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], H. Lee[1], J. D. Neelin[4], and D. E. Waliser[1]

[1]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

[2]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

[3]Joint Institute for Regional Earth System Science and Engineering, University of California Los Angeles, Los Angeles, CA, USA

[4]University of California Los Angeles, Los Angeles, CA, USA

*Corresponding author:
P.C. Loikith

Abstract

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.

1. Introduction

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.

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-20th century as well as a tendency towards 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.

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

69 occurring over a wide range of geographic and climatic settings. They further noted how the
70 details of the PDF tails significantly impact the estimation of threshold exceedances. For
71 example, under a warming-induced shift of the distribution, locations with high-side Gaussian
72 tails would experience a greater increase in a given warm threshold exceedance relative to
73 locations with a fat (e.g., exponential) tail.

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

82

83 **2. Data sets and methodology**

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

89 Additionally, T_s data is analyzed from the Modern Era-Retrospective Analysis for Research
90 and Applications [MERRA], developed by NASA's Global Modeling and Assimilation Office
91 and disseminated by the Goddard Earth Sciences Data and Information Services Center

92 [*Rienecker et al.,* 2011]. MERRA assimilates observations from multiple sources including
93 weather stations and balloons, satellite data, ships, buoys, and aircraft. This data product is
94 defined on a global, regular uniform grid at a spatial resolution of 0.5° latitude x 0.67° longitude,
95 coarser than NARR, but finer than other widely used reanalysis products. Because of the
96 different grid nests for NARR and MERRA, the latter covers more of the Earth at lower latitudes
97 than the NARR domain. While other reanalysis products cover this domain, these two were
98 chosen because the relatively high resolution allows for analysis of regional scale phenomena.
99 To construct PDFs, daily January T_s data are first de-seasosonalized by removing the daily
100 climatology over the 33-year (1979-2011) period; long-term linear trends are also removed for
101 individual grid points. Only January is considered here for demonstration purposes. Anomalies
102 are computed so that all grid points have a mean of 0, allowing for systematic comparison of
103 PDFs across the domain. Anomalies for all 1023 days are sorted into bins of 0.5 K width using
104 d=152 bins at all gridpoints and normalized by the total number of days. While this results in
105 many grid points having multiple bins with zero counts, d=152 was necessary to span the range
106 of temperature anomalies at all grid points. Next, k-means cluster analysis is applied to group the
107 PDFs. Clustering is performed on the log of probability (log10[bin count(i)/1023 where
108 i=1,2,…152) to increase the weight of distribution tails in the clustering. In other words, the
109 clustering algorithm seeks k sets among these vectors of length d of the log of probabilities, over
110 the data set of n spatial points, that minimizes the within-cluster sum of squares of the distance in
111 the d-dimensional space.
112 Here, k=5 clusters is used for demonstration purposes. While the choice of k = 5 clusters is
113 arbitrary, in a simple sensitivity analysis in which the number of clusters was varied from three
114 to eight, we found the results for k = 5 to be straightforward to interpret physically. The optimal

115 number of clusters and associated sensitivities will be explored in more detail in ongoing work
116 that applies this methodology to evaluation of climate models.
117
118 **3. Results and discussion**
119 Fig. 1 depicts the standard deviation (SD) and skewness of the daily T_s variability. In general,
120 MERRA has smaller SD along the margins of sea ice (Labrador Sea, Bering Sea) while NARR
121 has smaller values over western Canada and Alaska. The overall geographic pattern of skewness
122 is very similar in both products. Moreover, both products are able to capture local features such
123 as the band of positive skewness over the coastal waters adjacent to California and much of Baja
124 California caused by offshore winds that induce strong positive temperature excursions, e.g., the
125 Santa Ana winds in southern California [*Hughes and Hall, 2010*]. *Loikith and Broccoli* [2012]
126 document similar skewness structure using coarser resolution gridded daily T_s observations. In a
127 simple sensitivity analysis where the data were divided into two equal temporal intervals, the SD
128 and skewness values did not change appreciably in most of the domain, suggesting these patterns
129 and values are stable with respect to time period at least over the late 20th and early 21st
130 centuries.
131 Although the spatial patterns of 2nd and 3rd moment statistics in Fig. 1 capture important
132 features of daily T_s variability, PDF modality is not readily discernable in terms of a single
133 moment. In this sense, it may be instructive to consider diagnostics of the overall shape of the
134 PDFs, especially if such diagnostics are sufficiently limited, i.e., the number of shape categories
135 is small. To this end, we apply the k-means cluster method. Fig. 2 depicts the PDFs associated
136 with each of the clusters, showing cluster-mean PDFs (thick lines) and ±1 SD (shading;
137 calculated as the SD of all the points of the cluster within each temperature bin). Maps of the

138 pointwise cluster assignments are in Fig. 3. Here the colors plotted on the map correspond to the
139 individual PDFs that comprise the mean PDFs in Fig. 2, e.g., the red curve in Fig. 2 is the mean
140 of the PDFs for each red-shaded grid point in Fig. 3. The mean SD and skewness values,
141 computed as the average of all gridpionts within the cluster, are indicated in Fig. 2. The clusters
142 are numbered based on the mean SD of all gridpoints within the cluster from high (C1) to low
143 (C5) SD values.

144 The gridpoints falling into C1 consist largely of sub-Arctic regions and are characterized by
145 high temperature variance, as evident by the wide PDF, and relatively large spread within the
146 cluster, reflecting significant local variations. This region matches the band of high SD in T_s
147 (Fig. 1) and includes the transition zone from predominantly negative skewness to the south and
148 positive skewness to the north reflected in the symmetrical PDF. This region is subject to strong
149 anomalous T_s advection associated with synoptic-scale weather events [*Loikith and Broccoli*,
150 2012] across a gradient such that either colder or warmer air masses can be advected into the
151 region.

152 C2 exhibits relatively high variance and encompasses the Arctic as well as the continental
153 mid-latitudes. The Arctic is an area of predominantly positive skewness while a mixture of
154 negative and positive skewness occurs over the continental mid-latitudes (Fig. 1). This
155 combination is reflected in the symmetrical mean PDF. While the two regions described by this
156 cluster have little in common climatologically, different mechanisms may allow for similar PDF
157 characteristics, especially variance. The Arctic (C2) has lower variance than areas to the south
158 (C1) since the region is among the coldest in the hemisphere, thus precluding outbreaks of
159 extreme cold air (in an anomalous sense) that can occur at lower latitudes. The mid-latitude
160 region is within the main storm track, but has lower variance compared with areas immediately

161 to the north (C1) due in part to the modification of extreme cold airmasses as they move
162 equatorward.

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

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

183 The approach described here yields a first view of regional distributions of PDFs; however,
184 to emphasize differences in PDF shape, cluster analysis is applied to PDFs computed from
185 anomalies normalized by their SD. If all the distributions were Gaussian the normalization
186 would tend to collapse them into a single cluster, so this approach can be anticipated to give a
187 view of the prevalence of non-normality. Fig. 4 shows an example in which three clusters are
188 used to group PDFs of normalized temperature anomalies. The PDF cluster assignments reflect
189 the higher-order moments of skewness and kurtosis. While skewness appears to be the most
190 apparent characteristic for clustering, kurtosis is also influential with C1 having the highest
191 kurtosis.
192
193 **4. Summary and conclusions**
194 Variations in T_s PDFs over a large geographic area encompassing North America and
195 surrounding oceans are examined using simple k-means clustering. In both datasets, the cluster
196 analysis yields stable, spatially coherent patterns that can be understood in terms of distinct T_s
197 regimes, such as smaller variability over tropical oceans and larger variability over the high
198 latitude continental interior. The shape of the reconstructed PDF for each cluster, along with the
199 geographical distribution of the clusters, fit well with physical interpretations in terms of
200 temperature advection in the presence of a maintained background temperature gradient and
201 advection by synoptic-scale events. In general, temperature variances appear to be the leading
202 determinant in defining clusters. Skewness, also affects some cluster assignments, suggesting
203 cluster-based approaches are useful for identifying regions with common PDF shape. By
204 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.

Future work will focus on developing the cluster analysis approach outlined in this paper for categorizing PDF characteristics in regional climate model (RCM) 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 gridpoints 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 sub-regions. 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.

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: NOAA NA11OAR4310099 (JDN) and New Jersey Agricultural Experiment Station Hatch grant NJ07102 (BRL)

We thank Joyce Meyerson for her assistance with figure visualization.

230 **References Cited**

231

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

235

236 Diffenbaugh, N. S., J. S. Pal, R. J. Trapp, and F. Giorgi (2005), Fine-scale processes regulate the
237 response of extreme events to global climate change, *Proc. Natl. Acad. Sci. U. S. A.*, *102*,
238 15,774–15,778.

239

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

243

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

247

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

250

251 Hannachi A. (2006), Quantifying changes and their uncertainties in probability distribution of
252 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, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.

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., and co-authors (2006), North American Regional Reanalysis, *Bull. Amer. Meteor. Soc., 87*, 343–360.

274 Neelin, J. D., B. R. Lintner, B. Tian, Q. Li, L. Zhang, P. K. Patra, M. T. Chahine, and S. N.
275 Stechmann (2010), Long tails in deep columns of natural and anthropogenic tropospheric tracers,
276 *Geophys. Res. Lett., 37*, L05804, doi:10.1029/2009GL041726.
277
278 Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney (2007), Evaluation of the AR4
279 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and
280 Precipitation over Australia Using Probability Density Functions. *J. Clim.*, *20*, 4356-4376.
281
282 Rienecker, M.M., M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich,
283 S.D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, et
284 al., 2011. MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications.
285 *J. Clim., 24*, 3624—3648, doi:10.1175/JCLI-D-11-00015.1.
286
287 Ruff, T. W., and J. D. Neelin (2012), Long tails in regional surface temperature probability
288 distributions with implications for extremes under global warming, *Geophys. Res. Lett., 39*,
289 L04704, doi:10.1029/2011GL050610.
290
291 Sura, P., and P. D. Sardeshmukh (2008), A global view of non-Gaussian SST variability, J. Phys.
292 Oceanogr., *38*, 639–647.
293
294 Stechmann, S. and J. D. Neelin, (2011), A stochastic model for the transition to strong
295 convection, J. Climate, *68*, 2955–2970, doi:10.1175/JAS-D-11-028.1.
296

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.

Figure Captions

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

Figure 2. The mean PDF of each cluster for NARR (left) and MERRA (right). Each curve is the average of the PDFs from all gridpoints 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.

Figure 3. Maps of cluster assignments for NARR (top) and MERRA (bottom). 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 anomaly 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.

Figures

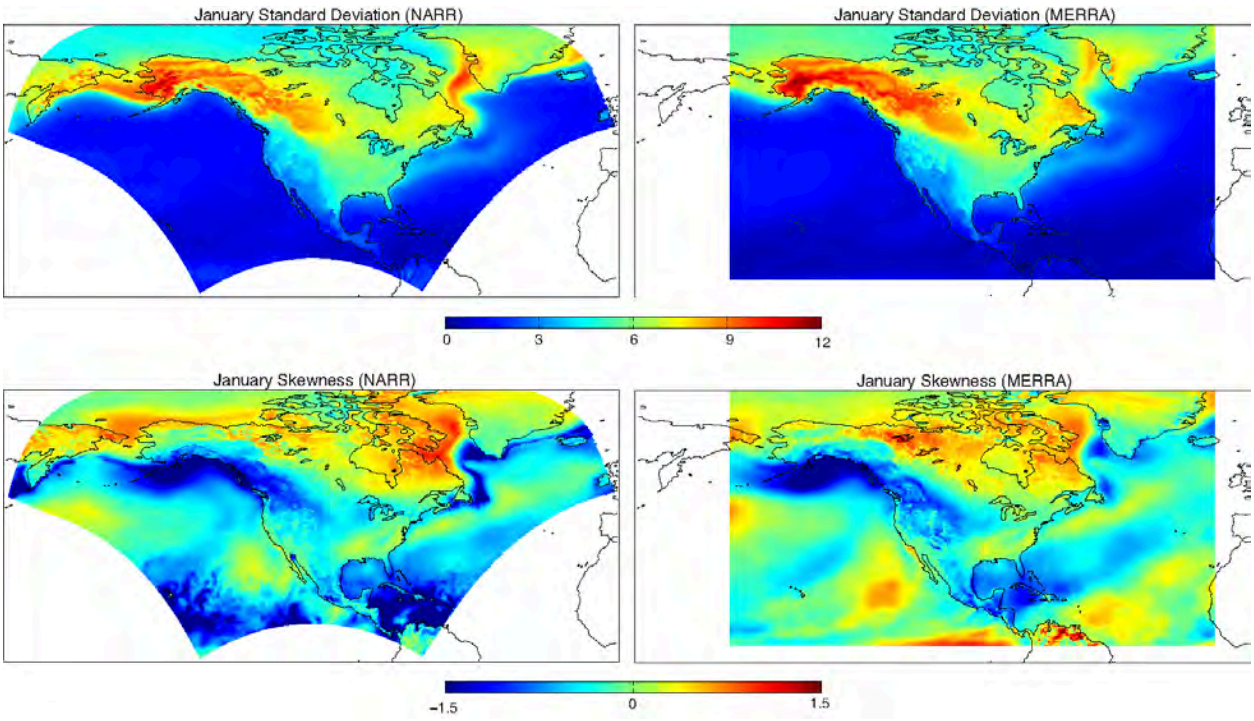

Figure 1.

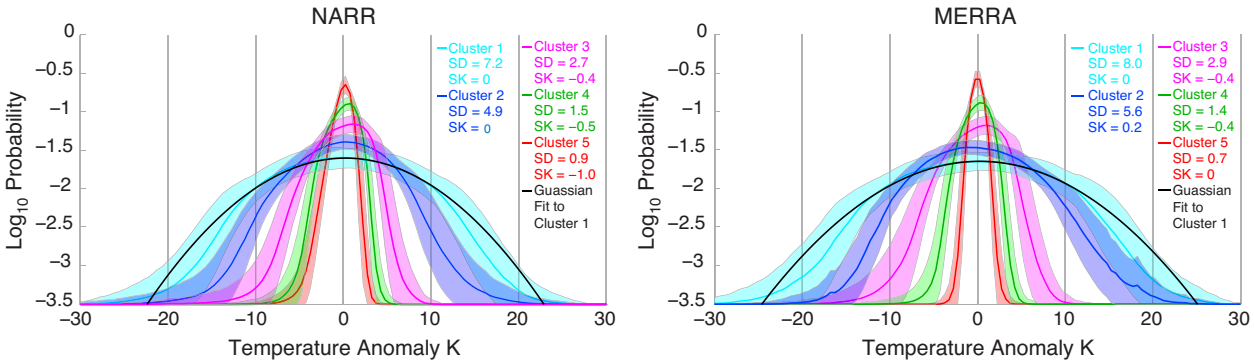
NARR
MERRA
Log10 Probability
Temperature Anomaly K
Cluster 1
SD = 7.2
SK = 0
Cluster 2
SD = 4.9
SK = 0
Cluster 3
SD = 2.7
SK = −0.4
Cluster 4
SD = 1.5
SK = −0.5
Cluster 5
SD = 0.9
SK = −1.0
Guassian Fit to Cluster 1
Cluster 1
SD = 8.0
SK = 0
Cluster 2
SD = 5.6
SK = 0.2
Cluster 3
SD = 2.9
SK = −0.4
Cluster 4
SD = 1.4
SK = −0.4
Cluster 5
SD = 0.7
SK = 0
Guassian Fit to Cluster 1

Figure 2.

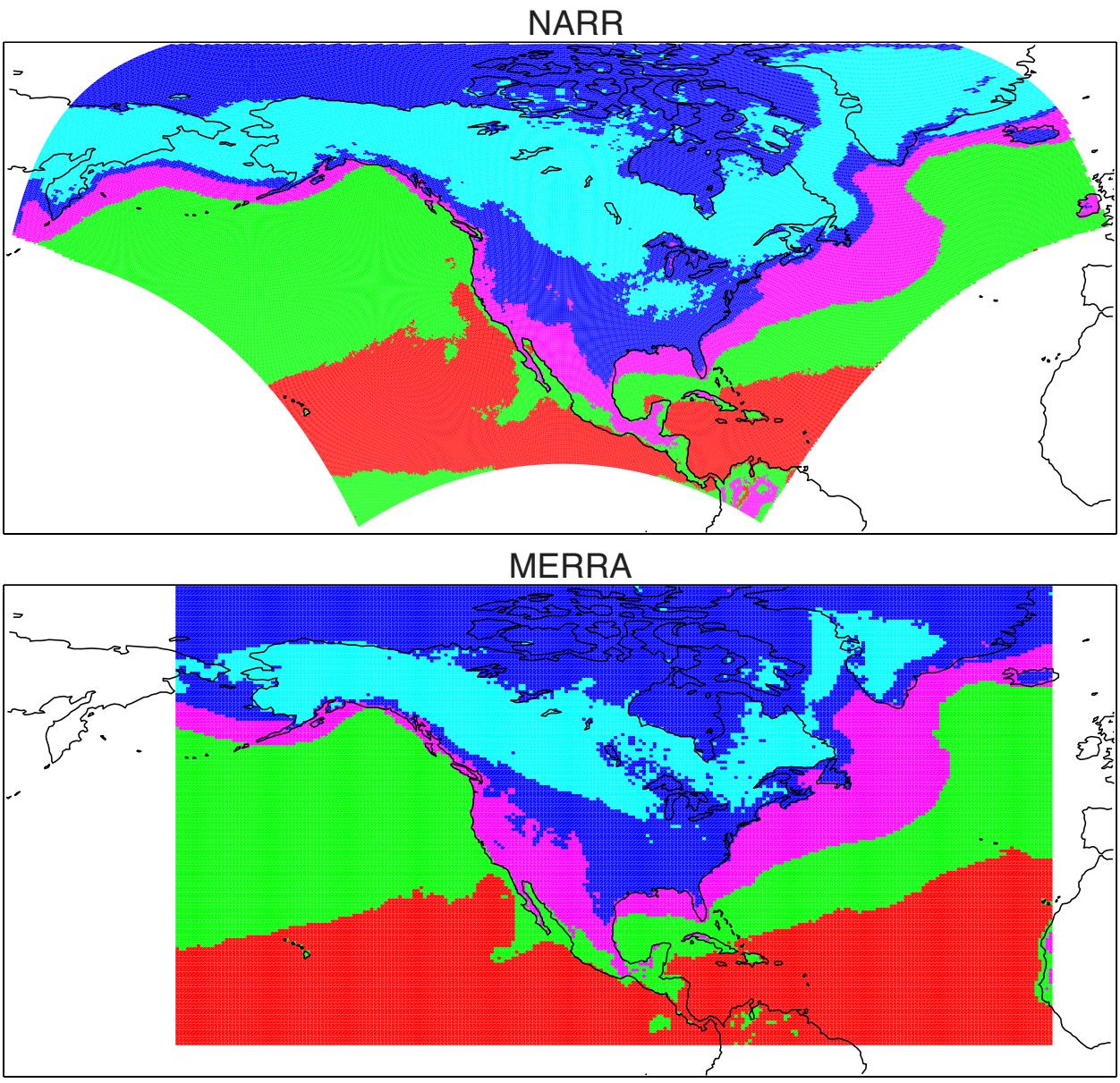

Figure 3.

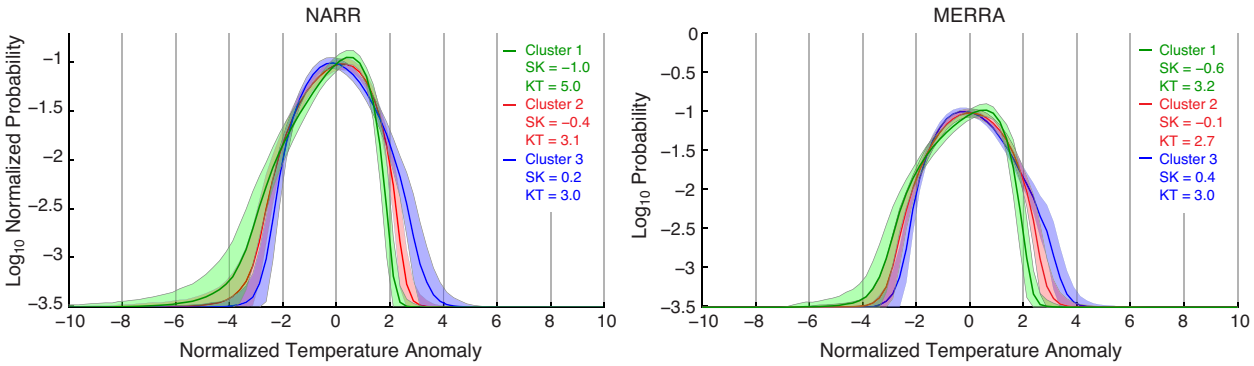
NARR
Log10 Normalized Probability
Normalized Temperature Anomaly
Cluster 1
SK = −1.0
KT = 5.0
Cluster 2
SK = −0.4
KT = 3.1
Cluster 3
SK = 0.2
KT = 3.0
MERRA
Log10 Probability
Normalized Temperature Anomaly
Cluster 1
SK = −0.6
KT = 3.2
Cluster 2
SK = −0.1
KT = 2.7
Cluster 3
SK = 0.4
KT = 3.0

Figure 4.