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

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 National Climatic Data Center Climate Divisions. Based on areal-averaging techniques, month-to-month precipitation statistics have been compiled, division by division, since 1895 (and just recently, utilizing new and improved techniques, a recalculation completed of the entire division-by-division precipitation statistics, by year). With such huge distances 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 significant associations with phenomena such as the various ENSO phases ("El Nino", "Neutral", or "La Nina) should make for interesting study.

To this end, the existence and relative frequencies of California Climate Division rain year anomaly variation 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 2013-14, some 119 seasons.

As applied to K-Means, the V-Fold Cross Validation Algorithm is an automated, iterative training sample type procedure that tends to optimize the number of resolved K clusters, depending on the choice of statistical distance metric and a specified percent distance improvement cutoff threshold, the latter measured by comparing, between successive n=K and n=K+1 cluster candidates, the percentage reduction in average training sample statistical distances from their respective cluster centroids.

The present study performs the cluster analysis on the 119 seasons' (normalized) data utilizing the Squared Euclidean distance metric combined with the (default) 5-percent distance improvement threshold. Then, with the "optimal" number of clusters determined, and through referencing of two lists from the NOAA Climate Prediction Center online site which identify past ENSO episode phases back through 1895-96, a Bayesian statistical analysis is performed that addresses the following questions: given an impending ENSO phase type, what are the conditional ("posterior") probabilities that each of the inter-divisional anomaly patterns will be realized for a given July-June rain season. Results are described and interpreted, with the Bayesian probabilities compared among episode types.

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).

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 internal automatic software feature, to reduce them to a common scale (between 0.0 and 1.0) and lessen the influence of outliers. Cluster results would be presented in pre-normalized data form.

Since the percent improvement threshold default setting (5 percent) can be changed, potentially resulting in a different "best" cluster size, an alternative graphical tool is available that can provide a different selection option. This tool, the Scree Plot, traces the actual (usually decreasing) mean training sample statistical distances over a range of increasing K's. Inflection points on the Scree Plot can be interpreted as "natural" cutoff points, the "best" cluster size corresponding to the inflection point's Kth position on the graph. The percent improvement cutoff K may differ (the iterations having stopped at K+1), so, alternatively, if one opts to choose the inflection point as the "right" K and it is different than the percent improvement threshold K, the program can be rerun, "forcing" the "optimal" cluster size and accompanying analysis and information to correspond to that at K, and K only. If one is interested in a more exhaustive analysis, the forcing could be done at a Scree K-value that exhibits essentially zero change in mean training sample statistical distance from the preceding K-1 level. At even higher K levels, of course,

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

the statistical distance curve might trend back upward, reflecting over-fitting.

In this study, the 5 percent default distance improvement cutoff threshold was utilized along the Squared Euclidean distance metric (default: Euclidean), together with Scree Plot inspection.

3. BAYESIAN ANALYSIS

From *Wikipedia*, Bayesian inference is a method of which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is acquired. In the context of this study, the initial hypothesis would be a probabilistic belief, or "Prior Probability", that a given anomaly pattern (cluster) would occur unconditionally (historical percent frequency of the pattern), updated by a processing of evidence that relates the occurrence of the pattern to ENSO phase. The latter could be referred to as "accounting for evidence" and the result, or "impact", multiplied by the "Prior Probability" would produce a "Posterior Probability" that incorporates this new conditional information (the ENSO phase) into a revised probabilistic belief that the given pattern will occur. A desirable outcome would be a marked contrast in magnitudes between the Posterior and Prior probabilities which would indicate that knowledge about the conditional variable "matters". The actual Bayesian expression will appear in a later section in which a case example is demonstrated on the California Climate Division precipitation data.

4. THE DATA

The raw data were downloaded via an NCDC online link which had the newly modified complete history for the July 1895 to June 2014 period of interest. 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 omitted.

Also, Figure 2 is a bar chart of the 119-year mean July-June precipitation figures, by division, and Figure 3 a similar type bar chart of the standard deviation statistics, by division. From Figure 2, there is a wide range of mean statistics, from nearly 50" in the North Coast division, to only 6" in the Southeast Desert. The standard deviation statistics in Figure 3 range from ~13" for the North Coast, to 2" in the Southeast Desert.

With such a wide division-to-division range in overall mean precipitation and variability across the State, it makes sense from an interpretation standpoint to express the individual cluster results, division-by-division, in terms of relative or standardized deviations from the overall averages in Figure 2, based on the overall standard deviation statistics depicted in Figure 3.

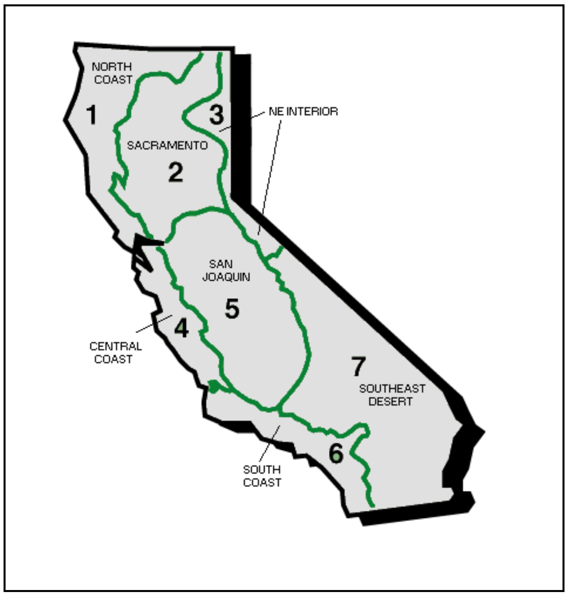


Figure 1 – Map of California Climate Divisions – from NCDC.

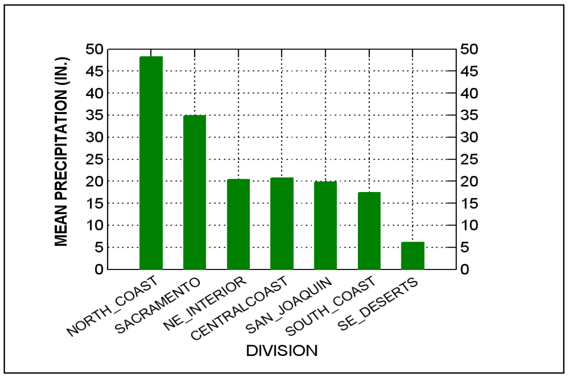


Figure 2 – Mean Seasonal (July-June) Precipitation (In.) for NCDC California Climate Divisions (1895-96 thru 2013-14 Period of Record

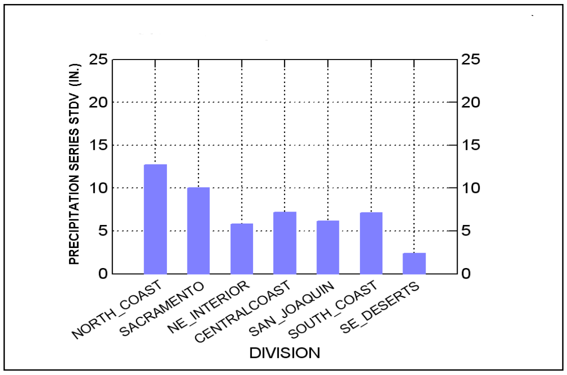


Figure 3 – Seasonal (July-June) Precipitation Series' Standard Deviations (In.) - NCDC California Climate Divisions (1895-96 thru 2013-14 Period of Record)

5. RESULTS

The K-Means/V-Fold algorithm produced six clusters, ranging in individual percent frequency from 23.5% to 11.8% for a pair of patterns.

5.1. – Scree Plot

Figure 4 is a Scree Plot of the iterative results. An inflection point is visible at K=6, matching the "Best" K determined by the 5% default improvement threshold setting - a reinforcing outcome. The curve is also essentially flat from K=6 to K=7, further reinforcement that there are essentially six inter-divisional anomaly modes in existence for California July-June total precipitation.

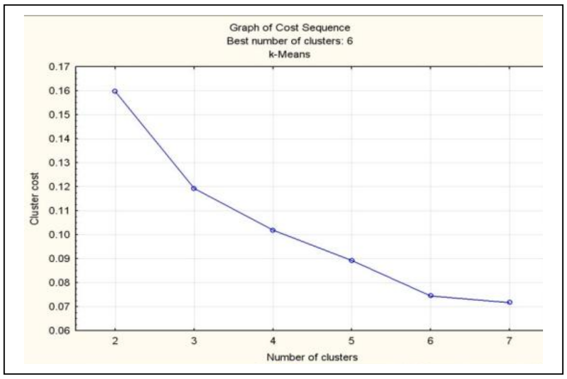


Figure 4 – Scree Plot of K-Means/V-Fold Cross Validation Algorithm Analysis of California Climate Divisions' Seasonal (July-June Total) Precipitation Anomalies.

5.2. – Standardized Mean Anomaly Charts for the Individual Patterns.

Figures 5 thru 10 present the division-by-division standardized mean anomalies for each of the six patterns, in descending order of importance.

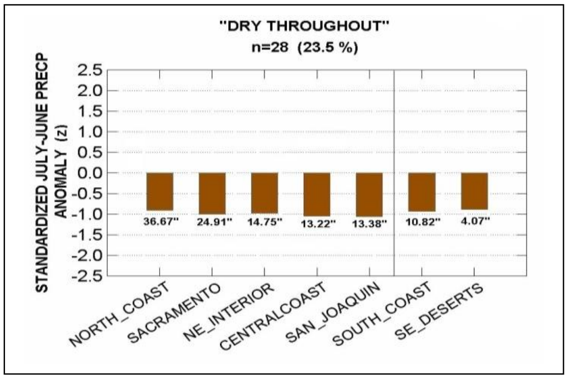


Figure 5 – Standardized Mean Division-by-Division Anomalies for the "Dry-Throughout" Pattern (Mode #3).

Figure 5 shows the most frequently represented pattern (23.5% incidence), titled "Dry-Throughout". The standardized division-to-division anomalies are quite uniform, about one standard deviation each below their respective climatologies. The actual cluster means are annotated immediately below the edges of the bars, ranging from 36.67" for the "North Coast" to 4.07" for the "SE Deserts". These can be compared with the overall averages depicted in Figure 2.

Ranking second is the "Dry North & Central, Near-Normal South" Pattern (21.0% incidence, see Figure 6), exhibiting relatively pronounced mean standardized anomalies (~-0.75) for the northernmost three divisions ('North Coast", "Sacramento", and "Northeast Interior") slightly less negative ones for the middle two (~-0.50) divisions ("Central Coast and San Joaquin"), and near zero ones for the southernmost two ("South Coast" and "SE Deserts")

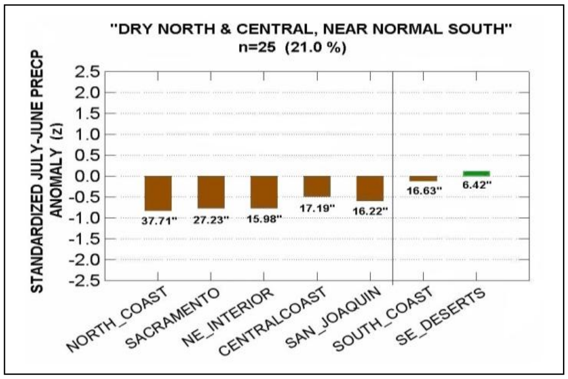


Figure 6 – Standardized Mean Division-by-Division Anomalies for the "Dry North & Central, Near Normal South" Pattern – Mode #4.

Third most frequent is the "Slightly Wet to Slightly Dry Trend" Pattern (18.5% incidence, see Figure 7). This shows a modest north to south "drift" from slightly wet conditions in the North to dry ones in the South.

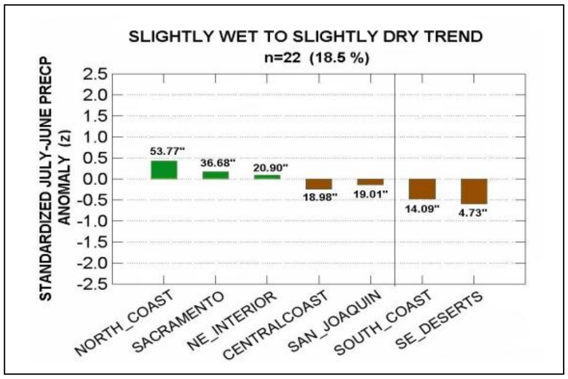


Figure 7 – Standardized Mean Division-by-Division Anomalies for the "Slightly Wet to Slightly Dry Trend" Relative Anomaly Pattern – (Mode # 5).

In fourth place is the “Progressively More Wet Relative Anomalies, N to S” Pattern (13.5 % incidence, see Figure 8). This depicts increasingly wetter character (in the relative sense) from north to south, particularly between the northernmost three divisions and the other four. Mean divisional precipitation for the North Coast (49.83”) is only slightly above overall climatology, while that for the South Coast (23.31”), and SE Deserts (8.15”) are each close to one standard deviation wetter than their overall norms.

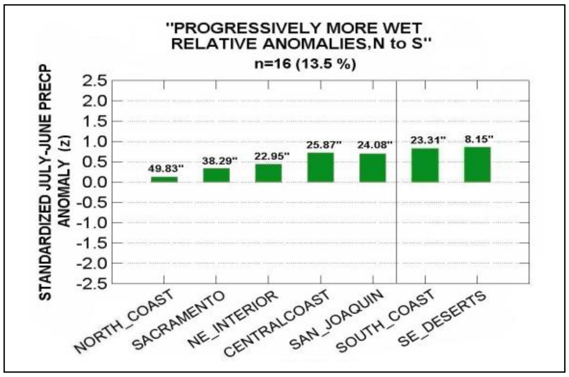


Figure 8 – Standardized Mean Division-by-Division Anomalies for the “Progressively More Wet Relative Anomalies, N to S” Pattern – Mode # 6.

Finally, tied for fifth place are the “Very Wet Throughout” and “Wet North & Central, Dry South” patterns (11.8% incidences each, see Figures 9 and 10, respectively).

Figure 9 exhibits exceptionally “wet” relative anomalies for all divisions, ranging from just under +1.5 standard deviations for the North Coast to nearly +2.0 for the South Coast; Figure 10 shows significantly positive ones for the northernmost five divisions but in a visibly sharp contrast, negative ones for the southernmost two, that for the SE Deserts approaching -0.5.

A few summary remarks can be made about the six patterns. First, Figures 5, 8 and 9, totaling 48.5% of the cases, have the same relative anomaly signs across all divisions; so, to generalize, it appears that in roughly half of the seasons, relative rainfall character across the State is the same (the figure goes up to about 70% if the results of Figure 6 are included, ignoring the slightly positive anomaly of the SE Deserts).

Also, there appears to be dichotomy of sorts in mean anomaly character (either in sign or magnitudes of the same sign) between the northernmost five and the southernmost two divisions, especially noticeable in Figures 6 and 10 (32.8% of the cases) and to a lesser extent in Figures 7 and 8 (another 32%), totaling about 2/3rds of the cases. This is probably due to the more southerly latitude of the South Coast and much of the SE Desert region, and the differing topography (e.g., the E/W oriented Transverse ranges, exhibiting an occasional “barrier” effect, along with the more southeasterly oriented coastline, etc.).

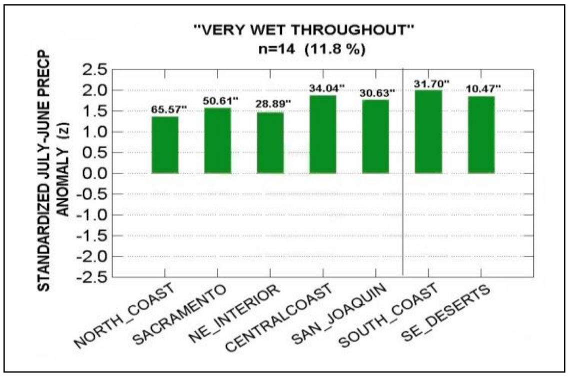


Figure 9 – Standardized Mean Division-by-Division Anomalies for the “Very Wet Throughout” Relative Anomalies” Pattern – Mode # 2.

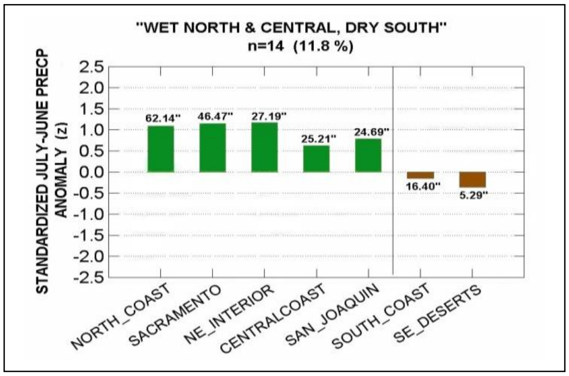
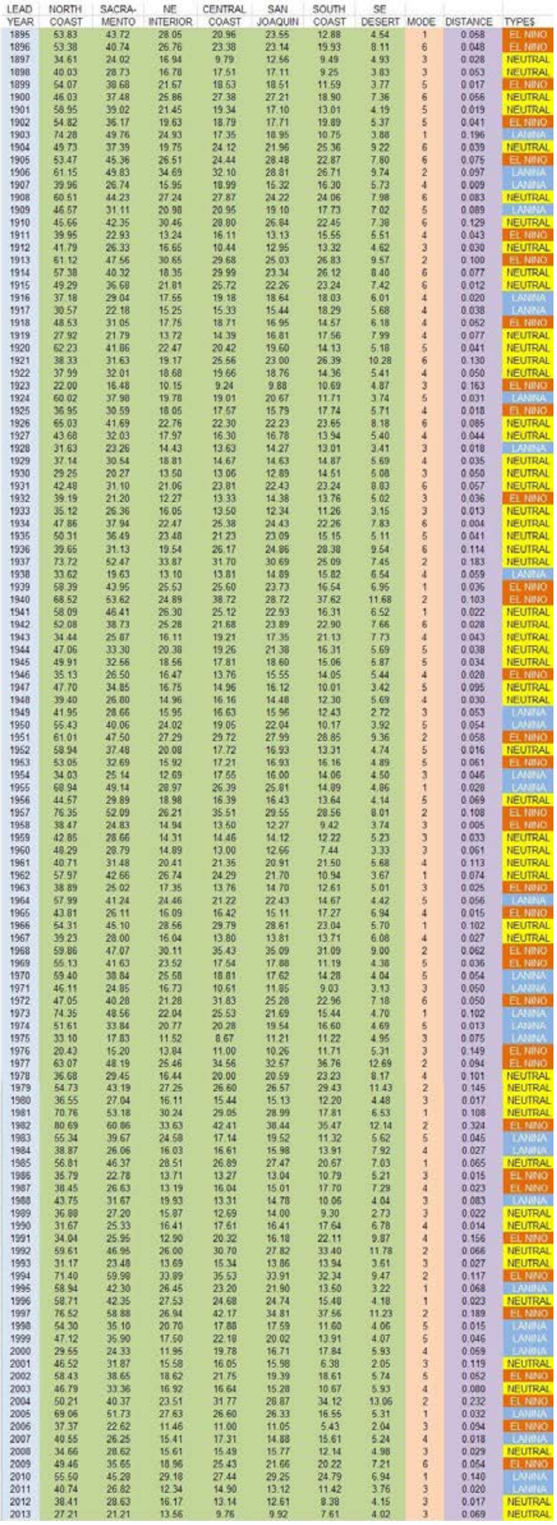


Figure 10 – Standardized Mean Division-by-Division Anomalies for the “Wet North & Central, Dry South” Relative Anomalies Pattern – Mode #1.

***5.3.** – Pattern Probabilities Conditioned on El Nino, Neutral, or LaNina occurrences – Bayesian Determinations*

While the percent frequencies of the above six patterns may be considered as probabilities that they may occur individually for a given July-June rain year, there are other climatic indicators that should provide additional, more refined probabilistic information on occurrence likelihoods. ENSO phase (“El Nino, “Neutral”, or” La Nina”) is one indicator known to influence California rainfall patterns, so the next step is to investigate the possible modifying influences of these three episode types on the “baseline” Prior probabilities above of the six patterns. This would be a conditional probability exercise, and the method of choice, already introduced, would be Bayesian Analysis.

First, the 119 seasons would have to be assigned ENSO episode classification. Identification of ENSO types is a not completely objective process, different researchers having compiled different lists, with likely more uncertainty for those years further back. For the purpose of this research, the lists utilized are those created by the NOAA Climate Prediction Center. The first covers the years 1877-2001, the second



LEAD YEAR	NORTH COAST	SACRA-MENTO	NE INTERIOR	CENTRAL COAST	SAN JOAQUIN	SOUTH COAST	SE DESERT	MODE	DISTANCE	TYPE$
1895	53.83	43.72	28.05	20.96	23.55	12.88	4.54	1	0.058	EL NINO
1896	53.38	40.74	26.76	23.38	23.14	19.93	8.11	6	0.048	EL NINO
1897	34.61	24.02	16.94	9.79	12.56	9.49	4.93	3	0.028	NEUTRAL
1898	40.03	28.73	16.78	17.51	17.11	9.25	3.83	3	0.053	NEUTRAL
1899	54.07	38.68	21.67	18.53	18.51	11.59	3.77	5	0.017	EL NINO
1900	46.03	37.48	25.86	27.38	27.21	18.90	7.36	6	0.056	NEUTRAL
1901	58.95	39.02	21.45	19.34	17.10	13.01	4.19	5	0.019	NEUTRAL
1902	54.82	36.17	19.63	18.79	17.71	19.89	5.37	5	0.041	EL NINO
1903	74.28	49.76	24.93	17.35	18.95	10.75	3.88	1	0.196	LANINA
1904	49.73	37.39	19.75	24.12	21.96	25.36	9.22	6	0.039	NEUTRAL
1905	53.47	45.36	26.51	24.44	28.48	22.87	7.80	6	0.075	EL NINO
1906	61.15	49.83	34.69	32.10	28.81	26.71	9.74	2	0.097	LANINA
1907	39.96	26.74	15.95	18.99	15.32	16.30	5.73	4	0.009	LANINA
1908	60.51	44.23	27.24	27.87	24.22	24.06	7.98	6	0.083	NEUTRAL
1909	46.57	31.11	20.98	20.95	19.10	17.73	7.02	5	0.089	LANINA
1910	45.66	42.35	30.46	28.80	26.84	22.45	7.38	6	0.129	NEUTRAL
1911	39.95	22.93	13.24	16.11	13.13	15.55	5.51	4	0.043	EL NINO
1912	41.79	26.33	16.65	10.44	12.95	13.32	4.62	3	0.030	NEUTRAL
1913	61.12	47.56	30.65	29.68	25.03	26.83	9.57	2	0.100	EL NINO
1914	57.38	40.32	18.35	29.99	23.34	26.12	8.40	6	0.077	NEUTRAL
1915	49.29	36.68	21.81	25.72	22.26	23.24	7.42	6	0.012	NEUTRAL
1916	37.18	29.04	17.55	19.18	18.64	18.03	6.01	4	0.020	LANINA
1917	30.57	22.18	15.25	15.33	15.44	18.29	5.68	4	0.038	LANINA
1918	48.53	31.05	17.75	18.71	16.95	14.57	6.18	4	0.052	EL NINO
1919	27.92	21.79	13.72	14.39	16.81	17.56	7.99	4	0.077	NEUTRAL
1920	62.23	41.86	22.47	20.42	19.60	14.13	5.18	5	0.041	NEUTRAL
1921	38.33	31.63	19.17	25.56	23.00	26.39	10.28	6	0.130	NEUTRAL
1922	37.99	32.01	18.68	19.66	18.76	14.36	5.41	4	0.050	NEUTRAL
1923	22.00	16.48	10.15	9.24	9.88	10.69	4.87	3	0.163	EL NINO
1924	60.02	37.98	19.78	19.01	20.67	11.71	3.74	5	0.031	LANINA
1925	36.95	30.59	18.05	17.57	15.79	17.74	5.71	4	0.018	EL NINO
1926	65.03	41.69	22.76	22.30	22.23	23.65	8.18	6	0.085	NEUTRAL
1927	43.68	32.03	17.97	16.30	16.78	13.94	5.40	4	0.044	NEUTRAL
1928	31.63	23.26	14.43	13.63	14.27	13.01	3.41	3	0.018	LANINA
1929	37.14	30.54	18.81	14.67	14.63	14.87	5.69	4	0.035	NEUTRAL
1930	29.25	20.27	13.50	13.06	12.89	14.51	5.08	3	0.050	NEUTRAL
1931	42.48	31.10	21.06	23.81	22.43	23.24	8.83	6	0.057	NEUTRAL
1932	39.19	21.20	12.27	13.33	14.38	13.76	5.02	3	0.036	EL NINO
1933	35.12	26.36	16.05	13.50	12.34	11.26	3.15	3	0.013	NEUTRAL
1934	47.86	37.94	22.47	25.38	24.43	22.26	7.83	6	0.004	NEUTRAL
1935	50.31	36.49	23.48	21.23	23.09	15.15	5.11	5	0.041	NEUTRAL
1936	39.65	31.13	19.54	26.17	24.86	28.38	9.54	6	0.114	NEUTRAL
1937	73.72	52.47	33.87	31.70	30.69	25.09	7.45	2	0.183	NEUTRAL
1938	33.62	19.63	13.10	13.81	14.89	15.82	6.54	4	0.059	LANINA
1939	58.39	43.95	25.53	25.60	23.73	16.54	6.95	1	0.036	EL NINO
1940	68.52	53.62	24.89	38.72	28.72	37.62	11.68	2	0.103	EL NINO
1941	58.09	46.41	26.30	25.12	22.93	16.31	6.52	1	0.022	NEUTRAL
1942	52.08	38.73	25.28	21.68	23.89	22.90	7.66	6	0.028	NEUTRAL
1943	34.44	25.87	16.11	19.21	17.35	21.13	7.73	4	0.043	NEUTRAL
1944	47.06	33.30	20.38	19.26	21.38	16.31	5.69	5	0.038	NEUTRAL
1945	49.91	32.56	18.56	17.81	18.60	15.06	5.87	5	0.034	NEUTRAL
1946	35.13	26.50	16.47	13.76	15.55	14.05	5.44	4	0.028	EL NINO
1947	47.70	34.85	16.75	14.96	16.12	10.01	3.42	5	0.095	NEUTRAL
1948	39.40	26.80	14.96	16.16	14.48	12.30	5.69	4	0.030	NEUTRAL
1949	41.95	28.66	15.95	16.63	15.96	12.43	2.72	3	0.053	LANINA
1950	55.43	40.06	24.02	19.05	22.04	10.17	3.92	5	0.054	LANINA
1951	61.01	47.50	27.29	29.72	27.99	28.85	9.36	2	0.058	EL NINO
1952	58.94	37.48	20.08	17.72	16.93	13.31	4.74	5	0.016	NEUTRAL
1953	53.05	32.69	15.92	17.21	16.93	16.16	4.89	5	0.061	EL NINO
1954	34.03	25.14	12.69	17.55	16.00	14.06	4.50	3	0.046	LANINA
1955	68.94	49.14	28.97	26.39	25.81	14.89	4.86	1	0.028	LANINA
1956	44.57	29.89	18.98	16.39	16.43	13.64	4.14	5	0.069	NEUTRAL
1957	76.35	52.09	26.21	35.51	29.55	28.56	8.01	2	0.108	EL NINO
1958	38.47	24.83	14.94	13.50	12.27	9.42	3.74	3	0.005	EL NINO
1959	42.85	28.66	14.31	14.46	14.12	12.22	5.23	3	0.033	NEUTRAL
1960	48.29	28.79	14.89	13.00	12.66	7.44	3.33	3	0.061	NEUTRAL
1961	40.71	31.48	20.41	21.35	20.91	21.50	5.68	4	0.113	NEUTRAL
1962	57.97	42.66	26.74	24.29	21.70	10.94	3.67	1	0.074	NEUTRAL
1963	38.89	25.02	17.35	13.76	14.70	12.61	5.01	3	0.025	EL NINO
1964	57.99	41.24	24.46	21.22	22.43	14.67	4.42	5	0.056	LANINA
1965	43.81	26.11	16.09	16.42	15.11	17.27	6.94	4	0.015	EL NINO
1966	54.31	45.10	28.56	29.79	28.61	23.04	5.70	1	0.102	NEUTRAL
1967	39.23	28.00	16.04	13.80	13.81	13.71	6.08	4	0.027	NEUTRAL
1968	59.86	47.07	30.11	35.43	35.09	31.09	9.00	2	0.062	EL NINO
1969	55.13	41.63	23.52	17.54	17.88	11.19	4.38	5	0.036	EL NINO
1970	59.40	38.84	25.58	18.81	17.62	14.28	4.04	5	0.054	LANINA
1971	46.11	24.85	16.73	10.61	11.85	9.03	3.13	3	0.050	LANINA
1972	47.05	40.28	21.28	31.83	25.28	22.96	7.18	6	0.050	EL NINO
1973	74.35	48.56	22.04	25.53	21.69	15.44	4.70	1	0.102	LANINA
1974	51.61	33.84	20.77	20.28	19.54	16.60	4.69	5	0.013	LANINA
1975	33.10	17.83	11.52	8.67	11.21	11.22	4.95	3	0.075	LANINA
1976	20.43	15.20	13.84	11.00	10.26	11.71	5.31	3	0.149	EL NINO
1977	63.07	48.19	25.46	34.56	32.57	36.76	12.69	2	0.094	EL NINO
1978	36.68	29.45	16.44	20.00	20.59	23.23	8.17	4	0.101	NEUTRAL
1979	54.73	43.19	27.25	26.60	26.57	29.43	11.43	2	0.145	NEUTRAL
1980	36.55	27.04	16.11	15.44	15.13	12.20	4.48	3	0.017	NEUTRAL
1981	70.76	53.18	30.24	29.05	28.99	17.81	6.53	1	0.108	NEUTRAL
1982	80.69	60.86	33.63	42.41	38.44	35.47	12.14	2	0.324	EL NINO
1983	55.34	39.67	24.58	17.14	19.52	11.32	5.62	5	0.045	LANINA
1984	38.87	26.06	16.03	16.61	15.98	13.91	7.92	4	0.027	LANINA
1985	56.81	46.37	28.51	26.89	27.47	20.67	7.03	1	0.065	NEUTRAL
1986	35.79	22.78	13.71	13.27	13.04	10.79	5.21	3	0.015	EL NINO
1987	38.45	26.63	13.19	16.04	15.01	17.70	7.29	4	0.023	EL NINO
1988	43.75	31.67	19.93	13.31	14.78	10.06	4.04	3	0.083	LANINA
1989	36.88	27.20	15.87	12.69	14.00	9.30	2.73	3	0.022	NEUTRAL
1990	31.67	25.33	16.41	17.61	16.41	17.64	6.78	4	0.014	NEUTRAL
1991	34.04	25.95	12.90	20.32	16.18	22.11	9.87	4	0.156	EL NINO
1992	59.61	46.95	26.00	30.70	27.82	33.40	11.78	2	0.066	NEUTRAL
1993	31.17	23.48	13.69	15.34	13.86	13.94	3.61	3	0.027	NEUTRAL
1994	71.40	59.98	33.89	35.53	33.91	32.34	9.47	2	0.117	EL NINO
1995	58.94	42.30	26.45	23.20	21.90	13.50	3.22	1	0.068	LANINA
1996	58.71	42.35	27.53	24.68	24.74	15.48	4.18	1	0.023	NEUTRAL
1997	76.52	58.88	26.94	42.17	34.81	37.56	11.23	2	0.189	EL NINO
1998	54.30	35.10	20.70	17.88	17.59	11.60	4.06	5	0.015	LANINA
1999	47.12	35.90	17.50	22.18	20.02	13.91	4.07	5	0.046	LANINA
2000	29.65	24.33	11.95	19.78	16.71	17.84	5.93	4	0.069	LANINA
2001	46.52	31.87	15.58	16.05	15.98	6.38	2.05	3	0.119	NEUTRAL
2002	58.43	38.65	18.62	21.75	19.39	18.61	5.74	5	0.052	EL NINO
2003	46.79	33.36	16.92	16.64	15.28	10.67	5.93	4	0.080	NEUTRAL
2004	50.21	40.37	23.51	31.77	28.87	34.12	13.06	2	0.232	EL NINO
2005	69.06	51.73	27.63	26.60	26.33	16.55	5.31	1	0.032	LANINA
2006	37.37	22.62	11.46	11.00	11.05	5.43	2.04	3	0.094	EL NINO
2007	40.55	26.25	15.41	17.31	14.88	15.61	5.24	4	0.018	LANINA
2008	34.66	28.62	15.61	15.49	15.77	12.14	4.98	3	0.029	NEUTRAL
2009	49.46	35.65	18.96	25.43	21.66	20.22	7.21	6	0.054	EL NINO
2010	55.50	45.28	29.18	27.44	29.25	24.79	6.94	1	0.140	LANINA
2011	40.74	26.82	12.34	14.90	13.12	11.42	3.76	3	0.020	LANINA
2012	38.41	28.63	16.17	13.14	12.61	8.38	4.15	3	0.017	NEUTRAL
2013	27.21	21.21	13.56	9.76	9.92	7.61	4.02	3	0.069	NEUTRAL

Figure 11 – Time Series of California Climate Division Precipitation, by Season and Division, with Cluster Assignments, Distance to Centroids, and ENSO designations (1895-96 through 2013-14 seasons).

1950-2013. Those years that overlap 1950-2013 are given the assignments of the latter list. From the TYPE$ column, 35 “El Nino’s”, 54 “Neutrals” and 30 “La Nina’s” are present.

Figure 11 is a table with the actual divisional precipitation data by season (lead years1895 to 2013), the mode (pattern) number assignments, statistical distances to the pattern centroids, and the ENSO designations. The mode numbers are those six originally assigned (in unranked order of importance) by the software upon execution of the K-Means/V-Fold algorithm. They are included in the Figure 5 to10 anomaly pattern graphs’ titles.

Next, the Bayesian conditional probabilities were calculated. Since there were six patterns and three different ENSO phases, there would be18 separate calculations. Figure 12 shows the Bayesian theorem along with the steps of a sample calculation, that for the conditional probability of Pattern #2 (“Very Wet Throughout” – see Figure 9) as associated with an imminent El Nino episode.

- BAYES THEOREM -

POSTERIOR PROB

PRIOR PROB

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B|A)P(A) + P(B|A_c)P(A_c)}.$$

$$P\left(\text{"VERY WET THROUGHOUT PATTERN"} \,\middle|\, \text{"EL NINO"}\right) = \frac{P\left(\text{"EL NINO"} \,\middle|\, \text{"VERY WET THROUGHOUT PATTERN"}\right) P\left(\text{"VERY WET THROUGHOUT PATTERN"}\right)}{P\left(\text{"EL NINO"} \,\middle|\, \text{"VERY WET THROUGHOUT PATTERN"}\right) P\left(\text{"VERY WET THROUGHOUT PATTERN"}\right) + P\left(\text{"EL NINO"} \,\middle|\, \text{OTHER FIVE PATTERNS}\right) P\left(\text{OTHER FIVE PATTERNS}\right)}$$

$$P\left(\text{"VERY WET THROUGHOUT PATTERN"} \,\middle|\, \text{"EL NINO"}\right) = \frac{(10/14) \text{ X } (14/119)}{\left[(10/14) \text{ X } (14/119)\right] + \left[(25/105) \text{ X } (105/119)\right]}$$

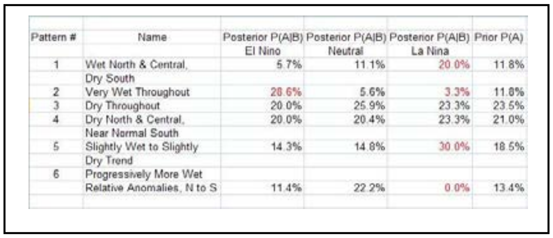
$$P\left(\text{"VERY WET THROUGHOUT PATTERN"} \,\middle|\, \text{"EL NINO"}\right) = \frac{.084}{.084 + .210} \text{ or } .286$$

Figure 12 – Bayes Theorem *(*from *Wikipedia)* and a Sample Calculation of the Conditional Probability of the “Very Wet Throughout Pattern” being realized, given an Impending or ongoing El Nino.

From Figure 12, the top expression shows the general Bayes Theorem, that immediately below it the expression adapted to the variables of the sample exercise. In the numerator on the right side of the equation, “P(A)” is the Prior Probability of the “Very Wet Throughout Pattern”, simply the original proportion of the 119 seasons that were so classified by the K-Means/V-Fold algorithm (14/119 or .118, or 11.8%). P(B|A) is the proportion of “Very Wet Throughout” cases that were associated with El Nino episodes (in this case, 10/14 or .586, a very high relative figure). P(A) and P(B|A) are then multiplied together, yielding .084, this result also copied into the denominator, to be added to the product of the proportional incidence of El Ninos in the other non-“Very Wet Throughout Pattern” years (25/105 or .238) times the converse of the Prior Probability (.882), yielding +.210. The final quotient (.084/(.084+.210) or 28.6% is the Posterior Probabilty, P(A|B): the likelihood that the “Very Wet Throughout

Pattern" will be ultimately be realized, given an impending El Nino. The Posterior Probability in this example is more than double than that of the Prior, indicating that an El Nino episode does "matter", in this instance, increasing the odds noticeably that the "Very Wet Throughout Pattern" will be expressed for the July-June rain season.

Table 1 lists the Posterior Probability results for all the 18 combinations of 3 ENSO types (columns) and 6 Patterns (rows). Some Posteriors of particular interest are shaded in red.



Pattern #	Name	Posterior P(A|B) El Nino	Posterior P(A|B) Neutral	Posterior P(A|B) La Nina	Prior P(A)
1	Wet North & Central, Dry South	5.7%	11.1%	20.0%	11.8%
2	Very Wet Throughout	28.6%	5.6%	3.3%	11.8%
3	Dry Throughout	20.0%	25.9%	23.3%	23.5%
4	Dry North & Central, Near Normal South	20.0%	20.4%	23.3%	21.0%
5	Slightly Wet to Slightly Dry Trend	14.3%	14.8%	30.0%	18.5%
6	Progressively More Wet Relative Anomalies, N to S	11.4%	22.2%	0.0%	13.4%

Table 1 – Posterior Probability Results for all combinations of ENSO Type vs. Pattern

To interpret, for example, the El Nino Posterior Probability column (third from the left), reading down, lists the conditional probabilities that each of the six patterns will be realized, given an El Nino episode. The 28.6% figure, shaded red for the "Very Wet Throughout" Pattern (shown in Figure 9 and having already served as the Bayesian computation example above) is the pattern most likely to happen of the six. As already discussed, this figure is much higher than the "Very Wet Throughout" pattern's "baseline" 11.8% Prior shown in Column 6. By the same token, if a La Nina is imminent, there is only a 3.3% chance that the "Very Wet Throughout" pattern will be expressed.

The most favored pattern for the La Nina (30.0% Posterior Probability) is the "Slightly Wet to Slightly Dry Trend", shown in Figure 7; this is markedly higher than the pattern's Prior (18.5%). Also, La Nina is the most frequent ENSO type associated with the Wet North & Central, Dry South pattern (See Figure 10), its frequency (20.0%) noticeably higher than the pattern's Prior (11.8%). Finally, there is a zero Posterior Probability associated with La Nina's and the "Progressively More Wet Relative Anomalies, N to S" pattern (See Figure 8). From these multiple results, it appears that La Nina's are not generally associated with wet seasons in the South.

So, in conclusion, conditioning the occurrence probabilities of the six patterns on ENSO phases did provide more refined insights on their likelihoods. The range of their Priors was 11.8% to 23.5%, that for the Posteriors 0.0% to 30.0%.

6. SUMMARY

Utilizing the clustering tool K-Means, integrated with the V-fold cross validation algorithm, the existence and character of seasonal (July-June total) precipitation modes were explored, collectively, for the seven NCDC California climate divisions, accessing the 1895-96 to 2013-14 period of record. Inputs were normalized, areal-averaged total precipitation statistics season-by-season, and division-by division.

Results resolved, unambiguously, six clusters (also "patterns" or "modes"), characterizing a variety of anomaly configurations across divisions, occasionally on a combined North & Central vs. South basis. Individual pattern frequencies ("Prior probabilities) ranged from 11.8% to 23.5%. Then, using Bayesian statistical methodology, conditional probability estimates (Posterior probabilities) were made of the occurrence likelihoods of the six patterns, given El Nino, Neutral, or La Nina episodes imminent or already in place. In roughly half of the 18 Posterior Probability calculations, the Posterior magnitudes differed significantly from the Priors (see Table 1), indicative that El Nino type was a useful predictive indicator. These figures ranged from 0.0% to 30.0%.

A combined Clustering/Bayesian analysis of this kind might prove similarly useful in other climatological-related applications.

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