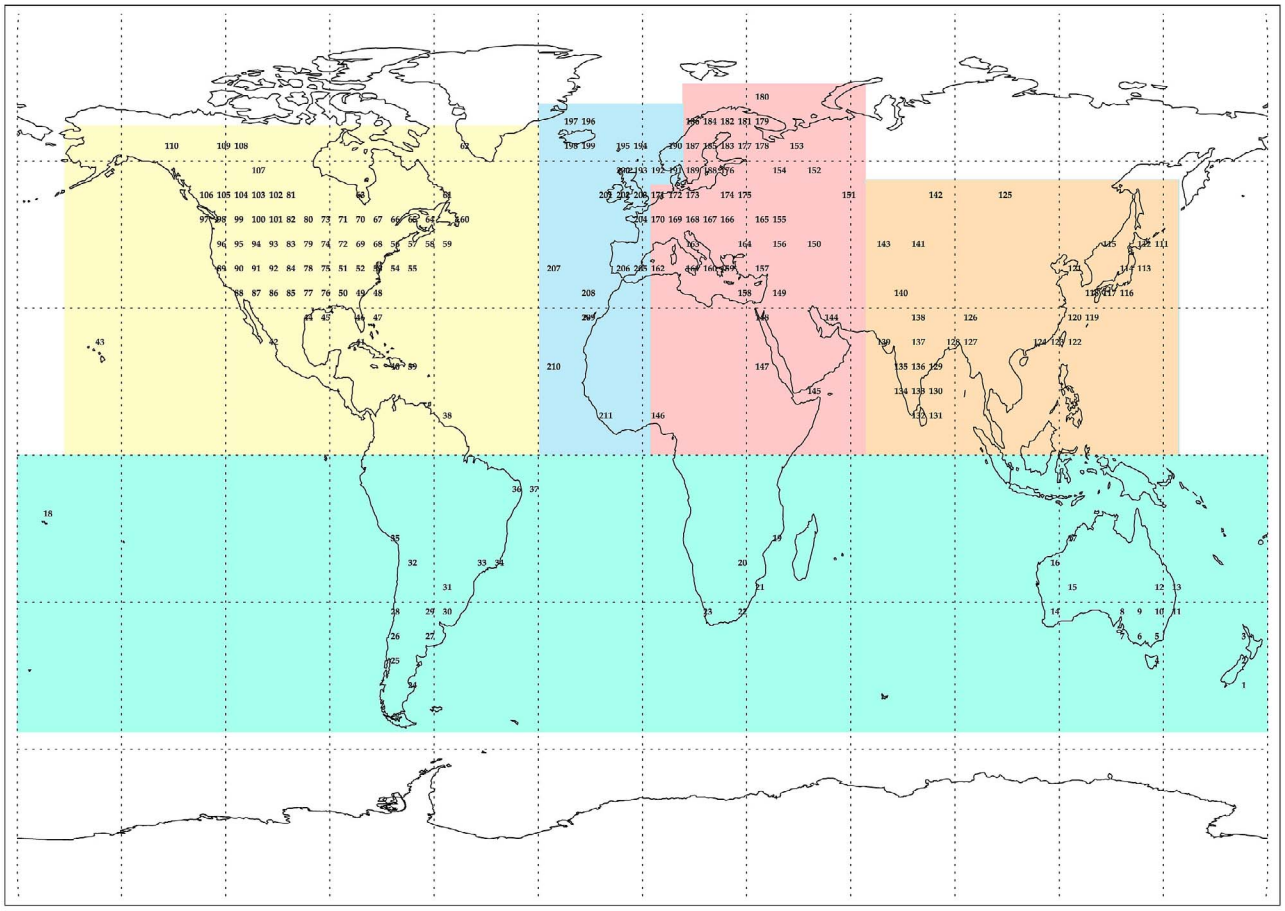

**Figure 1.** The 211 grid locations investigated, where the colors identify regions in Figures 4, 5, and 8–10.

(WCRP) Coupled Model Intercomparison Project phase 3 (CMIP 3) multimodel data set. The simulations investigated here included historical estimates of radiative boundary forcings but were initialized from arbitrary times in each model's preindustrial control run. Therefore, this study implicitly examines the degree to which observed IMD variability is reproducible via the inclusion of time-varying forcings. A methodology is developed that assesses model ability to reproduce observations while also accounting for variation in the magnitude of modeled unforced variability.

## 2. Data

[5] Observational surface temperature data was obtained from the HadCRUT3v combined land and marine gridded surface temperature data set [*Jones et al.*, 1999; *Rayner et al.*, 2003]. Only those grid points with completely uninterrupted monthly records between 1902 and 1999 were investigated (for details on the aggregation of raw station data to the regular 5° by 5° grid, see *Jones et al.* [1999, and references therein]). The 211 locations that met this criterion are numbered in Figure 1. The monthly temperature series at each location was averaged to form an annual temperature series. It should be noted that these grid areas represent a limited portion of the Earth's surface (∼10%) and it is unknown if the results discussed below would hold for a data set covering the entire globe. Additionally, these grid points have a strong spatial emphasis on the Northern Hemisphere, particularly the United States and Europe. The impact of this particular spatial distribution is investigated in section 4.2.

[6] The CGCM runs that were evaluated were those of the CMIP 3 climate of the twentieth century experiment (20C3M [*Meehl et al.*, 2007a]) as well as the associated preindustrial control experiment (PICNTRL). Models were considered that had at least three ensemble members available for the 20C3M experiment as well as at least 220 years available for the PICNTRL experiment. The 12 models that met these criteria are listed in Table 1.

[7] Before assessment was conducted, model output was bilinearly interpolated to the same 5° by 5° grid used in the HadCRUT3v data set (spatial weighing was also performed as is discussed in section 4.1). This interpolation was designed to minimize biases associated with grid resolution differences between models. Additionally, the 5° by 5° spatial scale was larger than any of the individual models' grid scales. Because of this, subgrid-scale influences (e.g., local topography) were unlikely to have introduced any significant biases when model output was compared with observations. Also, this spatial scale is near the minimum required for the influence of twentieth century external

**Table 1.** Information for the 12 Models Investigated[a]

| Model | Originating Group(s) | 20C3M Forcing | Number of 20C3M Ensemble Members | Number of 20C3M Versus 20C3M Pair Wise Comparisons | Number of 98 Year PICNTRL Segments |
|---|---|---|---|---|---|
| CCSM3 | National Center for Atmospheric Research | GHG, So, Su, V, $O_3$, H, B | 8 | 28 | 9 |
| CGCM3.1(T47) | Canadian Centre for Climate Modeling and Analysis | GHG, Su | 5 | 10 | 15 |
| CSIRO-Mk3.0 | CSIRO Atmospheric Research | GHG, Su, $O_3$ | 3 | 3 | 4 |
| ECHAM5/MPI-OM | Max Planck Institute for Meteorology | GHG, Su, H, $O_3$ | 4 | 6 | 6 |
| ECHO-G | Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model and Data group. | GHG, So, Su, V, $O_3$ | 5 | 10 | 4 |
| GFDL-CM2.0 | U.S. Department of Commerce, NOAA, Geophysical Fluid Dynamics Laboratory | GHG, $O_3$, Su, V, So, L | 3 | 3 | 6 |
| GFDL-CM2.1 | U.S. Department of Commerce, NOAA, Geophysical Fluid Dynamics Laboratory | GHG, $O_3$, Su, V, So, L | 3 | 3 | 6 |
| GISS-EH | NASA Goddard Institute for Space Studies | GHG, So, Su, V, $O_3$, H, L, B | 5 | 10 | 5 |
| GISS-ER | NASA Goddard Institute for Space Studies | GHG, So, Su, V, $O_3$, H, L, B | 9 | 36 | 6 |
| MIROC3.2(medres) | Center for Climate System Research (University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC) | GHG, So, V, $O_3$, L, Su, B, H | 3 | 3 | 6 |
| MRI-CGCM2.3.2 | Meteorological Research Institute | GHG, H, So, Su, V, $O_3$ | 5 | 10 | 4 |
| PCM | National Center for Atmospheric Research | GHG, So, Su, V, $O_3$ | 4 | 6 | 8 |

[a]The 20C3M forcing abbreviations are identified as follows: GHG, well-mixed greenhouse gases; H, halocarbons; Su, sulfate tropospheric aerosols; L, land use; V, volcanic aerosols; $O_3$, ozone; B, other aerosols; So, solar irradiance.

forcings to be apparent on surface temperature [*Karoly and Wu*, 2005].

## 3. Identification of IMD Variability in Time Series Data

[8] Modeled and observed time series were analyzed via the running Mann-Whitney Z (MWZ) method. This method has been used previously to identify significant IMD variation in observational time series [*Masiokas et al.*, 2010; *Mauget*, 2003, 2004; *Cordero et al.*, 2011]. Here, this work has been extended to comparisons between modeled and observed time series in a similar manner to *Mauget et al.* [2012]. Specifically, the running MWZ method is used to highlight IMD regimes of arbitrary onset and duration in both modeled and observed annual temperature time series (from 1902 to 1999) at each of the 211 grid locations investigated (Figure 1). This method is described in the six steps below and illustrated in Figure 2.

[9] 1. All the data values in the given annual temperature series at a particular grid location (e.g., Figure 2a) are ranked from lowest to highest.

[10] 2. The temperature series is sampled by a moving window of incrementally varying size from 6 to 30 years. For each window size, every possible sample in the time series is investigated. For example, when the window is 6 years in length, the first sample contains the years 1902–1907, the second sample contains the years 1903–1908 and the last sample contains the years 1994–1999. This same procedure is followed for all the remaining window sizes (7–30 years in duration).

[11] 3. A Mann-Whitney U statistic [*Mann and Whitney*, 1947] is calculated for each of the samples described in step 2. The U statistic is defined as the total number of years outside the sample that precede each year inside the sample in rank. Stated in another way, the number of nonsample years that were cooler than each year within a given sample is summed to obtain the U statistic. The U statistic can also be calculated using

$$U_I = R_I - \frac{n_I[n_I + 1]}{2}, \tag{1}$$

where $R_I$ is the sum of the ranks for the sample $I$ and $n_I$ is the size of the window in years [*Mendenhall et al.*, 1990; *Wilks*, 1995].

[12] 4. U statistics for each sample are normalized via a Z transformation,

$$Z = \frac{U - \mu}{\sigma}. \tag{2}$$

The Z transformation relies on the assumption that random sampling would produce a Gaussian distribution of U values between two extreme cases. The first extreme case would be that the given sample contains all the lowest ranking years in the temperature series, and the second extreme case would be that the given sample contains all the highest ranking years in the temperature series. Thus for a 98 year time series divided into a 10 year sample and an 88 year nonsample, the highest possible U statistic would occur when the sample contains the 10 highest-ranked years (U = 88*10). Conversely, the lowest possible U statistic would result from a sample containing the 10 lowest ranked years (U = 0*10). The mean ($\mu$) of the null distribution used in the Z transformation is simply the average of the two extreme cases,

$$\mu = \frac{n_I n_{II}}{2}, \tag{3}$$

where $n_I$ is the number of years inside the sample and $n_{II}$ is the number of years outside the sample. The standard

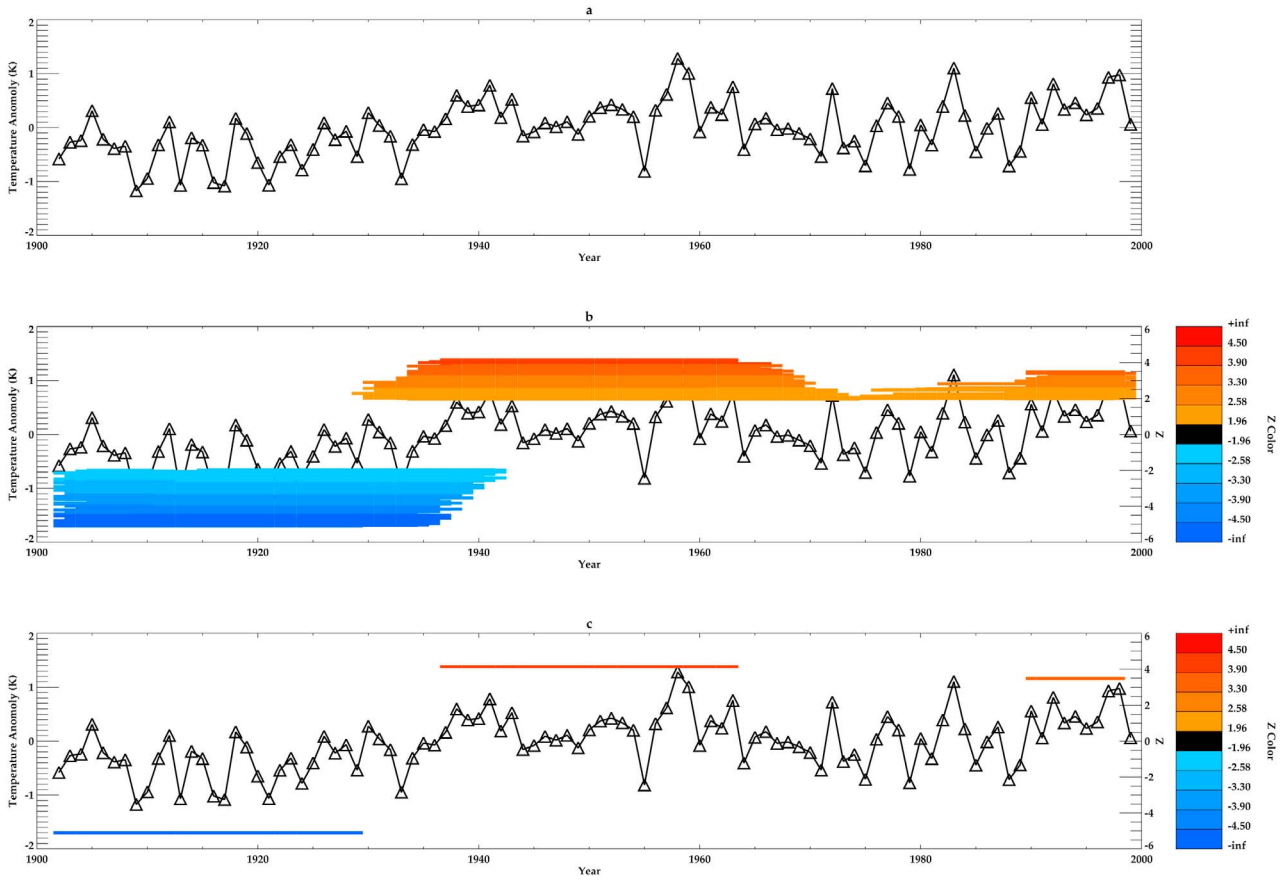


**Figure 2.** (a) Temperature series for HadCRUT3v data at location 42 in Figure 1. (b) All significant ($|Z| > 1.96$) runs of U statistics in the time series. (c) Original temperature series and the most significant periods that occurred over nonoverlapping samples plotted together. All gap years that are not part of a Z period are considered to have a Z value of zero.

deviation ($\sigma$) of this null distribution may be estimated by [*Mendenhall et al.*, 1990]

$$\sigma = \sqrt{\frac{n_I n_{II}[n_I + n_{II} + 1]}{12}}. \quad (4)$$

[13] 5. After all U statistics for each sample are Z transformed, all the periods significant at a 95% confidence level ($|Z| > 1.96$) are pooled (e.g., Figure 2b).

[14] 6. The significant periods are screened with the intent of identifying the samples with the highest absolute Z values, at all window lengths, that do not overlap in time. This is accomplished in two steps. First the period with the highest absolute Z value is identified and second all overlapping periods with lesser absolute Z values are deleted. This two-step process continues indefinitely until no remaining periods overlap in time (e.g., Figure 2c). Any year that is not included in one of the remaining significant Z periods is assigned a value of zero so that a continuous series of Z values can be produced.

[15] The resulting Z series highlights IMD variability in time series data. The central assumption of the method is that climate variations consist of simple, noncyclic, ranking regimes that occur over a range of time scales and have arbitrary onset times. This inclusive assumption gives the method an advantage over some filtering methods that may be considered simpler or more intuitive. For instance, spectral filters make assumptions about periodicity in the data that may be inappropriate for the study of temperature series affected by noncyclical radiative boundary forcings. The ranking-based nature of the method also allows for making observed versus modeled and model versus model comparisons with biased variance and/or means. Finally, the method is resistant to outliers that may be unrepresentative of the temperature regime being experienced at a given time (e.g., 1955 in Figure 2).

[16] As mentioned above, the process of Z transforming the U statistics was done by utilizing a normal null distribution, which assumed random and independent sampling. This is in contrast to how the method had been employed in some previously published works [*Masiokas et al.*, 2010; *Mauget*, 2003, 2004; *Cordero et al.*, 2011]. In these cases, the null distribution was created from Monte Carlo trials that attempted to embody a hypothetical climate characterized by year-to-year temperature persistence but no IMD variability. Thus, the purpose of these previous null distributions was to constrain the Z values that could be considered significant in

the traditional sense. In the present application, this method would be unfavorable because Monte Carlo trials would produce differing σ values for modeled and observed null distributions (equation (2)) of a given window length. Accordingly, identically ranked temperature series would result in slightly differing Z series. Because the overall goal of this analysis was to find dissimilarity between modeled and observed IMD variability, this would have been unacceptable. In order to guarantee that identical ranking sequences resulted in identical Z series, it was necessary to fix null parameters like those in equations (3) and (4) (see *Mauget et al.* [2012] for further discussion).

[17] Because observed and simulated temperature series display persistence, the assumption of random and independent sampling inevitably caused the detection of what would traditionally be considered "spurious" significance. Because of this issue, unusually large absolute Z values are seen in much of this analysis (e.g., $|Z| > 6.0$). It must be emphasized that this was not a problem in the current application of the method as the ultimate goal of the MWZ transformation was to highlight IMD variability in modeled and observed climate data so that they could be compared. In typical significance testing in climate analysis, the goal is to test observed variability against a null hypothesis that assumes a hypothetical stationary climate condition. As these hypothetical conditions are normally required to possess the interannual persistence of observed data, that hypothesis has to account for the persistence. Since the goal is not to test for nonstationarity in a time series but instead is to compare the ranking sequences of modeled and observed time series, consistent normalization of the U statistic trumps the use of realistic assumptions regarding the autocorrelation present in the data.

[18] The MWZ method uses 6 years as the minimum window length for the detection of significance because this is near the threshold where external forcings would likely be nearly undetectable through the noise of unforced variability. It is true that many external forcings (such as increases in the atmospheric concentrations of long-lived greenhouse gasses) are only expected to dominate unforced variability on the multidecadal scale and beyond [*Boer*, 2011]. Despite this, other external radiative forcings that are incorporated in the retrospective (20C3M) runs likely have a nonnegligible influence on the intradecadal scale [*Solomon et al.*, 2011]. In particular, volcanic eruptions as well as solar variability can influence surface temperature on time scales less than a decade in length. As the time scale approaches the interannual level, however, temperature changes would likely be dominated by ENSO as well as other modes of unforced variability.

## 4. Model Assessment Relative to Observations

[19] Models were assessed by comparing the Z series from the simulated and observed temperature, where the difference between modeled and observed Z series at each year t is defined by the Z error,

$$\mathrm{ZE}_t = Z_{Modeled(t)} - Z_{Obs(t)}. \tag{5}$$

Modeled Z series were scored on the basis of their mean absolute Z error (MAZE) over the 98 year period from 1902 to 1999,

$$\mathrm{MAZE} = \frac{1}{98}\sum_{t=1}^{98} |ZE_t|. \tag{6}$$

Smaller MAZE scores imply better agreement between the ranking sequences in observed and modeled time series. Therefore, because the MAZE metric is based on rankings, it does not incorporate information about the absolute magnitude of data values in a time series. Figure 3 shows an example of a modeled temperature series with a ranking sequence that matched observations relatively successfully, producing a low MAZE value of 1.00 (Figure 3a), as well as an example of a modeled temperature series with a ranking sequence that matched observations relatively poorly, producing a high MAZE value of 4.14 (Figure 3c). Figures 3b and 3d emphasize the degree to which these two modeled time series differed from observations by representing the magnitude of each year's Z error with a bar. The temperature series in Figure 3a had similar phasing (onset and duration) of IMD variability as was seen in observations (Figure 2). In contrast, the phasing of IMD variability shown in Figure 3c was a relative mismatch to observations.

### 4.1. Historical IMD Variability in Observations

[20] Figure 4 is a spatiotemporal representation of the Z series for the HadCRUT3v observational data set. In this plot, all 211 Z series in the domain have been collapsed onto a single vertical axis that is ordered according to Figure 1. This realization of the data illustrates that significant cool regimes dominated the first third of the twentieth century before warm regimes subsequently became more widespread. The significance and onset of regime changes is seen to vary widely by location. The most significant late century warm regimes were observed across the Southern Hemisphere (grid numbers 1–37) as well as eastern and central Asia (grid numbers 115–150). Some instances of anomalous late century cool regimes are also observed. The most significant late century cool regime was seen in the southeastern United States (grid numbers 49–52 and 75–79). This anomalous feature, often referred to as the U.S. "warming hole" has been well documented elsewhere [e.g., *Robinson et al.*, 2002; *Pan et al.*, 2004; *Kunkel et al.*, 2006] and is usually attributed to unforced variability or possibly local sulfate aerosol loading. In addition to the U.S. warming hole, some late century cool regimes can be seen in the eastern Mediterranean region (grid numbers 156–158) as well as in eastern and northern Europe (grid numbers 195–199).

[21] Overall model performance was assessed by measuring how well a given model's simulations matched the spatiotemporal Z series representation of observations. Mathematically, models were scored on the basis of their figure of merit (FM), which is defined as the area-weighted average of each Z series' MAZE over the 211 locations,

$$\mathrm{FM} = \frac{\sum_{\mathrm{igrd}=1}^{211} \mathrm{area(igrd)MAZE(igrd)}}{\sum_{\mathrm{igrd}=1}^{211} \mathrm{area(igrd)}}. \tag{7}$$

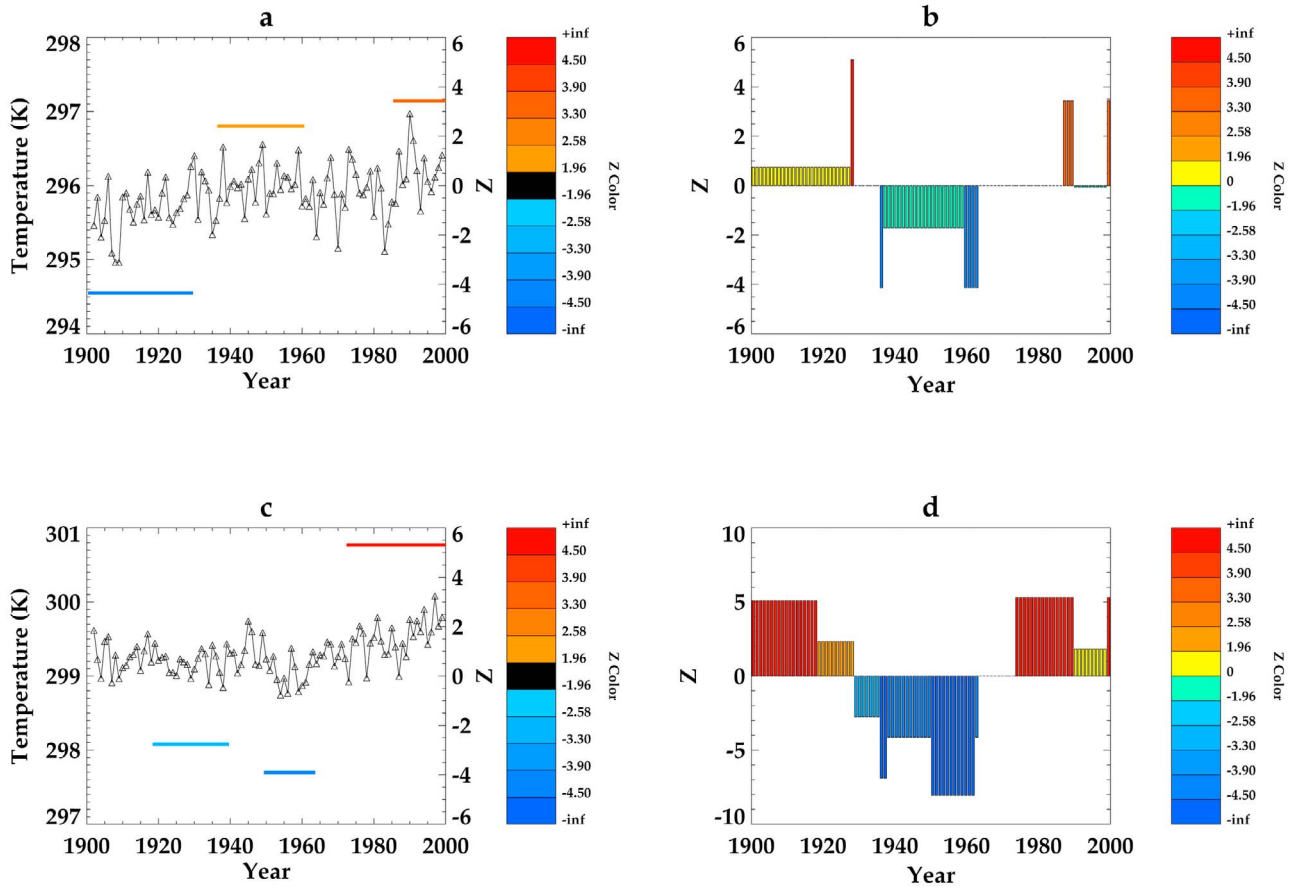


**Figure 3.** (a) Temperature time series and associated Z Series for Echo-G 20CM3 run 2 at location 42. (b) Z error for Echo-G 20CM3 run 2 at location 42. (c) Temperature time series and associated Z Series for Echam5/MPI-OM 20CM3 run 2 at location 42. (d) Z error for Echam5/MPI-OM 20CM3 run 2 at location 42. Note that the Z error plots are relative to the Z series associated with observations at location 42 shown in Figure 2c.

### 4.2. The 20C3M Versus Obs Experiment

[22] Models were ranked on the basis of their 20C3M ensemble members' ability to reproduce observations as quantified through the FM. In total, 57 20C3M simulations were examined from the 12 different models listed in Table 1. Figures 5a and 5c show the spatiotemporal representation of the Z series for two 20C3M simulations from different models. Figure 5a represents the single simulation with the best (lowest) FM, while Figure 5c represents the single simulation with the worst (highest) FM. These results are emphasized in Figures 5b and 5d, which show the spatiotemporal Z error between the modeled and observed Z series. It is apparent that even the best simulations disagree with observations. This is to be expected as unforced variability inevitably played a role in both the simulated and observed evolution of twentieth century climate.

[23] To investigate the effects of this unforced variability, FMs were calculated for every available ensemble member of each model. The 57 FMs associated with each individual ensemble member are plotted in Figure 6, where the models are ranked according to their ensemble mean FM (the simple average of the ensemble member's FMs for each model; model identities can be found in the auxiliary material).[1] For some models, like model 11, the FMs produced by their individual ensemble members were relatively consistent. These ensemble members were initialized with differing oceanic conditions and had only external radiative boundary forcings in common. Therefore, it was inferred that models with little spread in their FMs produced IMD variability that was heavily influenced by external radiative forcings. For these models, there was high confidence that their ensemble mean FM was representative of their ability to reproduce observations. Conversely, models with large spreads in their 20C3M versus observations (20C3M versus Obs) FM distributions, such as model 8, were likely to be more heavily influenced by unforced variability, and thus there was less confidence that the ensemble mean FM for these models was representative of their ability to reproduce observations. This concept was formalized with an application of the Student's t test (two tailed, assuming unequal variance) to each pairwise combination of models to test if their ensemble mean FMs were statistically distinguishable. This test was necessary to make any meaningful statements about a given

[1] Auxiliary materials are available in the HTML. doi:10.1029/2011JD016864.

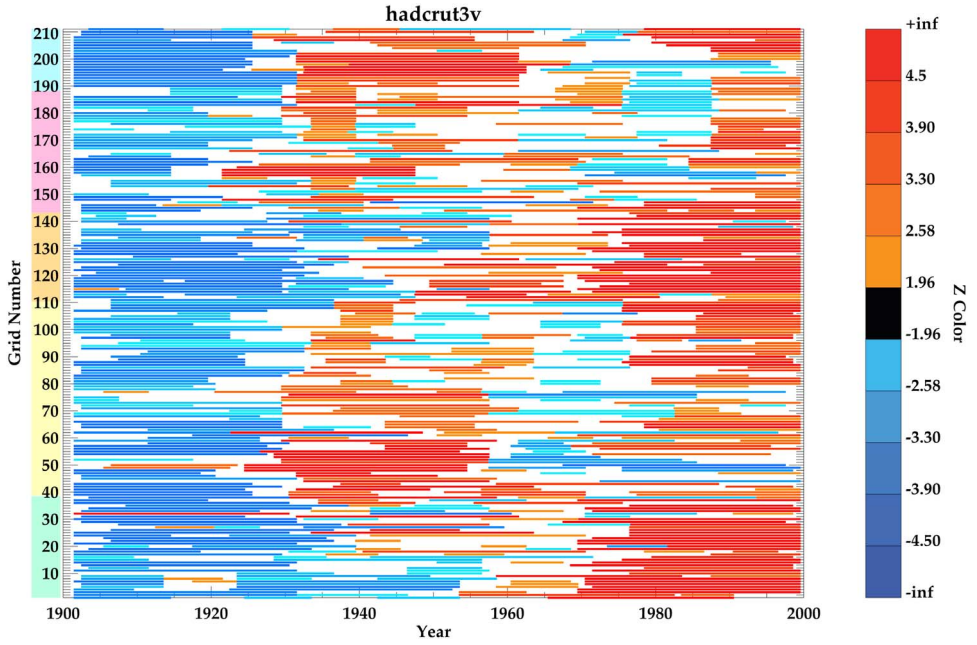


**Figure 4.** Spatiotemporal Z series plot for the HadCRUT3v data set at the 211 locations investigated. The grid numbers and colors shown in the left legend correspond to the numbered locations and colors in Figure 1.

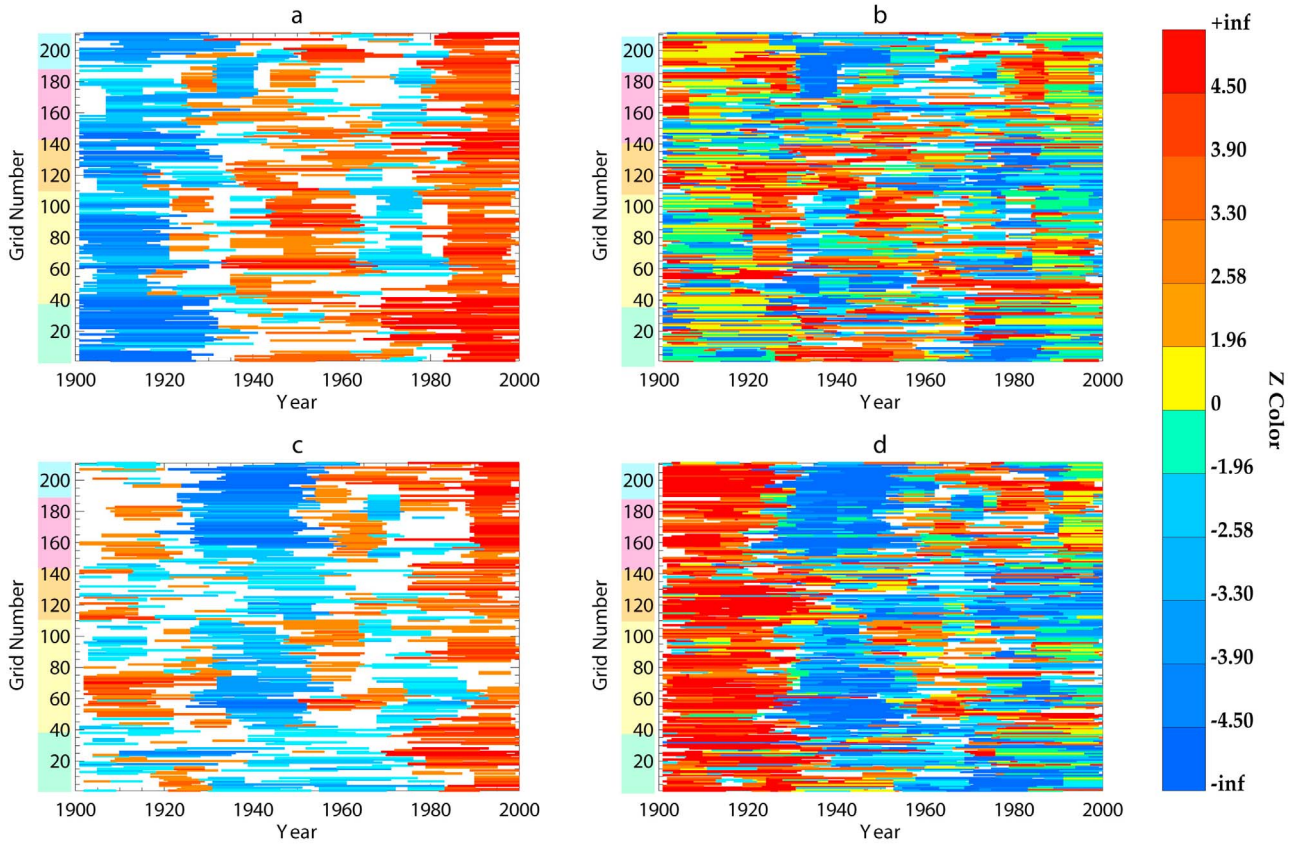


**Figure 5.** (a) Spatiotemporal Z series plot for the 20C3M simulation that was the closest match to observations (ECHAM5/MPI-OM run 3) and (b) its associated spatiotemporal Z error plot. (c) Spatiotemporal Z series plot for the 20C3M simulation that was the largest mismatch to observations (ECHO G run 2) and (d) its associated spatiotemporal Z error plot.

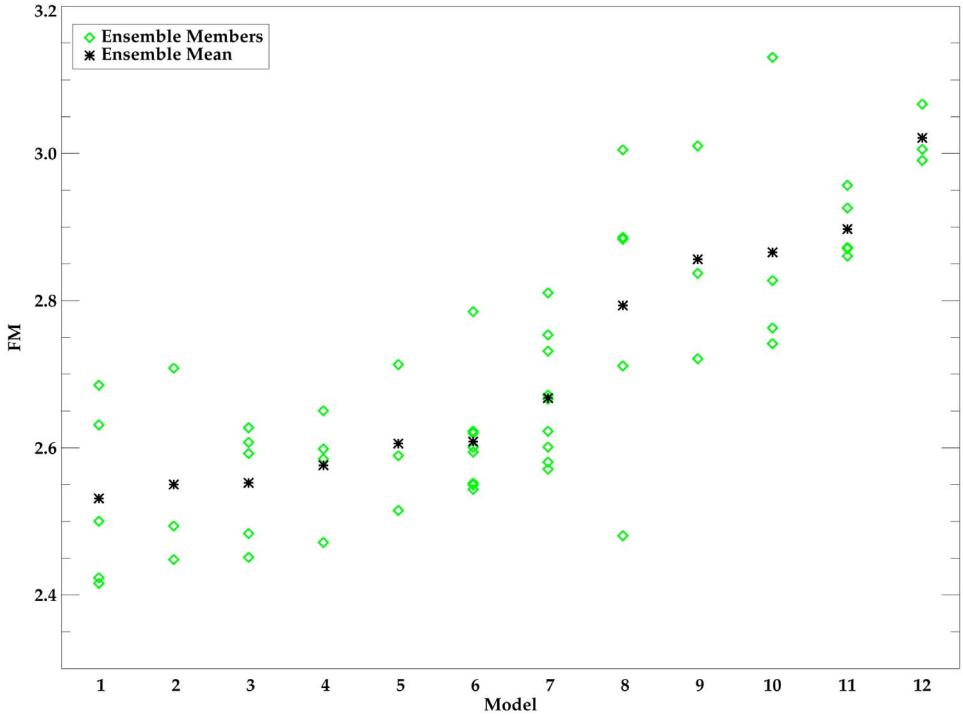


**Figure 6.** Figure of merit (FM) scores for each 20C3M versus Obs experiment (green diamonds) as well as the ensemble mean for each model (black stars). Models are ranked by their ensemble mean FM in ascending order from left to right. The identity of the models can be found in the auxiliary material.

model's ability to reproduce observations relative to any other model. Figure 7 shows the results of these tests. Of the 66 model-model comparisons made, 35 were judged to be statistically unique, while 31 were judged to be statistically indistinguishable (at the 90th percentile). Accordingly, relative model performance could only be assessed in those 35 statistically significant comparisons (colored cells in Figure 7). The models that were ranked 1st, 3rd, and 4th in Figure 6, outperformed the highest number of remaining models (6 each) at a statistically significant level. Conversely, the 12th ranked model was outperformed by 9 of the 11 models ranked ahead of it at a statistically significant level.

[24] The sensitivity of the above results to the spatial domain used was also investigated. Model rankings were recomputed using a subset of the spatial domain shown in

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | |
| 2 | 0.86 | | | | | | | | | | | |
| 3 | 0.75 | 0.98 | | | | | | | | | | |
| 4 | 0.52 | 0.79 | 0.66 | | | | | | | | | |
| 5 | 0.39 | 0.61 | 0.48 | 0.69 | | | | | | | | |
| 6 | 0.25 | 0.55 | 0.25 | 0.51 | 0.97 | | | | | | | |
| 7 | 0.07 | 0.28 | 0.03 | 0.10 | 0.41 | 0.15 | | | | | | |
| 8 | 0.05 | 0.09 | 0.06 | 0.08 | 0.13 | 0.11 | 0.25 | | | | | |
| 9 | 0.04 | 0.06 | 0.05 | 0.06 | 0.08 | 0.09 | 0.14 | 0.63 | | | | |
| 10 | 0.02 | 0.05 | 0.03 | 0.04 | 0.06 | 0.06 | 0.11 | 0.59 | 0.94 | | | |
| 11 | 0.00 | 0.04 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.32 | 0.68 | 0.75 | | |
| 12 | 0.00 | 0.02 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.07 | 0.18 | 0.18 | 0.01 | |

Significant at 99th Percentile
Significant at 95th Percentile
Significant at 90th Percentile

**Figure 7.** Results of Student's t tests between the 20C3M versus Obs FM distributions for each model. The value in each cell is the p value of the t statistic (rounded to the hundreds place) for a comparison between the models in the associated row and column. Statistically significant values are colored according to their level of significance and indicate that the ensemble FM means for those model combinations are distinct from one another.

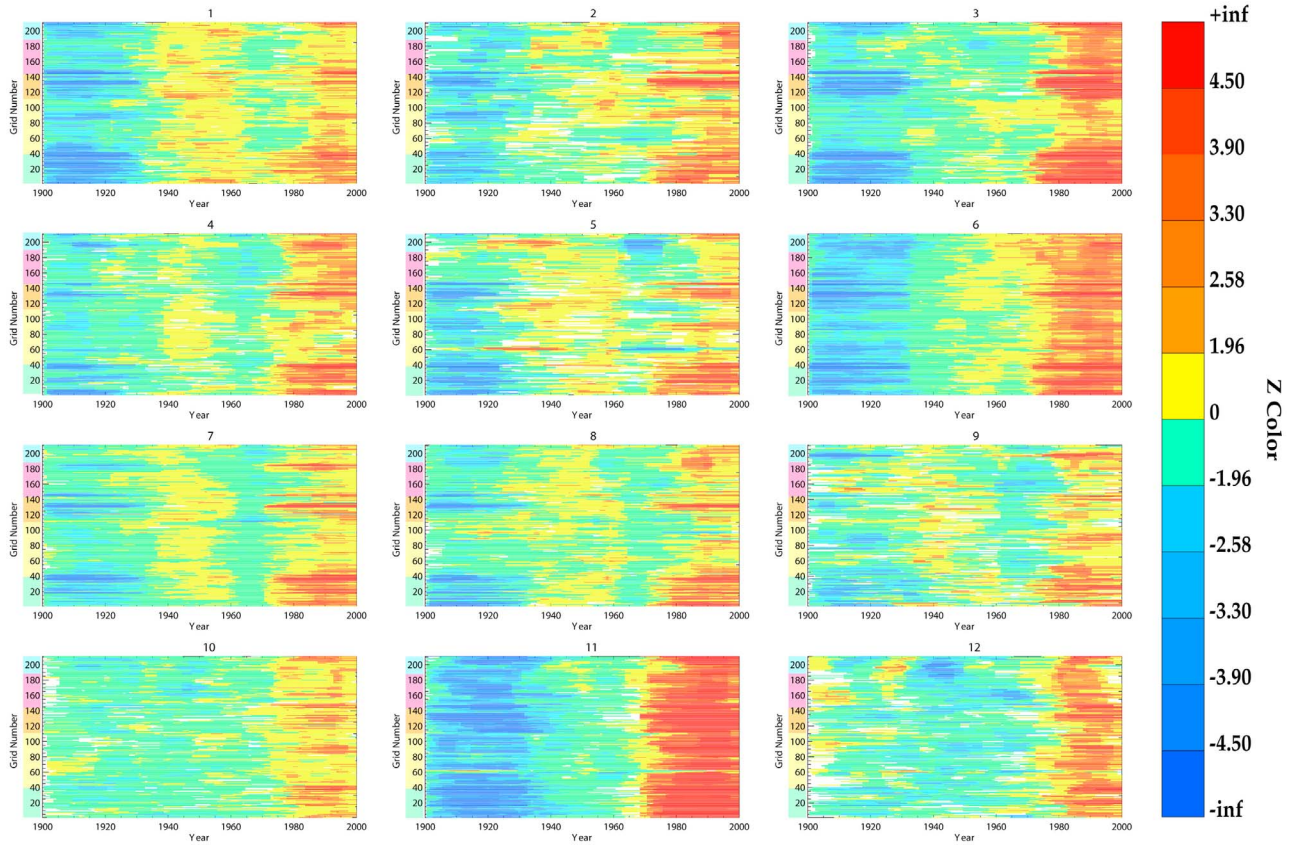


**Figure 8.** Ensemble mean spatiotemporal Z series plots for each model from their 20C3M experiment. Models are ordered by their rank in the 20C3M versus Obs experiment (Figure 6).

Figure 1. This subset consisted of 99 grid cells that were more evenly spaced, and had less of an emphasis on the United States and Europe. This domain resulted in model rankings that were slightly different than the ones shown above. Three models shifted two positions each while four models shifted one position each. The remaining five models remained in the same position that they had occupied previously. These results suggest that the model rankings in question are affected to some degree by the domain used and that this analysis favors those models that reproduce observations well over the United States and Europe. Despite this, the model rankings did not change significantly under a new domain which indicates that the rankings shown above are relatively robust.

## 5. Average Spatiotemporal IMD Variability in Models

[25] Ensemble mean spatiotemporal Z series plots for each model are shown in Figure 8, while Figure 9 shows the corresponding Z error plots. Figures 8 and 9 are averages of the spatiotemporal Z series plots of the models' individual ensemble members and are thus not the result of applying the MWZ methodology to ensemble mean temperature series. It should be noted that the plots associated with numerous ensemble members emphasize the forced signal more than the plots associated with fewer ensemble members.

[26] All the models investigated showed the broad warming pattern that was apparent in observations (Figure 4). Also, many of the models reproduced the general overall cool-warm-cool-warm pattern that was seen at many locations. Most models produced early century regimes that were generally too warm compared with observations. In the middle of the century, however, most models tended to produce Z values that were too cold compared with observations. Late century warm regimes were reproduced relatively successfully by most of the models. However, very few late century cool regimes were present in these ensemble means, which suggests that their presence in observations was either a result of unforced variability or a forced mechanism that was not well modeled.

[27] Despite the overall reproduction of a warming pattern, models differed considerably in their ability to reproduce the specific spatiotemporal signatures of IMD variability in the observed record. The analysis in section 4.2 suggested that the models ranked 1st, 3rd, and 4th were statistically better at reproducing observations than the bottom six models, and that the model ranked 2nd was statistically better at reproducing observations than the bottom five models. This is apparent in the spatiotemporal analysis as well. The top four models were consistent in their ability to reproduce the observed pattern that the most significant early century cool regimes were located in the Southern Hemisphere (grid numbers 1–37), east central Asia (grid numbers 115–150) and in proximity to the northeastern Atlantic Ocean (grid numbers 190–211). Models 5, 7, and 8 displayed some similar behavior but matched observations to a lesser degree. Models 9, 10, 11, and 12 did not

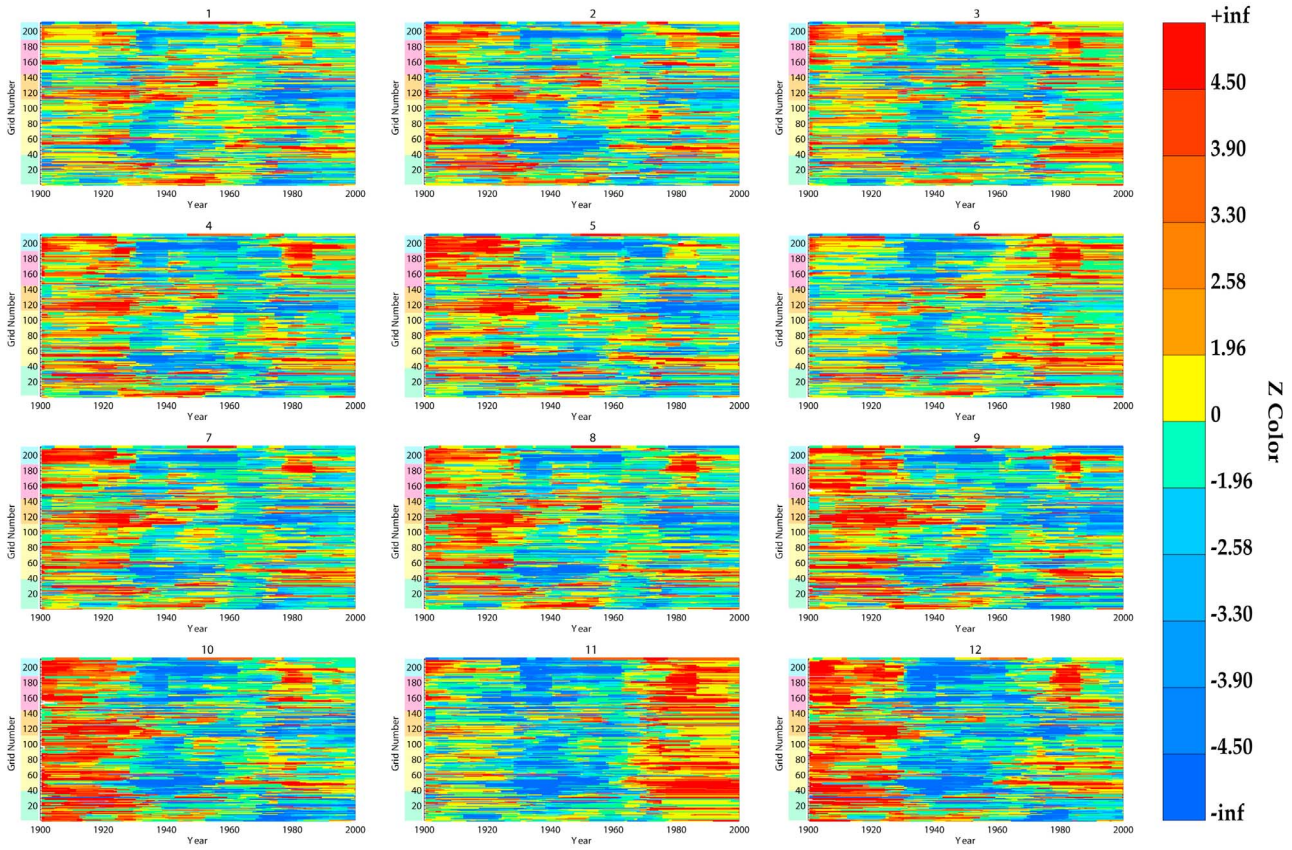


**Figure 9.** Spatiotemporal Z error plots for the 20C3M versus Obs experiment ensemble mean for each model. Models are ordered by their rank in the 20C3M versus Obs experiment (Figure 6).

produce this pattern at all. Model 6 as well as the models ranked 10th to 12th also failed to reproduce many of the warm regimes that appear in observations from the 1930s to the 1960s. The top three models did the best job of representing late century warmth as being most significant in the Southern Hemisphere (grid numbers 1–37) and east central Asia (grid numbers 115–150). Some of the poorer scoring models also produced this pattern but still produced high Z errors because of differences in the magnitude of significance.

## 6. Context for Model Performance

[28] Because the methodology outlined in section 4 has not been used in previous studies, there is no standard measure for what may be considered good or poor FMs. In order to give the results of the 20C3M versus Obs experiment some context, two additional experiments (PICNTRL versus Obs and 20C3M versus 20C3M) were conducted. The PICNTRL versus Obs experiment was intended to identify poor FM values associated with decorrelated IMD variability, while the 20C3M versus 20C3M experiment was intended to identify good FM values associated with a satisfactory reproduction of forced IMD temperature variability. The latter experiment was also used to form a relative performance metric that effectively handicapped model performance on the basis of each model's own potential to reproduce observations.

### 6.1. PICNTRL Versus Obs Experiment

[29] To determine the magnitude of FM values consistent with essentially random variability, each model's PICNTRL simulations were compared with observations of the twentieth century at the 211 grid locations. For each model, multiple 98 year segments were extracted from their PICNTRL experiment at 60 year lag intervals (e.g., 0–98, 60–158, 120–218, etc.). This worked to increase the number of 98 year segments that could be produced while still ensuring that run FMs were minimally dependent. The first segment for each model was excluded from evaluation so that any nonphysical adjustments associated with model spin-up would not contaminate the results. In total, seventy-nine 98 year PICNTRL runs were investigated from the 12 models (Table 1).

[30] In this scenario, high FM values were expected since there was no reason to anticipate similar IMD variability between an unforced simulation and the observed twentieth century climate. Figure 10 illustrates the general characteristics of the IMD variability in unforced simulations from different models. The coincident occurrences of significant warm (cool) regimes in given regions indicates that the unforced simulations produce IMD variability that is semi-consistent across space and time, suggesting an influence from natural oceanic oscillations.

### 6.2. The 20C3M Versus 20C3M Experiment

[31] To estimate good (low) FMs, an experiment was performed that calculated FMs between 20C3M ensemble

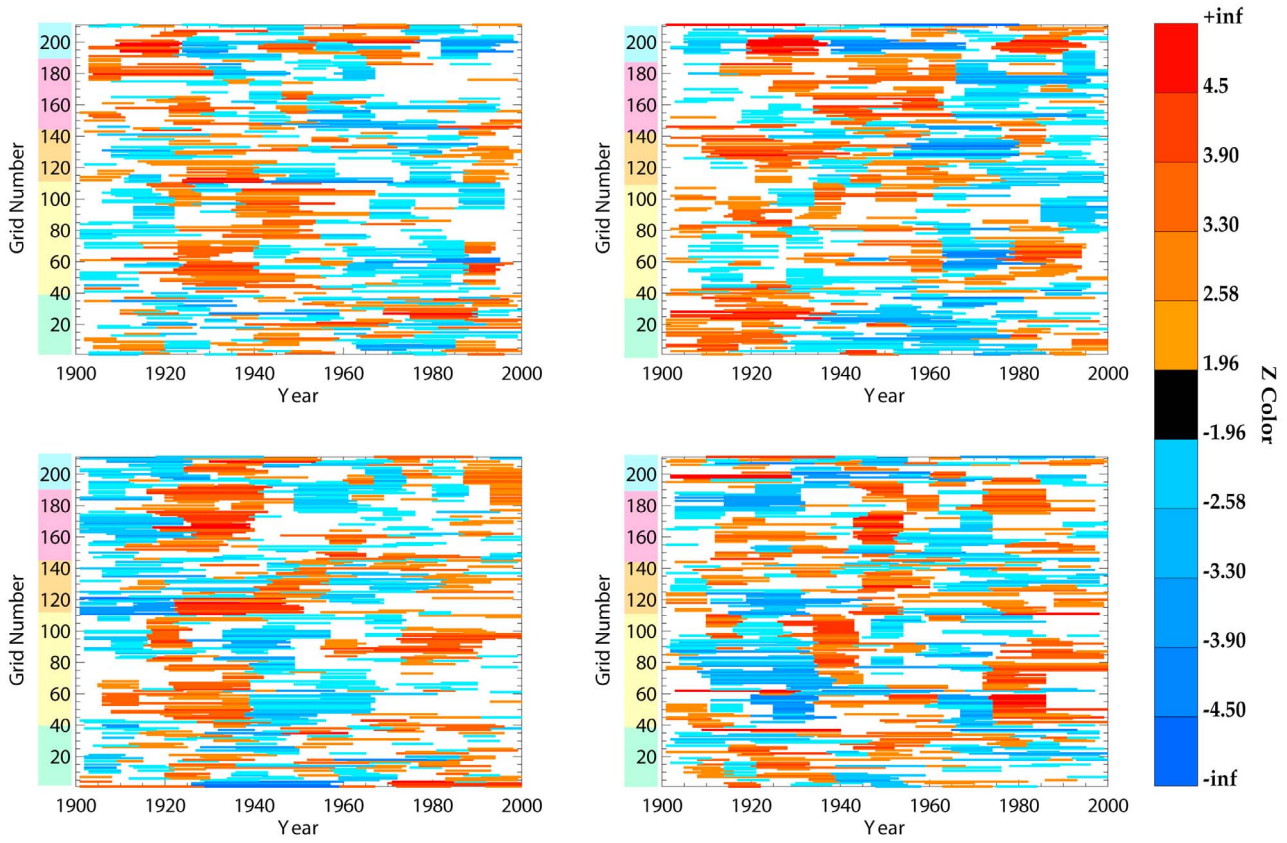


**Figure 10.** Spatiotemporal Z series plots for preindustrial control runs from four different models.

members of the same model. FMs for this experiment were calculated in the same way described in section 4, except that the Z error was redefined,

$$\mathrm{ZE}_t = Z_{Modeled(t)_i} - Z_{Modeled(t)_j}, \tag{8}$$

where each modeled Z value came from differing ensemble members within a given model (represented by the subscripts i and j, i ≠ j). FMs were calculated between each possible pairwise comparison of 20C3M ensemble members for each model (128 total comparisons as indicated in Table 1).

[32] IMD temperature variability can result from both external radiative boundary forcings as well as unforced variability [*Latif et al.*, 2010]. Because of the random initialization of a model's CMIP 3 ensemble members, the unforced variability among those ensemble members is essentially random. Consequently, even if a model were to represent the climatic response to external forcings perfectly, its 20C3M versus 20C3M FMs would still be nonzero because of the unforced component. Because ensemble members within each model incorporate identical radiative boundary forcings, any nonzero FM in this experiment can be attributed exclusively to differences in initial conditions and thus can act as a quantitative measure of the unforced variability for that model. As a result, these intraensemble FMs are consistent with the best possible scores that could be expected in the 20C3M versus Obs experiment.

### 6.3. FM Results for All Experiments

[33] The 20C3M versus Obs FMs are plotted along with the PICNTRL versus Obs and 20C3M versus 20C3M FMs in Figure 11a. Intuitively, we expect the PICNTRL versus Obs FMs to be higher than the 20C3M versus Obs FMs. This was true for many of the models although it was not universal as the models ranked 7th, 9th, 10th, and 11th all had at least one of their 20C3M versus Obs FMs score worse than one of their PICNTRL versus Obs FMs. This would indicate that these 20C3M versus Obs ensemble members did a particularly poor job of reproducing the observations. Despite this, Student's t tests indicated that for each model, the 20C3M versus Obs FM means were statistically distinguishable from their own PICNTRL versus Obs FM means at the 90th percentile or greater. This illustrated that external radiative boundary forcings acted to bring IMD temperature variability into a greater agreement with observations in every model investigated.

[34] Also in accordance with expectations, the 20C3M versus 20C3M FMs generally scored better than the 20C3M versus Obs FMs. Student's t tests indicated that for each model 20C3M versus Obs FM means were statistically distinguishable from their own 20C3M versus 20C3M FM means at the 95th percentile or greater. This illustrated that none of the models reproduced observations at their theoretical optimum (where a model's only source of error is due to unforced variability). Assuming that the observational data set used in this analysis was essentially correct, this

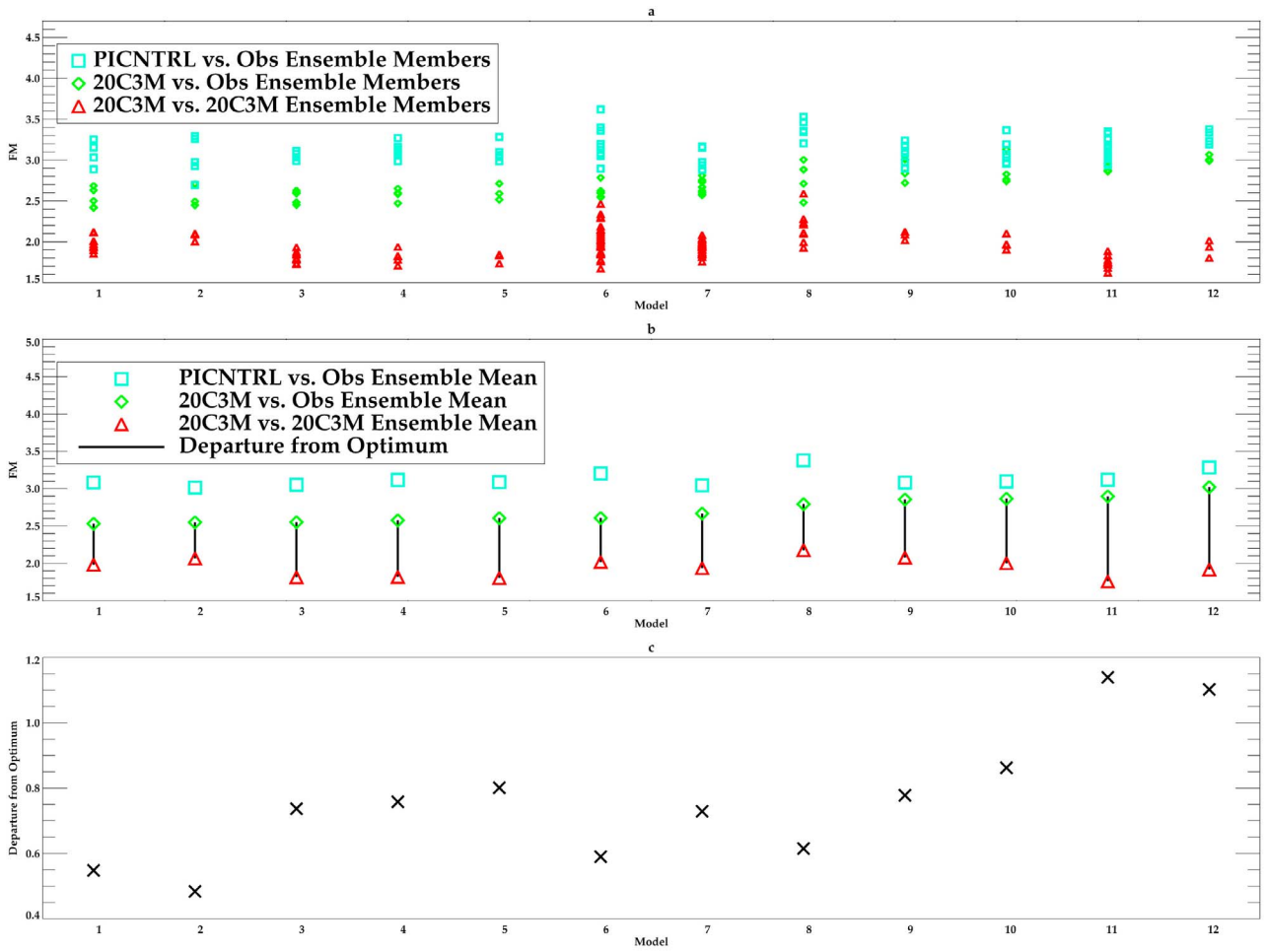


**Figure 11.** (a) FM scores for each model and experiment. (b) The mean FMs for each model and experiment. (c) Departure from optimum (DFO) for each model. In each plot models are ranked by their 20C3M versus Obs ensemble mean FM in ascending order from left to right. The identity of the models can be found in the auxiliary material.

would suggest that there was substantial room for improvement in the simulation of twentieth century IMD variability.

[35] The performance of the 20C3M versus 20C3M experiment also differed perceptibly from model to model. For example, the mean 20C3M versus 20C3M FM for model 8 was 2.18 compared with 1.76 for model 11. This suggested that there were distinct differences in the magnitude of unforced variability between the models. Because the spatial domain had a strong emphasis on the United States and Europe, these differences in internal variability might be traced to different realizations of the AMOC, which have been shown to be distinct between CMIP 3 models [*Meehl et al.*, 2007b]. More generally, differences could be attributed to different representations of ocean dynamics that can be affected by factors such as the model's oceanic grid resolution [*Swanson et al.*, 2009].

### 6.4. Departure From Optimum

[36] In addition to providing context to the 20C3M versus Obs experiment, the 20C3M versus 20C3M experiment was used to create the departure from optimum (DFO) metric, where

$$\mathrm{DFO} = \mathrm{avg}(\text{20C3M versus Obs FMs}) - \mathrm{avg}(\text{20C3M versus 20C3M FMs}). \quad (9)$$

More intuitively, the DFO can be thought of as the lengths of the lines in Figure 11b. The DFO indicates continuously better model performance as it approaches zero, or as the model's 20C3M versus Obs performance approaches its own 20C3M versus 20C3M performance.

[37] The DFO is relative in the sense that models are scored in relation to their own optimum FM expectations on the basis of their 20C3M versus 20C3M experiment. Physically, the DFO handicaps model performance on the basis of the amount of unforced variability present in each model. If a given model had a high average FM for the 20C3M versus 20C3M experiment (e.g., model 8), then this model incorporated a relatively high amount of unforced IMD variability in its representation of twentieth century climate compared with other models. Accordingly, the expectation

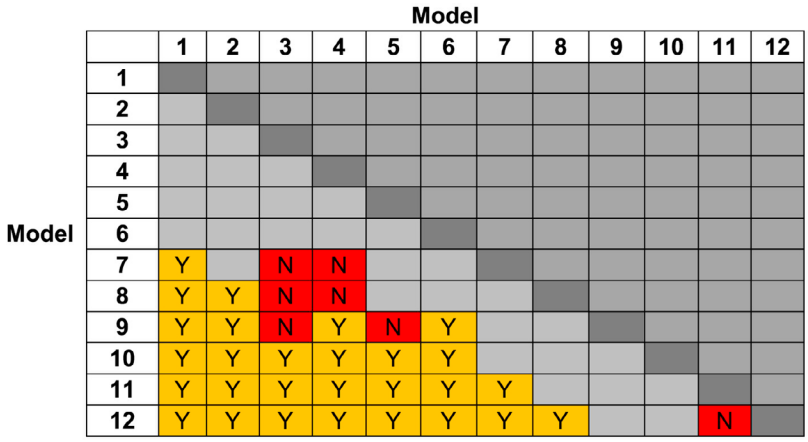


**Figure 12.** Model comparison summary chart illustrating where statistically distinguishable model-model comparisons (in the 20C3M versus Obs experiment) were also characterized by the lower-ranking model (in numerical value, e.g., 1 is ranked lower than 2) achieving a lower DFO. The Y's indicate that the lower-ranked model also had a lower DFO than the higher-ranked model, while the N's indicate that the higher-ranked model had a lower DFO. As an example, the model ranked 3rd was found to be statistically better than the models ranked 7th–12th in the 20C3M versus Obs experiment, but model 3 only had a lower DFO than the models ranked 10th–12th.

that this model could reproduce observed IMD temperature variability via the inclusion of external radiative forcings should be reduced. Alternatively, if a given model had a low average FM for the 20C3M versus 20C3M experiment (e.g., model 11) then this model incorporated a relatively low amount of unforced IMD variability and thus the model could be expected to reproduce more of the observed IMD variability using only external radiative forcings.

[38] DFOs for each model are shown in Figure 11c. Models that tended to score better in the absolute sense by performing better in the 20C3M versus Obs experiment also tended to produce better DFOs. In particular, the top two ranked models in the 20C3M versus Obs experiment were also the top two ranked models in the DFO metric. Additionally, the three bottom ranked models (in terms of the 20C3M versus Obs experiment) had DFOs ranked in the bottom three as well. This suggests that performance in the 20C3M versus Obs experiment was not heavily biased against models with intrinsically more unforced variability, and it corroborates the ranking based on the 20C3M versus Obs experiment alone. However, the DFO does illustrate that the influence of unforced variability should be considered when attempting to rank uninitialized models in their ability to reproduce observations. For instance, the models ranked 3rd to 5th in the 20C3M versus Obs experiment were outperformed in DFO by the models ranked 6th to 8th in the 20C3M versus Obs experiment. This was a result of the models ranked 3rd to 5th containing a lower amount of unforced variability, which effectively raised the expectations that they could reproduce observed patterns.

[39] The fact that every model had a positive DFO demonstrated that there is ample room for improvement in model performance. This does not necessarily mean that the improvement should be achieved by lowering the 20C3M versus Obs FMs to be in closer agreement with the 20C3M versus 20C3M FMs. It could very well be the case that these models systematically underestimate the amount of unforced variability in the climate system (as has been suggested by *Swanson et al.* [2009]). In this case, lower DFOs could be achieved through the raising of the 20C3M versus 20C3M FMs.

### 6.5. Combining 20C3M Versus Obs and the DFO

[40] Because the 20C3M versus Obs experiment and the DFO results do not produce identical model rankings, assessment of model performance relative to other models can be ambiguous. We would recommend that to be confident that a given model is performing better than another model, it should have a lower mean 20C3M versus Obs FM (at a statistically significant level), and it should also have a lower DFO. Figure 12 illustrates where this has occurred by indicating which of the statistically distinguishable model-model comparisons (in the 20C3M versus Obs experiment shown in Figure 7) were also characterized by the lower-ranking model (in terms of numerical value) having a lower DFO. Of the 35 statistically distinguishable model-model comparisons shown in Figure 7, 28 comparisons passed this additional constraint that the lower-ranking model have a lower DFO. In these 28 cases, we can be reasonably sure that the models being compared were characterized by distinct differences in skill.

[41] One potential explanation for the distinctions in model performance seen above is associated with differences in the incorporated external forcings. Much of the hemispheric-scale decadal climate variability of the past 1,000 years has been a result of solar and volcanic forcing [*Crowley*, 2000]. These two forcings are also thought to have had an influence over the twentieth century, particularly with regard to early century warming [*Hegerl et al.*, 2003]. All of the top nine ranked models (in both the 20C3M versus Obs experiment as well as the DFO) incorporated solar and volcanic variability among their forcings while the three bottom ranked models did not. This is evidence that twentieth century IMD temperature variability has been heavily influenced by these two factors. Also, *Tett et al.* [2002] attributes a portion of early century warming

to natural variability associated with the AMOC. AMOC variability, however, may be externally forced [*Ottera et al.*, 2010]. Therefore, the degree to which a model can simulate a realistic AMOC may also contribute to model performance in this analysis.

[42] Some of the noted differences in performance may also be attributed to the underlying construction of the models themselves. However, models that are considered to be more similar by construction did not necessarily perform similarly in this analysis. For example, the models ranked 1st, 3rd, 6th, and 9th (in the 20C3M versus Obs experiment) are considered to be relatively closely related [according to *Masson and Knutti*, 2011]. To attribute differences in model performance exclusively to model construction, it would be necessary to standardize the external radiative forcings included in the retrospective simulations.

## 7. Summary

[43] This work demonstrated that the CMIP 3 models investigated varied considerably in their capacity to reproduce the timing, significance, and location of historical IMD variability. This was revealed through the process of ranking models on the basis of their 20C3M versus Obs ensemble mean FMs and then testing the ensemble distributions for statistically different levels of performance. A spatiotemporal analysis was employed that allowed for a better understanding of the model rankings by identifying distinctions between higher- and lower-ranked models.

[44] Model ability to reproduce observations was given context by comparing preindustrial control runs to observations and by comparing ensemble members within each model to each other. The former experiment demonstrated that the inclusion of time-varying radiative forcings did indeed bring modeled IMD variability into closer agreement with observations. The latter experiment demonstrated that model performance was not at its theoretical optimum. This second experiment also allowed for a performance metric to be devised that assessed model skill relative to the degree of internal variability inherent in each model. The "departure from optimum" metric, in conjunction with the original 20C3M versus Obs metric, were then combined to highlight considerable distinctions in model performance.

[45] Because the external radiative boundary forcings differed between the models, differences in model performance could not be attributed exclusively to model construction. In particular, it appears that models must include solar and volcanic forcing in order to have a realistic chance of reproducing observed IMD variability over the twentieth century. Nevertheless, certain model/forcing combinations were found to outperform other model/forcing combinations by a wide margin in both the 20C3M versus Obs and the DFO metric. Therefore, future projections of temperature change, based on a given radiative forcing trajectory, may see an improvement in predictive power if they are asymmetrically weighted toward the better performing model/forcing configurations, as revealed by these two metrics.

[46] These results, however, do come with a number of caveats. Most importantly, it is unknown if a similar ranking would have been achieved if the entirety of Earth's surface were included in the spatial domain or if the spatial resolution had been altered. Additionally, when modeled and observed IMD variability differed, it was assumed that the model was in error, when in reality it was possible that the observations themselves were inaccurate. Finally, the ranking of models was complicated by the reality that each model included different estimates of historical radiative boundary forcings as well as a different number of ensemble members. Despite these uncertainties, our results demonstrate a useful methodology for comparing model ability to simulate IMD variability that may be helpful in upcoming model assessments.

[47] **Acknowledgments.** We acknowledge the modeling groups, Program for Climate Model Diagnosis and Intercomparison (PCMDI), and WCRP's Working Group on Coupled Modeling (WGCM) for their roles in making available the WCRP CMIP 3 multimodel data set. Support for this data set is provided by the Office of Science, U.S. Department of Energy. 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.

## References

Boer, J. G. (2011), Decadal potential predictability of the twenty-first century climate, *Clim. Dyn.*, *36*, 1119–1133, doi:10.1007/s00382-010-0747-9.

Compo, G. P., and P. D. Sardeshmukh (2009), Oceanic influences on recent continental warming, *Clim. Dyn.*, *32*, 333–342, doi:10.1007/s00382-008-0448-9.

Cordero, E. C., W. Kessomkiat, J. Abatzoglou, and S. Mauget (2011), The identification of distinct patterns in California temperature trends, *Clim. Change*, *108*, 357–382, doi:10.1007/s10584-011-0023-y.

Crowley, T. J. (2000), Causes of climate change over the past 1000 years, *Science*, *289*, 270–277, doi:10.1126/science.289.5477.270.

DelSole, T., M. K. Tippett, and J. Shukla (2011), A significant component of unforced multidecadal variability in the recent acceleration of global warming, *J. Clim.*, *24*, 909–926, doi:10.1175/2010JCLI3659.1.

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. Clim.*, *15*, 1141–1158, doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

Goosse, H., and H. Renssen (2004), Exciting natural modes of variability by solar and volcanic forcing: Idealized and realistic experiments, *Clim. Dyn.*, *23*(2), 153–163, doi:10.1007/s00382-004-0424-y.

Hegerl, G. C., T. J. Crowley, S. K. Baum, K.-Y. Kim, and W. T. Hyde (2003), Detection of volcanic, solar and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature, *Geophys. Res. Lett.*, *30*(5), 1242, doi:10.1029/2002GL016635.

Hegerl, G. C., T. J. Crowley, M. Allen, W. T. Hyde, H. N. Pollack, J. Smerdon, and E. Zorita (2007), Detection of human influence on a new, validated 1500-year temperature reconstruction, *J. Clim.*, *20*, 650–666, doi:10.1175/JCLI4011.1.

Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor (1999), Surface air temperature and its variations over the last 150 years, *Rev. Geophys.*, *37*, 173–199, doi:10.1029/1999RG900002.

Karoly, D. J., and Q. Wu (2005), Detection of regional surface temperature trends, *J. Clim.*, *18*, 4337–4343, doi:10.1175/JCLI3565.1.

Knutson, T. R., et al. (2006), Assessment of twentieth-century regional surface temperature trends using the GFDL CM2 coupled models, *J. Clim.*, *19*, 1624–1651, doi:10.1175/JCLI3709.1.

Kravtsov, S., and C. Spannagle (2008), Multidecadal climate variability in observed and modeled surface temperatures, *J. Clim.*, *21*, 1104–1121, doi:10.1175/2007JCLI1874.1.

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. Clim.*, *19*, 4137–4153, doi:10.1175/JCLI3848.1.

Latif, M., et al. (2010), Dynamics of decadal climate variability and implications for its prediction, in *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society 2* [CD-ROM], edited by J. Hall, D. E. Harrison, and D. Stammer, *ESA Publ. WPP-306*, Eur. Space Agency, Paris, doi:10.5270/OceanObs09.cwp.53.

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, doi:10.1214/aoms/1177730491.

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. Hydrometeorol.*, *11*, 822–831, doi:10.1175/2010JHM1191.1.

Masson, D., and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

Mauget, S. (2003), Multi-decadal regime shifts in U.S. streamflow, precipitation, and temperature at the end of the twentieth century, *J. Clim.*, *16*, 3905–3916, doi:10.1175/1520-0442(2003)016<3905:MRSIUS>2.0.CO;2.

Mauget, S. (2004), Low frequency streamflow regimes over the central United States: 1939–1998, *Clim. Change*, *63*, 121–144, doi:10.1023/B:CLIM.0000018502.86522.57.

Mauget, S. A., E. C. Cordero, and P. T. Brown (2012), Evaluating modeled intra- to multidecadal climate variability using running Mann-Whitney Z statistics, *J. Clim.*, *25*(5), 1570–1586, doi:10.1175/JCLI-D-11-00211.1.

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell (2007a), The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Meehl, G. A., et al. (2007b), Global climate projections, in *Climate Change 2007: The Physical Science Basis*, edited by S. Solomon et al., pp. 747–845, Cambridge Univ. Press, Cambridge, U. K.

Mendenhall, W., D. D. Wackerly, and R. L. Sheaffer (1990), *Mathematical Statistics With Applications*, PWS-Kent, Boston, Mass.

Neelin, D. J., D. S. Battisti, A. C. Hirst, F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak (1998), ENSO theory, *J. Geophys. Res.*, *103*, 14,261–14,290, doi:10.1029/97JC03424.

Ottera, O. H., M. Bentsen, H. Drage, and L. Suo (2010), External forcing as a metronome for Atlantic multidecadal variability, *Nat. Geosci.*, *3*, 688–694, doi:10.1038/ngeo955.

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. U. S. A.*, *106*, 8441–8446.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan (2003), Globally complete analyses of sea surface temperature, sea ice, and night marine air temperature, 1871–2000, *J. Geophys. Res.*, *108*(D14), 4407, doi:10.1029/2002JD002670.

Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Meteorol. Soc.*, *89*, 303–311, doi:10.1175/BAMS-89-3-303.

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*(D24), 4748, doi:10.1029/2001JD001577.

Solomon, A., et al. (2011), Distinguishing the roles of natural and anthropogenically forced decadal climate variability, *Bull. Am. Meteorol. Soc.*, *92*, 141–156, doi:10.1175/2010BAMS2962.1.

Swanson, K. L., G. Sugihara, and A. A. Tsonis (2009), Long-term natural variability and 20th century climate change, *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 16,120–16,123, doi:10.1073/pnas.0908699106.

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. Clim.*, *18*, 1524–1540, doi:10.1175/JCLI3363.1.

Tett, S. F. B., et al. (2002), Estimation of natural and anthropogenic contributions to twentieth century temperature change, *J. Geophys. Res.*, *107*(D16), 4306, doi:10.1029/2000JD000028.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, Academic, San Diego, Calif.

Wu, Z., N. Huang, J. Wallace, B. Smoliak, and X. Chen (2011), On the time-varying trend in global-mean surface temperature, *Clim. Dyn.*, *37*, 759–773, doi:10.1007/s00382-011-1128-8.

Zhang, R., T. L. Delworth, and I. M. Held (2007), Can the Atlantic Ocean drive the observed multidecadal variability in Northern Hemisphere mean temperature?, *Geophys. Res. Lett.*, *34*, L02709, doi:10.1029/2006GL028683.