Surface Temperature Probability Distributions in the NARCCAP Hindcast Experiment: Evaluation Methodology, Metrics and Results

Paul C. Loikith[1], Duane E. Waliser, Jinwon Kim[2,4], Huikyo Lee[1], Benjamin R. Lintner[3], J. David Neelin[4], Seth McGinnis[5], Chris A. Mattmann[1,2], Linda O. Mearns[5]

[1]*Jet Propulsion Laboratory/California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA, 91101*

[2]*Joint Institute for Regional Earth System Science and Engineering, University of Los Angeles, CA*

[3]*Department of Environmental Sciences, Rutgers, The State University of New Jersey, New Brunswick , NJ*

[4]*Department of Atmospheric and Oceanic Science, University of California, Los Angeles, California*

[5]*Institute for Mathematical Applications to the Geosciences, National Center for Atmospheric Research, Boulder, CO*

Contact: paul.c.loikith@jpl.nasa.gov

July, 2013

Abstract

Methodology is developed and applied to evaluate the characteristics of daily surface temperature probability distribution functions (PDFs) in a six-member regional climate model (RCM) hindcast experiment conducted as part of the North American Regional Climate Change Assessment Program (NARCCAP). The evaluation is based on two state-of-the-art high-resolution reanalysis products that provide the observational reference(s): the NCEP North American Regional Reanalysis and the NASA Modern Era-Retrospective Analysis for Research and Applications. Typically, the NARCCAP temperature biases for the tails and the medians of the PDFs are of the same sign, indicating a shift in the RCM-simulated PDFs relative to reanalysis. RCM-simulated temperature variance is often higher than reanalysis in both winter and summer. Temperature skewness is reasonably well simulated by most RCMs, especially in the winter, suggesting confidence in the use of these models to simulate future temperature extremes. To facilitate identification of model-reanalysis discrepancies and provide a regional basis for investigating mechanisms associated with such discrepancies, a k-means clustering approach is applied to sort model and reference data PDFs by PDF morphology. RCM cluster assignments generally match reanalysis cluster assignments with some discrepancy at high latitudes due to over-simulation of temperature variance by most models here. Model biases identified in this work will allow for further investigation into associated mechanisms and implications for future simulations of temperature extremes.

1. Introduction

As a result of anthropogenic warming, mean temperatures are expected to rise globally; however, changes in temperature extremes are expected to have the most substantial climate impacts (IPCC, SREX 2012). In particular, extreme warm events are expected to become more common and severe while extreme cold events are expected to become less frequent and severe (IPCC 2007, Meehl and Tebaldi 2004, Tebaldi et al. 2006, Meehl et al. 2007). Such changes will likely expose populations to extreme heat events that are unprecedented in the current climate (Meehl et al. 2009).

One particularly noteworthy example, the European heatwave of 2003, caused widespread heat-related illness and claimed tens of thousands of lives (Luber and McGeehin, 2008). Events of this magnitude, while virtually unprecedented in the current climate, are projected to become more frequent in the future due to climate warming (e.g. Beniston 2004, Schär et al. 2004, Stott et al. 2004). More recently, the 2011 Russian heatwave was also associated with drastically elevated mortality and morbidity due to heat stress and poor air quality associated with wildfires: some studies have speculated that the extreme nature of this event was related to a combination of natural variability and anthropogenic climate forcing (Dole et al. 2011, Rahmstorf and Coumou 2012, and Otto et al. 2012). Recent anomalous heat, including the hottest month on record in the United States (US), coupled with severe drought has had severe impacts on the United States agriculture sector (Karl et al. 2012).

77

78 Because the relationship between changes in mean temperature and its extremes is
79 often non-linear, relatively small changes in the mean may be associated with
80 disproportionately large changes in extremes (Hegerl et al. 2004, Griffiths et al
81 2005). Therefore, proper simulation of probability distribution function (PDF)
82 shape is essential for a realistic representation of extremes. Ruff and Neelin (2012)
83 analyzed surface temperature (T_s) PDFs from station data and documented several
84 examples of non-Gaussian, often asymmetric long-tailed distributions. They further
85 note the importance of daily T_s PDF shape, especially the distribution tails, and
86 related implications for future global warming in estimating threshold exceedances
87 with places exhibiting near Gaussian PDFs being more sensitive to incremental
88 warming than places with exponential PDFs.
89

90 Observational evidence points to a recent increase in temperature variance in the
91 tropics as well as a tendency towards more positive skewness globally (Donat and
92 Alexander 2012). On the other hand, Rhines and Huybers (2013) suggest that
93 observed changes in summertime extremes are primarily attributable to changes in
94 the mean rather than the variance. Lau and Nath (2012) demonstrate PDF shifts in
95 daily maximum temperature in two high-resolution general circulation models by
96 the middle of the 21st century with only small changes in PDF shape exhibited in
97 some places.
98

In order to quantify uncertainty in simulations of future climate, it is important to bring as much observational scrutiny as possible to historical climate model runs. Model evaluation is critical for identifying the range of error (magnitude, geographic distribution, sign) across models for the same region. Comprehensive evaluation of GCMs archived as part of the Coupled Model Intercomparison Project Phase 3 (CMIP3) was performed by Gleckler et al. (2008); however, the demand for more geographically specific climate projections has increased the prominence of limited domain RCMs. While the body of systematic RCM evaluation work is less mature than that for GCM evaluation, some studies have evaluated important variables in RCMs. Kjellström et al. (2011) analyze a suite of RCM hindcast and future projections driven by reanalysis and multiple GCMs over Europe. Kim et al. (2013a) evaluate mean surface temperature, precipitation, and insolation using monthly mean data over the conterminous US using models from the North American Regional Climate Change Assessment Program (NARCCAP).

Some work has focused on evaluation of model simulated PDFs. Perkins et al. (2007) introduced a PDF skill score to evaluate global models and applied this method over Australia, using climatologically homogenous sub-regions to compute PDFs of temperature and precipitation. Kjellström et al. (2010) used this method to evaluate temperature and precipitation PDF structure over Europe while also evaluating daily temperature at multiple percentiles of the distribution. Their results show that while some models perform better or worse than others, no model

121 is systematically better or worse in every region or season suggesting substantial
122 variability in the way RCM bias is manifested.
123
124 Comprehensive evaluation of PDF morphology is expected to provide information
125 regarding model representation of extremes and to enhance mechanistic
126 understanding of processes responsible for genesis of extremes. To this end, the
127 present study uses daily mean surface air temperature (T_s) from two reanalysis
128 products to evaluate model-simulated PDF characteristics over North America. The
129 remainder of this paper is organized as follows. Section 2 describes the data and
130 methodology used. Section 3 presents daily temperature bias at different
131 percentiles of the distribution and Section 4 evaluates model variance and
132 skewness. A clustering technique, used to compare model PDF structure across the
133 entire domain, is introduced and described in Section 5 followed by concluding
134 remarks in Section 6.
135
136 **2. Data and Methodology**
137 *2a. Data used*
138 All six models used in this paper are hindcast experiments performed for NARCCAP
139 (Mearns et al. 2009, 2012, http://www.narccap.ucar.edu). NARCCAP was designed
140 to serve the high-resolution climate modeling needs of the United States, Canada,
141 and Mexico and is comprised of RCMs nested within GCMs to form multi-model
142 ensemble scenarios for the region. In this work, all hindcast model simulations

were driven by large-scale forcing from the National Center for Environmental Prediction (NCEP) Reanalysis 2 (Kanamitsu et al. 2002).

While the official NARCCAP time period spans 1979 to 2004, the period 1980-2003 was used to span the longest possible time period for which all models have available T_s. The simulation domain covers most of North America and some of the adjacent Pacific and Atlantic Oceans. Each model is provided on a 50 km native curvilinear grid that was interpolated onto a regular 0.5° latitude by 0.5° longitude grid using the Python *griddata* function which linearly interpolates irregularly spaced data onto a regular rectangular grid. The analyzed data comprise daily means computed from 3-hourly model output.

Gómez-Navarro et al. (2012) demonstrate that observational uncertainty can be similar in magnitude to individual model errors emphasizing the need to quantify observational uncertainty in the evaluation process. To address this, two reanalysis products are used in this work: the NCEP North American Regional Reanalysis (NARR; Mesinger et al. 2006) and the NASA Modern Era-Retrospective Analysis for Research and Applications (MERRA; *Rienecker et al.,* 2011). Produced on a Lambert Conformal grid with 32 km resolution, the NARR data were interpolated to a 0.5° x 0.5° regular latitude/longitude grid in the same manner as the NARCCAP models. Developed by NASA's Global Modeling and Assimilation Office and disseminated by the Goddard Earth Sciences Data and Information Services Center (GES DISC), MERRA is originally on a global 0.5° x 0.67° latitude/longitude grid. It is coarser

than NARR, but still higher in resolution than most other global reanalysis products. The MERRA dataset was regridded to the same 0.5° x 0.5° degree latitude/longitude grid as the other datasets. The analysis domain was chosen to cover as much of North America as possible while ensuring that the domain had full coverage in all datasets.

2b. Methodology

The Regional Climate Model Evaluation System (RCMES: rcmes.jpl.nasa.gov), jointly developed by NASA's Jet Propulsion Laboratory and the University of California Los Angeles (Hart et al. 2011, Crichton et al. 2012), is a combined database-software package designed to aid in climate model evaluation by providing easy access to reference datasets and validation tools. Additionally, RCMES includes a toolkit designed to compute commonly used metrics e.g., bias, root-mean-square error, and correlation coefficients. Further details about RCMES and the scientific capabilities of the system can be found in Kim et al. (2013a,b). RCMES was used here to project all datasets onto a common temporal and spatial grid. Some of the metrics evaluated in this work will be incorporated into future versions of the RCMES toolkit.

Temperature anomalies, obtained by subtracting the daily climatological average from each daily value, are used in the computation of several metrics in this paper. Long-term trends were not removed as all datasets cover the same period and any influence from trend bias is inherently incorporated into other T_s biases. Bukovsky

189 (2012) evaluated temperature trends for all seasons over this period in NARCCAP
190 models and found reasonable trend agreement between the models and
191 observations. Evaluation was performed for the seasons of summer (June, July,
192 August; JJA) and winter (December, January, February; DJF). The multi-model
193 ensemble mean is calculated by concatenating the daily data from each of the six
194 models into one time series consisting of six data points (one for each model) at
195 each grid point at each time step.
196
197 **3. Percentile-based T_s Evaluation**
198 Temperature biases at three percentile thresholds (5^{th}, 50^{th}, and 95^{th}) were
199 calculated for each model with respect to NARR. The 5^{th} and 95^{th} percentiles are
200 chosen to represent the cold and warm tails of the distribution of daily-average total
201 temperature values respectively. Uncertainty in the reference data is shown in each
202 analysis using the difference between MERRA and NARR, i.e., MERRA is treated as a
203 model. For brevity, the discussion here emphasizes the median bias. Figure 1 shows
204 the bias in median temperature for DJF. While the errors differ in sign and
205 magnitude across models, all models exhibit a warm bias over the central and
206 northern Great Plains and a cold bias over Northern Mexico. In general, MERRA is
207 colder than NARR over much of the northern portion of the domain. In many cases,
208 the MERRA bias is comparable in magnitude to the model biases, especially at
209 higher latitudes. This suggests some uncertainty in quantification of model error at
210 high latitudes. HRM3 exhibits the largest warm biases with magnitudes exceeding
211 8°C while CRCM has a cold bias within one or two degrees of 5°C over most of the

212 domain. Large biases present over the Great Lakes and other inland bodies of water
213 are indicative of the way lake surface temperature is defined in this model
214 experiment. While these features are not directly related to RCM dynamics, such
215 biases could affect lake-effect precipitation downwind of the Great Lakes. Lake-
216 associated errors were also noted by by Kjellström et al. (2010) over Europe.
217
218 Figure 2 is the same as Figure 1, except for the northern hemisphere summer. Cold
219 biases are more common than in DJF across models. All models have an area of
220 positive median temperature bias over a portion of the Great Plains with HRM3
221 being the warmest (~6-8°). In general, the difference between reference datasets is
222 small with the western Gulf of Mexico coast and inland water bodies showing the
223 greatest disagreement. It is worth noting that the median temperature over the
224 Great Lakes is higher than surrounding land in MERRA (not shown) during summer
225 leading to the warm bias over the Lakes. In reality, the Lakes remain cool relative to
226 surrounding land due to the greater heat capacity of water. Overall, the biases
227 shown in both DJF and JJA are in qualitative agreement with other studies that
228 calculated bias in the mean (Kim et al. 2013a) and daily maximum and minimum
229 temperature (Rangwala et al. 2012) suggesting these biases are robust features of
230 the overall temperature distribution and throughout the diurnal cycle.
231
232 For additional perspective, the domain is further decomposed into four sub-regions,
233 chosen to be broadly representative of climate regimes and are defined as follows:
234 West, including the Pacific Ocean and Coast, Great Basin, and the US Rocky

Mountains; North, which includes southern Canada and Hudson Bay; Central, which is comprised of the US Great Plains, Midwest, and Gulf of Mexico; East, which includes the eastern US and the Atlantic Ocean and Coast. The portrait diagrams in Figure 3 show the spatial root mean squared of the bias, e.g. RMS error (RMSE), and the mean bias for the four sub-regions.

In general, RMSE is largest in the Northern sub-region and is comparable in the Western, Central, and Eastern sub-regions (with the exception of HRM3 in Central). The RMSE in JJA is smaller in most cases, due in part to lower temperature variability in the summer. The HRM3 model shows consistently high error relative to the other models in all percentiles, especially in the Central and Northern regions in DJF and JJA. In both seasons, ECP2 shows relatively small error in all regions. The ensemble mean generally has smaller error in median temperature compared to any individual ensemble member and in most cases MERRA exhibits smaller error than any individual model at all percentiles. There is some indication of DJF 5^{th} percentile RMSE being greater than the other percentiles suggesting asymmetry in model-simulated PDF bias.

The mean bias, plotted in the portrait diagrams in Figure 3, is useful for understanding where model PDFs are systematically shifted relative to NARR, regardless of shape, as opposed to where the PDF bias is not described by a simple shift. Models with a systematic shift in the PDF relative to NARR have the same sign in the biases at all three percentiles and are identified in Figure 3 as having the same

color (all red or all blue) in a given column of the portrait diagram. One notable example is HRM3 in DJF, showing a positive bias at all three percentile thresholds indicating a shift towards warmer values. MM5I shows different behavior in the Northern region in DJF where the mean bias at the cold tail is negative, the median is near zero, and the warm tail is positive leading to a wider PDF than NARR. It is interesting to note that no model has a PDF that is narrower than NARR at both tails. All cases with 5^{th} and 95^{th} percentile biases of opposite sign show cold bias at the 5^{th} percentile and warm bias at the 95^{th} percentile.

Figure 4 (top) shows the percentage of models that have the same sign of bias at all three percentiles, i.e., systematically cold- or warm biased. In general, the southern as well as much of the central and western portions of the domain have high percentages of models with a PDF shift in DJF while the same tendency occurs for the northern and western regions in JJA. In DJF, the Pacific Northwest and most of southern Canada have lower percentage values than other regions. This is consistent with the portrait diagram in Figure 3 for the Northern sub-region where several models have biases of opposing sign at the tails. The central portion of the domain shows mostly low percentage values in JJA, also consistent with Figure 3.

Figure 4 (bottom) show scatter plots of the mean bias for each model and each sub-region at the 5^{th} percentile versus the 95^{th} percentile. The diagonal black line indicates where the models would lie if they had the same bias at both percentiles, indicating a completely symmetrical shift of the PDF tails as estimated from these

thresholds. Models to the left of the diagonal line have a wider PDF than NARR while models to the right have a narrower PDF. In both DJF and JJA, most models have a net widening with fewer models showing a near systematic shift or a net narrowing. In DJF, the Northern sub-region mean biases (diamonds) are the farthest from the diagonal line, consistent with the low percentage values. The same is true for the Central region (squares) for JJA.

4. Evaluation of Variance and Skewness

Whereas the analysis in Section 3 only used three percentiles to estimate differences in model-simulated PDFs, this section considers higher-moment statistics and the shape of the distributions. Because of the important relationship between temperature variability, the length of the distribution tails, and extremes, the standard deviation (SD) and skewness of model simulated T_s are compared against NARR. In what follows, all analyses use temperature anomalies to allow for easy comparison of PDF shape between datasets (all have mean of 0) and to remove any influence from intraseasonal variability on higher moments (especially skewness).

4a. Standard Deviation

The ratios between the SD for each model and NARR in DJF are displayed in Figure 5. Values greater (less) than one indicate where the model has a higher (lower) SD than NARR. The SD values are statistically significant at the 5% confidence level where shaded, as determined by a two-sided F test. Models generally show higher SD than NARR in the northern portion of the domain and lower SD than NARR over

304 the southern Great Plains and southeastern US. Both WRFG and RCM3 have
305 elevated variance over the Great Lakes compared with NARR. These models also
306 have strong cold biases in median temperature over the lakes, (Figure 1) suggesting
307 that the air over the lakes has similar characteristics to land. In general the
308 differences between MERRA and NARR are smaller and less statistically significant
309 than any individual model; however, MERRA still shows higher variance over much
310 of the northern half of the domain compared to NARR.
311
312 In many examples (MM5I, CRCM, HRM3, and WRFG) the most striking areas of
313 positive bias are present to the north of the region of maximum SD in NARR. The
314 band of high variance in NARR (stretching from the northwest corner of the domain
315 southeast along the eastern edge of the Canadian Rockies and into the northern
316 Great Plains) is in an area highly influenced by large intraseasonal temperature
317 fluctuations due to synoptic-scale weather events associated with warm advection
318 from lower latitudes and cold advection from higher latitudes (Loikith et al. 2013).
319 Areas further north of this region show smaller variance where the availability of
320 extremely cold air relative to the local climate is lower due to the proximity of this
321 region to the coldest air in the hemisphere. For this reason, most variability in daily
322 temperature occurs only on the warm side of the PDF here. This is in contrast to the
323 band of higher SD to the south, which is characterized by a PDF that is more
324 symmetrical about the mean. The tendency for the models to have positive SD bias
325 north of this region of climatologically large variance indicates that models tend to
326 expand this high variance region substantially northward compared with NARR.

One possible mechanism for this feature is a northward expansion or shift of the main winter storm track. The northern region of positive SD bias is consistent with the band of low percentage values in Figure 4 where most models have a widening of the PDF.

Figure 6 shows the SD ratios for JJA. While the daily temperature variability is lower in the summer compared with winter, resulting in overall lower SD values, the ratio is generally higher for JJA than DJF. Overall, SD is higher in all six models over most of the domain with the coastal waters of the Pacific Ocean and the southern US showing the systematically largest ratios. MERRA SD is generally very similar to NARR with slightly smaller values throughout much of the domain (ratio of 0.75-1.0) with exceptions in the southwestern region and along the near-coastal waters of Hudson Bay.

It is interesting to note that all datasets have higher SD along and offshore of the Pacific Coast. Climate variability here is influenced largely by occasional offshore wind events producing anomalously warm T_s values (e.g., Santa Ana events in southern California (Hughes and Hall 2010)). It is possible that the positive SD bias is indicative of a tendency for more frequent and/or intense offshore wind events. MERRA and MM5I have notable positive SD biases over Hudson Bay, indicative of possible prolonged sea ice cover causing the surface of the bay to have physical characteristics similar to land instead of open water. In the case of MM5I, this feature encompasses the entire Bay and the Labrador Sea.

Results of the evaluation of SD are summarized for the four sub-regions in Figure 7 using portrait diagrams. The shaded values represent the spatial mean of the ratios as calculated and plotted in Figures 5 and 6. Only statistically significant grid points contribute to the computation of the mean. Except for CRCM and HRM3, no individual model has a mean ratio less than one in any sub-region in DJF. Overall, MERRA has ratio values consistently near one indicating strong similarity between the two reference datasets. In JJA, all RCMs have a mean ratio greater than one except for CRCM and RCM3 in the Northern sub-region. Only MERRA has a mean ratio substantially less than one in more than a single sub-region. Overall, most model SD ratios are within the range 0.6-1.4 indicating that the variance over- or under-estimates are not too severe for most regions.

4b. T_s Skewness

T_s skewness for all datasets is shown in Figure 8 for DJF. As opposed to variance, which primarily describes the width of the PDF, skewness is more directly related to extreme values as it describes the shape of the tails and the degree of symmetry of the PDF. All distributions are subjected to a skewness significance test (D'Agostino, 1970) and blank areas in Figure 8 indicate where skewness does not deviate from a normal distribution at the 10% significance level. For NARR, skewness is primarily positive in the northeastern portion of the domain while the rest of the domain has negative or near zero skewness. This generally occurs along the transition zone from the primarily positive skewness in the north to the negative skewness to the

373 south. Statistical significance is low in the region of weak skewness in the
374 southeastern part of the domain.
375
376 In general, the models capture the large-scale skewness pattern, with positive
377 skewness in the northeast, a large coherent region of relatively strong negative
378 skewness extending from the northwestern US across to parts of the Great Lakes
379 region, and modest skewness over the southeastern US. ECP2 and the multi-model
380 ensemble have the most realistic representation of skewness including sign,
381 magnitude, and areas of statistical significance. While MERRA captures much of the
382 same regional patterns, skewness is more negative over the Rocky Mountains and
383 more positive over the southeastern US compared to NARR. HRM3 has negative
384 skewness present much further north than in NARR which may have a physical
385 relation to the fact that it is the warmest model at all percentiles in DJF (e.g. Figure
386 1d). It is interesting to note that the transition zone from primarily negative (south)
387 to positive (north) skewness corresponds to the band where few models have bias
388 of the same sign at all percentiles (Figure 4). Models may have difficulties
389 accurately capturing this spatial shift in T_s regime, leading to errors in the simulated
390 PDF shape.
391
392 Model fidelity in simulating skewness in winter is likely indicative of differences in
393 simulation of large-scale climate mechanisms, including mechanisms associated
394 with extremes. Details of these mechanisms and their relationship to extremes
395 were extensively examined in an observational study by Loikith and Broccoli

(2012). For example, in winter the PDFs in the northern region have long warm tails resulting from advection of relatively warm air from lower latitudes. Advection of cold anomalies of comparable magnitude from the north rarely occurs because the air in this region is climatologically among the coldest in the hemisphere. Models that show more restricted regions of positive skewness (e.g. HRM3, ECP2) would generate extreme warm events less frequently than in NARR; models that show a larger area of strong positive skewness in this region (e.g. RCM3, WRFG) may simulate the occurrence of warm advection events too frequently in the region. In addition to having skewness that is more positive than NARR, WRFG also has a colder background climate in this region (Figures 1, 3) with a warm bias to the south. Under conditions of northward advection into the cold-biased region, extreme warm anomalies may occur that contribute to the positive skewness bias. RCM3 has similar skewness error as WRFG, but with warm biases over this region and cold biases to the south, making it more difficult to propose a mechanism here.

Another illustrative example in DJF is the region of negative skewness in the northwestern part of the domain encompassing Oregon, Washington, and British Columbia. Climate in this region is generally dominated by cool maritime air that suppresses the occurrence of extreme warm events, especially close to the coast. Extreme cold events occur when air originating from high continental latitudes is advected into the region. However, such events are rare because inland mountain ranges prevent cold, dense, and often shallow Arctic airmasses from advecting westward. Many models, as well as MERRA, extend this area of negative skewness

419 further south than NARR. This suggests that these datasets may generate more
420 frequent or severe extreme cold air outbreaks than NARR. HRM3 is an outlier with
421 weak negative skewness over much of the western United States, similar to NARR.
422 Here, HRM3 is the only model that generates predominantly warm biases at the 5th
423 percentile (Figure 3). All other models are colder than NARR in this region at the 5th
424 percentile further supporting the hypothesis that the more negative skewness
425 simulated by most models results from more frequent cold outbreaks. In all cases
426 the biases at the 95th percentile are less negative than at the 5th percentile, further
427 contributing to an asymmetry in model error that disproportionately affects days in
428 the cold tail. Figure 4 also shows a low percentage of models with bias values of the
429 same sign at all percentiles consistent with disagreement in PDF shape.
430
431 Figure 9 shows skewness for the JJA period. With the exception of MM5I, the
432 general pattern of skewness for all models and MERRA differs substantially from
433 NARR, especially at the lower latitudes. All models show positive skewness along
434 the Pacific Coast, although most models have skewness that is more positive than
435 NARR in this region. This same area also shows positive SD bias in Figure 6. This
436 feature, resulting from occasional extreme warm offshore wind events advecting
437 continental air over the relatively cool ocean, is most closely captured by ECP2.
438
439 NARR has a broad area of negative skewness over the southern half of the domain
440 becoming more negative over the Gulf of Mexico and tropical Atlantic Ocean. This
441 feature is not present in most models or MERRA. Here during summer, daily

442 temperature variability is low and the occurrence of synoptic-scale weather events
443 that are often associated with advection of anomalous T_s are rare. As a result, the
444 tails of the distribution are likely influenced largely by variations in insolation,
445 precipitation, and land surface conditions. For example, soil moisture has been
446 associated with the occurrence and implicated as a source of amplification and
447 persistence for heat waves (Hong and Kalnay, 2000; D'Odorico and Porporato, 2004,
448 Fischer et al. 2007, Loikith and Broccoli 2013). On the other hand, decreased
449 insolation due to clouds and evaporative cooling from rain can result in
450 anomalously cool temperatures and climatologically humid air originating from the
451 Gulf of Mexico may enhance latent heat flux sufficiently to limit extreme heat events
452 here. It is therefore plausible that this region exhibits negative skewness as there is
453 more opportunity for unusually cool days than for extreme warm days. Ruff and
454 Neelin (2012) show a wide cold tail using Houston, TX JJA station data. Loikith and
455 Broccoli (2012) also show negative skewness in this region in July using coarser
456 resolution, gridded temperature observations from the HadGHCND dataset (Caesar
457 et al. 2006). Peron and Sura (2013) show negative skewness over most of the
458 southern United States using NCEP/NCAR Reanalysis 1. In this case, the majority of
459 models and MERRA may have truncated cold tails due to improper simulation of
460 cloudiness or precipitation. Drought that is too frequent or severe over the region
461 could also act to widen the warm tail. It is also possible that higher statistical
462 moments are sensitive to slight differences in PDF shape when there is
463 climatologically low T_s variability (i.e. narrow PDF). These factors make intuitive

model evaluation of skewness difficult in this case because of the large spread in JJA skewness values across datasets.

Figure 10 summarizes the results in Figures 8 and 9 in the form of a scatter plot. The percentage of grid points in the domain that have positive and negative skewness are plotted on the x and y-axes respectively for each sub-region. DJF (JJA) is plotted with diamonds (circles) color coded according to dataset and the size is proportional to the mean skewness in that sub-region. Open (filled) markers are where mean skewness is negative (positive). Only grid points with statistically significant skewness at the 10% significance level are included making it possible for the sum of the percentages to be less than 100%.

As indicated in Figures 8 and 9, the models cluster around NARR (black markers) more closely in DJF than in JJA where skewness differs more. The spread in JJA is largest in the Central and Eastern and smaller in the Western and Northern sub-regions while the spread is smallest in the Western sub-region in DJF. NARR is more negative and has a larger negative skewness percentage in the Central and Eastern sub-regions compared with the other datasets as seen in Figure 9 in JJA. In general, most datasets agree on the proportion of negative skewness grid points over the Western sub-region in DJF with some disagreement in magnitude. Here all datasets are clustered close to each other; WRFG and ECP2 show more negative skewness than NARR and HRM3 shows less negative skewness than NARR.

4c. Individual cases

The PDFs for four individual grid points are plotted in Figure 11. Each case corresponds to an example used in Ruff and Neelin (2012, from now on RN2012) using station data (1950-2009). Here, skewness and variance are examined for each example using RN2012 as observational support for PDF asymmetry. All locations are chosen as the closest grid point to the actual observation station located at the major airport for each city. All PDFs are defined as frequencies of occurrence computed from temperature anomalies binned every 0.5 degrees. For reference, Gaussian PDFs are plotted with the same SD as the NARR and MERRA data. All PDFs are plotted on a log scale.

The top two panels in Figure 11 are DJF examples for (a) Seattle and (b) Chicago. All datasets exhibit a Gaussian-like warm tail and a long cold tail for both locations as supported by the negative skewness values. In both of these locations, RN2012 show long cold tails, with the asymmetry more pronounced in Seattle. NARR, MERRA, and the ensemble mean all have negative skewness with Seattle exhibiting skewness that is more negative than Chicago. In general, model variance is higher in Seattle than reanalysis, but similar in Chicago. Fig. 8 indicates that in all datasets, Seattle is positioned near the strongest (coastal) part of a long, large-scale feature of negative skewness that stretches from the West Coast to near Chicago. This suggests a substantial role of large scales in the air mass advection creating these long cold tails. While this may make it less surprising that the models do qualitatively well at

509 capturing the long tail in this region, it also helpa to boost confidence in using these
510 models to predict changes in this feature.
511
512 The bottom-left panel in Figure 11 is for Houston, Texas where RN2012 show a
513 wide cold tail. Station data and NARR both show negative skewness here while
514 MERRA shows positive skewness and the ensemble mean zero skewness. The cold
515 tail difference between NARR and MERRA is slightly greater than the difference in
516 the warm tail, consistent with the hypothesis that MERRA and the RCMs are not
517 properly simulating conditions associated with unusually cold days (see section *4b*).
518 Variance is similar between the two reanalysis prodcuts and slightly larger for the
519 model mean. It is apparent from this example that the disagreement here is largely
520 in the most extreme temperature days while the cores of the PDFs are generally
521 similar.
522
523 RN2012 show wide warm tails in Los Angeles and Long Beach, California station
524 data for JJA. In this region, the prevailing surface wind trajectory is from the
525 relatively cool Pacific Ocean, preventing large temperature excursions on the cold
526 side of the PDF while infrequent offshore wind events can cause large excursions on
527 the warm side (e.g. Santa Ana Winds). The bottom right panel of Figure 11 shows
528 positive skewness for NARR, MERRA, and the ensemble mean similar to the station
529 observations. The warm side deviation from Gaussian is greater for MERRA and the
530 models than NARR and there is a substantial difference in variance with NARR
531 having lower variance than all other datasets. Because this is a coastal location

where there is a sharp variance gradient (small over the ocean, large over land), some of the variance difference could be an artifact of regridding where marine and continental grid points are interpolated. The maps in Fig. 9 indicate that all models and both reanalysis data sets have a region of positive skewness along most of the West Coast, and that NARR differs mainly in the geographic width and strength of this feature, especially near Los Angeles. The fact that the models do better than NARR with respect to the station data at capturing this feature should thus be regarded as a quantitative rather than qualitative difference. It is encouraging that the models can reproduce this feature of the station data to a reasonable extent.

5. Cluster Analysis

Sections 3 and 4 have focused on evaluating PDFs moments individually while PDF shape is described by multiple moments. A methodology that is capable of evaluating more than one statistical moment of the PDF could provide a useful tool in holistically evaluating model-simulated PDFs. As an early step towards this goal, k-means clustering (Wunsch and Xu, 2008) is introduced as a tool for comparing model PDFs to reference data over a large geographic area. Loikith et al. (2013) introduced and demonstrated the efficacy of k-means clustering as a tool for characterizing T_s regimes based on PDF characteristic by clustering over PDFs using NARR and MERRA data. Here, k-means clustering is applied in a similar manner but to multiple data sets for the purpose of comparison in DJF.

554 Loikith et al. (2013) showed that k=5 clusters provide stable, easily interpretable
555 PDF groups. Here, k=4 clusters are chosen because the domain size in this work is
556 smaller so it follows that the number of climate regimes is smaller. In the application
557 demonstrated here, the choice of the number of clusters is arbitrary. Ongoing work
558 that further explores the applicability of the clustering technique as a tool for model
559 evaluation will more thoroughly address the issue of optimal cluster number.
560
561 Model cluster assignments in this case use the NARR cluster basis PDFs to allow for
562 intuitive comparison between datasets. First, k-means clustering is performed on
563 the NARR dataset. The basis PDF, defined as the mean PDF for each of the four
564 clusters, is then calculated for NARR (Figure 12). For each RCM, the RMS differences
565 between the PDF at each grid point and each of the four basis PDFs from NARR are
566 calculated. Finally, the cluster corresponding to the minimum RMS difference is
567 assigned to the RCM grid point. In other words, each grid point in an RCM is
568 assigned to a cluster based on which basis PDF most closely resembles the RCM
569 PDF.
570
571 Maps showing the cluster assignments for each grid point are shown in Figure 13
572 The clusters are assigned numbers based on monotonically decreasing mean cluster
573 variance with cluster 1 (C1) having the highest variance and cluster 4 (C4) having
574 the lowest. A scatter plot showing the mean skewness and SD for each cluster for
575 each dataset is presented in Figure 14.
576

577 As shown in Loikith et al. (2013), temperature variance is the most prominent
578 feature reflected in the cluster assignments when performed in this manner;
579 however skewness also appears to have some influence on cluster assignment,
580 especially the first two clusters (Figure 14). C1, which has the highest variance,
581 comprises a band in the northern portion of the domain in NARR and MERRA,
582 surrounded to the north and to the south by C2. In the models, C1 generally extends
583 further north and east than in NARR with the exception of the ensemble mean and
584 MERRA. This expansion is consistent with the positive SD bias seen in Figure 5. In
585 the reference data, the position of the region assigned to C1 is optimal for allowing
586 horizontal temperature advection (cold from the north and warm from the south) to
587 cause large excursions from the mean (see section *4a*). The width of the PDF here is
588 largely influenced by large-scale atmospheric circulation patterns within the main
589 storm track as well as possible orographic effects in the mountainous areas. In
590 models that extend C1 further north, it is possible that the storm track is wider than
591 in reanalysis. Figure 14 shows that skewness is negative in all datasets and all
592 models have mean skewness values that are more positive than NARR in C1. This
593 bias is likely due to the northward expansion of C1, relative to NARR, where
594 skewness is predominantly positive.
595
596 C2 surrounds C1 on the north and south and covers most of the eastern US and
597 Canada in NARR. The region of C2 that is to the south of C1, is dominated by
598 synoptic-scale weather patterns associated with strong horizontal temperature
599 advection. This results in temperature PDFs with relatively high variance, but still

600 lower than in C1. Whereas C1 is optimally positioned between the warmest and
601 coldest air in the hemisphere, C2 is marginally closer to these regions (cold on the
602 north side and warm on the south side) resulting in a reduced potential for
603 extremely large deviations in temperature. All datasets have positive or near 0
604 mean skewness for C2 with a generally even spread about NARR on both the
605 positive and negative side (Figure 14). In general the southern portion of C2 is
606 geographically similar in all models and reanalysis products suggesting a consistent
607 representation of synoptic-scale weather patterns here while the northern portion
608 is extended further north.
609
610 C3 encompasses the southern and western portions of land as well as coastal waters
611 of the Gulf of Mexico and Atlantic Ocean. This region has the lowest variance of the
612 four clusters over land while the portions over the ocean have the highest variance
613 relative to other ocean grid points. Over land, this area is on the southern edge of
614 the main winter storm track, which results in fewer and weaker strong advection
615 events compared with C1 and C2. In the west, terrain features also likely play a role
616 in the generally lower variance as horizontal temperature advection is impeded by
617 mountains and the horizontal resolution of the datasets is too low to capture local,
618 orographically-induced variability. The near-coastal waters are located downwind
619 of the continent and within the storm track causing temperature variance to be
620 elevated relative to surrounding ocean areas. The position of the Gulf Stream,
621 parallel to the coast, increases the horizontal temperature gradient in near-surface
622 temperature also likely contributing to locally large temperature variance. All

623 models capture this region with similar mean SD. All datasets show negative mean
624 skewness and there is no systematic skewness bias in Figure 14, with NARR in the
625 middle of the spread. C4 is comprised almost entirely of ocean grid points where
626 variance is smallest. All datasets capture these characteristics with a small spread
627 in mean SD and skewness.
628
629 This application of k-means cluster analysis provides an overall comparison
630 between model and reference data PDFs; however variations on this tool could
631 provide additional information. For example, Loikith et al. (2013) cluster over PDFs
632 of normalized temperature anomalies which assigns clusters based on higher-
633 moment statistics such as skewness and kurtosis.
634
635 **6. Summary and Conclusions**
636 Multiple methodologies are employed to evaluate daily surface temperature
637 distributions from a suite of six NARCCAP RCM hindcast experiments against NARR.
638 Model biases are identified and quantified with many models showing systematic,
639 and in some cases large, shifts in the temperature distribution at all percentiles. In
640 many cases, additional PDF structure biases are found. While temperature biases,
641 especially biases that are systematic across the entire probability distribution, can
642 be accounted for and corrected in model output, error in model-simulated PDF
643 shape is more problematic. In particular, error related to the tails of model-
644 simulated PDFs will impact the accuracy with which models simulate extremes.
645

Two reanalysis datasets, NARR and MERRA, were used in an attempt to gauge uncertainties in reference data. In general, the differences between NARR and MERRA were smaller in magnitude than any individual model; however, both datasets show large differences in summertime skewness in the southern portion of the domain and wintertime temperature in the northern regions. An intercomparison of other reanalysis and observational data sets could provide better constraint on observational uncertainty, however the relatively high resolution of the two reanalysis products used here provides detail on regional variations that other products cannot.

Variance is generally higher than NARR across all models in the northern portion of the domain in winter and throughout the domain in summer while in winter variance is smaller than NARR in the south (Figures 5 and 6). In some cases, models that exhibit positive variance biases also show positive temperature biases at all percentiles in the same regions indicating a shifted and wider PDF relative to NARR (Figure 3). Other models show a widening of the PDF without such a systematic shift.

The major patterns in skewness i.e. positive skewness in the northeastern part of the domain, negative to the south, are realistic in most models in the winter (Figure 7). Summertime skewness exhibits regions of substantial difference, especially along the Gulf of Mexico, across all datasets (Figure 8). Several factors may be related to these discrepancies including differing cloud and precipitation

669 representation while skewness may also be highly sensitive to slight changes in PDF
670 structure in these regions with low temperature variance.
671
672 Comparison of temperature PDFs for selected locations to those previously analyzed
673 from station data (Figure 11) can be particularly useful when interpreted in light of
674 these skewness maps. Long cold tails in the distribution of wintertime daily
675 temperature anomalies seen for locations such as Seattle and Chicago are
676 reasonably well simulated in the models. These are part of a coherent region of
677 negative skewness that stretches from the US Northwest to the Great Lakes region
678 that is likewise reproduced in the models with reasonable fidelity. Long warm tails
679 in the summer temperature distribution for the Los Angeles region are qualitatively
680 reproduced in the models and form part of a coherent positive skewness region that
681 stretches along most of the North American West Coast. For such features that
682 validate reasonably well, the models may be used in future work to further analyze
683 the dynamics yielding the long tails. Predictions of changes in extreme temperature
684 occurrences, for instance under global warming, may also be more reliable for these
685 regions where the tail characteristics for present climate are comparable to
686 observations. On the other hand, identifying regions such as along the Gulf of Mexico
687 in the summer where the skewness and tail characteristics do not validate well can
688 help pinpoint regions where confidence would currently be lower in statements
689 about extreme temperature occurrences, and where model development efforts
690 might productively be focused.
691

692 The application of k-means clustering for comparing model-simulated PDFs to
693 reference data is introduced using winter temperatures (Figures 12-14). In general,
694 clusters assignments reflect temperature variance; however skewness is also
695 reflected. The cluster assignments in the models tend to resemble NARR with some
696 differences primarily over the higher latitudes where the cluster with the highest
697 variance is expanded northward compared with NARR. This disagreement reflects
698 the positive variance biases found in this region. This application of k-means
699 clustering has potential to be a versatile evaluation tool and is the focus of ongoing
700 research and development.
701
702 An important future direction in understanding RCM PDF uncertainty, and the
703 inherent relationship this uncertainty has to temperature extremes, is to use this
704 information to investigate mechanisms that are linked to model error. While
705 evidence exists connecting extreme temperature events to larger-scale, low-
706 frequency modes of climate variability such as the El Niño-Southern Oscillation and
707 the Arctic Oscillation (Kenyon and Hegerl 2007), which largely occur outside of the
708 domain of these RCMs, Loikith and Broccoli (2013) show that in many places
709 extreme temperatures are associated with local, amplified, transient weather events
710 which could be examined on an RCM domain. Evaluation of such mechanisms will
711 further identify discrepancies in dynamical processes. Additional analysis of model-
712 simulated soil moisture, cloud cover, and precipitation will also be useful for
713 understanding error in summertime extremes.
714
715

Acknowledgements

Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Part of this research was funded by NASA National Climate Assessment 11-NCA11-0028 and AIST AIST-QRS-12-0002 projects and the NSF ExArch 1125798 (P.C.L., J.K., H.L., D.E.W., C.M.). Part of this research was funded by NOAA NA11OAR4310099 (J.D.N.).

References

Beniston, M., 2004: The 2003 heat wave in Europe: A shape of things to come? An analysis based on Swiss climatological data and model simulations. *Geophys. Res. Lett.*, **31**, L02202, doi:10.1029/2003GL018857.

Bukovsky, M. S., 2012: 2012, Temperature Trends in the NARCCAP Regional Climate Models. *J. Climate*, **25**, 3985-3991.

Caesar, J., L. Alexander, and R. Vose, 2006: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *J. Geophys. Res.*, **111**, D05101, doi:10.1029/2005JD006280.

Crichton, D.J., C.A. Mattmann, L. Cinquini, A. Braverman, D.E. Waliser, M. Gunson, A. Hart, C. Goodale, P.W. Lean, and J. Kim, 2012: Software and Architecture for Sharing Satellite Observations with the Climate Modeling Community. *IEEE Software*, **29**, 63-71.

D'Agostino, R. B., 1970: Transforming to normality of the null distribution of *g1*. *Biometrika*, **57**, 679-681.

D'Odorico P, Porporato A. Preferential states in soil moisture and climate dynamics. Proceedings of the National Academy of Sciences of the United States of America 101:8848-8851, 2004.

Dole, R., M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, T. Zhang, X.-W. Quan, T. Xu, and D. Murray, 2011: Was there a basis for anticipating the 2010 Russian heat wave? An analysis based on Swiss climatological data and model simulations. *Geophys. Res. Lett.*, **38**, L06702, doi:10.1029/2010GL046582.

762 Donat, M. G., and L. Alexander, 2012: The shifting probability distribution of global
763 daytime and night-time temperatures. *Geophys. Res. Lett.*, **39**, L14707,
764 doi:1029/2012GL052459.
765 Fischer, E. M., S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär, 2007: Soil
766 moisture-atmosphere interactions during the 2003 European summer heat
767 wave. *J. Climate*, **20**, 5081-5099.
768 Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate
769 models. *J. Geophys. Res.*, **113**, doi: 10.1029/2007JD008972.
770 Gómez-Navarro, J. J., J. P. Montávez, S, Jerez, P. Jiménez-Guerrero, and E. Zorita,
771 2012: What is the role of the o bservational dataset in the evaluation and
772 scoring of climate models? *Geophys. Res. Lett.*, **39**, L24701,
773 doi:10.1029/2012GL054206.
774 Griffiths, G. M., et al., 2005: Change in mean temperature as a predictor of extreme
775 temperature change in the Asia-Pacific region, *Int. J. Climatol.*, **25**, 1301–
776 1330.
777 Hart, A. F., C. E. Goodale, C. A. Mattmann, P. Zimdars, D. Crichton, P. Lean, J. Kim, and
778 D. Waliser, 2011: A cloud-enabled regional climate model evaluation system.
779 Waikiki, Honolulu, HI.
780 Hegerl, G. C., F. W. Zwiers, P. A. Stott, and V. V. Kharin, 2004: Detectability of
781 anthropogenic changes in annual temperature and precipitation extremes. *J.*
782 *Climate*, **17**, 3683-3700.

783 Hong, S.-Y. and Kalnay, E.: Role of sea surface temperature and soil-moisture
784 feedback in the 1998 Oklahoma–Texas drought, Nature, 408, 842–844,
785 doi:10.1038/35048548, 2000.
786 Hughes, M., and A. Hall (2010), Local and synoptic mechanisms causing Southern
787 California's Santa Ana winds, *Clim. Dyn.*, **34**, 847-857.
788 IPCC, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate*
789 *Change Adaptation. A Special Report of Working Groups I and II of the*
790 *Intergovernmental Panel on Climate Change*, edited by C. B. Field, V. Barros, T.
791 F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K.
792 Plattner, S. K. Allen, M. Tignor, and P. M. Midgley, Cambridge University Press,
793 Cambridge, UK, and New York, NY, USA, 582 pp.
794 Kanamitsu, M., W. Ebisuzaki, J. Woollen, S-K Yang, J. J. Hnilo, M. Fiorino, and G. L.
795 Potter, 2002: NCEP-DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.*,
796 **83**, 1631-1643.
797 Karl, T. R., B. E. Gleason, M. J. Menne, J. R. McMahon, R. R. Heim Jr., M. J. Brewer, K. E.
798 Kunkel, D. S. Arndt, J. L. Privette, J. J. Bates, P. Y. Groisman, and D. R.
799 Easterling, 2012: U. S. Temperature and Drought: Recent Anomalies and
800 Trends. *EOS*, **93**, 473-496.
801 Kenyon, J., and G. C. Hegerl, 2008: Influence of modes of climate variability on global
802 temperature extremes. *J. Climate*, **21**, 3872-3889.
803 Kim, J., D. E. Waliser, C. A. Mattmann, L. O. Mearns, C. E. Goodale, A. F. Hart, D. J.
804 Crichton, S. McGinnis, H. Lee, P. C. Loikith, M. Boustani, 2013: Evaluation of
805 the surface air temperature, precipitation, and insolation over the

conterminous U.S. in the NARCCAP multi-RCM hindcast experiment using RCMES. *J. Climate*, *in press*.

Kim, J., D. E. Waliser, C. Mattmann, C. Goodale, A. Hart, P. Zimdars, D. Crichton, C. Jones, G. Nikulin, B. Hewitson, C. Jack, C, Lennard, and A. Favre, 2013: Evaluation of the CORDEX-Africa multi-RCM Hindcast: Systematic Model Errors. *Climate Dynam.*, DOI 10.1007/s00382-013-1751-7.

Kjellström, E., F. Boberg, M. Castro, J. Hesselbjerg Christensen, G. Nikulin, E. Sánchez, 2010: Daily and monthly temperature and precipitation statistics as indicators for regional climate models. *Clim. Res.*, **44**, 135-150.

Kjellström, E., G. Nikulin, U. Hansson, G. Strandberg, and A. Ullerstig, 2011: 21st century changes in the European climate: uncertainties derived from an ensemble of regional climate model simulations. *Tellus*, **63A**, 24-40.

Lau, Ngar-Cheung, Mary Jo Nath, 2012: A Model Study of Heat Waves over North America: Meteorological Aspects and Projections for the Twenty-First Century. *J. Climate*, **25**, 4761–4784.

Loikith, P. C., and A. J. Broccoli, 2012: Characteristics of observed atmospheric circulation patterns associated with temperature extremes over North America. J. Clim., **20**, 7266-7281.

Loikith, P. C., and A. J. Broccoli: The Influence of Recurrent Modes of Climate Variability on the Occurrence of Extreme Temperatures over North America. *J. Climate*, *in review*.

Loikith, P. C., B. R. Lintner, J. Kim, H. Lee, J. D. Neelin, D. E. Waliser: Classifying reanalysis surface temperature probability density functions (PDFs) over

North America with cluster analysis. *Geophys. Res. Lett.*, **40**, doi:10.1002/grl.50688.

Luber, G., and M. McGeehin, 2008: Climate change and extreme heat events. *Am. J. Prev. Med.,* **35**, 429-435, doi: 10.1016/j.amepre.2008.08.021.

Mearns, L.O., et al., 2007, updated 2012. *The North American Regional Climate Change Assessment Program dataset*, National Center for Atmospheric Research Earth System Grid data portal, Boulder, CO. Data downloaded 2013-01-15. [doi:10.5065/D6RN35ST]

Mearns, L. O., W. J. Gutowski, R. Jones, L.-Y. Leung, S. McGinnis, A. M. B. Nunes, and Y. Qian, 2009: A regional climate change assessment program for North America. *EOS*, **90**, 311-312.

Mearns, L. O., and coauthors, 2012: The North American Regional Climate Change Assessment Program: Overview of Phase I Results. *Bull. Amer. Meteor. Soc.*, 93, 1337-1362.

Meehl, G.A., T.F. Stocker, W.D. Collins, P. Friedlingstein, A.T. Gaye, J.M. Gregory, A. Kitoh, R. Knutti, J.M. Murphy, A. Noda, S.C.B. Raper, I.G. Watterson, A.J. Weaver and Z.-C. Zhao, 2007: Global Climate Projections. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the FourthAssessment Report of the Intergovernmental Panel on Climate Change [Solomon, S.,D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Meehl, G. A., C. Tebaldi, G. Walton, D. Easterling, and L. McDaniel, 2009: Relative

852 increase of record high maximum temperatures compared to record low
853 minimum temperatures in the U.S. *Geophys. Res. Lett.*, **36**, L23701,
854 doi:10.1029/2009GL040736.
855 Meehl, G. A., and C. Tebaldi, 2004: More intense, more frequent, and longer lasting
856 heat waves in the 21st century. *Science*, **305**, 994-997.
857 Mesinger, F., and co-authors (2006), North American Regional Reanalysis, *Bull.*
858 *Amer. Meteor. Soc., 87*, 343–360.
859 Otto, F. E. L., N. Massey, G. J. van Oldenborgh, R. G. Jones, and M. R. Allen, 2012:
860 Reconciling two approaches to attribution of the 2010 Russian heat wave.
861 *Geophy. Res. Lett.*, **39**, L04702, doi:10.1029/2011GL050422.
862 Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the
863 AR4 Climate Models' Simualted Daily Maximum Temperature, Minimum
864 Temperature, and Precipitation over Australia Using Probability Density
865 Functions. *J. Climate*, **20**, 4356-4376.
866 Perron, M. and P. Sura, 2013: Climatology of Non-Gaussian Atmospheric Statistics. *J.*
867 *Climate*, **26**, 1063-1083.
868 Rangwala, I., J. Barsugli, K. Cozzetto, J. Neff, and J. Prairie, 2012: Mid-21st century
869 projections in temperature extremes in the southern Colorado Rocky
870 Mountains from regional climate models. *Clim. Dyn.*, DOI 10.1007/s00382-
871 011-1282-z.
872 Rahmstorf, S. and D. Coumou, 2011: Increase of extreme events in a warming world.
873 *Proc. Nat. Acad. Sci.*, doi:10.1073/pnas.1101766108.

874 Ruff, T. W., and J. D. Neelin, 2012: Long tails in regional surface temperature
875 probability distributions with implications for extremes under global
876 warming. *Geophys. Res. Lett.*, doi:10.1029/2011GL050610.
877 Rhines, A., and P. Huybers, 2013: Frequent summer temperature extremes reflect
878 changes in the mean, not the variance. *Proc. Nat. Acad. Sci*, **110**, E546.
879 Rienecker, M.M., M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G.
880 Bosilovich, S.D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A.
881 Conaty, A. da Silva, et al., 2011. MERRA: NASA's Modern-Era Retrospective
882 Analysis for Research and Applications. *J. Clim., 24*, 3624—3648,
883 doi:10.1175/JCLI-D-11-00015.1.
884 Ruff, T. W., and J. D. Neelin, 2012: Long tails in regional surface temperature
885 probability distributions with implications for extremes under global
886 warming, *Geophys. Res. Lett.,* **39**, L04704, doi:10.1029/2011GL050610.
887 Schär, C., P. L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. A. Liniger, and C. Appenzeller,
888 2004: The role of increasing temperature variability in European summer
889 heatwaves. *Nature*, **427**, 332-336.
890 Stott, P. A., D. A. Stone, and M. R. Allen, 2004: Human contribution for the European
891 heatwave of 2003. *Nature*, **432**, 610-614.
892 Tebaldi, C., K. Hayhoe., J. M. Arblaster, and G. A. Meehl, 2006: Going to the extremes,
893 an intercomparison of model-simulated historical and future changes in
894 extreme events. *Climatic Change*, **79**, 185-211.
895 D. Wunsch and R. Xu, *Clustering (IEEE Press Series on Computational Intelligence)*.
896 Washington, DC: IEEE Computer Society Press, 2008.

Figures

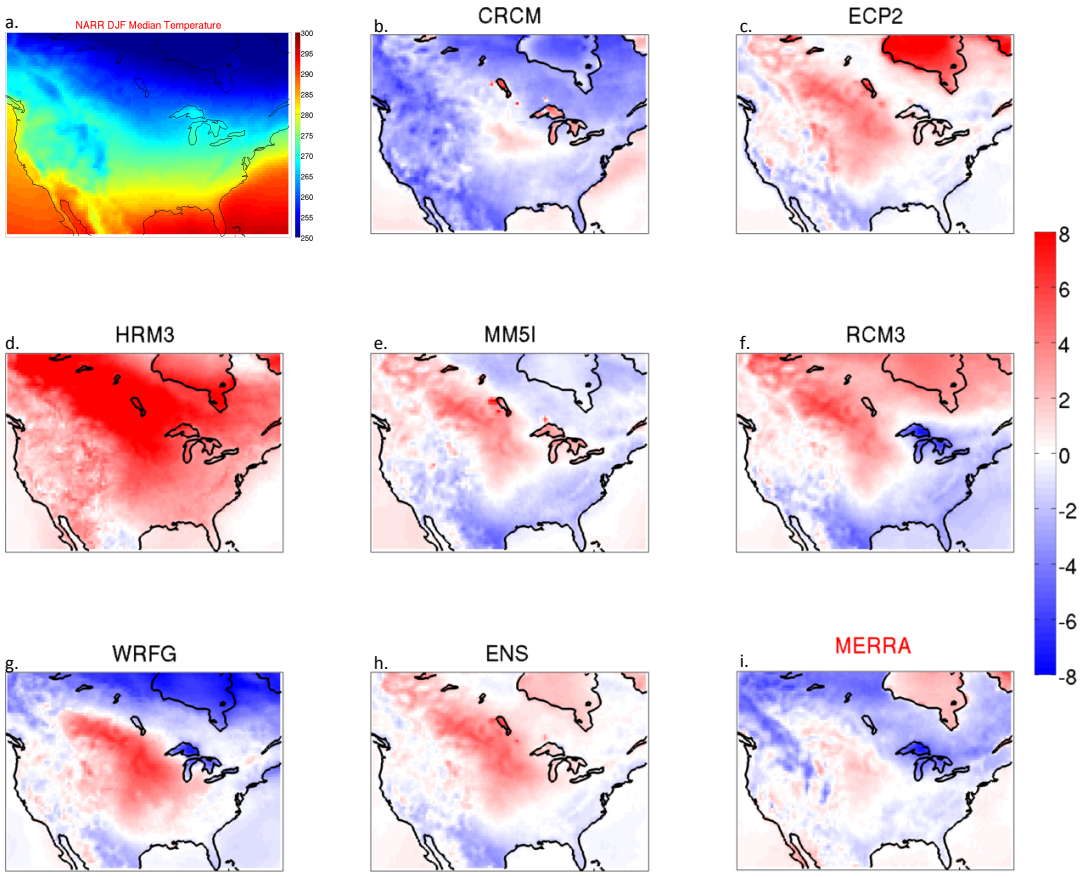


Figure 1. Median DJF temperature for NARR (a) and median DJF temperature bias for all models and the multi-model ensemble (ENS). The panel (i) is the bias in median DJF temperature for MERRA with respect to NARR.

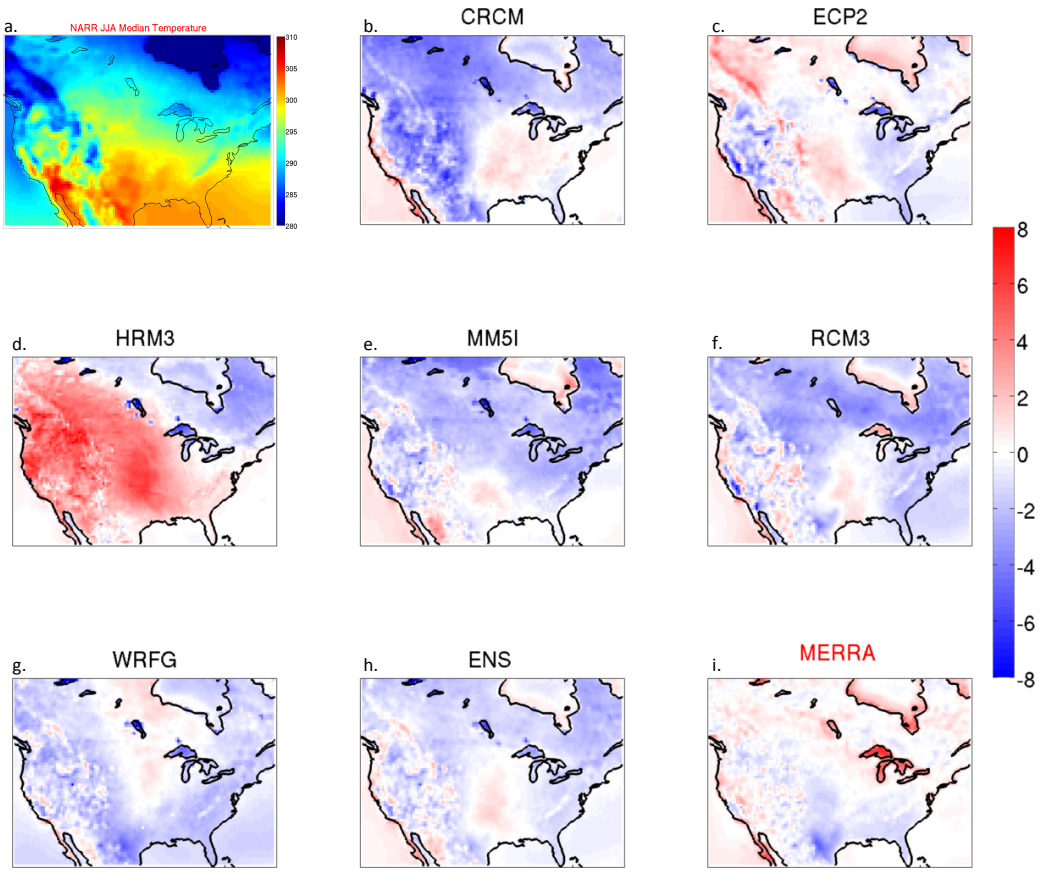


Figure 2. Same as Figure 1 except for JJA.

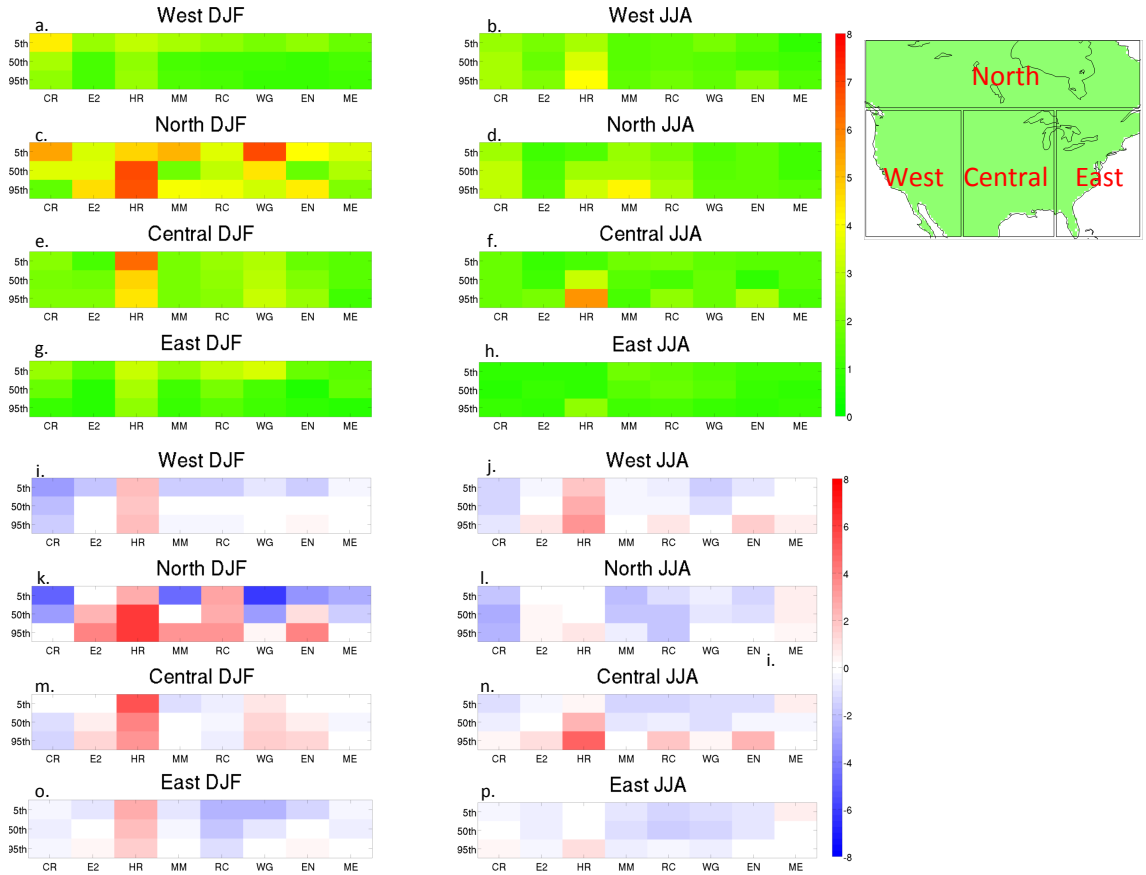
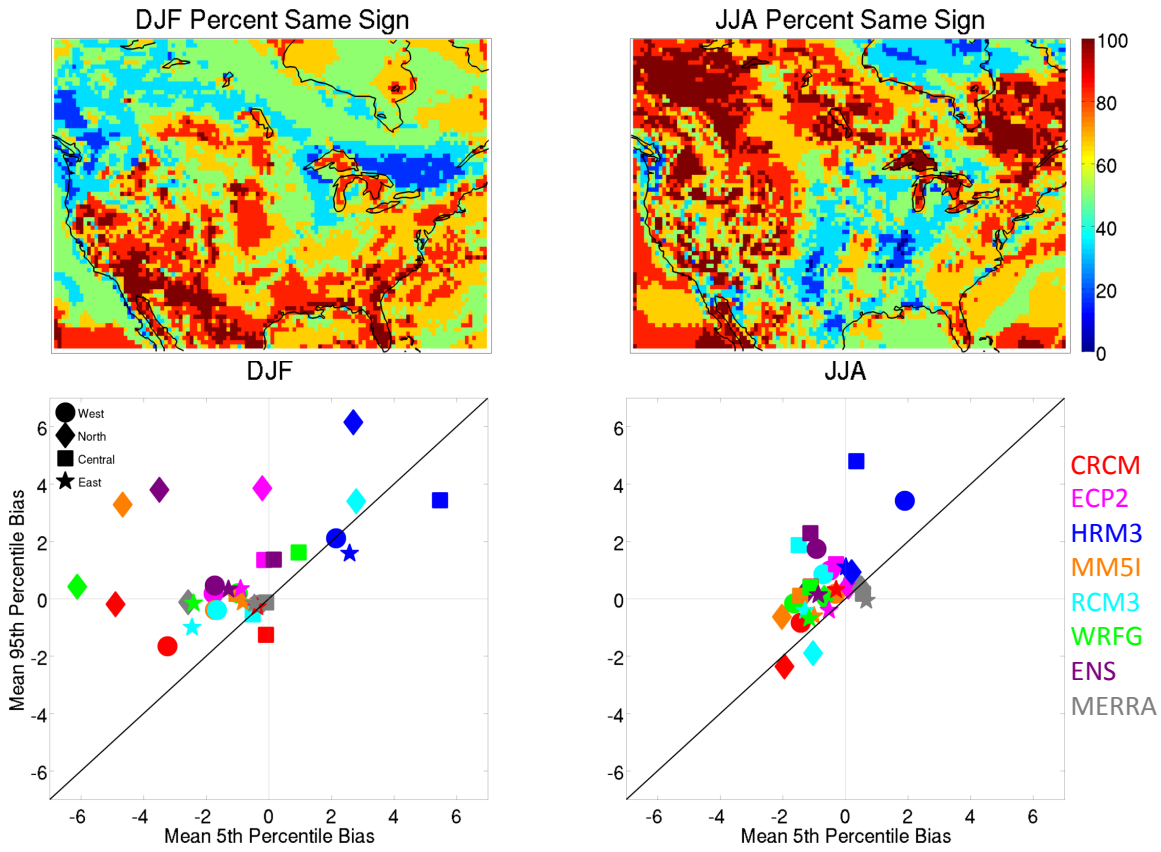


Figure 3. Portrait diagrams of (a-h) RMS bias and (i-p) mean bias over the four different domains outlined in the map for (left) DJF and (right) JJA. The models are CRCM (CR), ECP2 (E2), MM5I (MM), WRFG (WG), multi-model ensemble (ENS), and MERRA reanalysis (ME). The map at the top right outlines the domains of the sub-regions.



931

932 Figure 4. (top) Percentage of the six models that have temperature bias of the same
933 sign for the 5th, 50th, and 95th percentiles during (left) DJF and (right) JJA. Places
934 where there is a high percentage are indicative of a majority of models having a
935 systematic shift in the PDF, independent of any change in shape. (bottom) Scatter
936 plots showing the mean bias for the 5th and 95th percentiles for each sub-region as
937 defined in Figure 3. Each symbol represents a different sub-region and each color
938 represents a different dataset. Using these percentiles as benchmarks for PDF
939 width, if the dataset falls on the solid black line, the bias at the 5th percentile is the
940 same in magnitude and sign as the 95th percentile and the PDF width is the same as
941 in NARR. A model that falls to the right of the line has a net reduction in PDF width
942 and a model to the left has a net increase in PDF width.

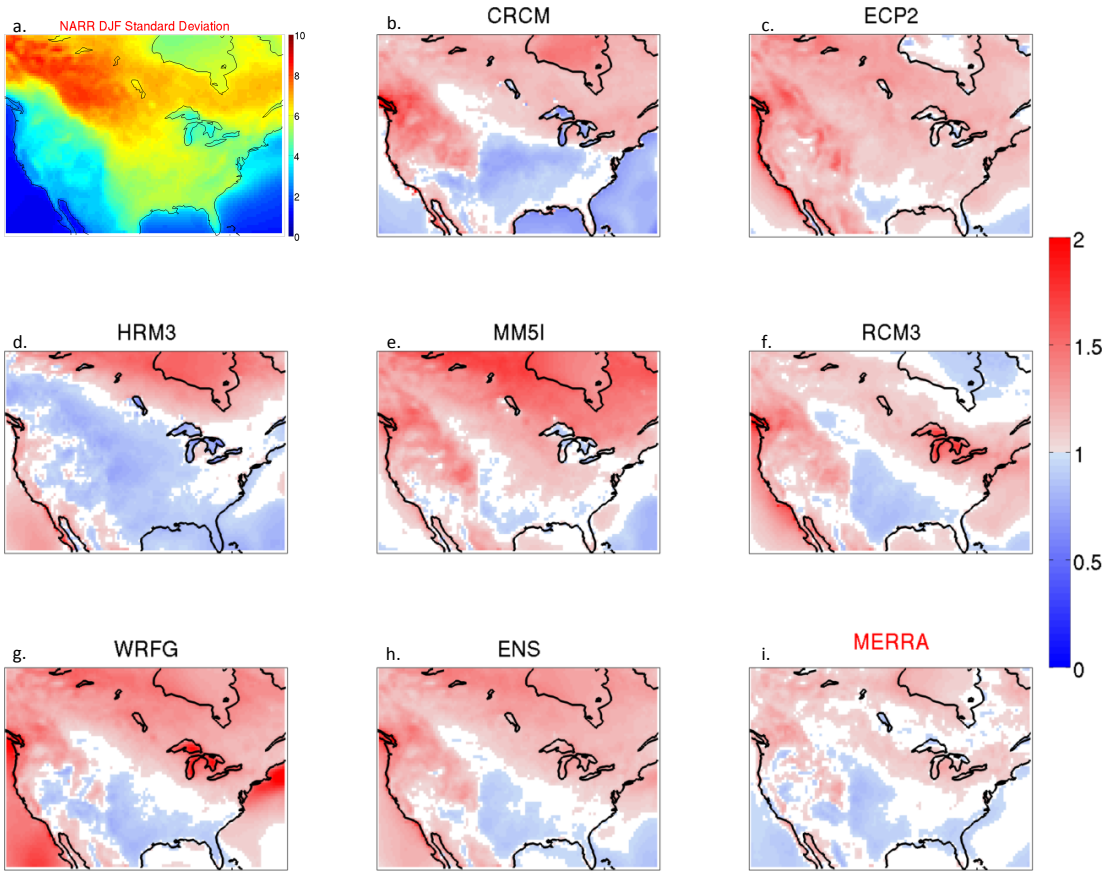


Figure 5. Ratio of model standard deviation to NARR standard deviation for DJF. Areas that are not shaded are where the difference between model and NARR standard deviation is not significant at the 5% confidence level as determined by a two-sided F-test. Panel (a) shows the actual standard deviation for NARR.

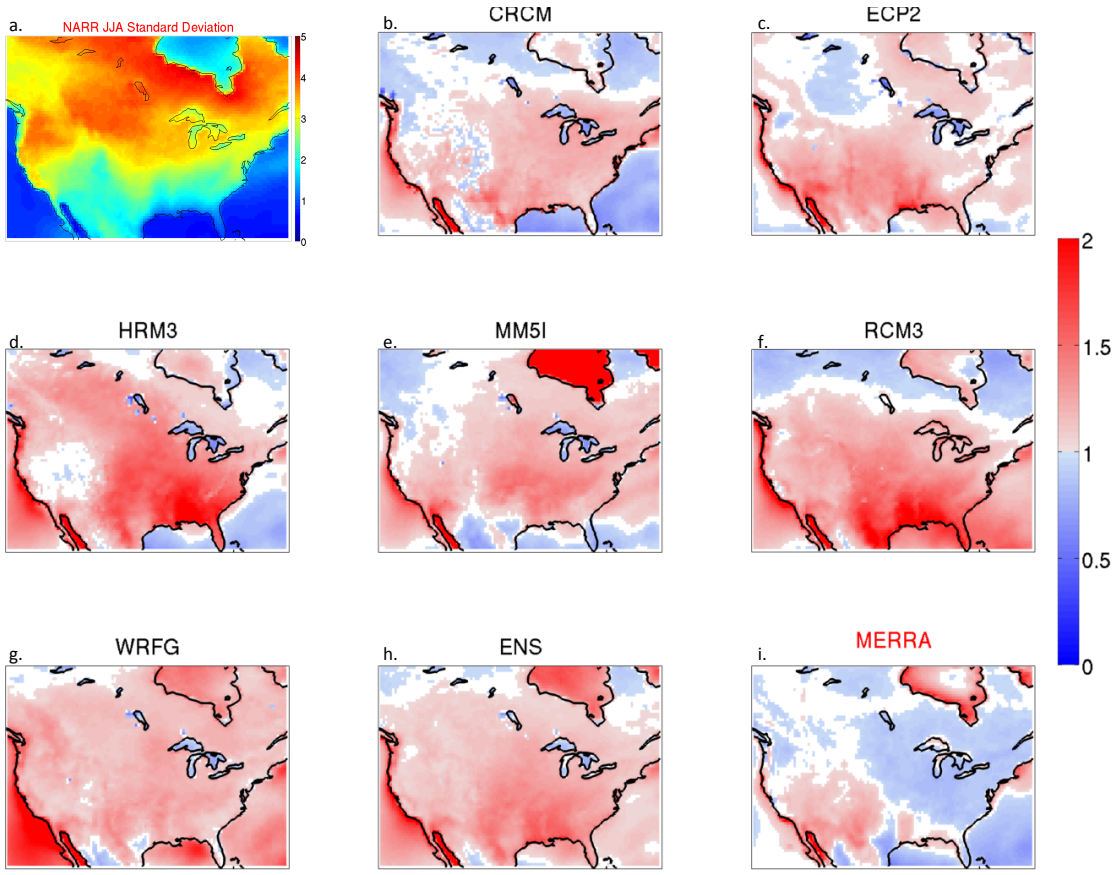


Figure 6. Same as Figure 5, except for JJA.

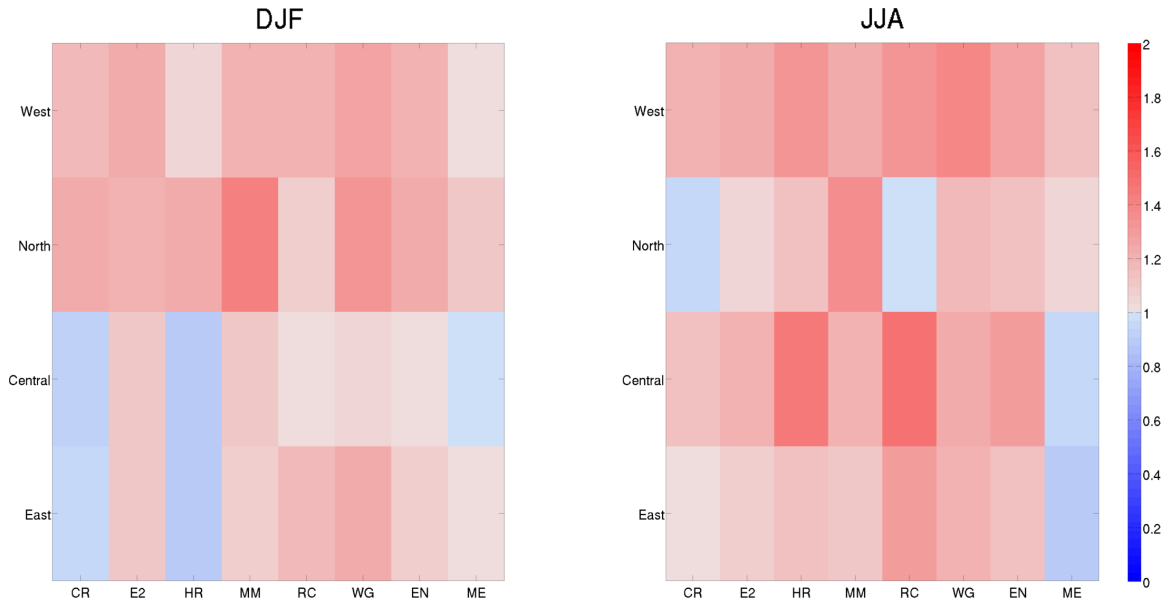


Figure 7. Portrait diagrams showing the mean ratio (as plotted in Figures 5 and 6) of model standard deviation to NARR standard deviation for each sub-region. Values greater (less) than one are indicative of a model with larger (smaller) standard deviation than NARR. Only grid points where the difference between model and NARR standard deviation is statistically significant at the 5% significance level as determined by a two-tailed F-test are included in the averaging.

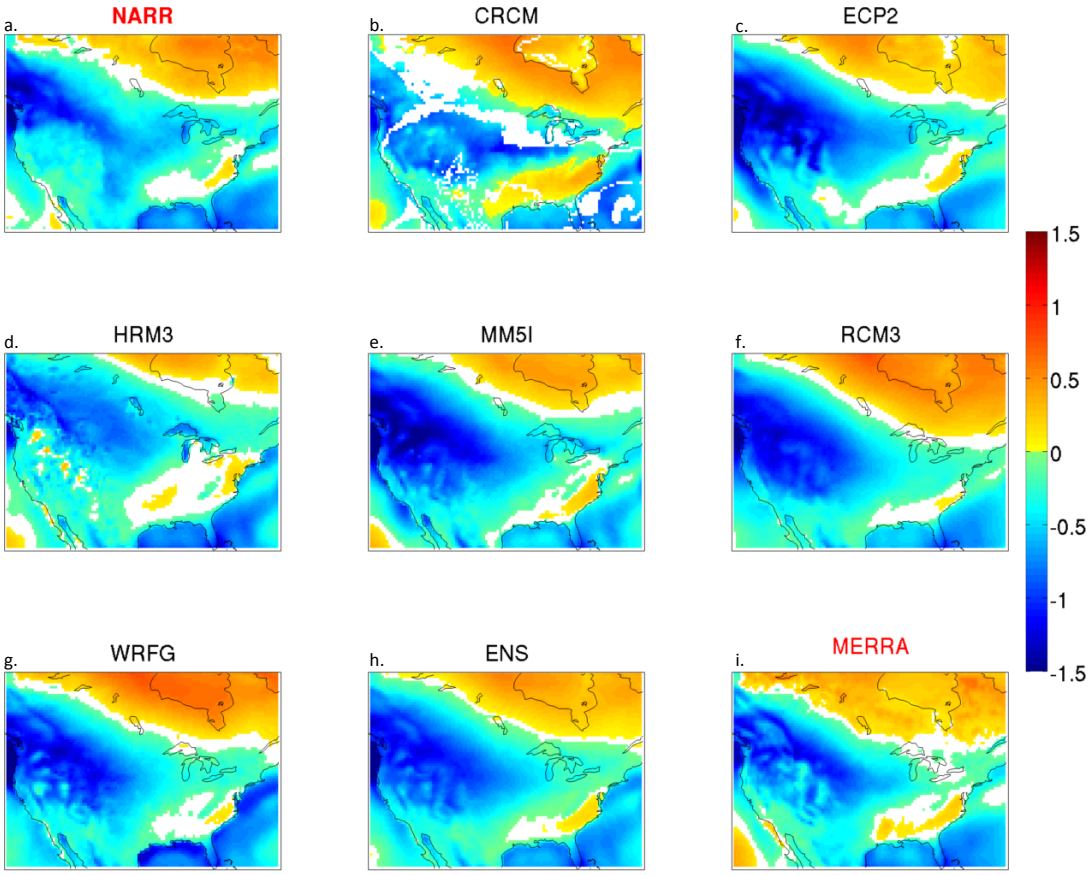
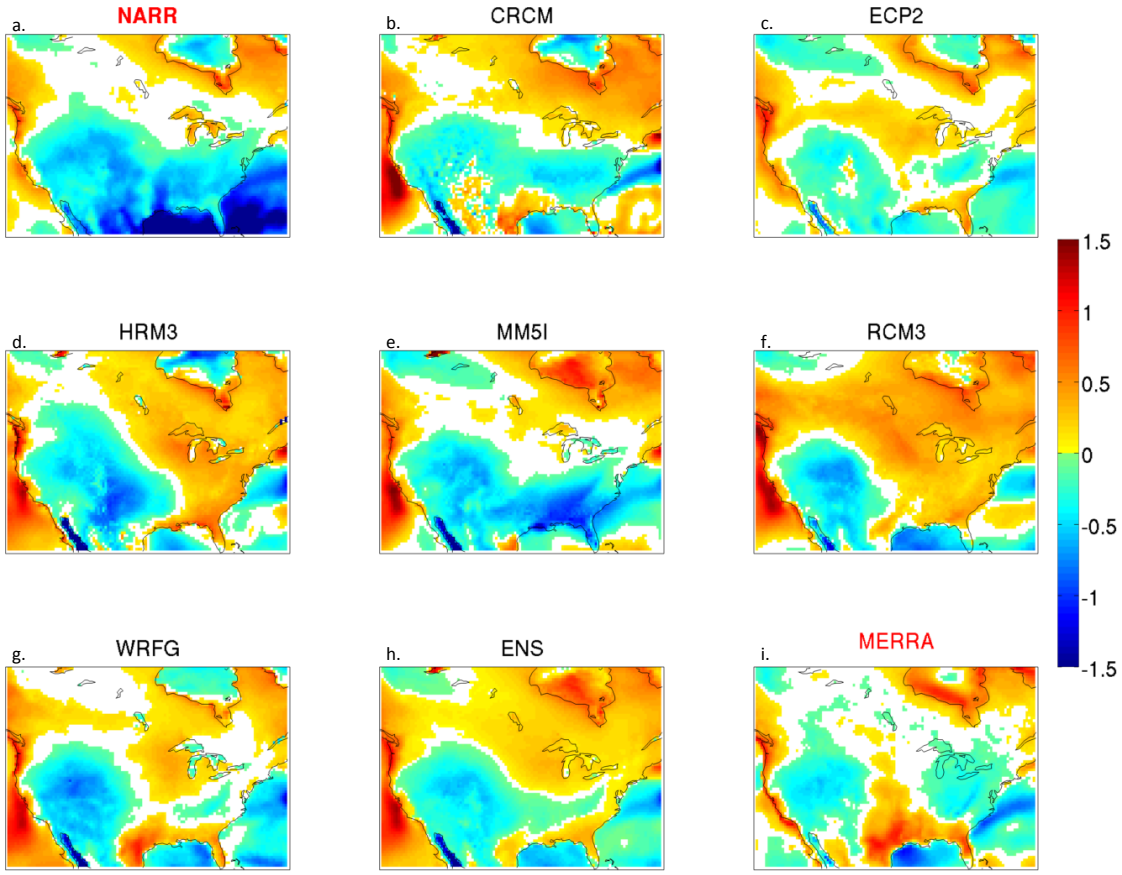


Figure 8. Skewness of January temperature. NARR is the top left panel and MERRA is at the bottom right. Grid points where skewness was not significantly different from a normal distribution at the 10% confidence threshold is not shaded.



990

991 Figure 9. Same as Figure 8 except for JJA.

992

993

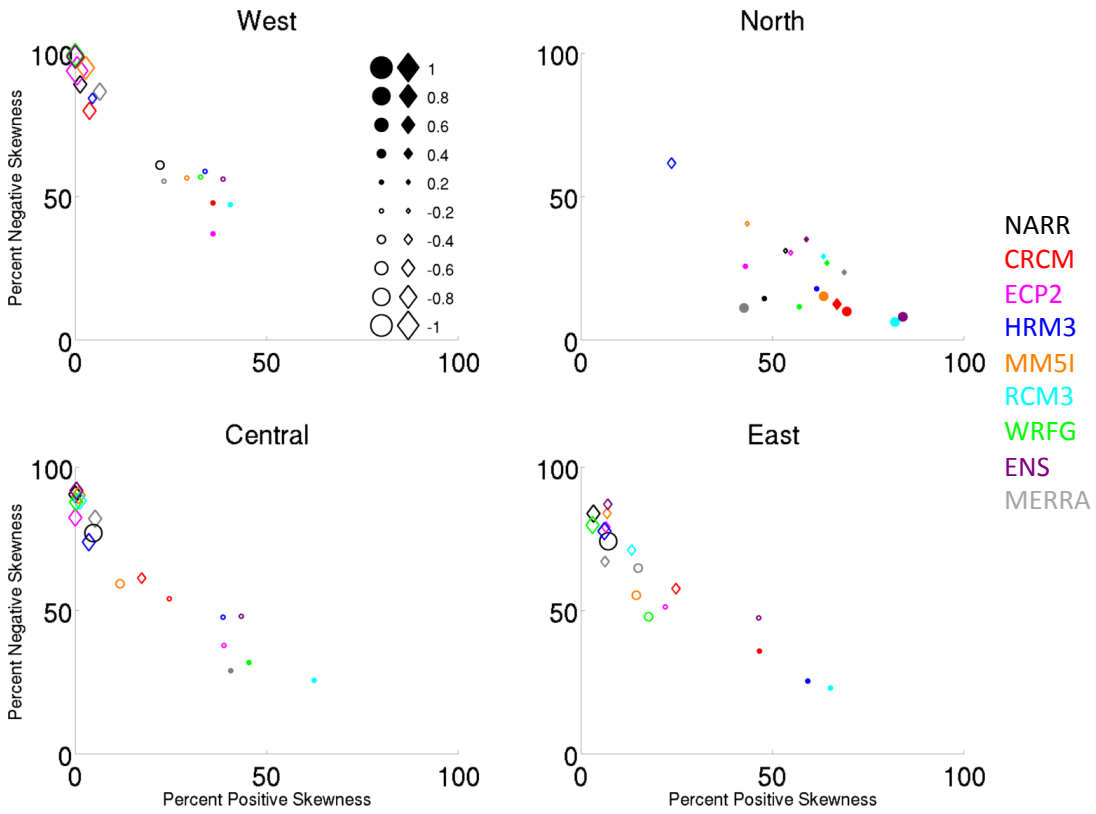
994

995

996

997

998



999

1000 Figure 10. Scatter plots showing the percent of grid cells within each sub-region
1001 that have positive skewness values versus those with negative skewness values.
1002 Only grid points with skewness that differs from a normal distribution at the 10%
1003 confidence threshold are counted as positive or negative skewness grid points
1004 (percentages of negative and positive skewness may not add up to 100). DJF
1005 skewness is plotted with a diamond and JJA with a circle. The size of the markers is
1006 proportional to the magnitude of the mean skewness of that region for each dataset.
1007 Open (filled) markers indicate mean negative (positive) skewness with larger
1008 markers indicating more negative (positive) skewness. The color of the marker
1009 corresponds to the dataset colors in the legend on the right.

1010

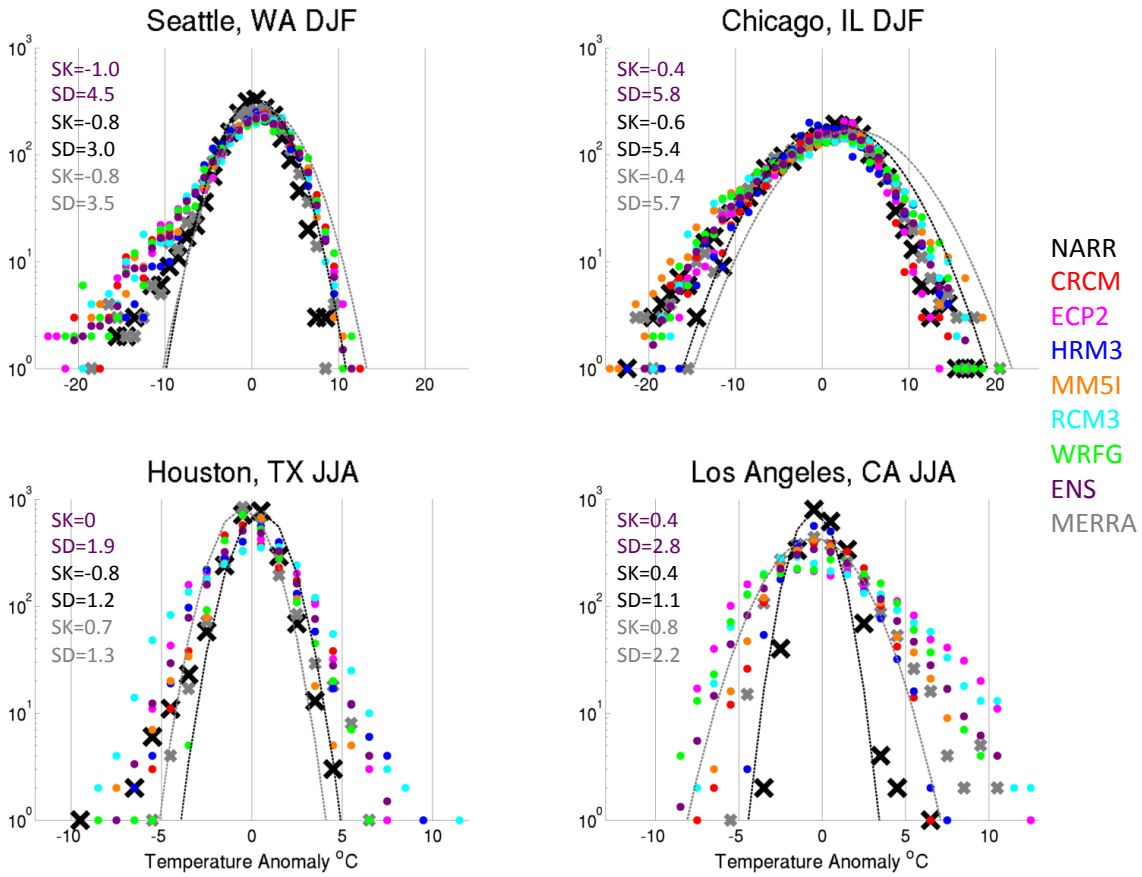
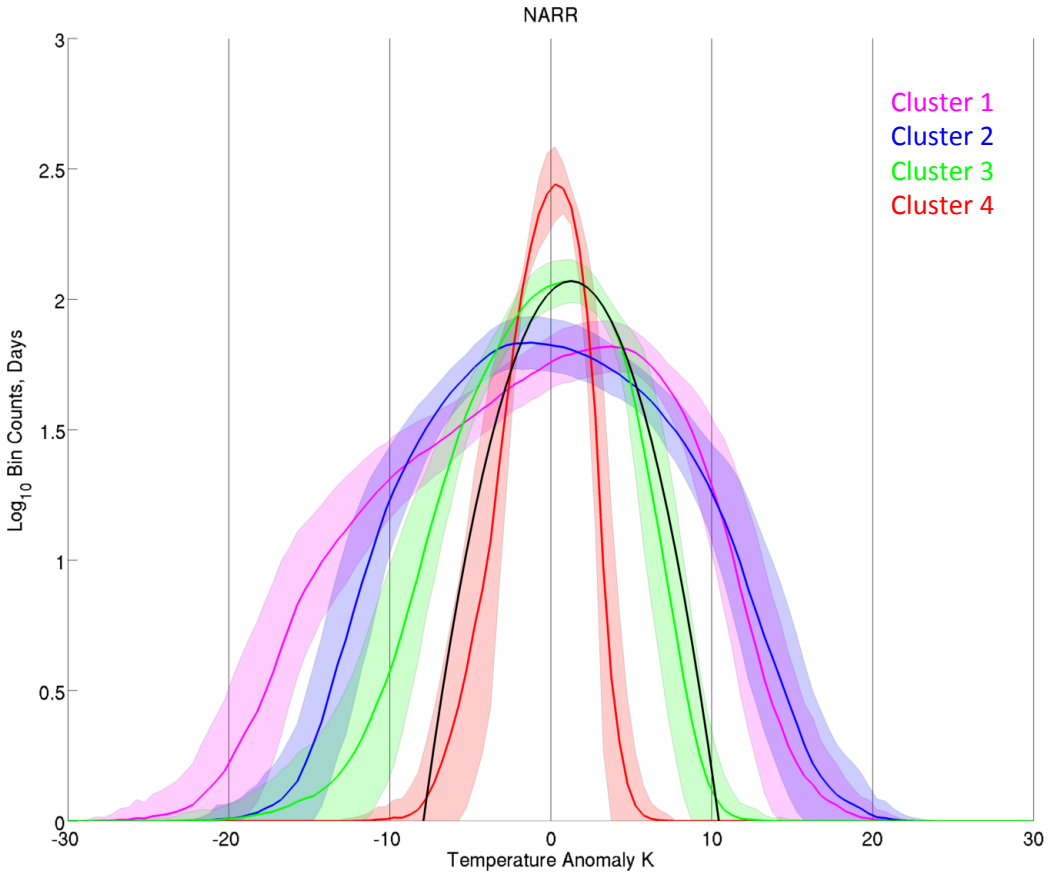


Figure 11. Actual PDFs, plotted on a log scale, of temperature anomalies for four locations corresponding to station data examples used by Ruff and Neelin (2012). Temperature anomalies are binned every one half degree. The black Xs are for NARR and the gray Xs are for MERRA. The dotted black curve is a Guassian fit with the same standard deviation as NARR and the dotted gray curve is the same for MERRA. The skewness and standard deviation for NARR is printed in back, MERRA is gray, and the skewness and standard deviation of the multi-model mean PDF is in black. Note the different x-axis scales between DJF and JJA.



1023

1024 Figure 12. Basis PDFs from NARR computed as the mean PDFs for each cluster.

1025 Each curve is the average of the PDFs from all gridpoints that were assigned to the

1026 indicated cluster for NARR. The shaded region surrounding each curve gives ± 1

1027 standard deviation within each temperature bin computed from the set of PDFs over

1028 all the spatial points in the cluster. The black curve is a Guassian fit to the core of

1029 the mean PDF for cluster 3, for reference. The y-axis is the log of the bin counts

1030 (plotted on a linear scale).

1031

1032

1033

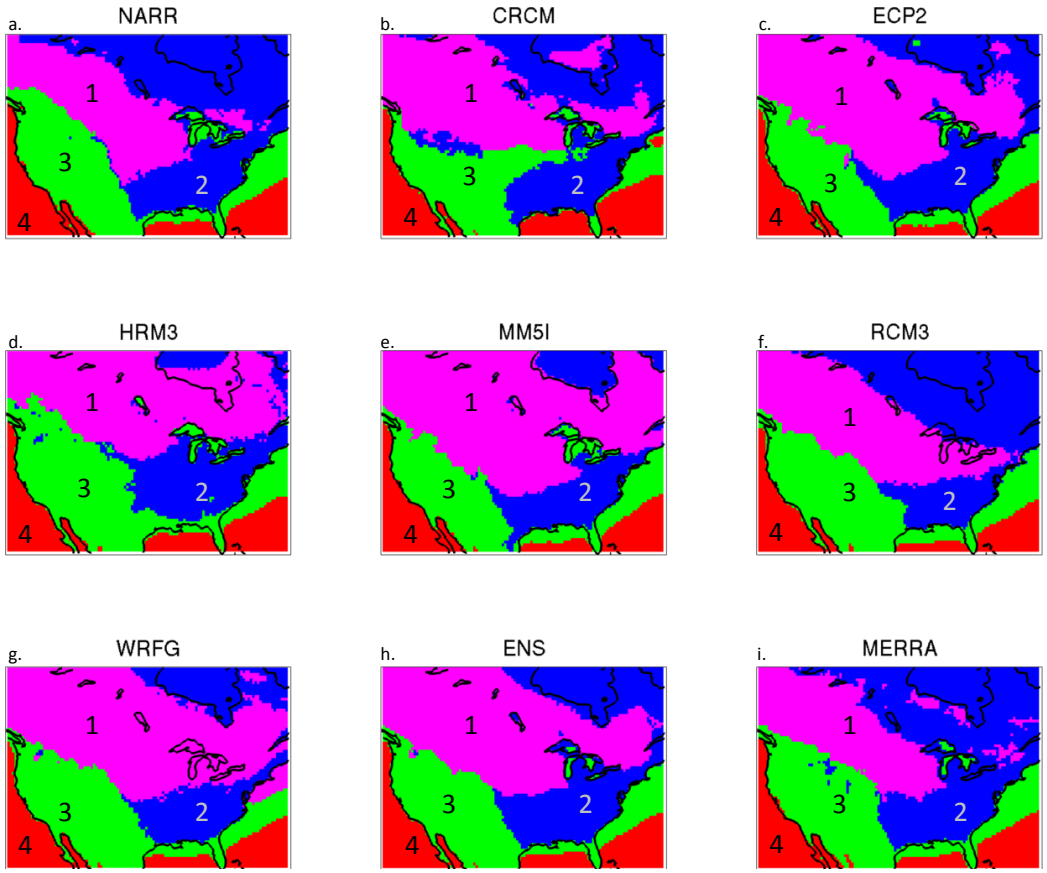
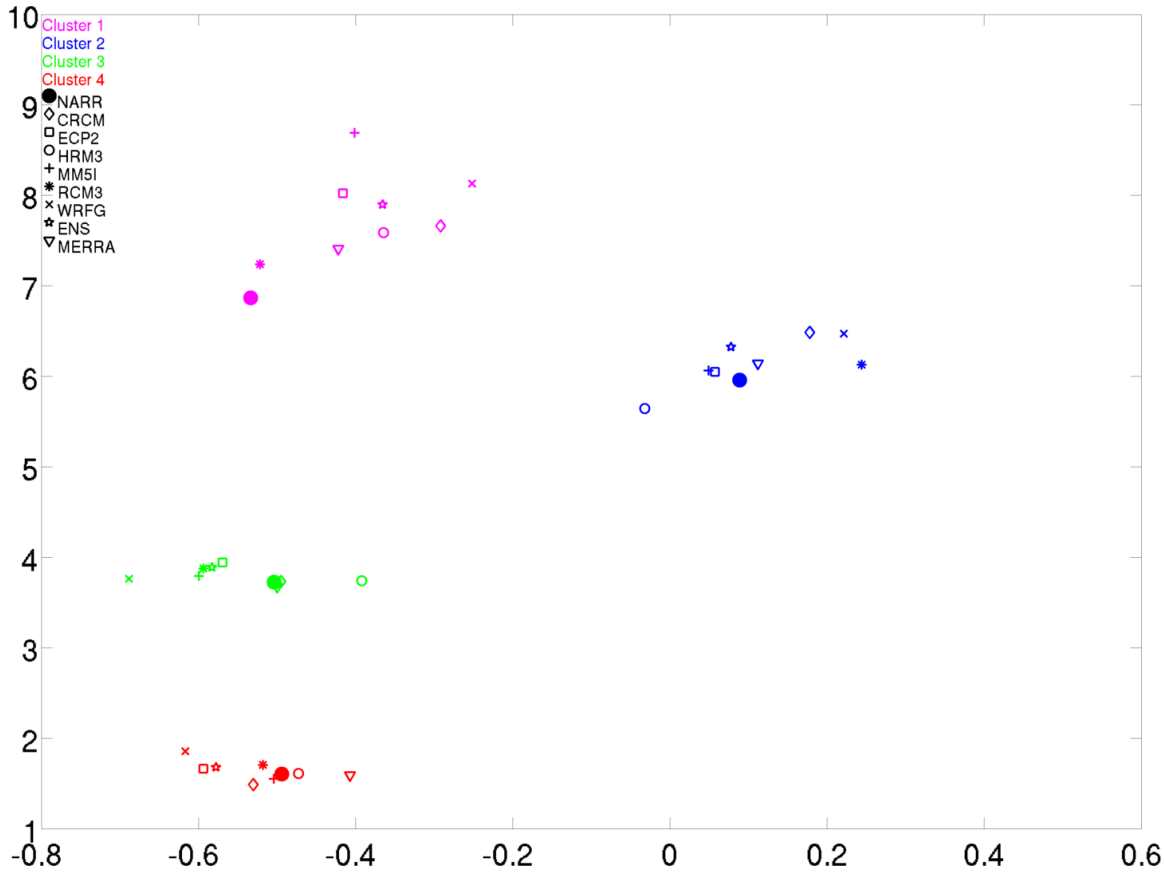


Figure 13. Maps of cluster assignments based on the NARR basis PDFs determined by the cluster assignments for (a) NARR. Cluster assignments are determined at each grid point by finding the minimum RMS difference between the model PDF and each of the four basis PDFs (Figure 12). The cluster associated with the basis PDF with the smallest RMS difference is assigned to the model PDF. The assignment is color coded to match the colors in Figure 12 (i.e. all green areas are assigned to cluster 2) and the associated cluster number is indicated on the map.



1046

1047 Figure 14. Scatter plot of mean standard deviation (y-axis) versus skewness (x-axis)

1048 for each cluster and each model for DJF.

1049
1050

1051