Classifying reanalysis surface temperature probability density functions (PDFs) over North America with cluster analysis

P. C. Loikith,[1] B. R. Lintner,[2] J. Kim,[3,4] H. Lee,[1] J. D. Neelin,[4] and D. E. Waliser[1,3]

Received 7 May 2013; revised 19 June 2013; accepted 20 June 2013; published 26 July 2013.

[1] An important step in projecting future climate change impacts on extremes involves quantifying the underlying probability distribution functions (PDFs) of climate variables. However, doing so can prove challenging when multiple models and large domains are considered. Here an approach to PDF quantification using k-means clustering is considered. A standard clustering algorithm (with k = 5 clusters) is applied to 33 years of daily January surface temperature from two state-of-the-art reanalysis products, the North American Regional Reanalysis and the Modern Era Retrospective Analysis for Research and Applications. The resulting cluster assignments yield spatially coherent patterns that can be broadly related to distinct climate regimes over North America, e.g., low variability over the tropical oceans or temperature advection across stronger or weaker gradients. This technique has the potential to be a useful and intuitive tool for evaluation of model-simulated PDF structure and could provide insight into projections of future changes in temperature. **Citation:** Loikith, P. C., B. R. Lintner, J. Kim, H. Lee, J. D. Neelin, and D. E. Waliser (2013), Classifying reanalysis surface temperature probability density functions (PDFs) over North America with cluster analysis, *Geophys. Res. Lett.*, *40*, 3710–3714, doi:10.1002/grl.50688.

1. Introduction

[2] Although global warming impacts on climate are often framed in terms of mean change, the potential changes in extremes arguably pose a greater concern for societal or ecosystem adaptive capacity [*Kharin et al*., 2007; *Trenberth et al*., 2007; *IPCC*, 2012]. Nonlinear relationships between changing means and extremes suggest that even small changes in the former can be associated with large changes in the latter [*Griffiths et al*., 2005]. Quantifying vulnerability to and risks associated with extreme climatic events—and more critically, projecting how such risks may change in the future—requires detailed knowledge of the underlying probability distribution functions (PDFs) of important climate variables. Furthermore, in order to constrain uncertainties in model simulations of future climate extremes, standardized metrics are required to evaluate model fidelity.

[3] Because climate change may alter multiple moments of the PDF of a climate variable [*Hannachi*, 2006], evaluation metrics should give insight on multiple aspects of PDF structure. For example, *Donat and Alexander* [2012] present evidence of increasing variance in the Tropics since the mid-twentieth century as well as a tendency toward more positive skewness. Additionally, there can be considerable spatial variation or dependence on small-scale processes in these PDFs [*Easterling et al*., 2000; *Diffenbaugh et al*., 2005]. While some theoretical guidance exists for relating the governing dynamics of a system to its PDF characteristics [e.g., *Bourlioux and Majda*, 2002; *Sura and Sardeshmukh*, 2008; *Neelin et al*., 2010; *Stechmann and Neelin*, 2011], further effort is needed to apply theoretical understanding to climate variables in observations and models.

[4] Analyzing surface temperature (T_s) PDFs from the Global Surface Summary of the Day product, *Ruff and Neelin* [2012] documented non-Gaussian, often asymmetric long tails occurring over a wide range of geographic and climatic settings. They further noted how the details of the PDF tails significantly impact the estimation of threshold exceedances. For example, under a warming-induced shift of the distribution, locations with high-side Gaussian tails would experience a greater increase in a given warm threshold exceedance relative to locations with a fat (e.g., exponential) tail.

[5] The size and scope of currently available observational and model data products present practical challenges for generalizing the diagnosis, interpretation, validation, and intercomparison of PDF characteristics. Consequently, evaluation and comparison of large observational and model data sets require the development of flexible yet standardized and readily applicable diagnostics to facilitate model evaluation, interpretation, and development. To that end, we present results of a cluster analysis applied to T_s PDFs obtained from two reanalysis products. Our results demonstrate that a few stable PDF categories can be obtained and related to some readily understood aspects of different climatic regimes.

2. Data Sets and Methodology

[6] The North American Regional Reanalysis (NARR) [*Mesinger et al.*, 2006] is a high-resolution reanalysis product covering North America. This data set is derived from a data assimilation scheme with near-surface observations ingested hourly and atmospheric profiles of temperature, winds, and moisture from rawinsondes and dropsondes ingested every 3 h. The native NARR data are available on a Lambert conformal grid (3-hourly, approximately 32 km).

[1]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA.
[2]Department of Environmental Sciences, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA.
[3]Joint Institute for Regional Earth System Science and Engineering, University of California, Los Angeles, California, USA.
[4]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, California, USA.

Corresponding author: P. C. Loikith, NASA Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA. (paul.c.loikith@jpl.nasa.gov)

0094-8276/13/10.1002/grl.50688

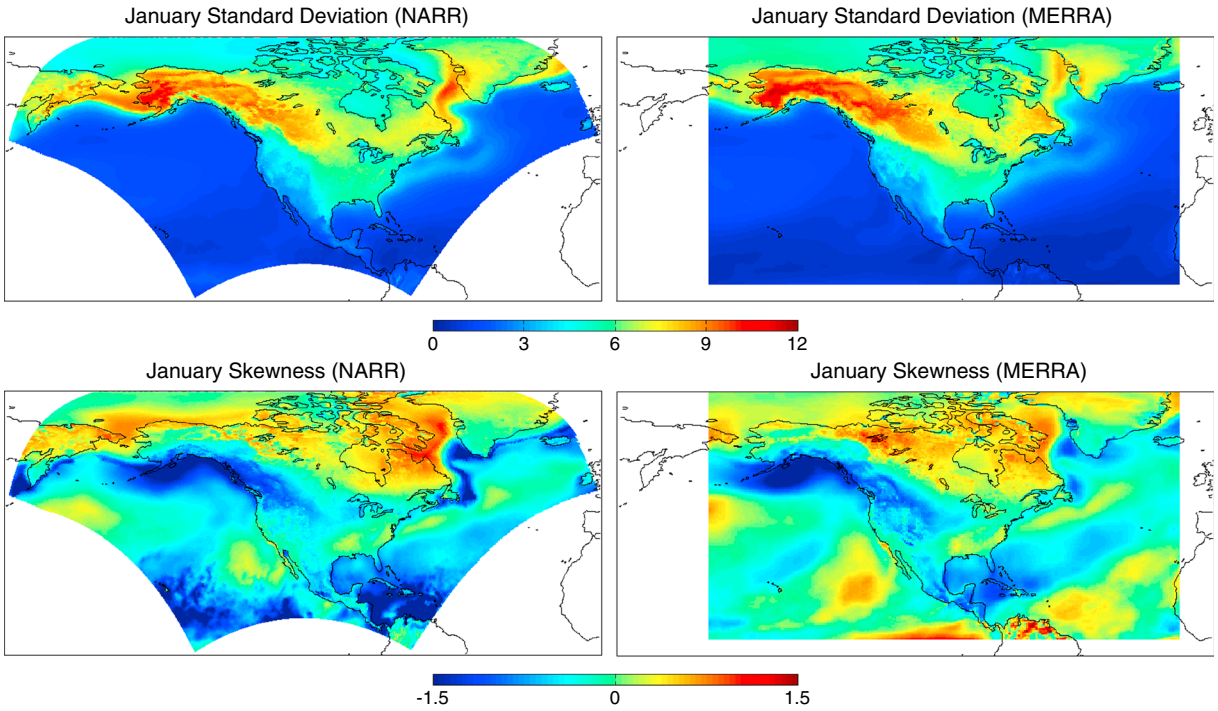

Figure 1. Maps of the (top) standard deviation of January temperature and (bottom) skewness of January temperature for (left) NARR and (right) MERRA.

[7] Additionally, T_s data are analyzed from the Modern Era Retrospective Analysis for Research and Applications (MERRA), developed by NASA's Global Modeling and Assimilation Office and disseminated by the Goddard Earth Sciences Data and Information Services Center [*Rienecker et al*., 2011]. MERRA assimilates observations from multiple sources including weather stations and balloons, satellite data, ships, buoys, and aircraft. This data product is defined on a global, regular uniform grid at a spatial resolution of 0.5° latitude × 0.67° longitude, which is coarser than NARR but finer than other widely used reanalysis products. Because of the different grid nests for NARR and MERRA, the latter covers more of the Earth at lower latitudes than the NARR domain. While other reanalysis products cover this domain, these two were chosen because the relatively high resolution allows for analysis of regional scale phenomena.

[8] To construct PDFs, daily January T_s data are first deseasonalized by removing the daily climatology over the 33 year (1979–2011) period; long-term linear trends are also removed for individual grid points. Only January is considered here for demonstration purposes. Anomalies are computed so that all grid points have a mean of 0, allowing for systematic comparison of PDFs across the domain. Anomalies for all 1023 days are sorted into bins of 0.5 K width using d = 152 bins at all grid points and normalized by the total number of days. While this results in many grid points having multiple bins with zero counts, d = 152 was necessary to span the range of temperature anomalies at all grid points. Next, k-means cluster analysis is applied to group the PDFs. Clustering is performed on the log of probability (log10[bin count(i)/1023] where i = 1, 2,…, 152) to increase the weight of distribution tails in the clustering. In other words, the clustering algorithm seeks k sets among these vectors of length d of the log of probabilities, over the data set of n spatial points, that minimizes the within-cluster sum of squares of the distance in the d-dimensional space.

[9] Here, k = 5 clusters is used for demonstration purposes. While the choice of k = 5 clusters is arbitrary, in a simple sensitivity analysis in which the number of clusters was varied from three to eight, we found the results for k = 5 to be straightforward to interpret physically. The optimal number of clusters and associated sensitivities will be explored in more detail in ongoing work that applies this methodology to evaluation of climate models.

3. Results and Discussion

[10] Figure 1 depicts the standard deviation (SD) and skewness of the daily T_s variability. In general, MERRA has smaller SD along the margins of sea ice (Labrador Sea, Bering Sea), while NARR has smaller values over western Canada and Alaska. The overall geographic pattern of skewness is very similar in both products. Moreover, both products are able to capture local features such as the band of positive skewness over the coastal waters adjacent to California and much of Baja California caused by offshore winds that induce strong positive temperature excursions, e.g., the Santa Ana winds in Southern California [*Hughes and Hall*, 2010]. *Loikith and Broccoli* [2012] document a similar skewness structure using coarser-resolution gridded daily T_s observations. In a simple sensitivity analysis where the data were divided into two equal temporal intervals, the SD and skewness values did not change appreciably in most of the domain, suggesting that these patterns and values are stable with respect to time period at least over the late 20th and early 21st centuries.

[11] Although the spatial patterns of the second and third moment statistics in Figure 1 capture important features of daily T_s variability, PDF modality is not readily discernable

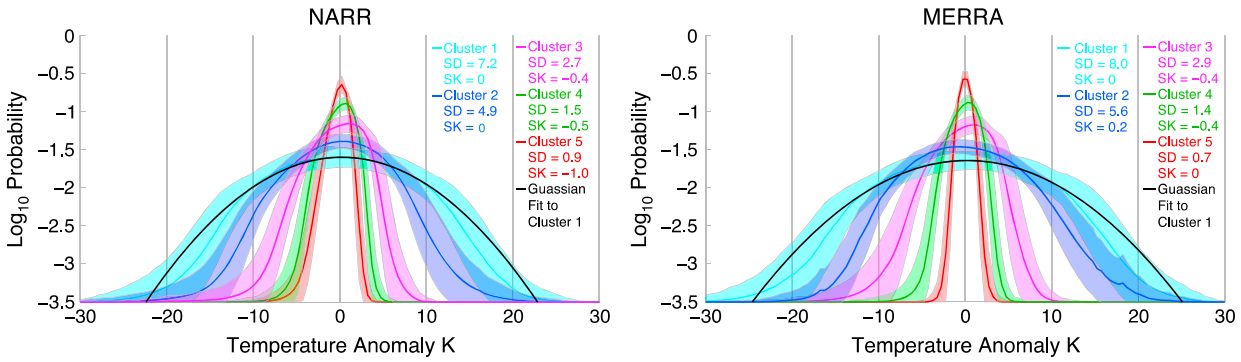

Figure 2. The mean PDF of each cluster for (left) NARR and (right) MERRA. Each curve is the average of the PDFs from all grid points that were assigned to the indicated cluster. The shaded region surrounding each curve gives ±1 standard deviation within each temperature bin computed from the set of PDFs over all the spatial points in the cluster. The black curve is a Guassian fit to the core of the mean PDF for cluster 1, for reference. The *y* axis is the log of the probability (plotted on a linear scale). The average standard deviation (SD) and skewness (SK) values for all the grid points assigned to each cluster are indicated in the legend.

in terms of a single moment. In this sense, it may be instructive to consider diagnostics of the overall shape of the PDFs, especially if such diagnostics are sufficiently limited, i.e., the number of shape categories is small. To this end, we apply the k-means cluster method. Figure 2 depicts the PDFs associated with each of the clusters, showing cluster-mean PDFs (thick lines) and ±1 SD (shading; calculated as the SD of all the points of the cluster within each temperature bin). Maps of the pointwise cluster assignments are in Figure 3. Here the colors plotted on the map correspond to the individual PDFs that comprise the mean PDFs in Figure 2, e.g., the red curve in Figure 2 is the mean of the PDFs for each red-shaded grid point in Figure 3. The mean SD and skewness values, computed as the average of all grid points within the cluster, are indicated in Figure 2. The clusters are numbered based on the mean SD of all grid points within the cluster from high (C1) to low (C5) SD values.

[12] The grid points falling into C1 consist largely of sub-Arctic regions and are characterized by high temperature variance, as evident by the wide PDF, and a relatively large spread within the cluster, reflecting significant local variations. This region matches the band of high SD in T_s

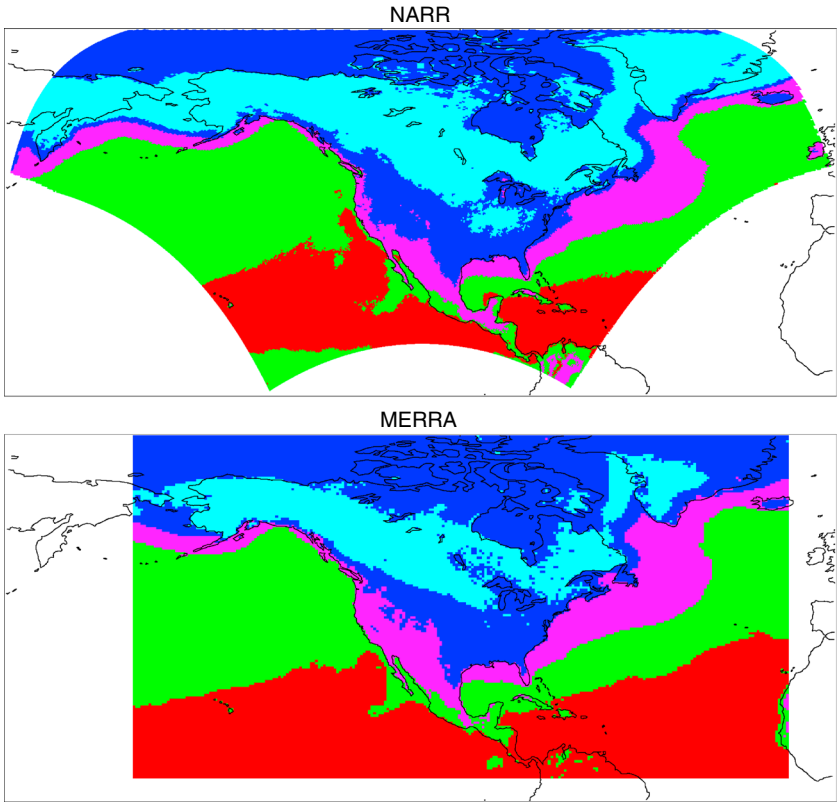

Figure 3. Maps of cluster assignments for (top) NARR and (bottom) MERRA. The assignment is color coded to match the colors in Figure 2, and the associated cluster number is indicated on the map.

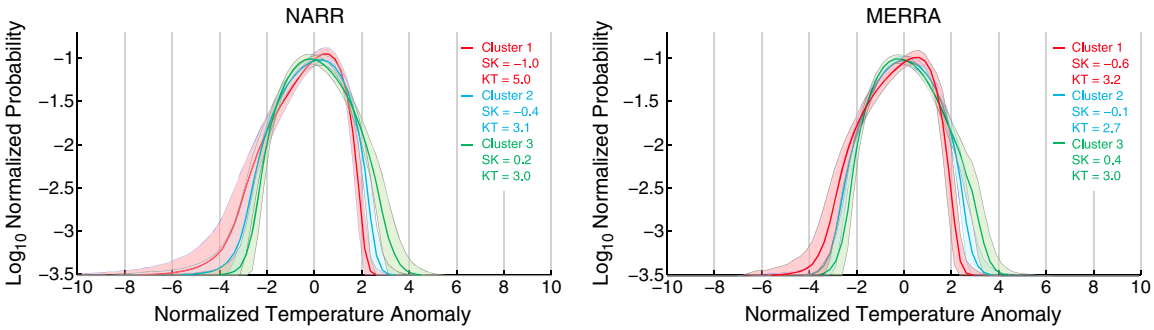

Figure 4. Same as Figure 2, except the cluster analysis is applied to PDFs of normalized temperature anomalies and only k = 3 clusters was used. The average skewness (SK) and kurtosis (KT) of all points in each cluster is indicated in the legend.

(Figure 1) and includes the transition zone from predominantly negative skewness to the south and positive skewness to the north reflected in the symmetrical PDF. This region is subject to strong anomalous T_s advection associated with synoptic-scale weather events [*Loikith and Broccoli*, 2012] across a gradient such that either colder or warmer air masses can be advected into the region.

[13] C2 exhibits relatively high variance and encompasses the Arctic as well as the continental midlatitudes. The Arctic is an area of predominantly positive skewness, while a mixture of negative and positive skewness occurs over the continental midlatitudes (Figure 1). This combination is reflected in the symmetrical mean PDF. While the two regions described by this cluster have little in common climatologically, different mechanisms may allow for similar PDF characteristics, especially variance. The Arctic (C2) has lower variance than areas to the south (C1) since the region is among the coldest in the hemisphere, thus precluding outbreaks of extreme cold air (in an anomalous sense) that can occur at lower latitudes. The midlatitude region is within the main storm track but has lower variance compared with areas immediately to the north (C1) due in part to the modification of extreme cold air masses as they move equatorward.

[14] C3 encompasses the southwestern United States, northern Mexico, and the coastal waters of southern Alaska, the northern Gulf of Mexico, and the western Atlantic Ocean. Included in C3 are coastal regions of high temperature gradient on the West Coast and regions of high oceanic temperature gradient off the East Coast of the U.S. Comparing to Figure 1, C3 includes some ocean regions with relatively high variance as well as the southwestern portion of the continent which has relatively low variance for a continental region. The mean PDF also exhibits negative skewness, especially evident over coastal Alaska. Here, the negative skewness is likely caused by extreme cold outbreaks associated with advection from the continental interior combined with a limited warm tail associated with the moderating effect of the ocean. The region over the Atlantic has high storm frequency in the winter, which elevates the temperature SD relative to other marine regions.

[15] C4 and C5, describing the midlatitude oceans and Tropics, respectively, have the smallest variance of the five clusters, associated with a smaller temperature gradient in the Tropics and with the moderation of advective effects by ocean heat capacity over C4. A substantial part of C4 is also to the south of the main storm track. C5 is south of the storm track and experiences smaller effects by midlatitude synoptic-scale weather variability. The mean PDF of C4 (and C5 for NARR) is characterized by a long cold tail, likely reflecting the occasional incursion of cold air masses from across the temperature gradient on the midlatitude side. The relatively short warm tail likely reflects the small gradient toward warmer tropical temperatures; as such, it is not possible to strongly increase temperatures by warm advection.

[16] The approach described here yields a first view of regional distributions of PDFs; however, to emphasize differences in PDF shape, cluster analysis is applied to PDFs computed from anomalies normalized by their SD. If all the distributions were Gaussian, the normalization would tend to collapse them into a single cluster, so this approach can be anticipated to give a view of the prevalence of non-normality. Figure 4 shows an example in which three clusters are used to group PDFs of normalized temperature anomalies. The PDF cluster assignments reflect the higher-order moments of skewness and kurtosis. While skewness appears to be the most apparent characteristic for clustering, kurtosis is also influential with C1 having the highest kurtosis.

4. Summary and Conclusions

[17] Variations in T_s PDFs over a large geographic area encompassing North America and surrounding oceans are examined using simple k-means clustering. In both data sets, the cluster analysis yields stable, spatially coherent patterns that can be understood in terms of distinct T_s regimes, such as smaller variability over tropical oceans and larger variability over the high-latitude continental interior. The shape of the reconstructed PDF for each cluster, along with the geographical distribution of the clusters, fits well with physical interpretations in terms of temperature advection in the presence of a maintained background temperature gradient and advection by synoptic-scale events. In general, temperature variances appear to be the leading determinant in defining clusters. Skewness also affects some cluster assignments, suggesting that cluster-based approaches are useful for identifying regions with common PDF shape. By normalizing the temperature anomalies by their SD, it is possible to use cluster analysis to group PDFs based on higher moment statistics, providing important information for characterizing regional sensitivity of temperature extremes to future warming.

[18] Future work will focus on developing the cluster analysis approach outlined in this paper for categorizing PDF characteristics in regional climate model simulations for the purpose of evaluating model data against observations/reanalysis. While other methodologies exist for systematic PDF evaluation, the ability of this tool to be used over

large domains or numbers of grid points makes it particularly versatile. For example, *Perkins et al*. [2007] developed and applied a PDF skill score for model evaluation over Australia using relatively homogenous subregions. Their technique provides a concise and standardized way to evaluate models; however, the clustering method has the advantage that it works over large inhomogeneous domains. Furthermore, this approach may serve to identify regions where future changes in T_s or other climate variables are likely to be relatively homogeneous. As such, this method may provide a foundation for elucidating changes in future climate extremes.

[19] **Acknowledgments.** 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., and D.E.W). Part of this research was funded by NOAA NA11OAR4310099 (J.D.N.) and New Jersey Agricultural Experiment Station Hatch grant NJ07102 (B.R.L.). We thank Joyce Meyerson for her assistance with figure visualization.

[20] The Editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References

Bourlioux, A., and A. J. Majda (2002), Elementary models with probability distribution function intermittency for passive scalars with a mean gradient, *Phys. Fluids*, *14*, 881–897, doi:10.1063/1.1430736.

Diffenbaugh, N. S., J. S. Pal, R. J. Trapp, and F. Giorgi (2005), Fine-scale processes regulate the response of extreme events to global climate change, *Proc. Nat. Acad. Sci. U.S.A.*, *102*, 15,774–15,778.

Donat, M. G., and L. V. Alexander (2012), The shifting probability distribution of global daytime and night-time temperatures, *Geophys. Res. Lett.*, *39*, L14707, doi:10.1029/2012GL052459.

Easterling, D. R., J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje (2000), Observed variability and trends in extreme climate events: A brief review, *Bull. Am. Meteorol. Soc.*, *81*, 417–425.

Griffiths, G. M., et al. (2005), Change in mean temperature as a predictor of extreme temperature change in the Asia–Pacific region, *Int. J. Climatol.*, *25*, 1301–1330.

Hannachi, A. (2006), Quantifying changes and their uncertainties in probability distribution of climate variables using robust statistics, *Clim. Dyn.*, *27*, 301–317.

Hughes, M., and A. Hall (2010), Local and synoptic mechanisms causing Southern California's Santa Ana winds, *Clim. Dyn.*, *34*, 847–857.

IPCC (2012), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, edited by C. B. Field et al., 582 pp., Cambridge Univ. Press, Cambridge, UK, and New York, NY, USA.

Kharin, V. V., F. W. Zwier, and G. C. H. X. Zhang (2007), Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations, *J. Clim.*, *20*, 1419–1444.

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, doi:10.1175/JCLI-D-11-00709.1.

Mesinger, F., et al. (2006), North American Regional Reanalysis, *Bull. Amer. Meteor. Soc.*, *87*, 343–360.

Neelin, J. D., B. R. Lintner, B. Tian, Q. Li, L. Zhang, P. K. Patra, M. T. Chahine, and S. N. Stechmann (2010), Long tails in deep columns of natural and anthropogenic tropospheric tracers, *Geophys. Res. Lett.*, *37*, L05804, doi:10.1029/2009GL041726.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney (2007), Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *J. Clim.*, *20*, 4356–4376.

Rienecker, M. M., et al. (2011), MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, *J. Clim.*, *24*, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Ruff, T. W., and J. D. Neelin (2012), Long tails in regional surface temperature probability distributions with implications for extremes under global warming, *Geophys. Res. Lett.*, *39*, L04704, doi:10.1029/2011GL050610.

Stechmann, S., and J. D. Neelin (2011), A stochastic model for the transition to strong convection, *J. Clim.*, *68*, 2955–2970, doi:10.1175/JAS-D-11-028.1.

Sura, P., and P. D. Sardeshmukh (2008), A global view of non-Gaussian SST variability, *J. Phys. Oceanogr.*, *38*, 639–647.

Trenberth, K. E., et al. (2007), Observations: Surface and atmospheric climate change, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 235–336, Cambridge Univ. Press, Cambridge, U. K.