IDENTIFICATION AND COMPARISON OF CALIFORNIA CLIMATE DIVISIONS' RAIN YEAR VARIABILITY MODES UTILIZING K-MEANS CLUSTERING ANALYSIS INTEGRATED WITH THE V-FOLD CROSS VALIDATION ALGORITHM (1895-96 THROUGH 2011-2012 SEASONS)

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

1. INTRODUCTION

The state of California, nearly 800 miles long and 250 miles wide, is divided into seven NOAA NCDC Climate Divisions. Based on areal-averaging techniques, single-valued month-to-month precipitation statistics have been compiled, division by division, since 1895. With such a huge distance between the northern to southern borders, and the great topographical variation, it would seem inevitable that the character of rain year (July-June) relative precipitation anomalies may not be consistent, division-to-division, from one year to the next. The degree and nature of these contrasts, and possible relationships to such phenomena as El Nino and La Nina should make for interesting study.

To this end, the existence and relative frequencies of California Climate Division rain year variability patterns (or "modes") is investigated using K-Means clustering analysis integrated with the V-Fold Cross Validation Algorithm. Period of record is 1895-96 thru 2011-12, a 117-year history.

One "sticking-point" associated with traditional K-means is that the number of clusters has to be guessed at in advance, the ultimate choice of how many there "are" requiring trial-and-error iterations combined with subjective judgment. Recent advances in the data-mining field, however, have resulted in adaptation of the V-Fold Cross-Validation algorithm, a "training-sample" type procedure which incorporated into K-Means allows for a more objective determination of the "right" number of clusters.

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).

As applied to K-means, the V-fold cross-validation scheme 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 the 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. Preliminary to the analyses, the Climate Division data were normalized, an automatic software feature, to reduce them to a common scale (between 0.0 and 1.0) and lessen 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 the present study, the 5 percent default distance improvement cutoff threshold is retained with the Squared Euclidean distance metric selected (default: Euclidean).

3. DATA

The raw data were downloaded from the NCDC online website, U.S. National/State/Divisional Data link, which has the complete history back to 1895. Issues have been raised in recent years about biases in the Climate Division data set, in particular concerning temperature (averaging methods, time-of observation, instrument placements, heat-island effects, etc.,). It seems less likely that these factors would have an impact on precipitation, and since no official caveats appear in the NCDC website, the regional precipitation data are analyzed as reported.

Figure 1 is a map of the California Climate Divisions. Their full titles, in numeric ordering are, 1.) "North Coast Drainage, 2.) "Sacramento Drainage", 3.) "Northeast Interior Basins", 4.) "Central Coast Drainage", 5.) "San Joaquin Drainage", 6.) "South Coast Drainage", and 7.) "Southeast Desert Basin". In the results' discussions below, the titles appear in shortened fashion, with the "Drainage" and "Basin" portions removed.

---

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

Figure 1 – Map of California Climate Divisions – from NOAA NCDC.

4. RESULTS

The K-Means/V-Fold algorithm produced six clusters, the normalized means, by cluster and division, depicted in Figure 2. Higher (lower) normalized means within a cluster for a given division indicated, of course, a wetter(drier) year, but only in relative terms, as the normalized scale (0.0 to 1.0) was identical for each.

4.1. - Normalized Precipitation Patterns by Cluster and Division

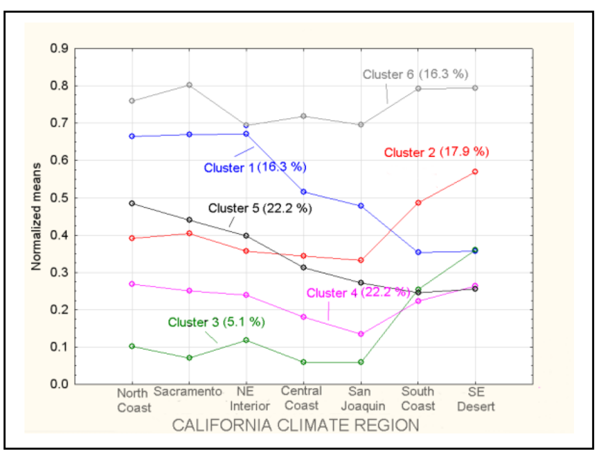

Figure 2 – Plot of Mean Normalized July-June Precipitation Patterns, by Cluster and Division - 1895-6 thru 2011-12 Seasons

From Figure 2, no clear predominance in high frequency exists for any individual cluster (range: 22.2 % to 16.6%), but one (Cluster 3) is significantly lower in percent occurrence (5.1%).

Maximum incidence (22.2%) is a tie between Clusters 4 and 5. Cluster 4 depicts a drought-year pattern for all seven divisions, the normalized means all in the .20's, except those for the "Central Coast" and "San Joaquin", which dip into the .10's. Cluster 5, in contrast, shows a near-normal to dry gradient from north to south - the northernmost divisions ("North Coast", 'Sacramento" and "Northeast Basin") approximately "normal" or slightly below (0.50 to 0.40 means), the other four more categorically dry, with figures of 0.30 for "Central Coast" 0.28 for "San Joaquin", 0.25 for "Southeast Desert", and 0.24 for "South Coast".

In third place is Cluster 2 (17.9%). This presents a dry to wet pattern transition between the northernmost five divisions and the southernmost two. For the north group, normalized means are at moderately dry levels (around .40 to the low .30's), but they elevate to around 0.50 for "South Coast" and to near 0.60 for "Southeast Desert".

Tied for fourth place (16.6%) are Clusters 1 and 6. Cluster 1 shows another north to south transition: very wet for the northernmost divisions (normalized means approaching 0.70 for "North Coast", "Sacramento", and "Northeast Basin"), near normal for "Central Coast", and "San Joaquin" (normalized means near 0.50), and relatively dry for "South Coast" and Southeast Desert (figures around 0.35). Cluster 6 displays the wettest pattern of all, normalized means for each division between 0.70 and 0.80, the highest magnitudes encountered for any cluster.

In sixth place with just a 5.1 % frequency (fortunately) is Cluster 3. This depicts intense drought for all but the two southernmost divisions. For the five to the north, the normalized statistics run from .05 to .10, the lowest normalized magnitudes encountered for any cluster. For "South Coast" the figure increases to a still quite "dry" 0.25 magnitude, but a bit higher to 0.35 for "Southeast Desert".

In sum, three of the curves depict the same character for all divisions: cluster 6 (all wet), cluster 4 (all dry), and cluster 3 (intense drought for northernmost five, dry for southernmost two). The other three indicate varying transitions (wetter to drier and vice-versa) from north to south. The contrasts of "South Coast" and "Southeast Desert" with the other five also seem more pronounced in certain instances, large dissimilarities apparent for three of the six clusters (clusters 1, 2, and 3).

Figure 3 below is a (zoomable) tabular summary of the Climate Regions' actual rain-year precipitation totals, by season, by division, along with their cluster assignments ("MODE"), statistical distances to cluster centroids ("DISTANCE"), and ENSO designations ("CLASS$"). The 117-year mean seasonal precipitation statistics for the seven climate divisions are as follows: "North Coast" (41.24"), "Sacramento" (37.35"), "Northeast Interior" (21.05"), "Central Coast" (21.16"), "San Joaquin" (20.06"), "South Coast" (16.91"), and Southeast Desert (7.64").

LEAD YEAR	NORTH COAST	SACRA-MENTO	NE INTERIOR	CENTRAL COAST	SAN JOAQUIN	SOUTH COAST	SE DESERT	MODE	DISTANCE	CLASS$
1895	43.13	38.43	22.46	20.95	20.41	15.01	6.79	5	0.039	NEUTRAL
1896	42.30	36.59	20.19	18.57	17.61	15.77	8.23	2	0.051	WARM
1897	23.23	18.26	12.30	8.36	11.08	8.67	4.59	3	0.047	NEUTRAL
1898	29.32	23.41	15.71	12.59	15.10	11.02	5.46	4	0.010	NEUTRAL
1899	44.03	40.12	20.36	20.51	17.98	11.54	4.81	5	0.013	WARM
1900	43.39	40.18	23.57	23.28	23.49	20.09	8.67	2	0.087	NEUTRAL
1901	39.23	35.00	20.13	18.77	19.24	12.18	5.93	5	0.007	NEUTRAL
1902	41.30	36.61	23.22	21.03	20.75	15.42	6.46	5	0.041	WARM
1903	53.51	44.73	24.67	22.80	22.88	12.15	4.94	1	0.059	COLD
1904	43.98	39.46	17.56	23.17	21.05	30.18	15.30	2	0.324	NEUTRAL
1905	48.16	49.29	27.56	31.84	31.29	29.00	11.63	6	0.042	WARM
1906	56.62	55.03	32.46	33.52	33.68	29.79	13.18	6	0.032	COLD
1907	37.02	34.34	18.46	19.23	17.05	13.71	6.13	5	0.018	COLD
1908	65.06	60.96	28.29	35.64	27.87	26.62	10.69	6	0.094	NEUTRAL
1909	49.63	44.24	23.68	24.76	22.07	16.77	8.10	1	0.043	COLD
1910	50.11	52.02	27.80	32.14	29.56	27.41	11.69	6	0.026	NEUTRAL
1911	40.82	31.89	14.81	17.39	15.64	13.77	5.69	5	0.049	WARM
1912	40.28	34.22	18.04	17.35	15.58	12.91	6.67	5	0.025	NEUTRAL
1913	61.83	62.81	35.69	39.00	34.65	30.47	13.35	6	0.157	WARM
1914	58.65	57.10	21.82	33.92	26.31	26.76	11.15	6	0.101	NEUTRAL
1915	60.59	56.08	27.98	31.99	27.60	28.43	11.67	6	0.026	NEUTRAL
1916	41.66	38.10	19.15	21.12	18.13	18.11	9.04	2	0.024	COLD
1917	33.28	27.85	15.30	15.11	15.18	13.16	5.84	4	0.009	COLD
1918	40.92	34.83	18.00	19.53	18.43	17.81	8.98	2	0.016	WARM
1919	27.15	26.00	14.06	16.00	16.68	18.33	8.88	2	0.119	NEUTRAL
1920	60.92	55.46	28.52	29.44	24.77	16.38	7.17	1	0.076	NEUTRAL
1921	44.43	42.86	21.61	26.00	23.99	21.80	10.07	2	0.117	NEUTRAL
1922	42.50	39.76	22.26	22.39	20.28	15.31	7.10	5	0.053	NEUTRAL
1923	16.87	13.97	8.19	8.94	10.04	11.51	6.21	3	0.038	WARM
1924	47.84	41.98	22.63	22.27	21.01	13.31	5.48	5	0.058	COLD
1925	36.48	33.25	16.47	20.23	16.58	18.50	8.76	2	0.024	WARM
1926	57.41	50.12	25.16	27.05	23.45	20.51	8.71	1	0.073	NEUTRAL
1927	43.32	36.68	19.37	18.99	17.57	11.21	5.08	5	0.004	NEUTRAL
1928	31.10	27.97	14.62	15.50	14.71	11.61	4.86	4	0.005	COLD
1929	33.86	31.85	16.44	18.56	17.38	15.72	7.26	2	0.069	NEUTRAL
1930	22.37	19.62	12.77	12.13	13.65	16.15	8.56	3	0.040	NEUTRAL
1931	31.89	33.61	20.92	19.36	21.94	21.01	12.52	2	0.069	NEUTRAL
1932	29.44	22.48	12.34	12.57	14.04	13.70	6.55	4	0.031	WARM
1933	26.00	26.78	14.31	11.92	11.94	11.86	5.26	4	0.020	NEUTRAL
1934	36.29	39.44	19.70	20.30	23.77	21.38	10.13	2	0.036	NEUTRAL
1935	37.66	39.65	21.74	18.40	22.50	14.55	6.96	5	0.059	NEUTRAL
1936	31.09	33.48	18.26	22.85	23.80	28.23	11.90	2	0.139	NEUTRAL
1937	57.34	54.17	30.65	27.73	29.37	23.82	11.80	6	0.066	NEUTRAL
1938	26.27	20.69	12.55	12.54	14.37	15.01	8.13	3	0.039	COLD
1939	46.21	44.96	22.99	25.40	23.63	19.23	10.11	1	0.115	WARM
1940	56.81	53.55	21.48	34.10	27.38	36.05	15.45	6	0.158	WARM
1941	48.82	46.48	23.70	25.22	22.60	14.65	7.77	1	0.034	NEUTRAL
1942	43.76	38.02	24.11	21.11	24.09	22.39	11.08	2	0.097	NEUTRAL
1943	29.57	26.57	14.35	17.30	16.78	20.30	10.78	2	0.090	NEUTRAL
1944	39.13	33.07	19.71	19.88	21.45	16.01	8.04	2	0.036	NEUTRAL
1945	40.65	32.69	18.82	17.76	18.95	14.94	8.26	2	0.046	NEUTRAL
1946	28.30	27.45	16.30	13.27	15.14	14.72	7.69	4	0.043	WARM
1947	38.15	35.61	14.09	15.34	17.47	9.64	4.80	5	0.066	NEUTRAL
1948	31.26	27.49	13.59	16.02	14.54	12.02	6.36	4	0.011	NEUTRAL
1949	34.87	30.88	18.45	16.96	16.21	12.18	3.77	5	0.052	COLD
1950	48.00	43.55	22.79	21.97	22.61	9.56	3.53	5	0.107	COLD
1951	51.00	50.71	26.89	30.55	27.21	27.24	11.88	6	0.042	WARM
1952	48.39	39.00	19.98	18.87	16.72	12.24	5.14	5	0.023	NEUTRAL
1953	42.72	33.59	14.48	16.03	16.66	15.21	6.16	5	0.058	WARM
1954	30.00	26.25	13.25	16.48	15.84	12.72	5.62	4	0.014	COLD
1955	54.63	49.75	31.35	27.44	24.95	13.88	5.50	1	0.034	COLD
1956	34.42	30.50	19.90	16.41	15.85	12.95	5.63	5	0.044	NEUTRAL
1957	60.97	52.40	27.24	34.79	27.99	26.63	10.92	6	0.047	WARM
1958	30.16	25.73	16.66	12.42	11.57	8.54	4.57	4	0.027	WARM
1959	34.27	30.62	16.25	15.71	14.08	11.56	5.68	4	0.014	NEUTRAL
1960	37.78	28.38	13.95	12.79	11.85	6.33	3.80	4	0.074	NEUTRAL
1961	33.20	33.35	22.00	19.00	18.77	18.59	7.95	2	0.037	NEUTRAL
1962	47.70	46.79	30.19	25.46	20.20	9.40	4.57	1	0.093	NEUTRAL
1963	31.61	26.07	18.19	13.29	13.46	12.72	6.49	4	0.018	WARM
1964	46.49	44.18	26.64	20.63	21.47	13.75	6.00	1	0.053	COLD
1965	35.64	28.21	17.76	15.23	14.07	18.98	9.24	2	0.076	WARM
1966	46.81	47.03	29.26	28.67	26.28	21.20	9.70	1	0.099	NEUTRAL
1967	32.28	28.10	17.36	14.37	13.04	13.11	7.14	4	0.025	NEUTRAL
1968	51.51	49.63	31.49	29.74	32.74	28.04	13.37	6	0.045	WARM
1969	45.04	26.53	23.43	17.79	17.34	9.92	4.92	5	0.062	WARM
1970	48.90	37.88	24.18	18.40	16.79	12.98	5.62	5	0.043	COLD
1971	37.82	23.79	17.82	10.14	10.86	8.45	5.03	4	0.069	COLD
1972	41.54	41.35	21.08	30.24	23.82	20.36	8.16	1	0.152	WARM
1973	62.28	48.89	23.41	25.76	20.35	13.58	5.21	1	0.106	COLD
1974	43.15	35.95	22.00	18.54	18.54	15.31	5.27	5	0.015	COLD
1975	26.65	16.91	11.63	8.52	10.75	10.19	5.22	3	0.031	COLD
1976	17.36	15.55	12.62	11.76	10.56	14.22	8.04	3	0.026	WARM
1977	52.67	49.34	22.71	31.35	30.70	34.48	15.17	6	0.117	WARM
1978	31.08	31.03	16.01	19.35	20.18	21.66	9.49	2	0.030	NEUTRAL
1979	46.53	44.42	29.09	25.41	26.01	28.61	12.15	6	0.142	NEUTRAL
1980	29.60	27.06	16.21	15.11	14.92	11.01	4.43	4	0.009	NEUTRAL
1981	60.07	56.96	37.89	31.96	29.50	15.96	6.53	1	0.245	NEUTRAL
1982	68.27	63.85	36.98	40.47	39.03	32.00	12.58	6	0.320	WARM
1983	48.25	42.02	28.74	18.59	20.64	11.03	6.41	1	0.098	COLD
1984	33.16	26.39	17.32	17.22	15.98	13.31	8.55	4	0.073	COLD
1985	49.06	49.56	34.33	26.04	27.94	18.85	7.35	1	0.079	NEUTRAL
1986	30.07	23.79	13.18	13.12	13.06	9.86	5.01	4	0.015	WARM
1987	31.64	27.83	12.98	14.57	14.56	15.57	7.48	4	0.045	WARM
1988	34.54	33.07	22.63	14.45	14.23	9.40	4.06	5	0.083	COLD
1989	30.19	29.72	18.97	13.41	14.69	8.96	3.06	4	0.061	NEUTRAL
1990	26.47	26.88	19.43	15.98	15.48	15.82	7.06	4	0.062	NEUTRAL
1991	29.02	27.49	16.80	18.95	15.62	19.65	10.56	2	0.072	WARM
1992	50.07	49.41	32.15	27.55	28.18	31.50	13.99	6	0.078	NEUTRAL
1993	26.08	23.81	12.51	14.97	13.76	12.18	4.27	4	0.028	NEUTRAL
1994	60.08	62.95	36.87	33.21	33.32	28.42	10.81	6	0.144	WARM
1995	48.08	42.67	27.49	24.57	21.75	11.57	3.96	1	0.080	COLD
1996	47.16	45.00	31.40	24.17	23.40	13.69	4.42	1	0.057	NEUTRAL
1997	64.30	60.52	26.26	42.55	34.21	34.06	13.76	6	0.176	WARM
1998	44.91	35.58	23.01	19.73	16.36	9.65	4.18	5	0.033	COLD
1999	38.29	37.94	19.42	23.98	19.70	11.59	4.84	5	0.033	COLD
2000	24.61	24.66	11.11	18.60	15.73	15.39	5.95	4	0.069	COLD
2001	39.47	32.82	16.66	18.57	16.51	5.32	2.12	5	0.129	NEUTRAL
2002	48.60	39.38	18.95	22.28	18.16	17.26	6.29	5	0.062	WARM
2003	38.93	33.49	16.70	18.15	15.02	9.30	6.03	5	0.040	NEUTRAL
2004	42.31	41.41	23.79	30.43	27.08	31.67	15.19	6	0.219	WARM
2005	57.25	51.69	34.25	29.15	26.62	13.55	6.05	1	0.090	COLD
2006	30.29	22.58	12.20	12.68	11.01	4.64	1.91	4	0.149	WARM
2007	32.58	26.76	16.93	17.59	14.09	13.40	5.49	4	0.016	COLD
2008	28.12	30.76	16.23	17.08	15.63	10.12	4.96	4	0.024	NEUTRAL
2009	41.20	36.64	18.83	24.62	20.80	18.61	6.99	2	0.068	WARM
2010	44.15	47.39	29.26	26.56	27.81	22.75	7.52	1	0.091	COLD
2011	31.70	26.30	14.73	14.67	11.89	10.68	4.04	4	0.018	COLD

Figure 3 – Time Series of California Climate Division Precipitation, by Season and Division, with Cluster Assignments, Distance to Centroids, and ENSO designations (1895-96 through 2011-12 seasons).

4.2. - *Cluster Incidence related to ENSO (El Nino, Neutral, and La Nina occurrences)*

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 1950-2011. Those years that overlap (1950-2011) are given the designations assigned by the latter list.

From the CLASS$ column, 34 “El Nino's” or “WARM”, 53 “Neutrals” and 30 “La Nina's” or “COLD” were tabulated.

4.2.1. *El Nino's* **-**

For the “El Nino's” 11 of the 34 or 32.4% are affiliated with mode 6 or the decidedly wet pattern (all regions with normalized values >.70), not a completely surprising result based on recent memory of the two great El Nino's of the late 20th century (1982-83 and 1997-98) which were wet throughout California (add also the more recent 2004-05 El Nino season to that group).

Ranking second was mode 4, (n=7 or 20.6%) depicting the dry pattern throughout the state, especially the middle regions (“Central Coast” and “San Joaquin”). The most recent season in this category was 2006-07, in which Downtown Los Angeles, in particular, set a record for least July-June rainfall in history. Other relatively recent El Nino seasons assigned to mode 4 are 1987-88, 1986-87, and 1963-64.

Tying for third place, six cases (or 17.6% each) were modes 2 and 5. Mode 2, characterized by seasonable to wetter than normal falls for the southernmost two regions (“South Coast” and “Southeast Desert”), but slightly drier than normal ones farther north, has the 2009-10 and 1991-92 as recent cases. Mode 5, exhibiting a north to south trend to greater relative dryness, includes 2002-03 and 1969-70 as recent examples.

Tying for last place, Modes 1 and 3 had just two cases (5.9%) each. Mode 1, showing relatively heavy falls in the far north (normalized values for “North Coast”, “Sacramento”, and “Northeast Interior” just under .70), below normal ones in the far south (corresponding values for “South Coast” and Southeast Desert about .35) has 1939-40 and 1972-73 as cases. Mode 3, reflecting severe drought over most of the State, includes 1923-24 and 1976-77. The latter season set a record low rain-year total for Downtown San Francisco.

From these results, on a State-wide basis one might conclude that if it is known that an El Nino is imminent, not taking into account strength, there is a 32% chance that it will bring widespread above average rains (mode 6), about a 27% probability of dry to droughty conditions throughout (modes 3 and 4), and about a 41% chance of mixed character falls division to division (modes 1, 2, and 5).

4.2.2. *Neutrals* -

"Neutrals", with a total count of 53, displayed a dissimilar distribution of pattern preferences compared to the El Nino's. The three most frequent were mode 2 (n=14 or 26.4%), and modes 4 and 5 (n=11 each, or 20.8%). Mode 1 (wet to relatively dry transition) had eight cases (most recent example: 1996-97), Mode 6 ("all wet") had seven, the most recent case being 1992-93, and mode 3 (intense drought for the northernmost five, dry for the southernmost two regions) had two cases, 1897-98 and 1930-31.

Using the same summarization scheme, on a state-wide basis, if "Neutral' conditions are anticipated, there is about a 13% chance of widespread decidedly above average rains (mode 6), a 25% probability of dry to droughty conditions (modes 3 and 4), and a 62% likelihood of mixed anomaly character results (modes 1, 2, and 5).

4.2.3. *La Nina's* -

La Nina's, with a 30 total count, had all but 4 assigned to three modes. Nine cases each (30.0%) were assigned to mode 1 and to 5 respectively, and eight to mode 4 (26.7%). Mode 1, as previously described, portrays a very wet to slightly drier than normal transition from north to south, Cluster 5 exhibiting a drying trend also, but from levels modestly below normal to those more droughty, and Mode 4 a dry pattern throughout, especially for "Central Coast" and "San Joaquin".

Two were assigned to mode 3 (severe drought over most of the state; cases: 1938-39 and 1975-76), one to mode 2 (dry to wet transition between the northernmost five zones and the southern-most two; case: 1916-17) and one to mode 1 (wet throughout; case: 1903-04). The recent 2010-2011 season, which experienced torrential rains over much of the State in December, is assigned to mode 1, the 2011-12 season just passed to mode 4.

Summarizing the La Nina results, given an expected episode, the probability of a wet pattern state-wide is just 3% (mode 1), for dry to droughty conditions (modes 3 and 4), the likelihood is 33%, and for a mixed character of relative anomalies it is 64%.

Comparing the probability distributions among the three types, it appears that "La Nina's" are more similar to "Neutrals" than "El Nino's" are to "Neutrals", and without a great deal of exaggeration, "La Nina's" might even be considered as "Dry Neutrals", especially with regards to the similar "dry to droughty" and "mixed anomalies" frequencies.

4.3. - *Identification of the Most Extreme Patterns*

An additional interesting side-application of a cluster analysis of this kind is identification of extreme patterns, utilizing the statistical distance information. In this application, to reiterate, cluster memberships of (normalized) individual rain-year observations were determined by comparing their squared Euclidean distances to each of six different 7-D cluster centroids; the cluster associated with the least distance would be that to which the observation was assigned.

Large statistical distances within a cluster would reflect individual cases that displayed unusually amplified patterns peculiar to the pattern described by that cluster, and extending this to the entire data set, ranking all the distances irrespective of mode, could be a means of assessing in a relative way the most extreme of these extremes.

Based on the STATISTICA output, Figure 4 is a histogram of the distances-to-centroid statistics for each of the 117 individual rain seasons' 7-D observations.

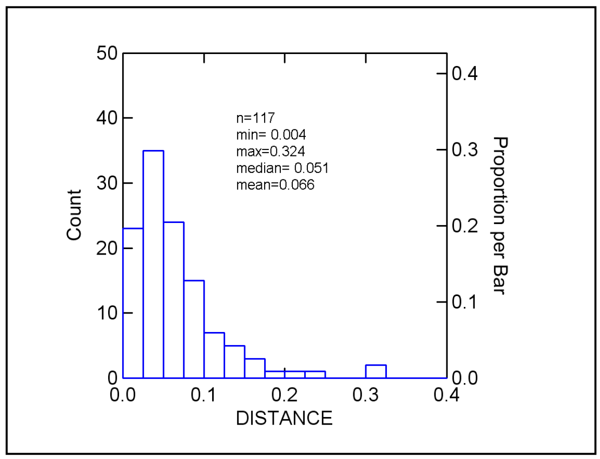

Figure 4 – Histogram of Statistical Distances of Individual Observations to Respective (Six) Centroids, California Climate Division Precipitation Data

In general, the distances from the parent centroids are quite close to zero (mean: 0.066), indication that the K-Means/V-Fold algorithm performed well in delineated clusters and assigning individual observations to them. Nonetheless, there are a few relative outliers, the most of extreme of which are of interest here.

Ranking highest is the 0.324 distance generated for the 1904-05 season, a "Neutral" that was assigned to Mode 2. As previously discussed, this mode describes a shift from a slightly dry character for the northernmost five zones to a wet one for the southernmost two. In 1904-05 the amplification to wetter than normal was unusually pronounced, the precipitation figure for "South Coast" (30.18") the sixth highest in the record (corresponding to a 0.81 normalized statistic), and that for "Southeast Desert" (15.30") the second highest (normalized statistic: 0.99).

Ranking second is the Great El Nino episode of

1982-1983 (distance: 0.320; Mode 6 classification). The high distance magnitude in this case can be attributed to the sheer heavy amounts that were received that season, record high areal-averaged totals noted for each of the five northernmost divisions. Departing only slightly from these rankings, “South Coast” received the 4rd highest amount in its history (32.00”; 0.87 normalized statistic), and “Southeast Desert” the 9th most (12.58”; 0.79 normalized statistic).

It should of course be repeated that these two extremes, while interesting in their own right, are based on distances to their own cluster centroids, not the *overall* data set centroid (presumably a 7-D array with 0.5 normalized statistics, each). Distances calculated to the latter might or might not show identical results.

5. SUMMARY

Utilizing the clustering tool K-Means, accompanied by the V-fold cross validation algorithm, the existence and characterization of seasonal (July-June total) precipitation modes were explored, collectively, for the seven California climate divisions, accessing the complete 1895-96 to 2011-12 period of record. The inputs were normalized, areal-averaged total precipitation statistics season-by-season, and division-by division.

Results resolved six clusters (or “modes”), characterizing a variety of anomaly patterns across divisions, mostly on a north-to-south basis. Two of the modes (3 and 4), encompassing 27.3% (or n=32) of the seasons, reflected a statewide pattern of below normal to drought conditions (less than 0.50 normalized statistics for all seven divisions). Mode 6, in contrast, comprising 16.3% (or n=19), reflected a wet pattern throughout. Eleven of these were associated with El Nino’s, seven with Neutrals, and one with a La Nina. The other four modes (1, 2, and 5), covering 56.4% (or n=66) of the seasons, displayed mixed anomaly patterns across the divisions. This last result confirms the notion in the introduction that division to division precipitation anomalies in California likely have an inclination (albeit slight) to be varied rather than uniform, and when they are uniform tend more to reflect dry or drought conditions.

In addition, making use of ranked individual seasons’ statistical distances from cluster centroids, two particularly extreme individual patterns were identified and described.

6. REFERENCES

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

http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp

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