K-MEANS CLUSTER ANALYSIS IDENTIFICATION OF IDEALIZED RELATIVE ANOMALY PATTERNS IN ANNUAL TOTAL PRECIPITATION ACROSS NEW ENGLAND STATES' NCDC CLIMATE DIVISIONS ENCOMPASSING THE 1895-2018 PERIOD OF RECORD

By Charles J. Fisk
Naval Base Ventura County, Point Mugu, CA.

1. INTRODUCTION

The New England states, defined as Connecticut, Rhode Island, Massachusetts, New Hampshire, Vermont, and Maine, are composed of 15 climatic sub-regions, part of a nationwide arrangement of 344 Climate Divisions organized state-by-state by the NOAA National Centers for Environmental Information (NCEI), formerly the National Climatic Data Center (NCDC). Based on area-averaging techniques, single-valued month-to-month precipitation statistics have been compiled, division-by-division, since 1895, and several years ago, using new, improved areal averaging techniques, the entire division-by-division statistics were redone, by year.

While the geographical extent of the New England states is not great, there are contrasts in topography and proximity to the Atlantic Ocean, which could conceivably induce contrasts in annual precipitation relative anomaly character across the divisions. To explore this possibility, the nature of idealized annual precipitation anomaly modes are resolved using K-Means Clustering Analysis, the K-Means treatment further enhanced by an optimizing data mining training/testing capability - the V-Fold Cross Validation Algorithm. Period of record is 1895 thru 2018, some 124 years.

Traditional K-Means is a trial-and-error procedure, and the V-Fold Algorithm is an automated, iterative, and enhancing training sample/testing sample procedure that rapidly produces in ascending order, 2 to K cluster sets. The iterations cease at some "optimal" selection number K, depending on a choice of statistical distance metric (Euclidean, Squared Euclidean, etc.), number of folds (or training sample subsets), and percent improvements between successive v-fold iterations of individual observations vs. centroid mean statistical distance magnitudes. The software that offers the K-Means/V-Fold capability, however, also gives the option of fixing the number of clusters, still utilizing the V-fold algorithms' processing speed and efficiency , and superior means of cluster membership selection.

Previous studies along these same methodological lines investigated variations in seasonal (July-June) total precipitation modes across the seven California Climate Divisions (Fisk, 2015), and also for twelve Climate Divisions for California, Oregon, and Washington, near to or bordering to the Pacific Ocean (Fisk, 2016).

---

* *Corresponding author address:* Charles J. Fisk, e-mail: cjfisk@att.net

The New England area analysis produced six clusters of annual precipitation variability across the 15 climate divisions, to be depicted below in the coming sections along with accompanying explanations. Also, as a special supplementary analysis, individual years with the most extreme individual statistical distances from cluster centroids (irrespective of cluster) are identified, presented, and described. The most extreme four years of the 124 are highlighted.

2. THE DATA

The raw data were downloaded and processed via an NCEI online link which has complete month-to-month precipitation histories for each of the fifteen divisions of interest, by year, suitable for summations into annual totals.

Figure 1 is a map of the New England region climate divisions along with their names. With the exception of Rhode Island, which is a division onto itself, the other five states are partitioned into two or three divisions.

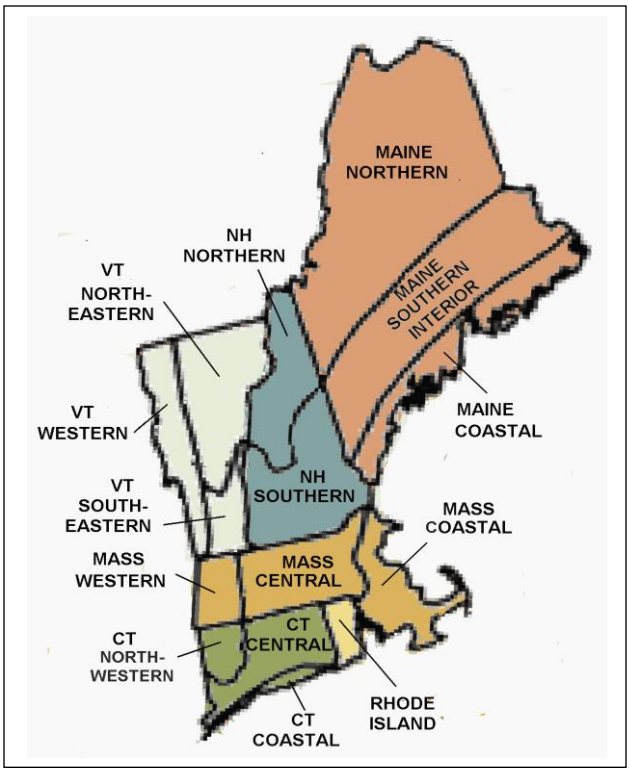

Figure 1 – Map of New England States Climate Divisions (Map Courtesy of NCEI)

Figure 2 is a bar chart of the 124-year annual mean precipitation figures, by division, Figure 3 a similar chart with the standard deviation statistics, by division. From Figure 2, there is a noticeable contrast of mean statistics, from 48.41" for the Connecticut "Northwestern" division to just over 40.95" for the Maine "Northern"" one. The standard deviation statistics in Figure 3 range from 7.26", also for "Northwestern Connecticut", to about 5.27" for "Northeastern Vermont". The divisional data will be reduced to a common scale prior to the clustering through normalization, but the final results will be expressed in standardized anomalies (z-cores), applicable to the respective climatological means and standard deviations for each division.

.

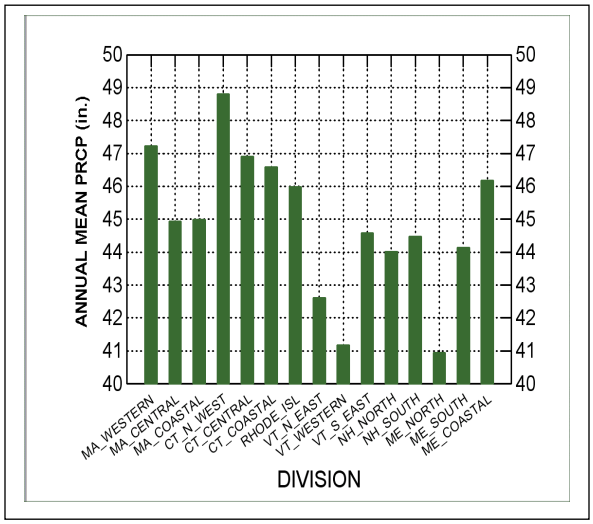

Figure 2 – Mean Annual Precipitation (in.) for Fifteen New England States' Climate Divisions.

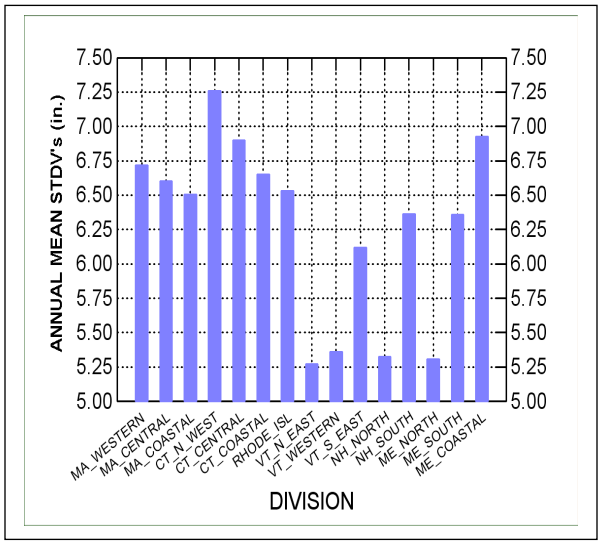

Figure 3 – Standard Deviations (in.) of New England Climate Divisions' Annual Precipitation Totals

3. RESULTS

For the clustering procedure, the default settings of ten folds and a five-percent improvement cutoff threshold were left in place, the distance metric changed to Squared Euclidean. Six clusters or "modes" were realized as the "optimal" K, and Figure 4 is a Scree Plot of the progressive decreases in mean observational distances from cluster centroids ("costs") after each training and testing iteration (i.e., increments in number of clusters, K). The downward trend in cost flattens after K=6, actually rising a bit for K=7, indicative of a trend to overfitting, and the iterations are"cutoff" at K=6, the "best" K.

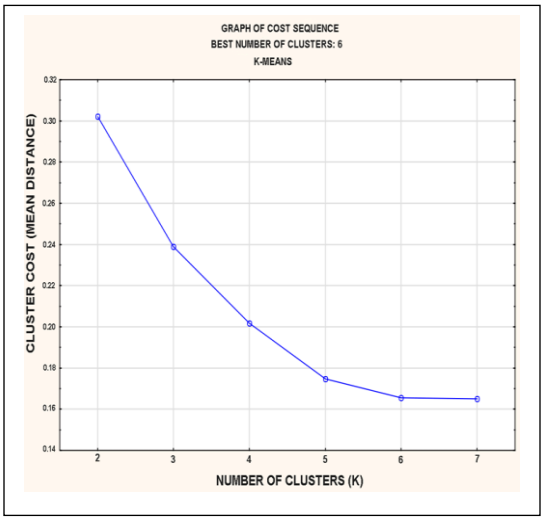

Figure 3 – Scree Plot for New England Climate Divisions' Annual Precipitation K- Means Cluster Analysis

Cluster relative frequencies ranged from 30.6 % to 7.3%. and as the graphs will show, five of the six displayed mean standardized anomaly pattern dichotomies of varying degrees between the "South" (Connecticut, Rhode Island, and Massachusetts) and the "North" (New Hampshire, Vermont, and Maine).

3.1 New England Ranking Mode #1 – 30.6 % Percent Frequency

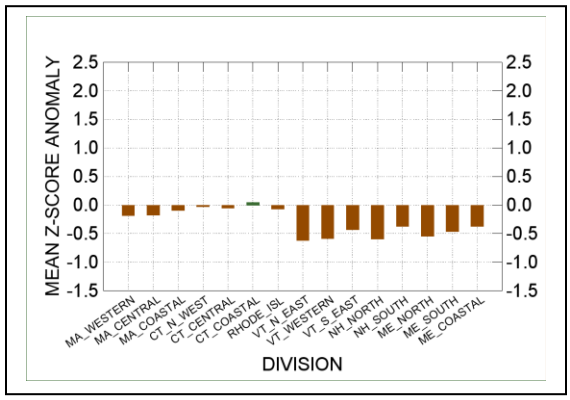

Figure 5. New England Mode #1 – "Near or Slightly Below Normal Means Anomalies "South", Noticeably More Pronounced Below Normals "North""

Figure 5 shows the mean standardized anomaly patterns for ranking mode #1, that which incorporated 30.6%,or 38 of the years. The southern states' divisions of Massachusetts, Connecticut, and Rhode Island all display slightly negative standardized mean departures, save for Coastal Connecticut with a very slightly positive statistic, those for the northern states Vermont, New Hamsphire, and Maine indicating more elevated, negative ones. Northeastern and Western Vermont (-.616z and -.583z, respectively) plus Northern New Hampshire (-.592z) and Northern Maine (-.544z) are all belowf -0.50z, translating into absolute mean annual deficits in precipitation between 2.88" (Northern Maine) to 3.25" (Northeastern Vermont).

3.2 New England Ranking Mode #2 – 18.6 % Percent Frequency

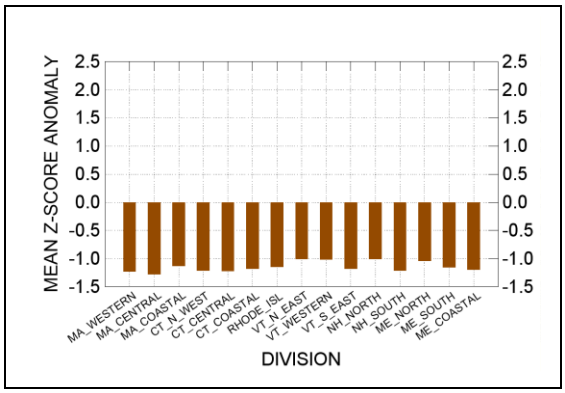

Figure 6. New England Mode #2 – "Extreme Below Normal Anomalies Throughout"

Figure 6 shows the results for Mode 2, ranking second of the six overall (18.6% frequency – encompassing 23 years of the history). This depicts a remarkably uniform pattern of considerably below normal annual standardized anomalies for all 15 divisions. The figures range from -1.27z (Central Massachusetts) to -1.01z (Northern New Hampshire & Northeast Vermont), the absolute mean deficits ranging from -5.32" (Northeast Vermont) to -8.79" for Northwestern Connecticut) .

3.3 New England Ranking Mode #3 – 16.9 % Percent Frequency

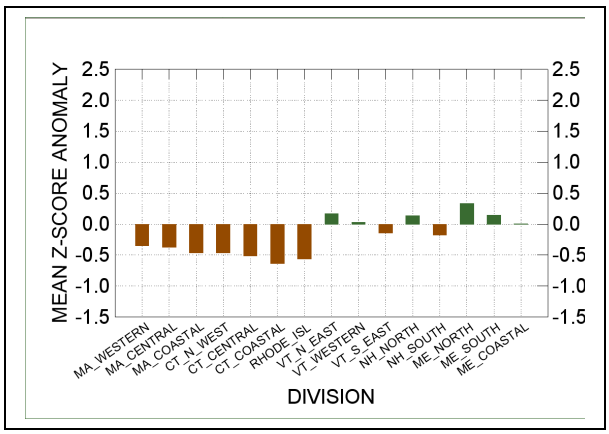

Figure 7. New England Mode #3 – "Noticeably Below Normal Anomalies South, mixed but Lesser Anomalies North"

Figure 7 presents the results for Mode 3, ranking third most important of the six, with a 16.9% percent frequency. The south to north dichotomy in mean anomaly character is striking, all of the southern divisions displaying negatives, the northern ones of a mixed character. The former's statistics range from -.347z for Western Massachusetts to -.637z for Coastal Connecticut, the latter translating into a -4.24" mean deficit in absolute terms. Northern Maine has the largest positive anomaly (+.336z).

3.4 New England Ranking Mode #4 – 15.3 % Percent Frequency

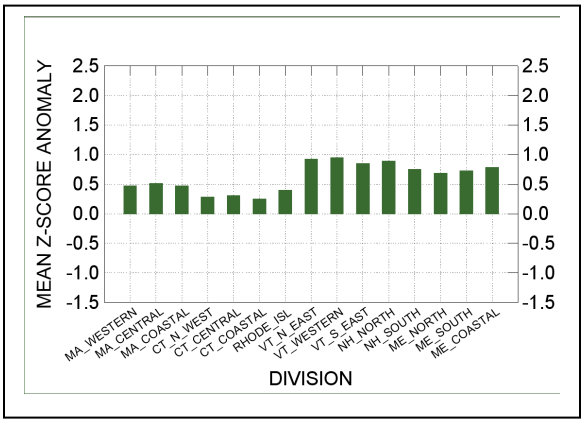

Figure 8, New England Ranking Mode #4 – "Wetter than Normal South, More Anomalously so North"

Figure 8 shows the results for Mode 4, fourth most important of the six (overall incidence: 15.3 %). The standardized precipitation anomalies are positive for all divisions, but again, there is a noticeable contrast between north and south in their collective degrees of departure. All of the northern division's anomalies are more positive than those in the south, the most extreme magnitudes being +0.920z and +0.947z for Northeastern and Western Vermont, respectively, corresponding to absolute mean departures of +4.85" and +5.08", respectively. Those for the three Connecticut divisions are the least positive of the fifteen, ranging from +.248z for the Coastal division (+1.65" absolute mean departure) to +.308 for the Central (+2.13" absolute departure).

3.5 New England Ranking Mode #5 – 11.3 % Percent Frequency

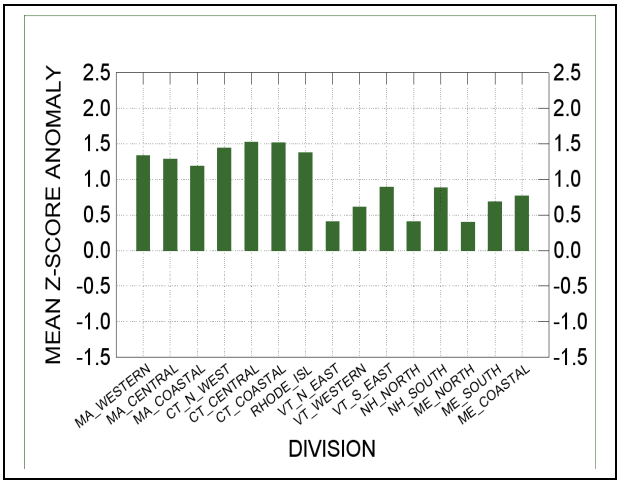

Figure 9 – New England Mode #5 – "Excessively Wet South, Wet but Less Anomalously So North"

Mode 5 (Figure 9) also displays a wetter than normal pattern for all fifteen divisions, but in this case the most extreme positive relative anomalies are shifted to the south, the south/north contrast being more pronounced than it was with Mode 4. All of the south divisions' anomalies are greater than +1.0z,,none of the north's divisions as high as that figure; three of them are less than +0.5z. In direct opposite contrast to Mode 4, the most extreme positive anomalies are observed for the three Connecticut divisions: +1.523z for the Central, +1.516z for the Coast, and +1.442z for the Northwest, these translating into absolute positive departures in excess of 10".

3.6 New England Ranking Mode #6 – 7.3 % Percent Frequency

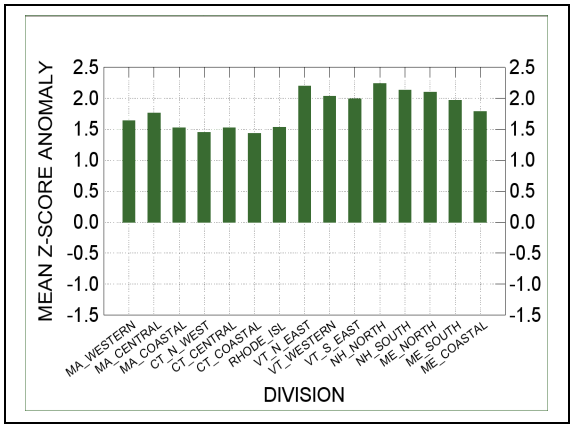

Figure 10 – New England Mode #6 – "Excessively Wet Throughout, but More So North"

Finally,Mode 6, the last ranking pattern at 7.3 % incidence (9 years so-classified) shows an excessively wet annual pattern for all fifteen divisions, every division but two with standardized departures at 1.5z or higher. In this pattern, the north division/south division dichotomy switches back to a north "favored" one, all of the north's z-values higher than the south's. Five of the north's divisions display z-scores at higher than +2.0z, including a +2.239z magnitude statistic for New Hampshire-North. The +2.132z figure for New Hampshire-South produces a mean absolute departure of 13.56". Mean absolute departure for the eight north divisions (New Hampshire, Vermont, and Maine) is +12.02", for the seven south (Massachusetts, Connecticut, and Rhode Island), 10.45"; mean overall is 11.29"

3.7 Identification of New England Most Extreme Individual Years' Region-to-Region Annual Precipitation Patterns Utilizing Ranked Statistical Distances from Centroids

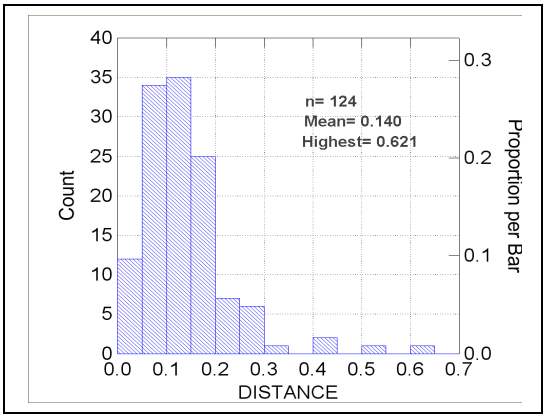

Figure 11: Histogram of Individual Years' Statistical Distances to Respective Cluster Centroids.

An interesting spinoff of a cluster analysis of this kind is that one can get an indication of the most extreme individual years' division-to-division annual patterns that have occurred over the history of the period of record through the ranking of statistical distances from cluster centroids. Combining the distances, irrespective of mode, into a single data set and then ranking the distances in order of lowest to highest, one could identify the most extreme of the extremes. The rationale of combining the distances from different clusters to arrive at the overall most anomalous cases is that a given divisional pattern of normalized annual precipitation patterns would certainly be more "distant" from centroids of the other clusters of which it was not a member.

Figure 11 above is a histogram of the New England divisional data set's distances, by year. Mean overall statistic was 0.140, but the extreme maximum distance was 0.621, more than four times higher, and thus, as an exercise to throw light on the "outermost" New England patterns, the four with the most extreme distances are plotted in Figures 12-15.

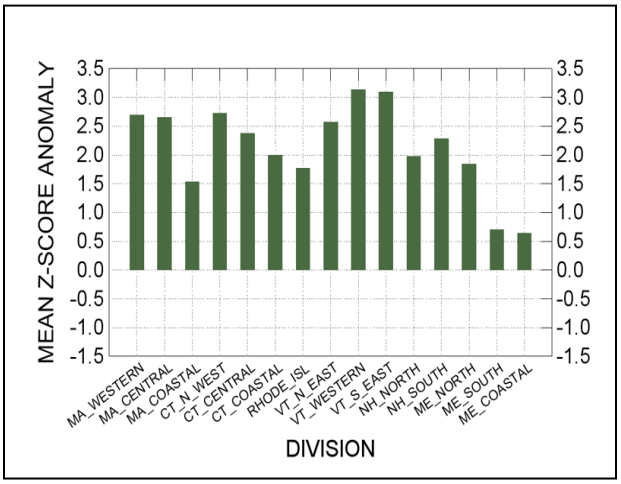

Figure 12. New England Climate Divisions' Standardized Precipitation Anomaly Patterns for 2011 (Distance: 0.400 - Fourth greatest distance)

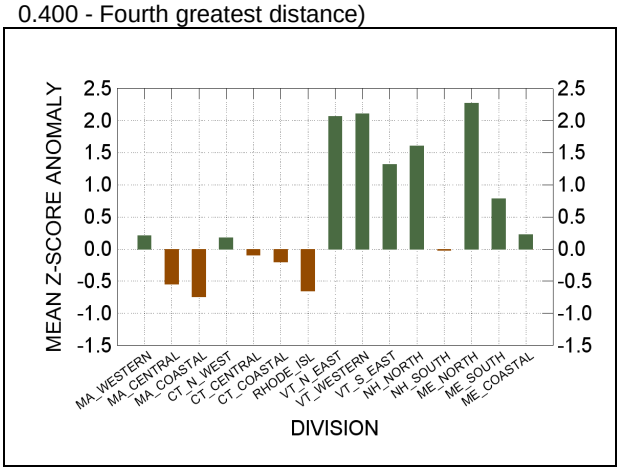

Figure 13. New England Climate Divisions' Standardized Precipitation Anomaly Patterns for 1976 (Distance: 0.427 – Third greatest)

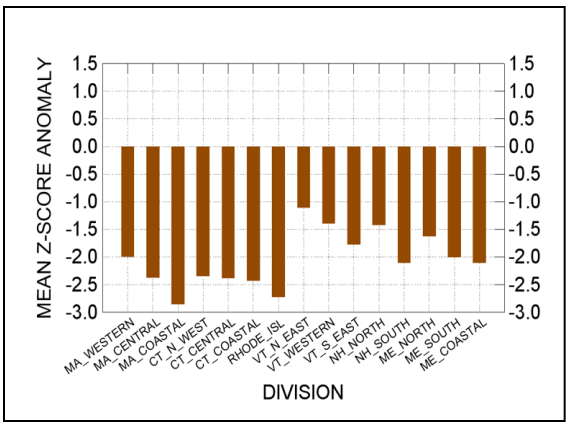

Figure 14 – New England Climate Divisions' Standardized Precipitation Anomaly Patterns for 1965 (Distance: 0.547 – Second greatest)

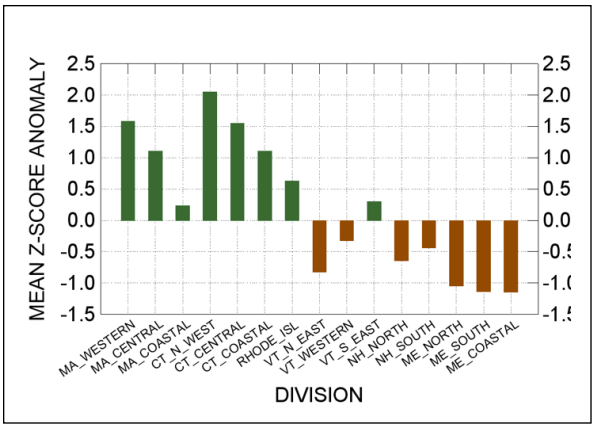

Figure 15 – New England Climate Divisions' Standardized Precipitation Anomaly Patterns for 1955 (Distance: 0.621 – Greatest Distance

The four extremes' charts show a variety of patterns, the north/south dichotomy still evident to at least some extent for three. The dichotomy is absent for the year 2011 (Figure 12), a very wet year for nearly all the divisions except Southern Maine and Coastal Maine, although these latter two were still more than 0.5z wetter than normal. Six of the divisions had annual totals more than 2.5z above normal, and two divisions, Western Vermont and Southeast Vermont were more than +3.0z above. For the year 1976, the dichotomy is very amplified, the south mostly drier than average, but the north significantly wetter than normal in most cases. Severe drought throughout the divisions contributed to 1965's standing as the second most "anomalous" year pattern-wise. The dichotomy is quite evident with most of the extreme-most negative departures evident in the south. Coastal Massachusetts and Rhode Island were each more than 2.5z below normal. Finally, 1955, the most "extreme" shows a clear dichotomy between the south (all divisions wet) and the north (all but one dry). Three of the south divisions had annual falls above +1.5z, but three in the north (all for Maine, were more

than 1.0z below. The two major hurricanes that hit the southern portions of New England in August 1955 (Connie and Diane) likely amplified the ultimate contrasts in annual precipitation totals between the southern divisions affected by the storms vs. the others which were not.

4. SUMMARY

Revisting the above six clusters' results, the following closing generalizations can be made about the nature of New England area annual precipitation character across the fifteen NCDC/NCEI climate divisions.

With the exception of severe drought years, which occur about once every five to six years (based on the 18.6% percent frequency statistic for Cluster #2). the New England climate divisions show annual precipitation anomaly dichotomies of various intensities between the three-state groupings: Connecticut, Rhode Island, and Massachusetts (south) vs. New Hampshire, Vermont, and Maine (north). The dichotomy contrasts are relatively insignificant for Clusters #1 and #3 (combined incidence: 47.5 %), the remaining 33.9 % (Clusters, #4, #5, and #6) reflecting wetter than average years of progressively increased anomaly character, which the clustering algorithm resolved as distinct modes. This north-to-south/south-to-north latitudinal separation likely reflects year-to-year variations in positions of the mean storm track.

5. REFERENCES

Fisk, C., 2016, Identification of California, Oregon, and Washington Coastal Area Climate Divisions' Relative Anomaly Modes in Total Seasonal Precipitation With Characterization of Their Occurrence Probabilities Relative to El Nino, Neutral and La Nina Episodes: AMS 28th Conference on Climate Variability and Change Intelligence, New Orleans, LA, 13 January, 2016. https://ams.confex.com/ams/96Annual/webprogram/Paper284304.html

Fisk, C., 2015, Identification of California Climate Division Rain Year Precipitation Anomaly Patterns (1895-96 to 2013-14 Seasons) with Bayesian Analyses of Occurrence Probabilities Relative to El Nino, Neutral, or La Nina Episodes: AMS 13th Conference on Artificial Intelligence, Phoenix, AZ, 6 January, 2015. https://ams.confex.com/ams/95Annual/webprogram/Paper260219.html

Nisbet, R., Elder, J., and Miner, G., 2009: Handbook of Statistical Analysis & Data Mining Applications Elsevier, 824 pp.