ANALYSIS OF MONTH-TO-MONTH PRECIPITATION VARIABILITY PATTERNS FOR LOS ANGELES, SAN DIEGO, AND SAN FRANCISCO (1877-78 THROUGH 2012-3 SEASONS) UTILIZING K-MEANS CLUSTERING ANALYSIS INTEGRATED WITH THE V-FOLD CROSS VALIDATION ALGORITHM

Charles J. Fisk *
NAVAIR-Point Mugu, CA

1. INTRODUCTION

Long-term monthly averages are a traditional means of characterizing climatological precipitation variability over the course of a rain year. Frequently based on the 30-year period of record, they serve as monthly precipitation "normals" which are a basis for anomaly calculations.

Such "normals", however, are only statistical idealizations, and actual individual years' month-to-month rainfall patterns invariably depart from mean climatology in some fashion. Inherent climatological tendencies may exist, for example, for occasional month-to-month clustering of positive or negative anomalies, or pronounced wet anomalies for certain months being succeeded by below normal ones several months later. These patterns might also be influenced by ENSO-type (i.e., "La Nina", "Neutral", or "El Nino"). Information on such tendencies, to whatever extent they are real, would represent a useful complement to the more conventional climatological characterizations.

To explore these possibilities, the following study investigates the existence and relative frequencies of month-to-month precipitation anomaly patterns for three California localities with lengthy periods of record. These are the downtown stations of San Francisco, Los Angeles, and San Diego, CA. The K-means clustering analysis methodology integrated with the V-Fold Cross Validation Algorithm is applied. The latter, a "training-sample" data mining procedure, allows for a more objective determination of the optimal number of clusters when incorporated into K-Means. Preliminary to the analysis, the raw monthly precipitation data are normalized by month, and to create and characterize the clusters, two distance methodologies, the Euclidean and Squared Euclidean, are utilized with their results compared. The nature of the patterns and their frequencies relative to El Nino, Neutral, and La Nina ENSO episodes are described and related.

Periods of record examined for each station are 1877-78 through 2012-13 July-June rain years. San Francisco's and San Diego's periods of record extend further back to 1849-50 and 1875-76, respectively, Los Angeles' to 1877-78, but for consistency the former two stations' records are shortened to conform to Los Angeles' historical length.

. Given the stations' winter rainfall maximum/summer drought character, the calendar period selection includes October-November, December, January, February, March, and April-May. Thus, the analyses become clustering exercises in six-dimensional space.

The K-Means/V-Fold methodology has been utilized in several other climatic-related studies, such as Coastal Southern California stations' diurnal wind patterns in summer (Fisk, 2012), single station (La Guardia, NY) 24-hour wind patterns for all months of the year (Fisk, 2013), and California Climatic Division rain year anomaly patterns (Fisk, 2013).

2. THE K-MEANS AND V-FOLD CROSS VALIDATION METHODOLOGIES

The original K-means methodology was introduced by Hartigan (1975), and the basic methodology consists of assigning observations to a designated number of K clusters such that the multivariate means across the clusters are as different as possible. The differences can be measured in terms of Euclidean, Squared Euclidean, City-Block, and Chebychev statistical distances (Nisbet, et. al., 2009).

The V-fold cross-validation scheme, as applied to K-means clustering involves dividing the overall data sample into V "folds", or randomly selected subsamples. K-means analyses are then successively applied to the observations belonging to the V-1 folds (training sample), and the results of the analyses are applied to sample V that was not used in estimating the parameters (the testing sample) to assess the predictive validity or the average distances of the training sample arrays from their cluster center centroids. The procedure is repeated for cluster sizes K+1. K+2, …, etc., until the incremental improvement in the average distances is less than some threshold, at which time the "optimal" cluster size is considered attained (Nisbet, et. al., 2009).

The STATISTICA Data Miner Clustering module was utilized to employ this technique. Normalization (an automatic software feature), reduces the data to a common scale and lessens the influence of outliers.

As the distance threshold can be changed, generation of the "optimal" number of clusters is not completely automatic. Nonetheless, the V-fold cross-validation algorithm enhances the methodological objectivity of a clustering technique like K-means.

In this study, the 5 percent default distance improvement cutoff threshold is retained in concert with the Euclidean (default) and Squared Euclidean distance metrics.

* *Corresponding author address:* Charles J. Fisk, NAWCWPNS, Point Mugu, CA. 93042: e-mail: charles.fisk@navy.mil

3. DATA AND PROCEDURES

Data for the three stations were secured from various online sites, including those of the National Climatic Data Center and National Weather Service. The precipitation histories of each station include a number of station moves locally, but for the purposes of this analysis, the moves are assumed to have negligible influence on results' outcomes.

Also, identification of ENSO episodes is a not completely objective or definitive process, different researchers have composed different lists, and there is likely more uncertainty with years further back than closer to the present. For the purpose of this research, the lists utilized are those formulated by the NOAA Climate Prediction Center. The first covers the years 1877-2001, the second 1950-2012. Those years that overlap (1950-2012) are given the designations assigned by the latter.

4. RESULTS

4.1 - Downtown San Francisco Results

Figure 1 is a bar graph depicting mean overall San Francisco precipitation figures for the six calendar periods under consideration. Individual period figures range from 4.38" for January, to 2.13 for April-May. Total average Oct-May precipitation is 20.98", the median 20.32"

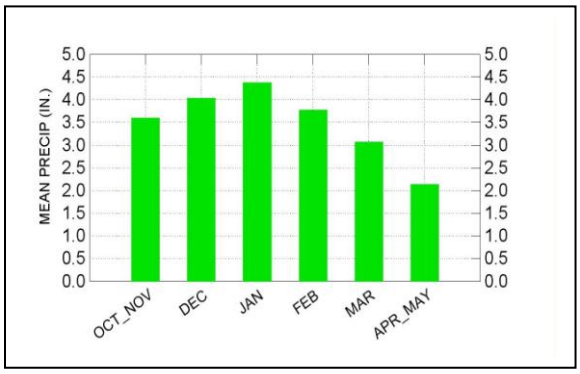

Figure 1. Mean Downtown San Francisco Precipitation for October-November, December, January, February, March, and April-May calendar periods, 1877-78 to 2012-13 Period of Record.

4.1.1- Downtown San Francisco Cluster generation with Euclidean Distances

Five clusters were created utilizing the Euclidean distance metric, combined with the 5 percent default distance improvement cutoff threshold. Mean statistical training errors of the normalized individual observations relative to their respective cluster centroids was .349. The "optimal" number of clusters in this instance is matched by the results of the STATISTICA scree plot in Figure 2, which shows an inflection point at cluster 5. The Scree plot is used in this manner as a graphical device for selecting a "best" number, and in this instance the Euclidean Distance Metric /5 percent cutoff threshold and Scree "bests" are the same.

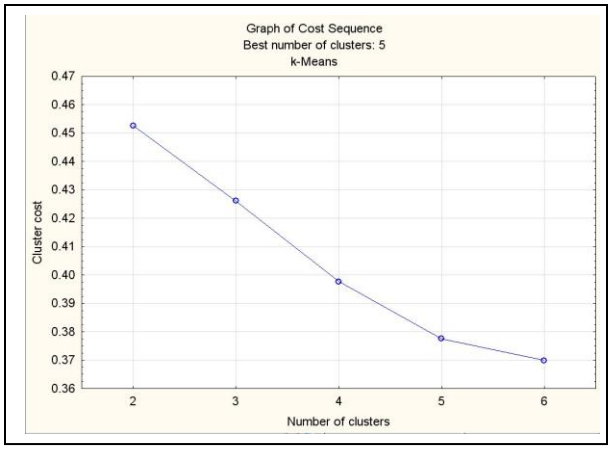

Figure 2. Scree Plot of Sequential Cluster Generation Cost Statistics for San Francisco Period-to-Period Precipitation Anomaly Patterns – Euclidean Distance Metric.

Figures 3 to 7 below are bar graphs of the mean monthly precipitation anomalies for each of the five clusters, arranged in rank order of importance. Annotations within each chart denote the percent occurrence of the pattern along with the percentage mix of La Nina, Neutral and La Nina episodes among cluster memberships. Overall, La Nina's, Neutral's, and El Nino's made up 25%, 46%, and 29%, respectively, of the 136 rain years.

From inspection of the charts, by far the most frequent idealized pattern for Downtown San Francisco (incidence: 40.4% - see Figure 3), is one of persistent moderate to slight dry anomalies for all six monthly periods. Average collective Oct-May precipitation for this cluster is 15.43" for 55 seasons, 26.4% below the overall mean (20.98"). The ENSO mix (23%, 44%, and 33%) does not depart appreciably from the overall makeup; such will be the case for nearly all of the five San Francisco Euclidean modes.

The second through fourth ranked patterns (Figures 4 to 6) display single-month highly positive anomalies for December, October-November, and March, respectively. Collectively, these made up 48.5% of the 136 rain seasons. In both the "Wet December" and "Wet October-November" charts, there is a noticeable dry anomaly three months downstream from each of the "wet" bars, suggestive of a climatological trough to ridge progression.

Mode 5 ("Wet January/February" – incidence 11.0%) is the only wet pattern involving contiguous months and also the only one in which El Nino's predominate the ENSO mix. Some 47% of the years belonging to this cluster are El Nino's, higher than either the Neutral or La

Nina figures, and much higher than the El Nino 29% overall average. Total mean October-May rainfall for this mode (28.28") is by far the highest of the five, and 37.4% wetter than the 20.98" overall average.

In sum, it appears that the great majority of San Francisco month-to-month idealized precipitation variability consists either of fairly non-descript near or slightly below normal rainfall one month after another, or episodes of heavy falls confined to a single month. There is, however, an apparent inclination for heavy falls to persist in both January and February during some El Nino's.

From this it did not appear that there was an all-encompassing association between San Francisco month-to-month precipitation variability and ENSO designation, and a Chi-Square test of uniformity of cross-tabulation frequencies confirmed this, the null-hypothesis being rejected only the .669 level (Chi-square:5.81, degrees of freedom: 8).

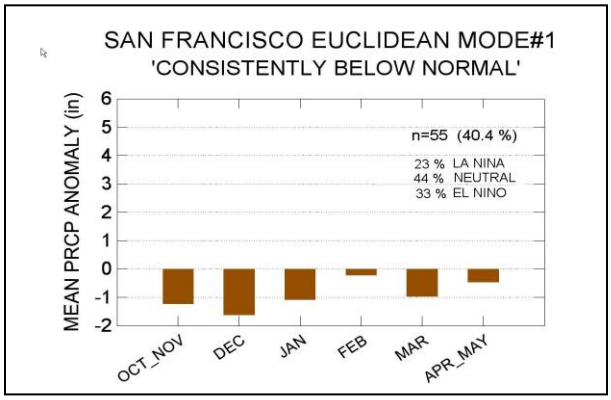

Figure 3. Monthly Precipitation Anomaly Chart for Downtown San Francisco "Consistently Below Normal" Pattern- 40.4% Incidence – Euclidean Distance Option

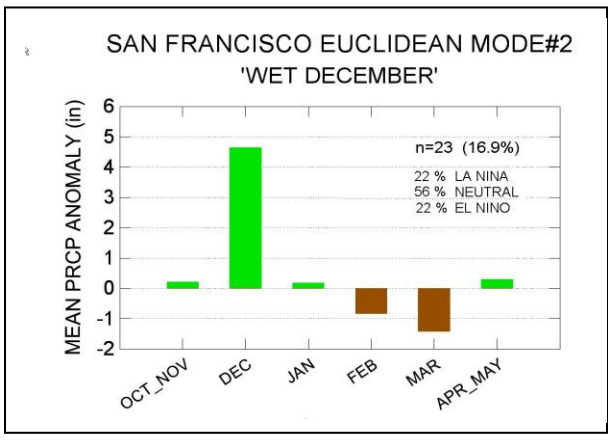

Figure 4. Monthly Precipitation Anomaly Chart for Downtown San Francisco "Wet December" Pattern – 16.9 % Incidence – Euclidean Distance Option

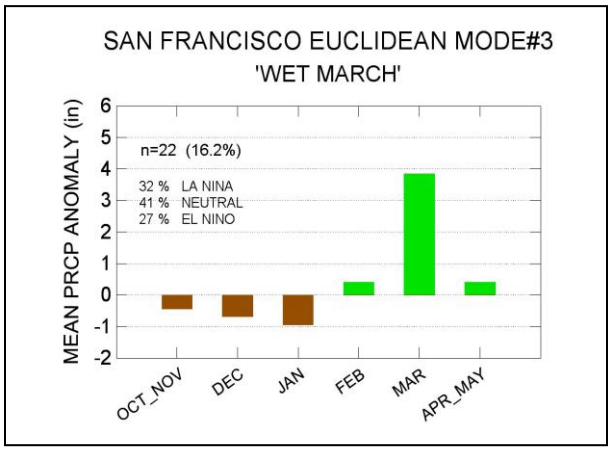

Figure 5. Monthly Precipitation Anomaly Chart for Downtown San Francisco "Wet March" Pattern – 16.2 % Incidence - Euclidean Distance Option

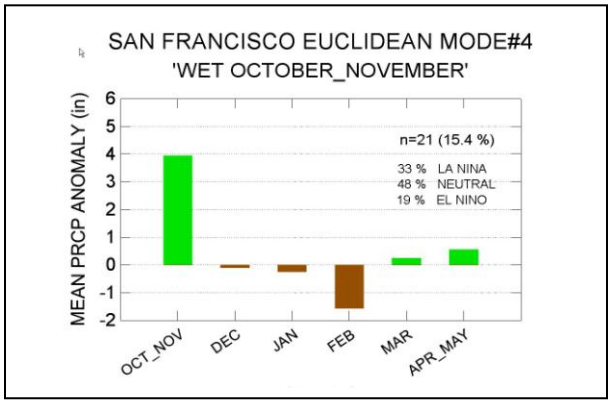

Figure 6. Monthly Precipitation Anomaly Chart for Downtown San Francisco "Wet October_November" Pattern- 15.4% Incidence – Euclidean Distance Option

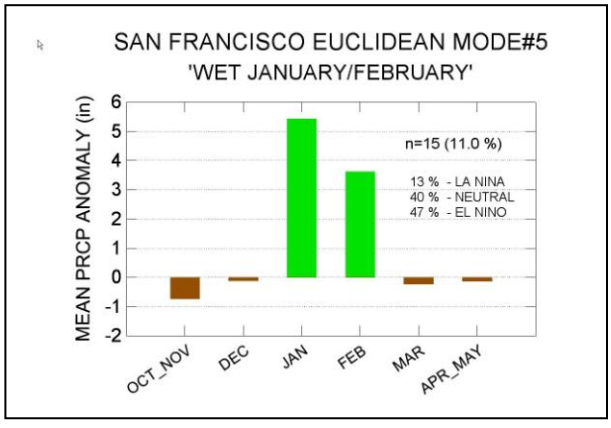

Figure 7. Monthly Precipitation Anomaly Chart for Downtown San Francisco "Wet January/February" Pattern- 11.0% Incidence – Euclidean Distance Option

4.1.2- Downtown San Francisco Cluster generation with Squared Euclidean Distances

The Squared Euclidean Distance Metric is utilized in clustering analysis when the goal is to "place progressively greater weight on objects that are farther apart" [Wikipedia,2013], " fleshing out", as it were, observations whose distances from others in n-dimensional space would be otherwise less distinct, and in the process promoting the generation of more clusters.

Six clusters for Downtown San Francisco were generated with this option, combined with the 5 percent cutoff threshold. Mean training error was a much improved (.130), a little more than a third of that for the Euclidean option (.349). The "optimal" number of clusters in this instance was more or less "confirmed" again by the STATISTICA Scree plot in Figure 8, although in this case, the percentage drop in "cost" between clusters 5 and 6 was too pronounced to produce a cutoff. The latter did result between clusters 6 and 7 as the cost of the latter was virtually unchanged from the former.

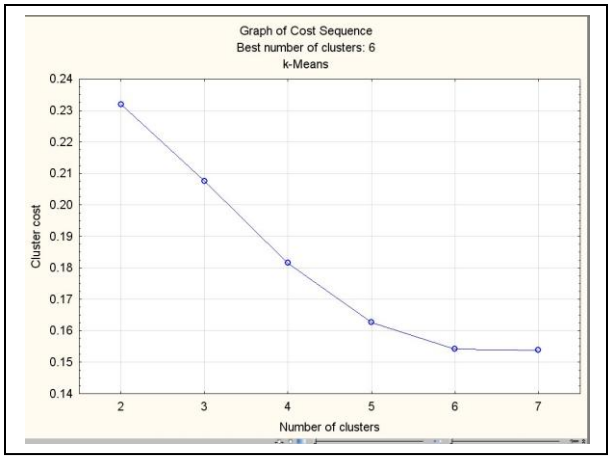

Figure 8. Scree Plot of Sequential Cluster Generation Cost Statistics for San Francisco Month-to-Month Precipitation Anomaly Patterns – Squared Euclidean Distance Metric.

Figures 9 to14 below are bar graphs of the mean monthly precipitation anomalies for each of the six patterns, arranged in rank order of prominence. Compared to the Euclidean method there is scarcely any contrast in their patterns and rank orderings, the only change of significance being the addition of a sixth cluster, and that only having two members.

Again, by far the most frequent pattern for Downtown San Francisco (incidence: 40.4% - see Figure 9), is that of the persistent moderate to slight negative anomalies for all the six monthly periods. As no membership changes resulted, average collective Oct-May precipitation for the 55 seasons remained at 15.43", more than 5" inches below average, to go with the same ENSO mix displayed: (23%, 44%, and 33%).

The second through fourth ranked modes (Figures 10 to 12), plus the new mode ("Very Wet April/May" – see Figure 14) each display single-month highly positive anomalies. These comprise 49.3% of the years. Also, the three months' downstream dry anomaly feature seen previously in the Euclidean "Wet December" and "Wet October-November" charts is present again (Figures 10 and 12, respectively)

Mode 5 (Figure 13 - "Wet January/February" – incidence 10.3%) is again the only pronounced wet pattern involving contiguous months. In slight contrast to the Euclidean method which had El Nino's being the most prominent (47%) relative to Neutrals and La Nina's, the El Nino incidence is slightly reduced here to 43%, matching the figure for Neutrals.

Again, subjectively, it did not seem that there was an overall association between San Francisco month-to-month precipitation variability and ENSO designation, and the Chi-Square test of uniformity established this to an even stronger degree, the null-hypothesis being rejected at just the .808 level (Chi-square:6.09, degrees of freedom: 10)

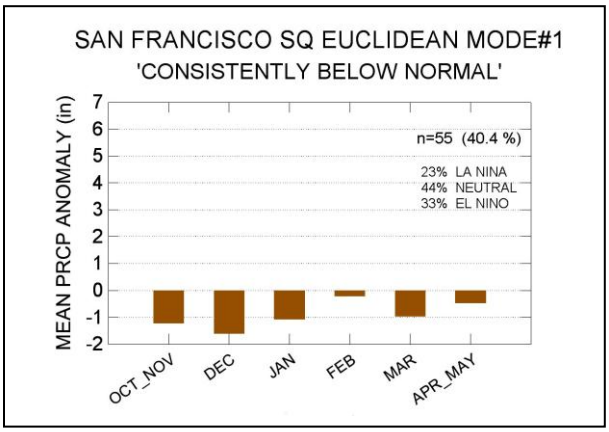

Figure 9. Monthly precipitation Anomaly Chart for Downtown San Francisco "Consistently Below Normal" Pattern- 40.4% Incidence – Squared Euclidean Method

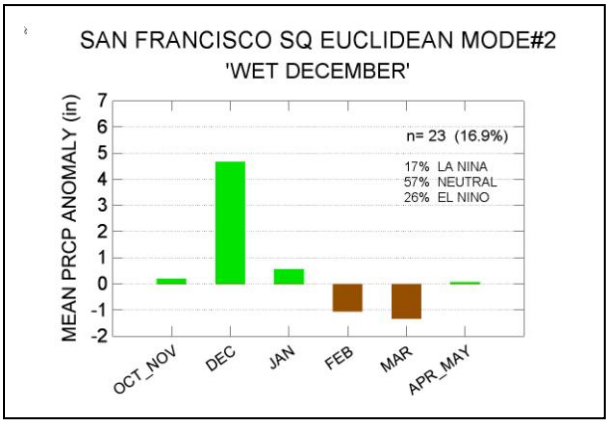

Figure 10. Monthly precipitation Anomaly Chart for Downtown San Francisco "Wet December Pattern" – 16.9 % Incidence – Squared Euclidean Method

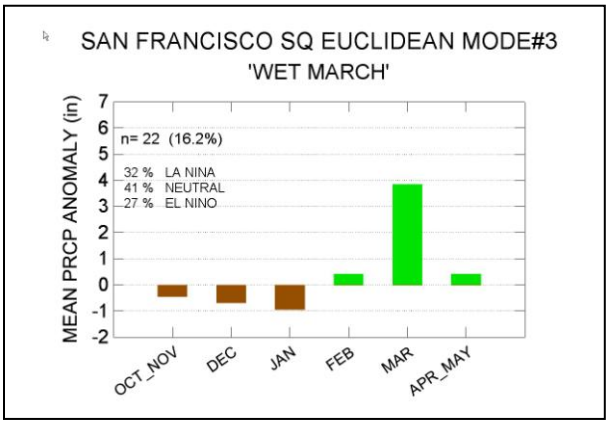

Figure 11. Monthly precipitation Anomaly Chart for Downtown San Francisco "Wet March Pattern" – 16.2 % Incidence - Squared Euclidean Method

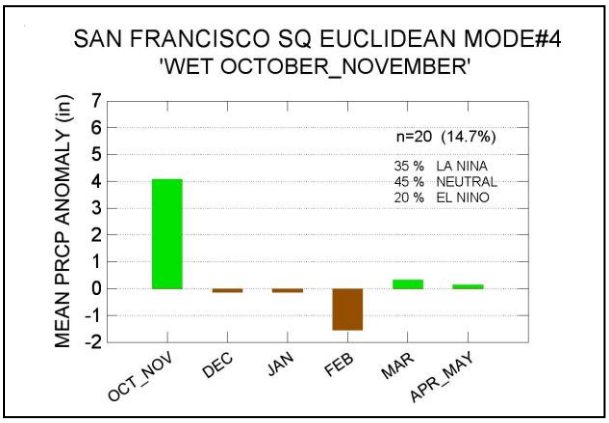

Figure 12. Monthly precipitation Anomaly Chart for Downtown San Francisco "Wet October-November" Pattern- 14.7% Incidence - Squared Euclidean Method

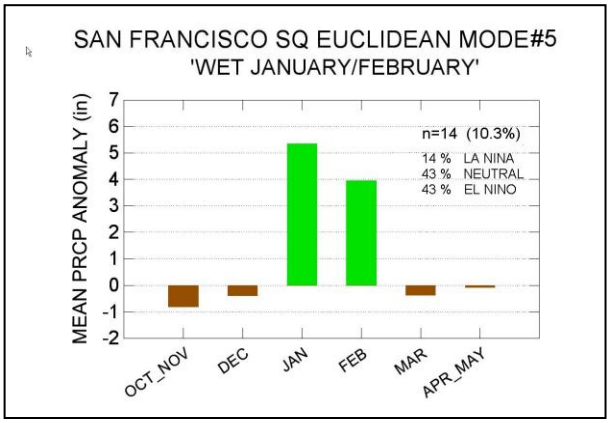

Figure 13. Monthly precipitation Anomaly Chart for Downtown San Francisco "Wet January/February Pattern- 10.3% Incidence - Squared Euclidean Method

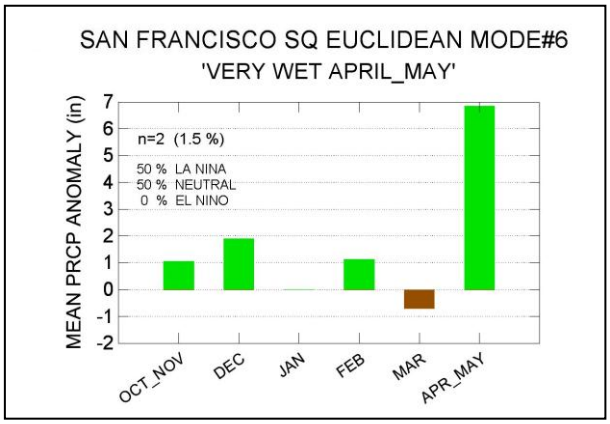

Figure 14. Monthly precipitation Anomaly Chart for Downtown San Francisco "Very Wet April/May" Pattern- 1.5% Incidence – Squared Euclidean Method.

4.2 - Downtown Los Angeles Results

Figure 15 is a bar graph depicting mean overall Downtown Los Angeles precipitation figures for the six calendar periods. Compared to San Francisco, Los Angeles' mean October-May rainfall (14.53") is 31% less, and its individual monthly maximum (3.26") is a February one rather than January. Median Los Angeles October-May rainfall (12.73") is nearly 2" less than the mean, indicative of an appreciable positive skewness in its136-year statistical distribution.

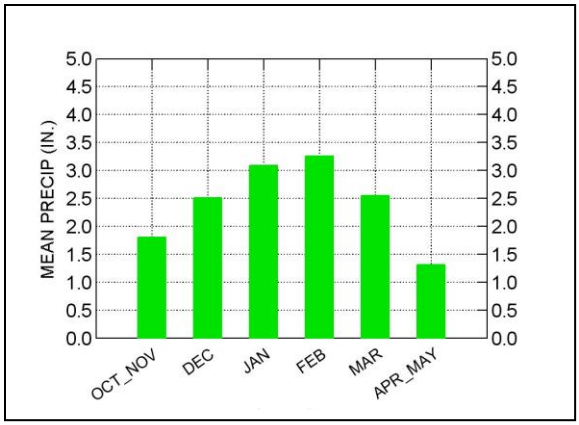

Figure 15 Mean Downtown Los Angeles Precipitation for October-November, December, January, February, March, and April-May calendar periods, 1877-78 to 2012-13 Period of Record.

4.2.1- Cluster generation with Euclidean Distances

The Euclidean option produced three clusters for Downtown Los Angeles, compared to five for San Francisco, but the patterns are more or less similar to San Francisco's, in form if not frequency. Mean statistical cluster centroid training error was .369, compared to San Francisco's .349. The "optimal" number of cluster figure was attained at an n=3 cutoff, as the cost statistic declined by less than 5 percent to n=4 (see Figure 16 Scree Plot below).

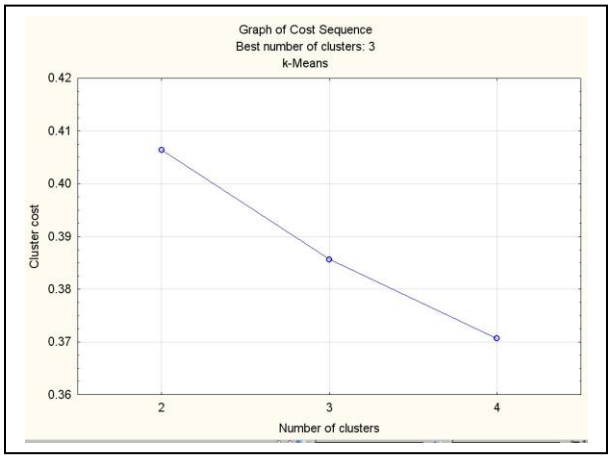

Figure 16. Scree Plot of Sequential Cluster Generation Cost Statistics for Los Angeles Month-to-Month Precipitation Anomaly Patterns – Euclidean Distance Metric

Figures 17 to 19 below are the rank ordered bar graphs of the mean monthly precipitation anomalies for each of the three clusters. Annotations within each chart are the same as those displayed in the San Francisco charts. To repeat, overall La Nina, Neutral, and El Nino incidence was 25%, 46%, and 29%, respectively.

By far the most frequent idealized Euclidean pattern for Downtown Los Angeles (incidence: 66.9% - see Figure 17), is one of persistent mostly slight negative anomalies for all six monthly periods, essentially a repeat of San Francisco's, but significantly more predominant in frequency. Average collective Oct-May precipitation for this cluster is 11.32" for 91 seasons, 22.1% below the overall mean (14.53"). The ENSO mix (31%, 45%, and 24%) is not markedly different from overall climatology. With the possible exception of February, the mean departures in Figure 17 are so slight that the values on an individual period basis could be considered essentially "near normal". It should be mentioned that since most *median* monthly rainfall figures for Los Angeles are appreciably less than mean amounts, most individual years' monthly precipitation totals are "below normal" (i.e. less than average).

Interpretation of this pattern and its high frequency leads to the conclusion that most rain seasons (~two-thirds) in Los Angeles are likely absent of excessive wet episodes, allowing, of course, for individual member years' variation.

Ranking second (incidence: 20.6%) is the "Wet February" pattern (See Figure 18). While February stands out with its pronounced mean positive anomaly (+5.42"), the mean positive anomalies of January and March (slightly greater than 1", are respectable also, considering the relatively low Los Angeles climatological mean figures for those months (3.09" and 2.55") respectively. Mean Los Angeles Oct-May Precipitation for this cluster was 21.89", some 151% of average. Inspecting the ENSO mix, there is a decided imbalance of frequencies - just 7% of the cluster members were La Nina's (expected: 25%) and 43% El Nino's (expected: 29%). By way of comparison with San Francisco, the pronounced February feature of Figure 18 was a somewhat less prominent element of the San Francisco "wet January/February" chart (Figure 7). "Wet February" is the favored Los Angeles mode for El Nino's, "wet January/February" the favored one for San Francisco.

The third and last mode (incidence: 12.5% - see Figure 19) is "Wet October-November/Dry February". Similar to San Francisco's "Wet December" pattern in Figure 4, the latter lacking, however, the relatively pronounced December positive anomaly. It displays the three-month trough to ridge "propagation" feature which conceivably could be associated with a climatological inclination to transition to a dry regime by February.

While the Los Angeles "Wet-February" pattern above did exhibit a clear indication of non-uniform frequencies for El Nino vs. La Nina episodes, a Chi-Square test of overall uniformity of the cross-tabulated frequencies for the three modes versus the three ENSO designations still fell "short" of high significance, rejecting the null hypothesis at the .101 level (Chi-square:7.765, degrees of freedom: 4).

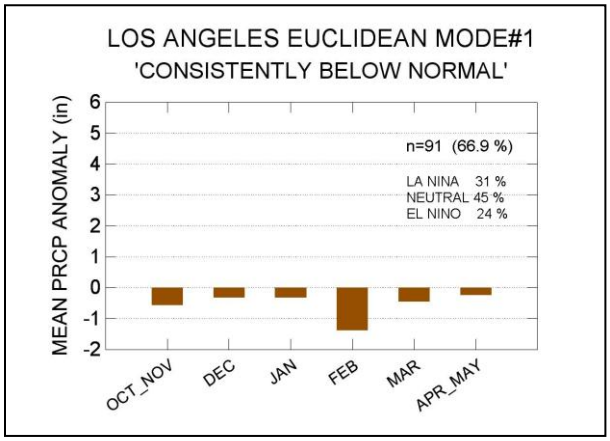

Figure 17. Monthly precipitation Anomaly Chart for Downtown Los Angeles "Consistently Below Normal" Pattern- 66.9% Incidence – Euclidean Method

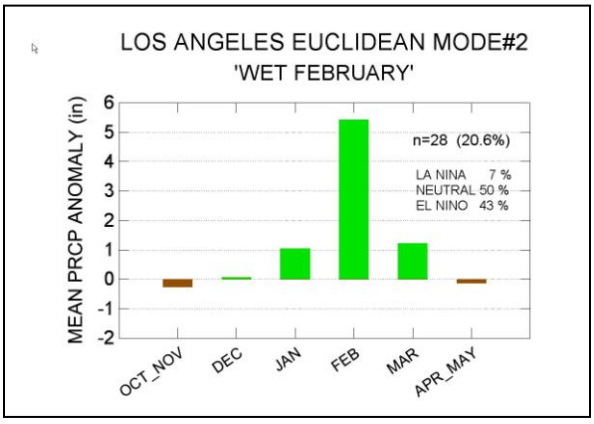

Figure 18. Monthly precipitation Anomaly Chart for Downtown Los Angeles "Wet February" Pattern- 20.6% Incidence – Euclidean Method

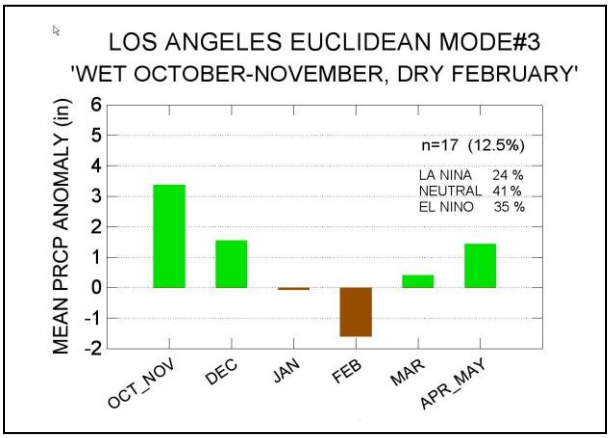

Figure 19. Monthly precipitation Anomaly Chart for Downtown Los Angeles "Wet February" Pattern- 12.5% Incidence – Euclidean Method

4.2.2- Cluster generation with Squared Euclidean Distances for Downtown Los Angeles

In contrast with the San Francisco results, which produced virtually no Squared Euclidean vs. Euclidean differences in patterns and rankings, save for the addition of a minor new cluster for the latter, the Los Angeles squared Euclidean method produced contrasts in outcomes.

Five clusters were generated utilizing the 5 percent cutoff threshold. Mean training error was a much improved .121, less than a third of that of the Euclidean option (.369). From the Scree plot in Figure 20, the modest falloff in "cost" from n=5 to n=6 resulted in the cutoff occurring at the former.

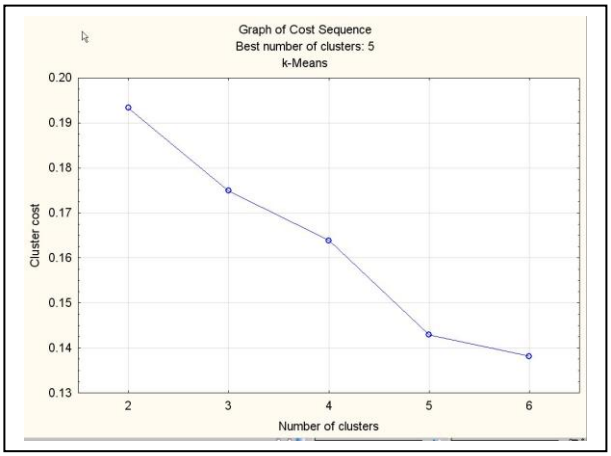

Figure 20. Scree Plot of Sequential Cluster Generation Cost Statistics for Downtown Los Angeles Month-to-Month Precipitation Anomaly Patterns – Squared Euclidean Distance Metric.

Figures 21 to 25 below are bar graphs of the mean monthly precipitation anomalies for each of the five pattern modes, arranged in rank order of importance.

Again, the most frequent pattern for Downtown Los Angeles (incidence: 44.1%) is the "Persistently Below Average" one shown in Figure 21, that with consistently negative anomalies for all six periods. Compared to its Euclidean counterpart, the frequency is down (from 66.9%), but the dry anomalies are slightly more substantial, especially for December, January, and February. Average collective Oct-May precipitation is reduced to 9.41", more than 5" inches below average. The ENSO mix (27%, 45%, and 28%) differs only slightly from Climatology.

Ranking second, as before, with some changes in configuration, is the "Wet February" pattern (incidence: 17.6% - see Figure 22). Four of the six anomalies are negligible, and aside from February's +5.74" departure, only March has an appreciable magnitude (+1.54 – 60.3% above average"). This mode probably could have named the "wet February/March" pattern without any loss of generality. The disproportionate mix of ENSO designation frequencies is again present, only 8% (expected: 25%) of the members being La Nina's versus 42% (expected: 29%) being El Nino's. Mean October-May precipitation for the 24 rain years in this cluster is 21.55", or 148% of average.

Third in rank is the "Wet January" cluster (incidence: 14.0% - see Figure 23). Like "Wet February" most of the other period anomalies are slight. January's mean anomaly is +5.35", none of the others exceeding 1". El Nino's make up 32% of the membership, compared to 21% for La Nina's, but this is a relatively modest distinction compared to "Wet February's" disparity.

Tied with "Wet January" in 14.0% frequency is the "Wet Oct-Dec, Dry Jan-May mode" (see Figure 24) so called as it consists of positive anomalies for October-November, & December, succeeded by negatives for

each of the remaining periods January, February, March, and April-May. This seems to be a “prototypical” La Nina mode, as some 47% of cluster members were of this designation, compared to only 16% El Nino’s.

Ranking fifth is the “wet-April-May” cluster (incidence 10.3%, see Figure 25). In addition to the large 2.90” positive anomaly for April-May (222% of average), the anomalies exhibit a period-to-period positive to negative fluctuation.

While the “Wet-February” pattern again exhibited a decided preponderance of El Nino’s over La Nina’s, and the “Wet Oct-Dec, Dry Jan-May” cluster had La Nina’s predominating El Nino’s, the Chi-Square test of overall frequency uniformity fell “short” again of major significance, rejecting the null hypothesis at only the .286 level (Chi-square:9.706, degrees of freedom: 8).

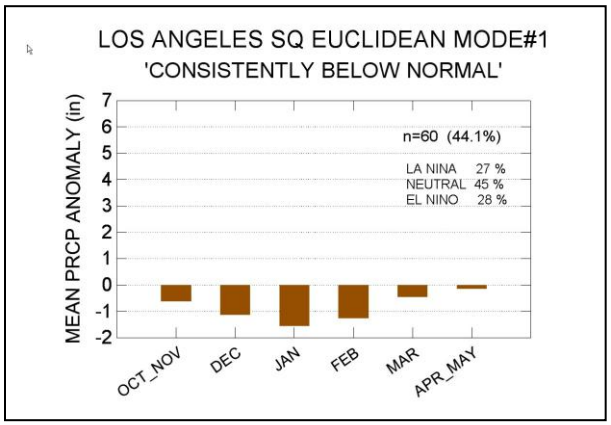

Figure 21. Monthly precipitation Anomaly Chart for Downtown Los Angeles “Consistently Below Normal” Pattern- 44.1% Incidence – Squared Euclidean Method

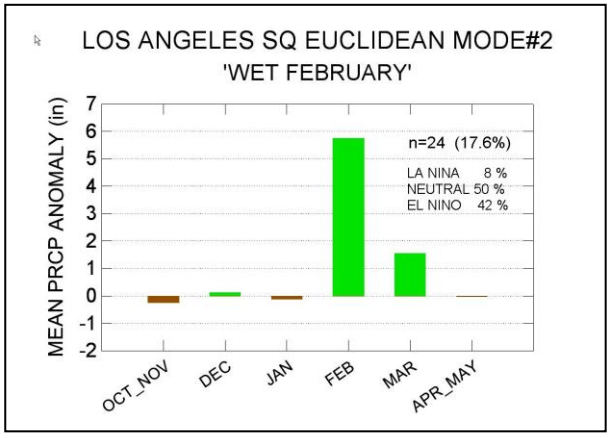

Figure 22. Monthly precipitation Anomaly Chart for Downtown Los Angeles “Wet February” Pattern- 17.6% Incidence – Squared Euclidean Method

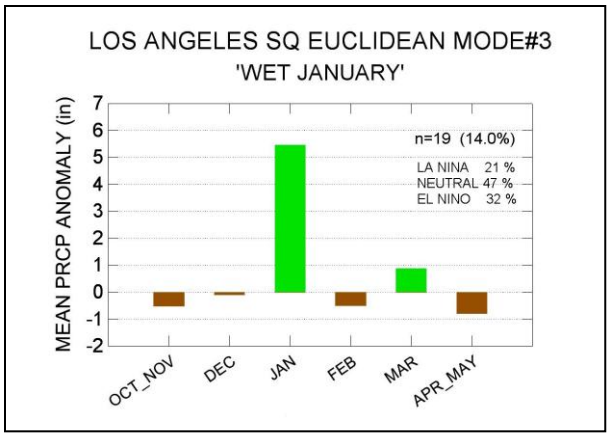

Figure 23. Monthly precipitation Anomaly Chart for Downtown Los Angeles “Wet January” Pattern- 14.0% Incidence - Squared Euclidean Method

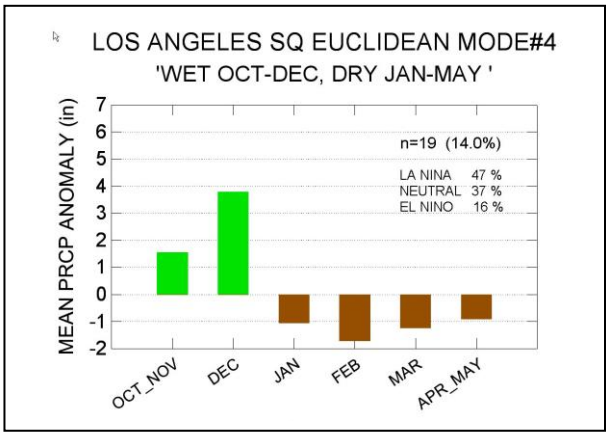

Figure 24. Monthly precipitation Anomaly Chart for Downtown Los Angeles “Wet Oct-Dec, Dry Jan-May” Pattern- 14.0% Incidence - Squared Euclidean Method

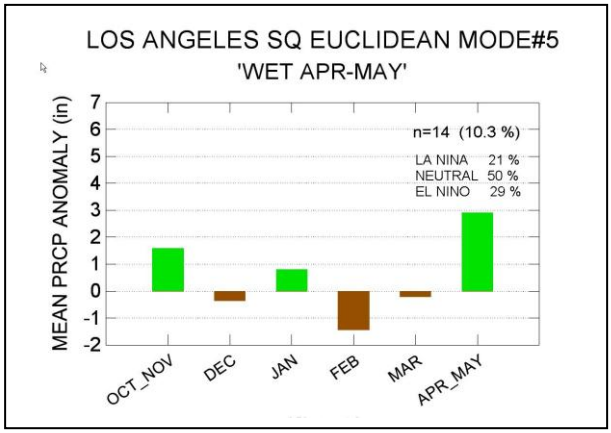

Figure 25. Monthly precipitation Anomaly Chart for Downtown Los Angeles “Wet Apr-May” Pattern- 10.3% Incidence - Squared Euclidean Method

In sum, similar to San Francisco, period-to-period idealized precipitation variability for Los Angeles consists of near or slightly below normal rainfall one period after another (See Figure 21). Aside from this, however, there is a stronger contrast exhibited by the Los Angeles data in the preferred timings of heavy rains during La Nina and El Nino episodes - La Nina's having a stronger early-season tendency (see Figure 24), El Nino's a later- season proclivity (See Figures 22 and 23).

4.3 - Downtown San Diego Results

Figure 26 is a bar graph depicting mean overall San Diego precipitation figures for the six calendar periods. Individual figures range from 1.99" for January and February each, to 1.00 for April-May. Total average Oct-May precipitation is 9.85", 32 % less than Los Angeles's (14.53") and 53% less than San Francisco's (20.98"). Median San Diego figure is 9.21"

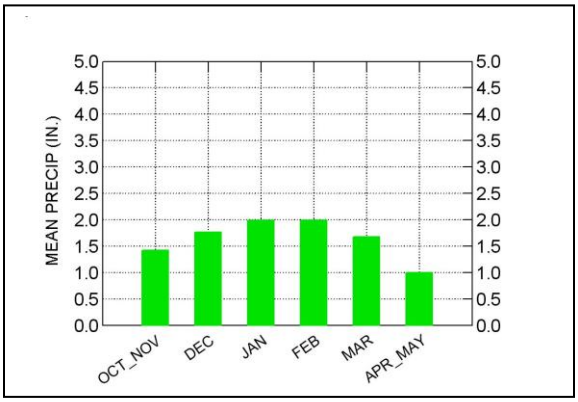

Figure 26. Mean Downtown San Diego Precipitation for October-November, December, January, February, March, and April-May calendar periods, 1877-78 to 2012-13 Period of Record.

4.3.1- Downtown San Diego Cluster generation with Euclidean Distances

The Euclidean option produced three clusters for Downtown San Diego, the same as for Los Angeles, but aside from the primary mode, displaying a similar persistent relative dryness pattern, they were more different than alike. Mean statistical cluster centroid training error was .379, compared to Los Angeles's .369 and San Francisco's .349. The "optimal" cluster number was attained at an n=3 cutoff, this coinciding with a noticeable inflection in the line trace (see Figure 27 Scree Plot below).

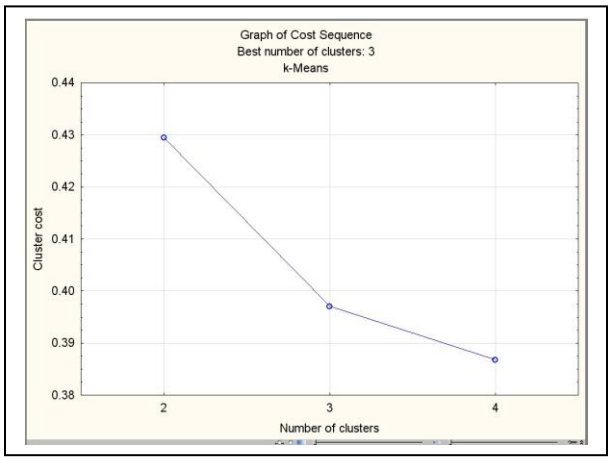

Figure 27. Scree Plot of Sequential Cluster Generation Cost Statistics for San Diego Month-to-Month Precipitation Anomaly Patterns – Euclidean Distance Metric

Figures 28 to 30 below are the rank ordered bar graphs of the mean monthly precipitation anomalies for each of the three clusters. Annotations are the same as those in the San Francisco and Los Angeles charts. Once more, overall La Nina, Neutral, and El Nino incidence was 25%, 46%, and 29%, respectively.

The primary idealized Downtown San Diego pattern (incidence: 55.9% - see Figure 28), like those for San Francisco and Los Angeles, is the familiar period-to-period succession of slightly negative anomalies – the San Diego exception being the very slightly positive one for January). Average collective Oct-May precipitation is 7.77", some 2.08" below average, and the ENSO mix is 30%, 46%, and 24%. The remaining two patterns display little similarity to Los Angeles's with contrasting configurations. The "Wet October-December, Dry January-March" pattern, ranking second in importance (incidence: 24.3% - see Figure 29) shows a wet October-November and December, the December anomaly being substantial, but with negative departures for the others.

The third ranked pattern "Wet February-March" (incidence: 19.9% - see Figure 30) has its period-to-period anomaly signs flip-flopped relative to Figure 29, exhibiting a very strong El Nino presence. Some 48% of cluster members are El Nino's compared to just 11% La Nina's. Mean Oct-May precipitation for this pattern is 13.93", more than 4" above average. While the disproportionality of El Nino vs. La Nina frequencies is striking, the Chi-Square test of overall frequency uniformity falls "short" again of major significance, rejecting the null hypothesis at the .123 level (Chi-square: 7.256, degrees of freedom: 4)

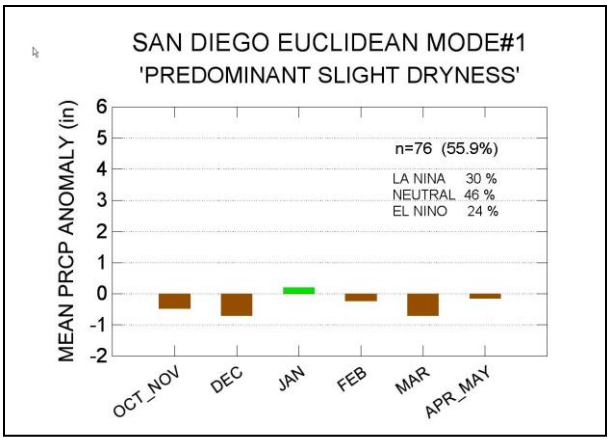

Figure 28. Monthly precipitation Anomaly Chart for Downtown San Diego "Predominant Slight Dryness" Pattern- 55.9% Incidence - Euclidean Method

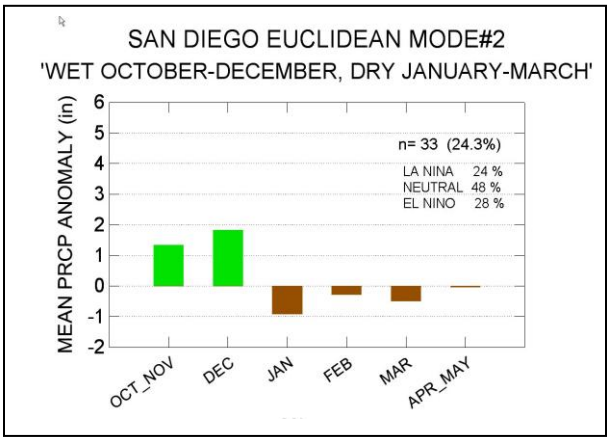

Figure 29. Monthly precipitation anomaly chart for Downtown San Diego "Wet October-December, Dry January-March" Pattern – 24.3% Incidence – Euclidean Method

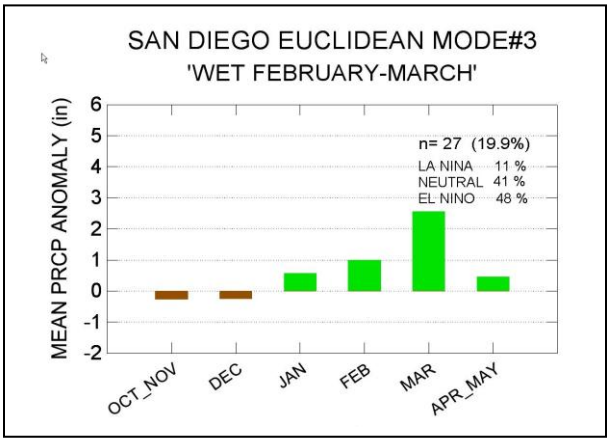

Figure 30. Monthly precipitation anomaly chart for Downtown San Diego "Wet February-March' Pattern" - 19.9% Incidence – Euclidean Method

4.3.2- Downtown San Diego Cluster generation with Squared Euclidean Distances

Repeating the results for Downtown Los Angeles, the Squared Euclidean method generated five clusters for Downtown San Diego. Mean training error was down to .136, some 36% of that for the Euclidean set (.379). Pattern-wise, significant similarities as well as differences were noted between the cluster sets of the two stations.

The Scree Plot in Figure 31 shows a major inflection point at n=5, corresponding again to the cutoff number as determined by the 5% reduction methodology.

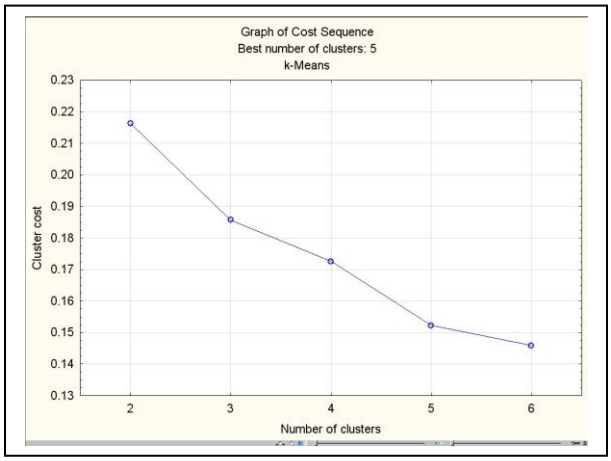

Figure 31. Scree Plot of Sequential Cluster Generation Cost Statistics for San Diego Month-to-Month Precipitation Anomaly Patterns – Squared Euclidean Method.

Figures 32 to 36 below are bar graphs of the mean precipitation anomalies for each of the five clusters, arranged in rank order of importance.

Once more, the most frequent pattern for Downtown San Diego (incidence: 49.3%) is the "Persistently Below Average" one shown in Figure 32, that with consistently negative anomalies for all six periods.

Relative to its Euclidean counterpart, the frequency is down slightly (from 55.9%), but the dry anomalies are more pronounced in negative magnitude. Average collective Oct-May precipitation is reduced to 6.98", nearly 3" below average (9.85"). The ENSO mix (34%, 41%, and 25%) shows a slight favoring of La Nina's over El Nino's (34% to 25%).

Ranking second is the "Wet Oct-Dec" pattern (incidence: 16.2% -see Figure 33). This bears some resemblance to Los Angeles' Squared Euclidean configuration in Figure 24 ("Wet Oct-Dec, Dry Jan-May" cluster), but the Figure 33 anomalies downstream of December are not all negative, nor as pronounced when they are negative. Moreover, there is no preponderance of La Nina memberships in "Wet Oct-Dec" like there is in "Wet Oct-Dec, Dry Jan-May"). In the former, Neutrals are the most frequent, with a 59% membership statistic (Figure 33); the latter has 47% La Nina's (Figure 24).

In third place is the “Wet March” pattern (incidence 14.0% - See Figure 34). The March positive anomaly stands out considerably in this chart, most of other mean anomalies, both positive and negative, being negligible. An analogous squared Euclidean pattern, emphasizing March to the exclusion of other periods is not present for Los Angeles. The San Diego “Wet March” mode displays a definitive El Nino membership preponderance with El Nino’s exceeding La Nina’s 47% to 16%, reinforcing the notion of El Nino’s being associated with late winter heavy rainfalls in Southern California. Mean Oct-May precipitation is 11.15”

Fourth in rank is the “Wet January” pattern (incidence: 13.2% - see Figure 35), having an identically named Los Angeles counterpart cluster represented in Figure 23. In the case of San Diego, Neutrals overwhelmingly predominate cluster membership (72%). Mean Oct-May precipitation is 12.77”.

In fifth place is “Wet Feb-May” (Incidence: 7.4% - See Figure 36). This is somewhat similar to the Los Angeles’ “Wet February” pattern (Figure 22), both exhibiting late-season positive anomalies encompassing multiple periods (February & March for “Wet February”, February & April-May for “Wet Feb-May”), and both have strong preponderances of El Nino cluster memberships (42% for the former, 70% for the latter). Allowing for the small sample size of the San Diego “Wet Feb-May” cluster (n=10), mean Oct-May rainfall is 15.35”, 156 percent of the 136-year average, and by far the highest figure for an individual San Diego cluster.

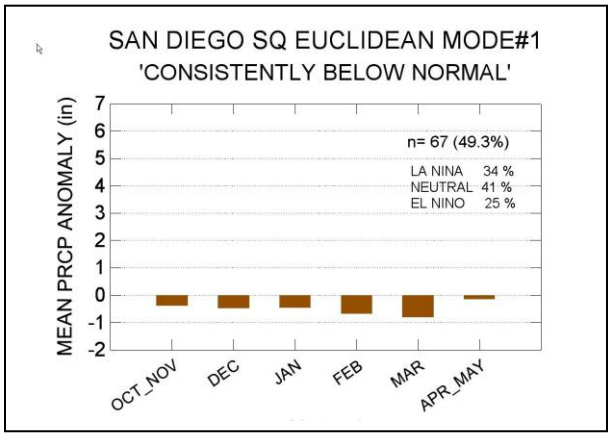

Figure 32. Monthly precipitation Anomaly Chart for Downtown San Diego “Consistently Below Normal” Pattern- 49.3% Incidence – Squared Euclidean Method

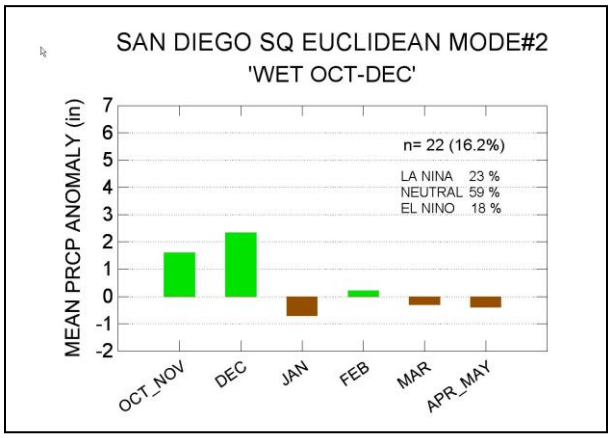

Figure 33. Monthly precipitation Anomaly Chart for Downtown San Diego “Wet Oct-Dec”” Pattern- 16.2% Incidence – Squared Euclidean Method

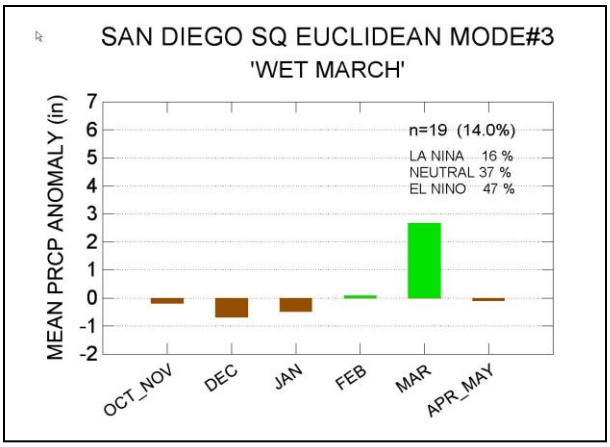

Figure 34. Monthly precipitation Anomaly Chart for Downtown San Diego “Consistently Below Normal” Pattern- 49.3% Incidence – Squared Euclidean Method

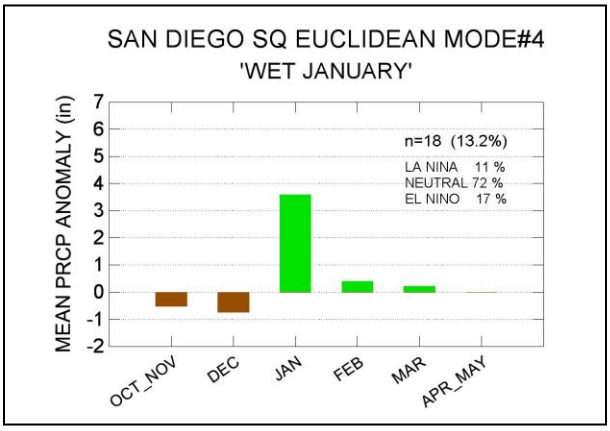

Figure 35. Monthly precipitation Anomaly Chart for Downtown San Diego “Consistently Below Normal” Pattern- 49.3% Incidence – Squared Euclidean Method

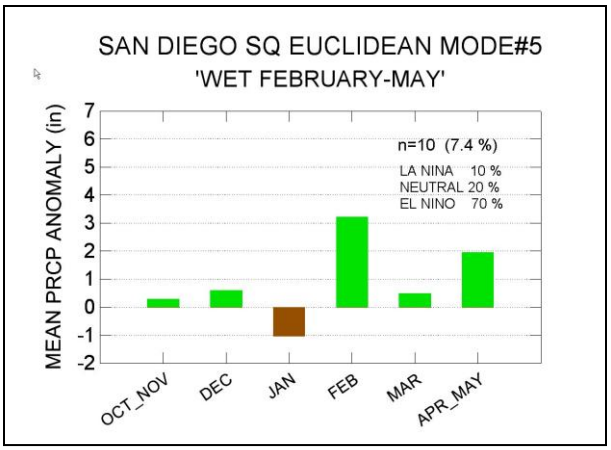

Figure 36. Monthly precipitation Anomaly Chart for Downtown San Diego "Consistently Below Normal" Pattern- 49.3% Incidence – Squared Euclidean Method

The overwhelmingly non-uniform frequencies of Neutral memberships in the "Wet-January" cluster and El Nino ones in "Wet February-May" led a null-hypothesis rejection of Chi-Square test of overall frequency uniformity at the .007 level (Chi-square: 21.151, degrees of freedom: 8).

In conclusion, not unlike San Francisco and Los Angeles, most San Diego month-to-month idealized precipitation variability consisted of near or slightly below normal rainfall one month after another, except for January in the Euclidean case (See Figures 32 and 28, respectively). For the lesser modes, like Los Angeles, San Diego showed more Euclidean vs. Squared Euclidean pattern contrasts, but although the two stations are only a little more than 100 miles apart, there were some individual one-on-one configuration contrasts, The greater tendency for El Nino's vs. La Nina's to bring heavy late-season rains was expressed in both some of the Los Angeles and San Diego charts, but the greater proclivity for La Nina's versus El Nino's to be associated with early season falls was not present in the San Diego results. Perhaps some of the discordance could be explained by the relatively low sample sizes associated with many of these lesser modes (frequently less than 20 in the Squared Euclidean cases) and the resulting sampling variability effects.

.

4.4 - Another Application – Combined Downtown Los Angeles and San Diego Results

Addressing this San Diego/Los Angeles Squared Euclidean discordance issue, and demonstrating another use of the K-Means/V-Fold methodology in the process, the two stations' data are combined into a single set with a squared Euclidean cluster analysis performed. This makes the application one in 12-dimensional space. Three clusters were generated, showing a .355 mean training error, comparable to the that for two stations' Euclidean option.

Figures 37 to 39 show the graphical results. In general, the patterns are much smoother, with both stations showing mean anomalies very much alike, period-to-period, a more sensible result for two stations so close in distance with a known similar climate. The most frequent mode "Consistently Below Average (Except Jan." – See Figure 37) makes up 59.6% of the cases. The Los Angeles departures (both positives and negatives) are a bit more pronounced on an individual basis than San Diego's - to be expected as Los Angeles in wetter climatologically. To reiterate, La Nina's, Neutral's, and El Nino's made up 25%, 46%, and 29%, respectively, of the 136 rain years, and the mix for this mode, as indicated by the annotations in Figure 37 is 27%, 46%, and 27% - scarcely different.

The second ranking mode, "Wet Early Season/Dry Jan-Mar" (incidence 20.6%- see Figure 38) displays a run of positive mean anomalies for Oct-Nov and December (both stations), a run of negatives for January-March (both stations), and a positive one for April-May (both stations). The mix of ENSO designations favors La Nina's over El Nino's by 32% to 21%, but since Neutrals also make up 46%, a suitable overall generalization would be that rainfall character for both Los Angeles and San Diego tends to progress from an early season mean wet regime to a drier one for all ENSO designations, slightly more so on a relative basis for La Nina's compared to El Nino's.

The third mode, "Wet February-March" (incidence 19.9% - see Figure 39) shows highly positive February and March anomalies for both stations, those for the other periods essentially negligible. In this "Wet February-March" case, the El Nino incidence is much higher than La Nina's (44% to 11%).

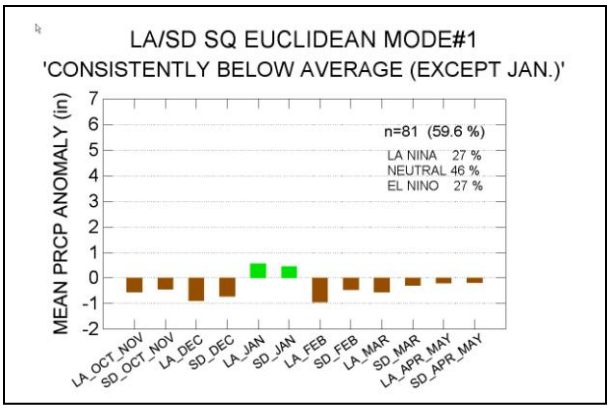

Figure 37. Monthly precipitation Anomaly Chart for Downtown San Diego "Consistently Below Normal" Pattern- 49.3% Incidence – Squared Euclidean Method

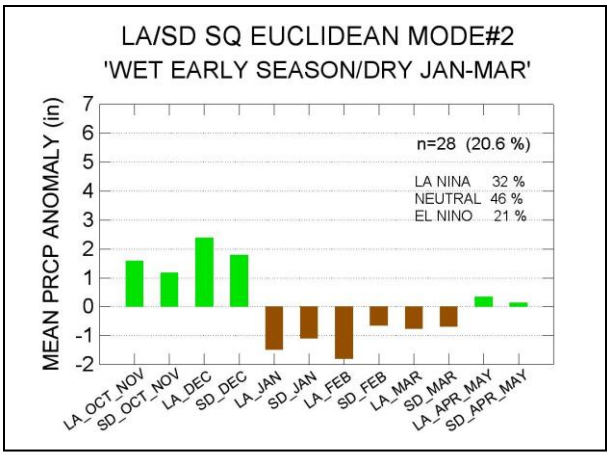

Figure 38. Monthly precipitation Anomaly Chart for Downtown San Diego "Consistently Below Normal" Pattern- 49.3% Incidence – Squared Euclidean Method

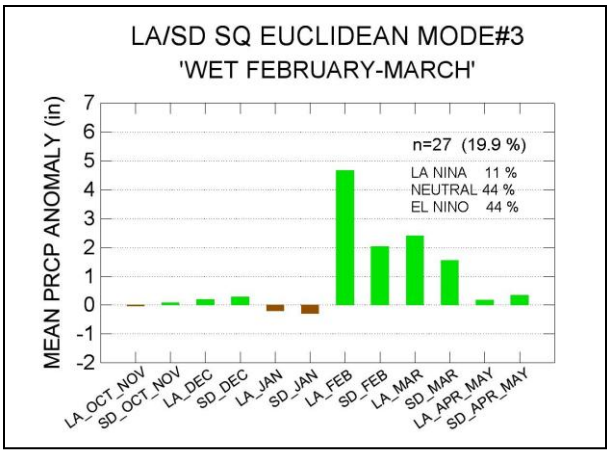

Figure 39. Monthly precipitation Anomaly Chart for Downtown San Diego "Consistently Below Normal" Pattern- 49.3% Incidence – Squared Euclidean Method

5. SUMMARY AND CONCLUSION

Utilizing the K-Means/V-Fold Cross-Validation clustering methodology, the foregoing investigated the existence and relative frequencies of month-to-month precipitation anomaly modes for the San Francisco, Los Angeles, and San Diego October-May rain year histories, each covering the 1877-78 through 2012-13 period of record. Two distance metrics, the Euclidean and Squared Euclidean, were applied to create the clusters. Results' were also examined in relation to ENSO designation ("La Nina", "Neutral", and "El Nino").

For each of the three stations, and with either method, The primary pattern (with one minor exception) was one of slightly negative (dry) anomalies for the six periods (October-November and April-May combined into one), at least in part due to the fact that precipitation series are generally skewed to the high values (i.e., the lower median values are more representative central tendency statistics than means). These primary modes comprised between 40 to 66% of the cases, and a generalization might be made that the percentages characterized the frequency of years that were relatively without wet precipitation extremes. .

For San Francisco, five and six clusters, respectively, were generated by the two metrics, with only minor pattern distinctions realized between them. The primary mode had a 40% frequency for each method. The rest of the patterns, save for one, reflected single-period "spikes" in positive (wet) anomalies, the exception being a wet contiguous period (January-February) signal which also was the only one clearly associated with El Nino's. There were also two clusters that showed wet to dry anomaly progressions covering three periods' extent (November- December to February, and December to March).

Los Angeles and San Diego, hundreds of miles to the south of San Francisco, and only 120 miles apart themselves, had five and three clusters each, respectively, generated by the Euclidean and Squared Euclidean methods. The primary mode (again, consistently negative anomalies, period-to-period) encompassed 67% and 44%, respectively, of the Los Angeles cases, 56% and 49%, respectively, of the San Diego ones. For both stations, the lesser modes showed more contiguous periods' positive (wet) anomaly successions, along with indication of an El Nino late-season heavy precipitation signal, but there were also some pattern dissimilarities.

Combining the Los Angeles and San Diego histories into one data set and performing a Squared Euclidean Cluster Analysis resulted in three clusters of quite smooth and distinct patterns; the primary mode (60% incidence) showing the usual period-to-period slightly negative anomalies (January the exception), followed by two lesser patterns roughly opposite in character, reflecting early vs. later-season presence of contiguous period positive anomalies (each about 20% incidence). The late-season wet pattern showed a clear inclination of El Nino occurrences over La Nina's, the early-season wet mode with a lesser La Nina over El Nino excess.

In conclusion, the K-Means/V-Fold methodology provided some useful insights into the deeper-level climatological period-to-period precipitation variability of San Francisco, Los Angeles, and San Diego. Results carried some weight as 136 seasons were analyzed. The V-Fold cross-validation algorithm removed some of the subjectivity involved in selecting the "right" number of K-Means clusters, but the cutoff threshold magnitude is still open to user-selection, as is the type of distance metric.

6. REFERENCES

Fisk, C.J., 2013: "Identification of California Inter-Climate Division Modes of Seasonal Precipitation Variability (1895-96 through 2011-12 Seasons)", 27th Conference on Hydrology, American Meteorological Society, 6-8 January 2013, Austin, TX
https://ams.confex.com/ams/93Annual/webprogram/Paper219161.html

Fisk, C.J., 2013: "Identification of Midnight-to-Midnight Hourly Wind Pattern Modes Utilizing the V-Fold Cross-Validation Algorithm Applied to K-Means Clustering, 20th Conference on Applied Climatology, American Meteorological Society, 5-10 January 2013, Austin, TX
https://ams.confex.com/ams/93Annual/webprogram/Paper215874.html

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

http://en.wikipedia.org/wiki/Euclidean_distance

http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears_1877-present.shtml

http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml