IDENTIFICATION OF MULTI-STATION EXTREME-MOST DAILY MAXIMUM/MINIMUM HISTORICAL TEMPERATURE PATTERNS USING PRINCIPAL COMPONENTS ANALYSIS

Charles J. Fisk *
Newbury Park, CA

1. INTRODUCTION

Floating-bar or Hi-Lo charts are a standard visual means of portraying daily maximum/minimum temperature time-series, for example, those appearing on the cover pages of NCDC Local Climatological Data (LCD) Annual Summaries since 1984; also the more recently introduced yearly/monthly versions in the "Climate Charts" sections of NWS stations' online sites. The depictions of the varied diurnal, synoptic, long-wave, and seasonal influences on daily temperatures over time can be sometimes quite interesting to look at, both from physical and pure pattern perspectives. Most inspections, though, are likely momentary with any interpretations subjective, but for those cases in which the configurations appear particularly irregular or uncharacteristic, the notion may arise on how anomalous they are on a statistical basis, relative to other years in a station's history. In this regard, as a sort of pure science complement to the simpler, more conventional statistics like overall means, standard deviations, extremes, and mean daily ranges, it would seem useful to have a methodology available which could objectively (or at least formally) characterize year-to-year patterns relative to climatology. Some of the most atypical patterns which may have occurred many years previous and never before graphed (much less identified) could be revealed for the first time (and then plotted). In the same manner that an extreme warmest or coolest year can be distinguished, in likewise fashion a year that has the most extreme daily max/min *configurations* could be brought to light.

Utilizing daily data for a single station (Downtown Los Angeles), a previous exploratory analysis delved into this topic (Fisk, 2004), employing two familiar statistical measures (the linear correlation coefficient, or "shape" attribute) and (the linear covariance coefficient, or "spread" property). These were an adaptation of concepts originally described by Yarnal (1993) in a textbook describing analytical techniques applicable to synoptic climatology, among them Linear Principal Components Analysis. It turned out that required calculations in the 2004 investigation could be performed as an unrotated PCA exercise, a somewhat unconventional, but labor-saving and valid application of the technique. The approach identified years that qualified as the most "extreme" in pattern, through referencing of first component correlation and covariance loadings' statistics, both individually and in the 2-D sense (bivariate confidence ellipsoids). Follow-up exploratory analyses, also using the unrotated PCA approach, examined Downtown Los Angeles daily mean temperature modes for selected calendar months (Fisk, 2007), and extreme patterns in LAX midnight-to-midnight hourly temperatures, also for certain calendar months (Fisk, 2012).

Returning to daily max/min calendar year data, this study expands the scope to six stations in the U.S., identifying and comparing the most extreme calendar year patterns in the combined 2-D shape/spread sense (as evaluated by comparing their relative point positions within or outside 2-D confidence ellipsoids). In a typical calendar year max/min application, first component PCA results, as indicated by the very high eigenvalue magnitudes and percent of variance explained, describe an overwhelming portion of the variance, but for a few stations, recently identified, second, third, and even higher component results display eigenvalue confidence interval bands that encompass the eigenvalue magnitude of 1.0, the minimum threshold for "original variable" status. In a daily max/min temperature application, first components' statistics describe adherence to patterns that are first harmonic in form. Second, third, and higher component patterns, in those few significant cases, may conform in a general but less precise sort of way to second, third, and higher quasi-harmonic forms, the agreement most visible for certain sub-portions of the year. Results will include floating-bar graphs, by station, for those years that qualified as the most "extreme" in the first component sense, and also, when applicable, for those higher component cases. Many of these extreme daily temperature series patterns have never before been graphed.

2. THE CORRELATION ("SHAPE") AND COVARIANCE ("SPREAD) METRICS IN IN THE CHARACTERIZATION OF CALENDAR YEAR DAILY MAX/MIN TEMPERATURE PATTERNS

What pattern makeups do the correlation and covariance capture and emphasize? Viewed from a linear regression standpoint, the correlation coefficient measures the linear association between two variables represented in a best-fit equation in the form a+bX, where a is a constant, b a regression coefficient, and X the "dependent" variable (can be either the climatological series' statistics or those of the individual year – there's no effect on the correlation). It turns out that if one adds a constant to each of the given year's daily maximum and minimum temperature recordings, refits the regression equation, and then recalculates the correlation coefficient, the latter will be exactly the same as before – the only change to the regression equation (of no particular interest in this application) will be the

----

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

incremented constant “a”. As the pattern configuration has remained unchanged, an interpretation thus would be that the correlation coefficient is influenced by pattern “shape”, being not influenced in a direct way to overall average departure from the climatological mean.

The covariance, in contrast, measures how an individual series and that of climatology vary in concert on a cross-product basis. If this correspondence is high (reflected in both in daily temperature ranges and contrasts in seasonality), the covariance statistic would be relatively high also. However, if the daily temperature ranges and/or seasonal contrasts were amplified, reduced, or out of phase with corresponding climatology in some fashion then the covariance statistic would be affected accordingly, depending on the exact nature and mix of the deviations. Thus, a descriptive label for the covariance might be an overall relative spread compared to climatology, or just “spread”. The “spread” statistic is usually positively correlated with the average daily range, and the “shape” and “spread” statistics are also almost always positively correlated themselves, indicative of a meshing of “shape” and “spread” information.

3. APPLICATION OF NON-ROTATED PRINCIPAL COMPONENTS ANALYSIS

In a daily max/min, calendar-year pattern analysis such as the one to follow, an interesting property is that the calculations can be performed as an unrotated Principal Components Analysis problem. The only input required is the array of individual years’ daily max and min temperatures, the rows being the calendar-day ordered daily max/min temperatures and the columns the years. The software produces “climatology” internally and automatically (advantageous when succeeding years’ data are periodically incorporated into the data base), except that the summary PCA statistics are in a rescaled form, either component “scores” or factor “scores”. The “scores”, either in raw (“components”) or standardized (“factors”) form, will have different magnitudes than those of the actual mean daily max/min climatology figures, but on a year-to-year basis, their correlations (or “loadings”) versus reference climatology will be exactly the same – identical also to the correlations generated between given years’ daily max/min’s and reference climatology. In such a calendar year application the first component’s scores would describe an idealized daily pattern of 2*n terms (“n” coefficients attached to the daily maxima and minima each) which is strongly first harmonic (seasonal) in overall shape, the seasonality effect reinforced by the fact that both max’s and min’s are being considered collectively. Thus, first component correlation loadings are a perfect analog for the correlation coefficient or “scale” metric, reinforcing the case for the PCA option. First component PCA covariance loading statistics, while also scaled differently internally than those of the conventional covariance statistics, are nonetheless perfectly correlated one-on-one also with their ordinary covariance counterparts, so they serve with complete validity as the “spread” metric. In this application, the PCA covariance loadings’ outputs were based on standardized (“factor”) scores (climatology) multiplied by the individual years’ daily maxima and minima basis. It should be mentioned that for rectangular matrix consistency sake (PCA calculations make use of rectangular matrices), leap year data are excluded.

Given the advantages of the PCA alternative: ease of input, the exact analog relationship of the correlation and covariance loadings stats to the correlation “shape” and covariance “spread” metrics, respectively, the computing ease/power, the option to analyze higher components, and the supplementary diagnostics output, it is considered highly desirable and practical to hand over the required calculations to a PCA module. The transference produces identical interpretations compared to those produced by the “traditional” correlation and covariance coefficients, not to mention all the additional information available from the PCA output.

As will be seen with the charts to follow, the lower-end magnitude “shape” and “spread” statistics, and to a lesser extent the higher-end “spread” figures tend to depict patterns/configurations that could be perceived visually as “anomalous” (or more precisely, “non-linear”).

Two-dimensional plots (Confidence Ellipsoid graphs) are constructed to assess the most extreme patterns in the bivariate sense, taking in account “shape” and “spread” simultaneously, and beyond Downtown Los Angeles, these will serve as the primary mode of presentation for all higher order components’ cases as well as all components for the other stations.

4. RESULTS

4.1. - *Downtown Los Angeles*

Downtown Los Angeles’ available digitized daily temperature history dates back to1921. As previously noted, first component correlation loadings (or the "shape" metric), represent the linear correlation coefficient of the individual year's daily max/min's with "climatology", either first component PCA scores, first component PCA factor scores, or reference climatology [i.e., mean daily max’s and min’s). These three produce the same linear correlation coefficients and are thus perfectly correlated one-on-one (r=+1.00), Figures 1 to 3 show their respective plots, the red colored points indicating "daily maxima", the blues, "daily minima". Figure 1 presents the raw (or "component") scores, Figure 2 the "component" scores in standardized form ("factors"), and Figure 3 reference climatology (or 1921-2016 mean daily maxima and minima). Note that their configurations are identical.

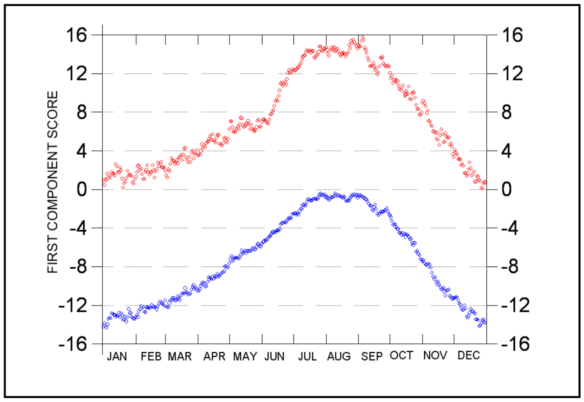

Figure 1: First Component Daily Max/Min PCA Scores for Downtown Los Angeles (1921-2016 data)

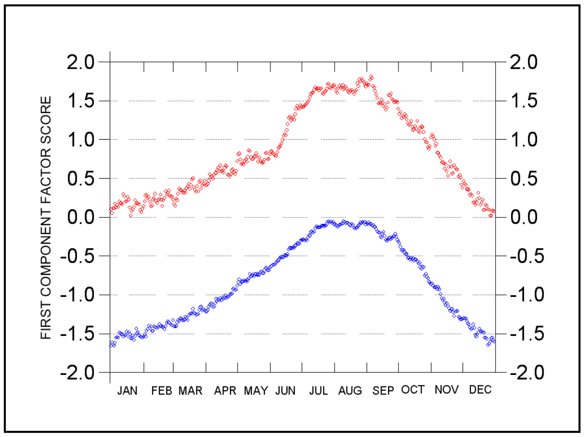

Figure 2: First Component Daily Max/Min PCA Factor Scores for Downtown Los Angeles (1921-2016 data)

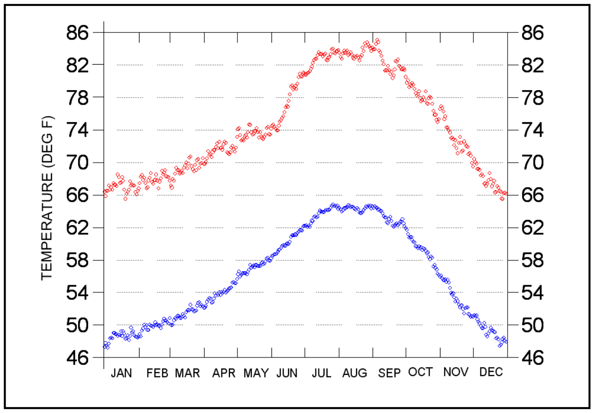

Figure 3: Mean Daily Max/Min Temperatures (Deg F) for Downtown Los Angeles (1921-2016 data)

For Downtown Los Angeles, highest correlation loading (+.915) was realized for the year 1948, the lowest (+.851) for 1930. Mean figure for the 96 years was +.886. Correlation of the loadings with annual means was +.067, indicative that annual mean temperature magnitudes were essentially transparent to "shape", but average daily ranges were linearly associated to a slightly positive degree (r= +.285).

4.1.1 – First Component Correlation Loadings or “Shape” Statistics for Downtown Los Angeles

Figures 4 and 5 show the actual max/min temperature patterns (upper chart) for these two extreme years along with their daily mean departures (lower chart).

While the year 1948's pattern does not appear strikingly irregular, the mean daily temperature range for the year was 21.1 F (climatology: 18.3 F), highest of any of the 96 individual years. Given 1948's high average daily range stat and the positive correlation in general of this metric with "shape" perhaps this was enough to elevate 1948's correlation loading figure to the extreme highest level, the difference, however, between it and many other years ranking just below it very slight. It would seem that the closer a given year's configuration conforms to the climatological "shape", the less anomalous it will appear, although such extreme conformance is in itself, "anomalous", statistically.

The year 1930, however, seems more irregular, displaying many short-term fluctuations in daily max/min temperatures, particularly over late October into November. Mean daily temperature range and annual mean were 17.9 F and 64.9 F, respectively.

The less irregularity evident for high correlation loading 1948 compared to 1930 seems to be a general property of the former type metric, this comparison also valid relative to the extreme highest and lowest covariance loadings patterns, respectively, to be shown later below representing 1949 and 1928.

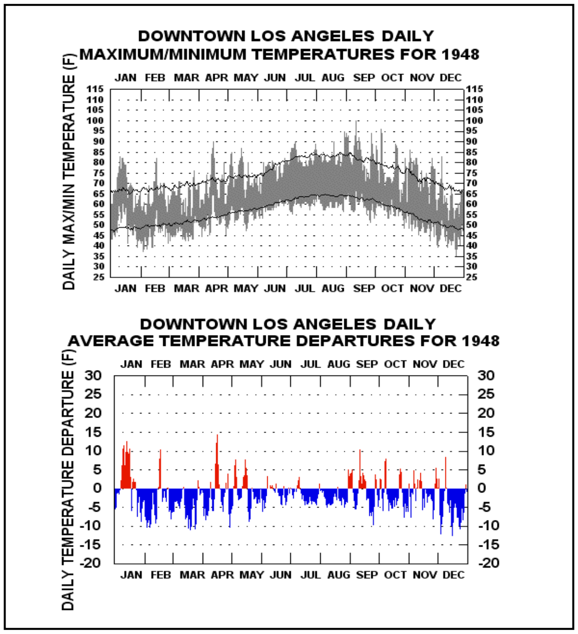

Figure 4: Daily Max/Min Pattern for Downtown Los Angeles (1948) – Highest First Component "Shape" Statistic (+.915)

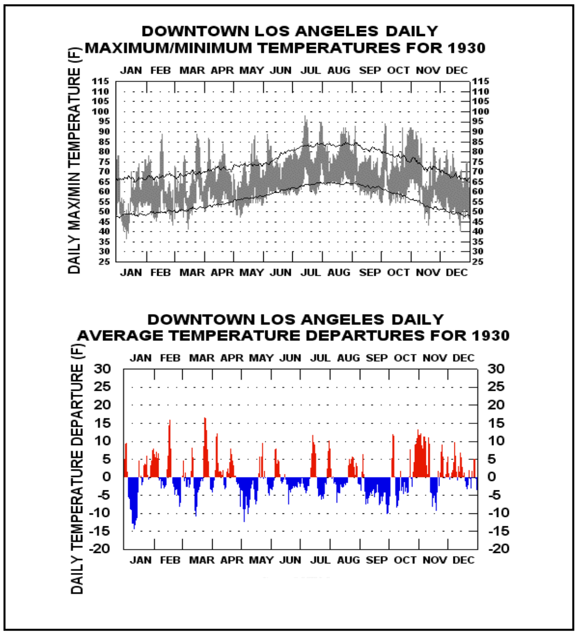

Figure 5: Daily Max/Min Pattern for Downtown Los Angeles (1930) – Lowest First Component "Shape" Statistic (+.851)

From a historical time-series perspective, Figure 6 is a graph of the Downtown Los Angeles correlation loadings statistics, by year. There seems to be a tendency for lower magnitudes in the earlier years of the record, thru about 1940, suggestive of a possible observational heterogeneity in the station history.

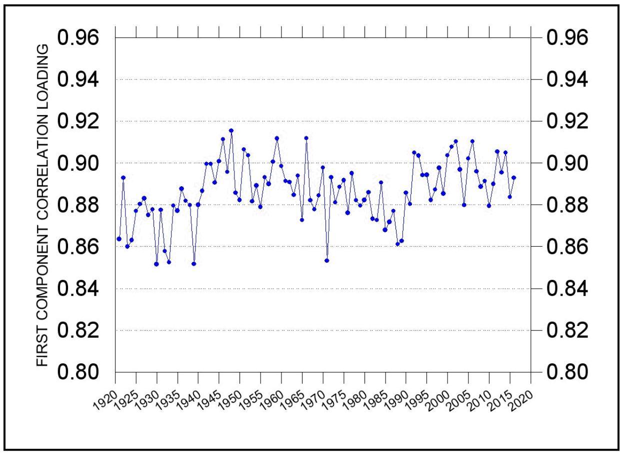

Figure 6: Time-Series Plot of Downtown Los Angeles First Component Correlation Loadings ("Shape") Statistics (1921-2016 data – Mean: +.886)

4.1.2. – First Component Covariance Loadings or "Spread" Statistics for Downtown Los Angeles

First component covariance loadings (or the "spread" metric) represent cross product calculations of the individual year's daily max/min's with their corresponding first component "factor" scores (see Figure 2). In the case of Downtown Los Angeles, highest covariance loading (+12.554) was produced for 1949, the lowest (+9.656) for 1928. Mean loading figure for the 96 years was +10.88 F. Correlation of the covariance loading statistics with annual mean temperatures was +.137, that with the average daily ranges an appreciably positive +.654, and that with the correlation loading statistics +.409, the latter indicative that "shape" and "spread" are not mutually exclusive measures in this application. Figures 7 and 8 show the actual daily max/min temperature patterns for these years.

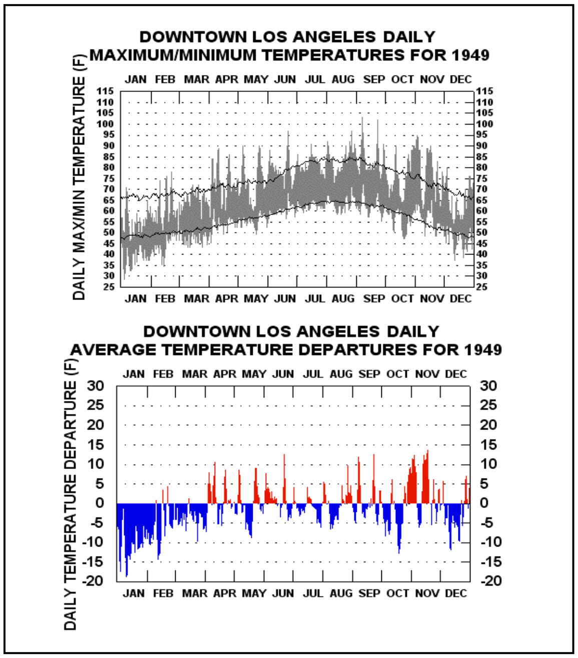

Figure 7: Daily Max/Min Pattern for Downtown Los Angeles (1949) – Highest First Component “Spread” Statistic (+12.554)

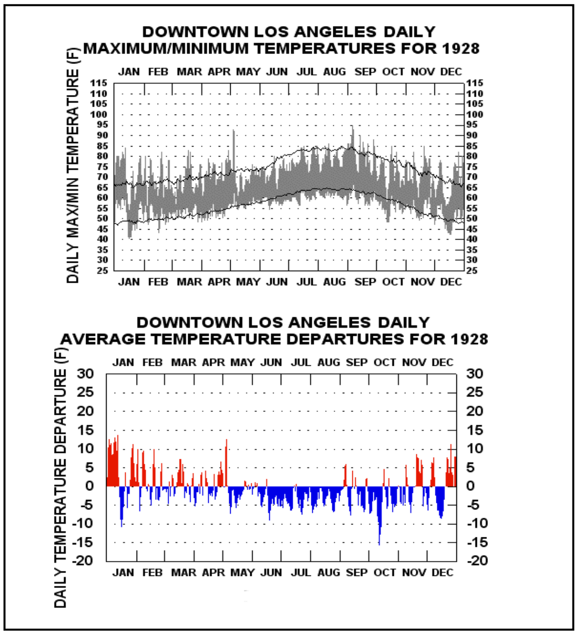

Figure 8: Daily Max/Min Pattern for Downtown Los Angeles (1928) – Lowest First Component “Spread” Statistic (+9.656)

As previously discussed, covariance loading stats are influenced by diurnal as well as seasonal contrasts.

The year 1949 displays a pronounced (for Los Angeles) seasonality, set up most dramatically by the long spell of colder than normal temperatures covering January thru March. In addition, average daily temperature range for the year as a whole was high, the mean figure (19.9 F) the third highest in the record. The year 1928, in contrast, displayed very modest (even for Los Angeles) seasonality, daily temperatures for much of January, for example, warmer than much of the summer, with winter and most of Spring warmer than average compared to summer which was predominantly cooler than climatology. Average daily temperature range (18.2 F), however, was about average.

Figure 9 is a time-series graph of the Downtown Los Angeles first component covariance loading statistics by year. Again, there is a tendency for lower magnitudes over the early years of the history, in this case through about 1945.

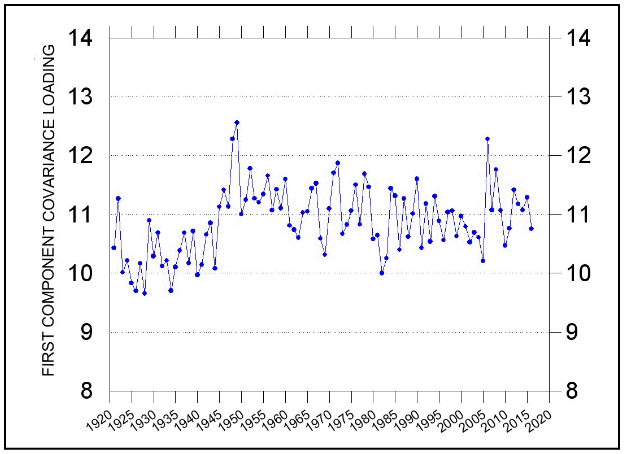

Figure 9: Time-Series Plot of Downtown Los Angeles First Component Covariance Loadings (“Shape”) Statistics (1921-2016 data – Mean: +.10.88)

4.1.3. – Combined First Component “Shape” and “Spread” (2-D Confidence Ellipsoid) Assessments

Next, as a summarization device, the Downtown Los Angeles shape and spread statistics, by year, were plotted in a 2-D confidence ellipsoid graph to isolate and identify the most aberrant configurations in the bivariate sense (see Figure 10). Through a trial and error process, confidence level ellipsoids of progressively higher significance levels were overlain until just one of the 96 points remained completely beyond the ellipsoid bound, this pinpointing the year with the most "anomalous" pattern. For Downtown Los Angeles, the year 1971's point was the only one completely outside the boundary, is this case at the 99th percentile level. Its shape measure, +.853, was just a shade higher than 1930's +.851, its spread measure, 11.704, the 7th highest. Evident from the chart, though, 1971's outlier position was only slightly more so than that of 1949.

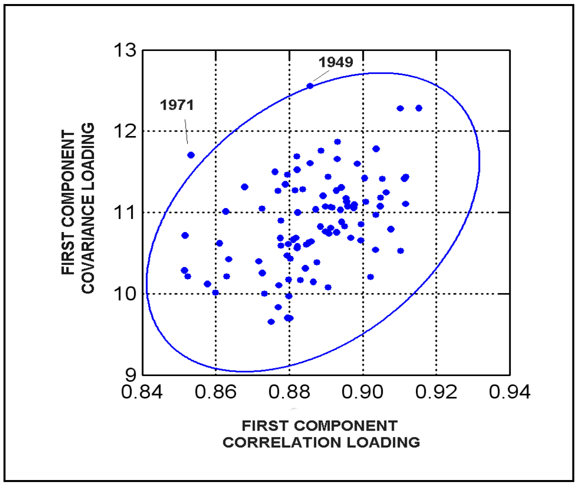

Figure 10 – 2-D Confidence Ellipsoid Plot for Downtown Los Angeles First Component “Shape” and “Spread” Statistics

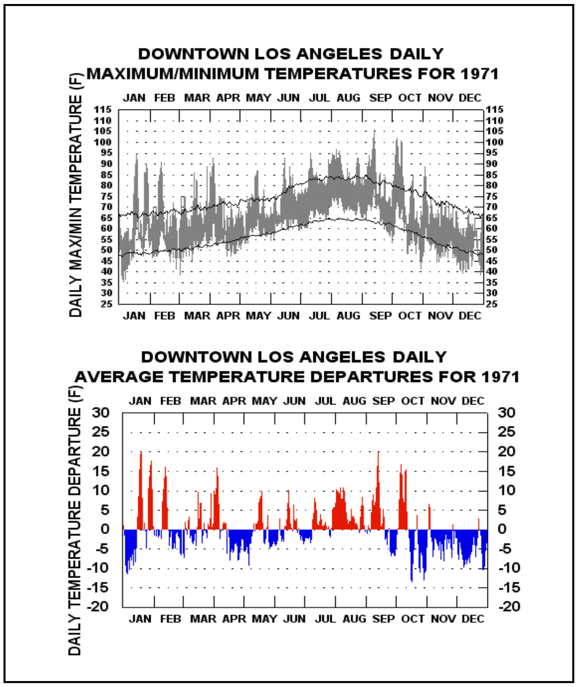

Figure 11 - Daily Max/Min Temperature Pattern for Downtown Angeles (1971). Most Anomalous Pattern in the joint “Shape”/”Spread sense.

Figure 11 is a plot of 1971’s daily max/min temperature pattern plus the daily mean departures. It displays a very irregular configuration of daily max/min temperatures, including short-period above average “spikes”, especially over late-January through mid-February, and early September plus early October. The latter was followed by a long spell of predominantly colder than average temperatures from mid-October through year-end.

4.1.4 – Higher Level Component “Shape” and “Spead” Statistics for Downtown Los Angeles

In addition to the overwhelmingly predominant seasonal signal, as represented by the first component scores in Figures 1 to 3, the analysis examines additional components which exhibit eigenvalues greater than or equal to 1.0 (satisfying the "original variable" criterion), or alternatively, just below 1.0 but still close enough to contain the 1.0 threshold within a confidence level band. Eigenvalue information produced by a PCA routine is of course merely a sample statistic based on an observational period-of-record, so it’s possible that a given component with a sample eigenvalue not quite as high as 1.0 could be acquire “original value” status based on a confidence interval consideration.

While by this principle, such lesser qualifying components are considered statistically "significant", compared to the first component, their scores are not interpretable in as strictly a meaningful a way. Indeed, occasionally some days’ scores for the maxima are less than those for the minima. In addition, their year-to-year correlation and covariance loadings' figures can be of either sign and relatively low in absolute magnitudes, reflecting the fact that the component's most distinctive loadings’ enhancing calendar days encompass only a relatively small sub-portion of the calendar year, and in the particular case of the covariance loadings, the positive or negative signs reflect the possibility that the particular sub-portion can be either of below or above average temperature character.

Still, the scores do provide useful means-to-an-end information, identifying in a semi-rigorous way contiguous sub-periods of significant length that may be susceptible to real, minor modes of variation, subsidiary to the seasonal cycle. The 2-D confidence ellipsoid method is applicable here also in flagging those individual years that express these modes in the most pronounced fashion.

In the case of Downtown Los Angeles, the first eigenvalue was 69.345, the second, 0.923, the latter by virtue of its proximity to 1.0 a potential “original variable” through confidence interval consideration.

Larson and Warne [2010] provide a formula for determining an eigenvalue confidence level band, given three pieces of information: the sample eigenvalue magnitude, the sample size, and a selected z-value. Figure 12 shows the expression:

$$l_i \pm z^* \left(\sqrt{\frac{2l_i^2}{n}}\right).$$

Figure 12 – Formula for constructing Confidence Interval (CI) about a true Eigenvalue (after Larson and Warne, 2010)

In the Figure 12 formula, ***l**_i* is the observed eigenvalue, ***z**** the chosen z-value, and ***n*** the sample size. Applying this with a nominal 99% one-tail level of confidence (2.33 z-value), the upper-level boundary is calculated at 1.233; thus the 1.0 critical value is within the confidence band, and the 2nd component is utilized. The third component's eigenvalue for Downtown Los Angeles (0.679), when substituted into the formula at the same level of confidence falls short, the upper one-tail bound only reaching 0.907.

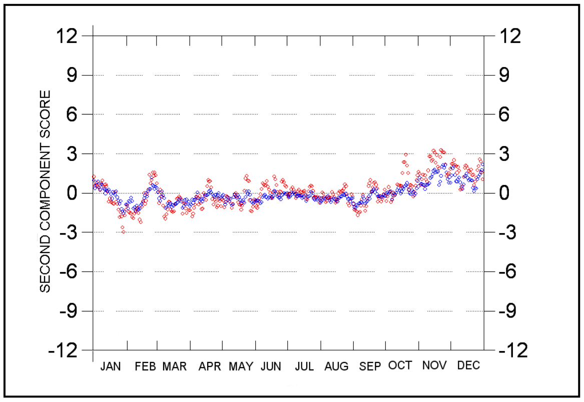

Figure 13. Second Component Daily Max/Min PCA Scores for Downtown Los Angeles (1921-2016 data)

Figure 13 is a plot of the Downtown Los Angeles second component scores. Compared to their first component counterparts in Figure 1, the Figure 13 scores are much lower in absolute magnitude and for the most part hover close to zero. The only major extended period exception is the tendency for a long run of late October thru year-end stats being of the same sign (positive), suggestive of a secondary minor mode of variability apart from the seasonal cycle.

The relative insignificance, however, of this second mode is further represented by Figures 14 and 15, time-series plots of the correlation loading ("shape") and covariance loadings ("spread") statistics, by year.

In Figure 14, the correlations range from -0.229 to +0.249 (mean: 0.00) and in Figure 15 the covariances vary from -3.330 to 3.328 (mean: +0.012). Curiously, the two time-series are virtually undistinguishable in pattern, the correlation coefficient between the two a very high; +.993; in the case of the first component this was only +.409. Thus, for the second component application, the 2-D confidence ellipsoid will be very narrow and elongated in form.

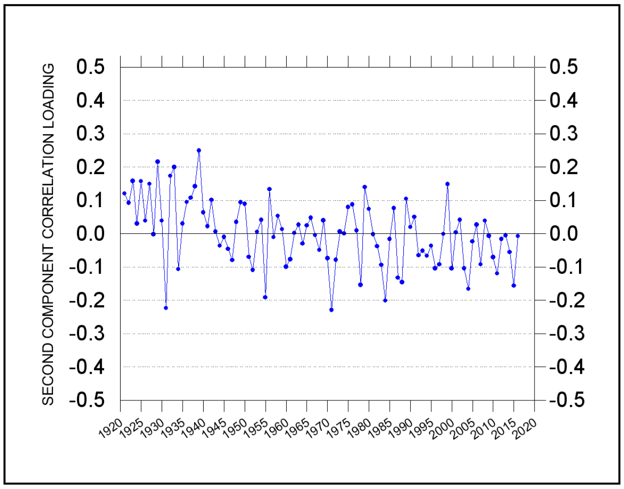

Figure 14: Time-Series Plot of Downtown Los Angeles Second Component Correlation Loadings ("Shape") Statistics (1921-2016 data – Mean: 0.00)

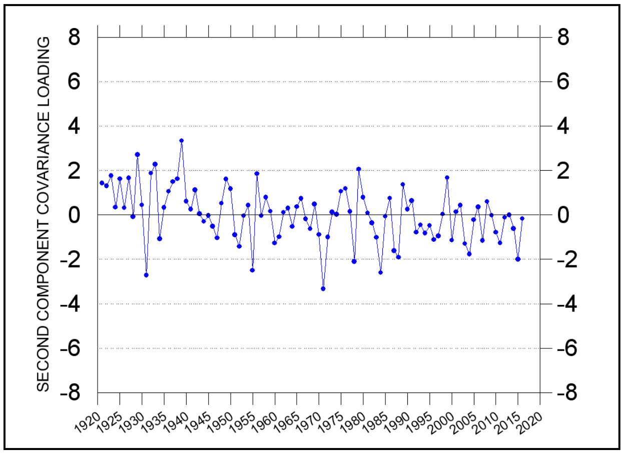

Figure 15: Time-Series Plot of Downtown Los Angeles Second Component Covariance Loadings ("Spread") Statistics (1921-2016 data – Mean: +0.12)

4.2.2 – Extreme Downtown Los Angeles Configurations Associated with the Second Component “Shape” and “Spread” Statistics

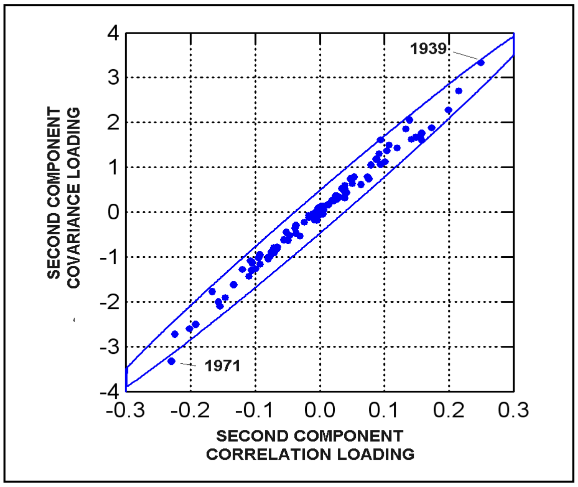

Figure 16 – 2-D Confidence Ellipsoid Plot for Downtown Los Angeles Second Component “Shape” and “Spread” Statistics

Repeating the iterative 2-D elimination process disussed in section 4.1.3, the year 1971 again (Figure 11), is isolated out as a most extreme outlier beyond the confidence ellipsoid boundary, this time at the .995 level (see Figure 16). The “extreme” nature of 1971’s position on the chart reflects it is extreme in an analogous way to 1948, which was unusual in its conformance to the seasonal cycle pattern of climatological mean daily max/min temperatures.

Comparing its configuration with the score magnitudes in Figure 13, the long stretch of below normal temperatures covering late-October thru year end conforms well in form if not sign with the scores, and hence, its covariance loading is negative (-3.330).

A case that illustrates high positive loading (although its point is slightly inside the confidence boundary) is that for 1939 (see Figure 17). This depicts a major heat wave covering mid to late September, succeeded by a predominantly above normal temperature pattern from mid-October through year-end; its covariance loading statistic is +3.328.

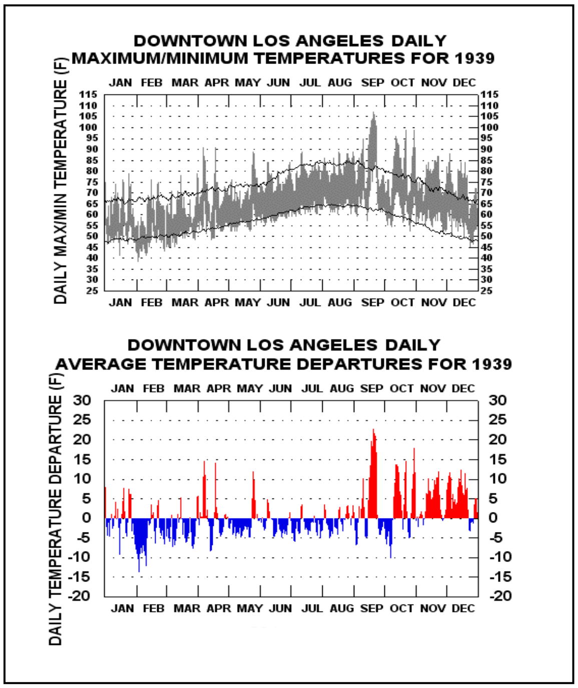

Figure 17- Daily Max/Min Temperature Pattern for Downtown Los Angeles (1939).

4.2. – San Diego

San Diego’s temperature record extends back to 1875, reflecting observations made at various locations downtown until 1927 when Lindbergh Field became the official source.

First component correlation loadings (“shape”) results showed some contrasts with those of Downtown Los Angeles, the highest figure +.927, the lowest +.822; mean overall statistic for the 142 years was +.891. The correlation with annual means was r= -0.053, similarly reflecting essentially a “transparent” association, but that with average daily ranges was elevated to a more positive (r=+.572).

First component covariance loadings (“spread”) statistics ranged from 6.415 to 10.465, the mean +8.44, considerably lower than Downtown Los Angeles’ +10.88, probably relating to the greater proximity of San Diego to the Pacific Ocean. Correlation with the annual means was r=+0.095, that with the mean daily ranges a significantly higher r=+.843.

4.2.1. – Combined First Component “Shape” and “Spread” (2-D Confidence Ellipsoid) Assessments

Repeating the procedure followed for Downtown Los Angeles, the 142 first component shape and spread statistics, by year, were plotted in a 2-D confidence ellipsoid graph to flag the year with most “anomalous” combined shape/spread configuration. Following iterations to the .995 level, the year 1899 was isolated out. Figures 18 and 19 display the 2-D confidence ellipsoid graph and the daily temperature floating bar plot for 1899, respectively.

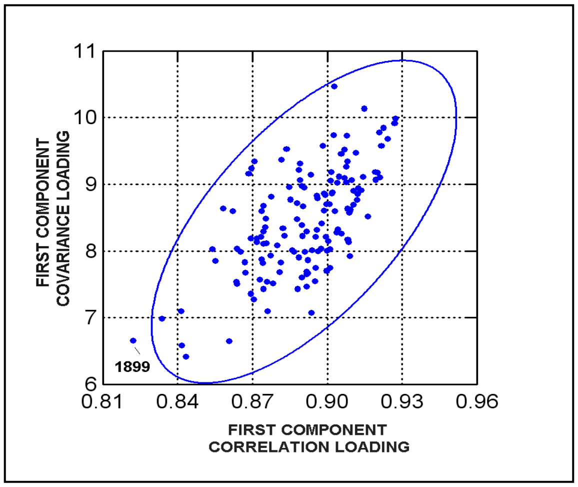

Figure 18 – 2-D Confidence Ellipsoid Plot for San Diego First Component “Shape” and “Spread” Statistics

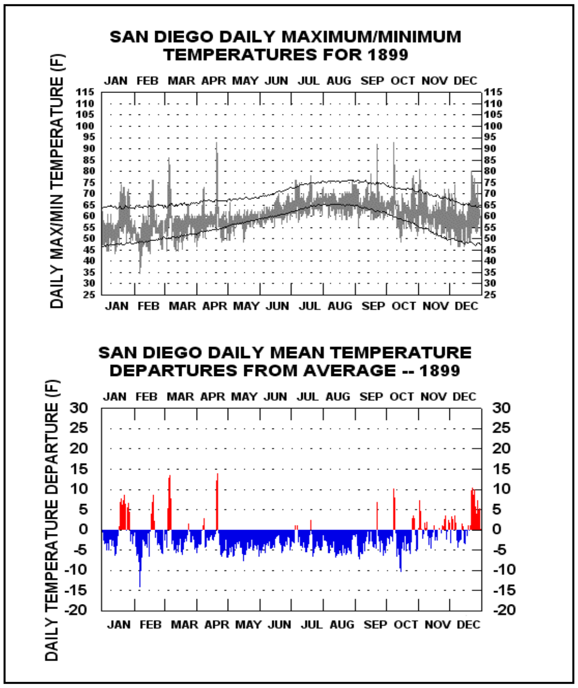

Figure 19 - - Daily Max/Min Temperature Pattern for San Diego (1899).

As reflected in the steeper positive orientation of the ellipsoid in Figure 18), the association between the correlation and covariance loadings for San Diego is significantly greater than Downtown Los Angeles, the former’s correlation between the two measures: r=+.647, compared to the latter’s r=+.409.

From Figure 19, the year 1899 was characterized by cool temperatures overall (ninth coolest at 60.2 F) accompanied by consistently low daily temperature ranges (second lowest mean annual figure at 10.65 F). There were a few significant warm spikes scattered through the year, none, however, after mid-April or before mid-September. From a configuration standpoint, this produced a very low shape statistic (r=+.822), lowest in the record, and the fourth lowest spread metric (6.656).

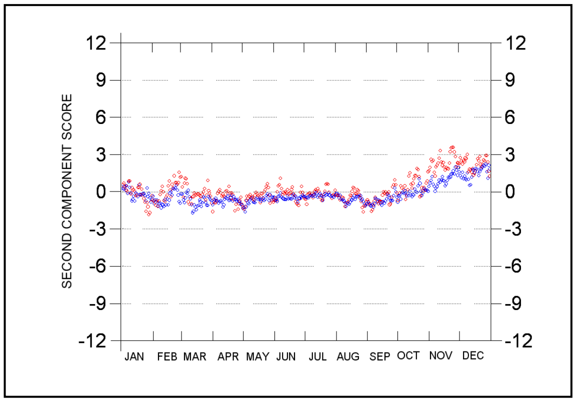

Figure 20. Second Component Daily Max/Min PCA Scores for San Diego (1875-2016 data)

The San Diego PCA output also identified the second eigenvalue as an “original” value outright, given its magnitude of 1.102. Interestingly, the scores’ pattern in Figure 20 above resembles those very closely with Los Angeles’ in Figure 13, exhibiting the late year uptrend in positive magnitudes; in fact, the overall correlation between the two is r=+.818 (the correlation between the first component scores is r=+.984). As will be recalled, Los Angeles’ second component information was rendered “usable” only through confidence interval interpretation and resulting elevation to “original value” status. This raises the point that with a larger sample size (San Diego’s period of record was more than 40% longer than Downtown Los Angeles’) the likelihood of higher order components reaching the eigenvalue threshold of one is enhanced.

Figure 21 below is a plot of the daily max/min temperatures for the year 1900, that which through the 2-D ellipsoid methodology was identified as the most conforming to Figure 20’s idealized configuration. The graph (not shown) had a narrow, elongated ellipsoid, similar to that shown in Figure 16. Correlation between the second component “shape” and “spread” statistics depicted in the graph was r=+.981.

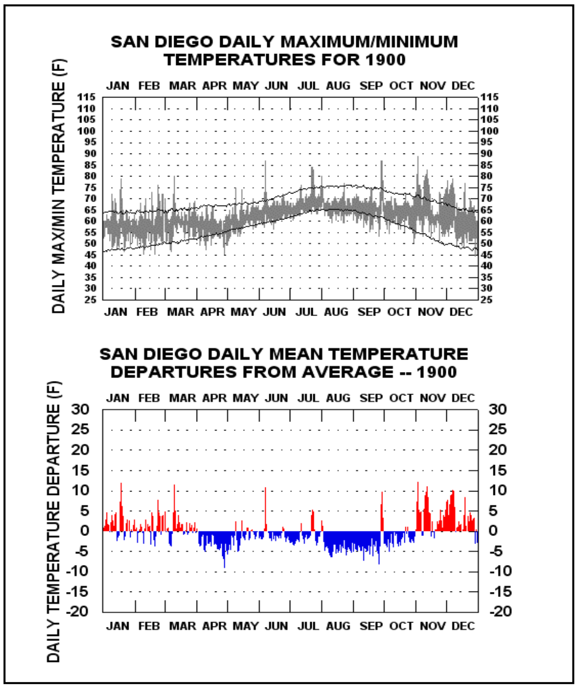

Figure 21 - - Daily Max/Min Temperature Pattern for San Diego (1900)

4.3. – Downtown San Francisco

Downtown San Francisco's available digitized period of record, like that for Downtown Los Angeles, dates back to 1921. First component correlation loadings ("shape") results were somewhat lower than either the Los Angeles or San Diego, the highest figure +.895, the lowest +.754; mean overall statistic for the 96 years was +.834. Linear correlation with the annual means was r= +0.361, reflecting a significantly positive association not seen with Los Angeles or San Diego - that with the average annual daily ranges also significantly positive at (r=+.534).

First component covariance loadings ("spread") stats ranged from 6.056 to 9.567, the mean +7.299, lower than either Downtown Los Angeles or San Diego. Correlation with the annual means was r=+0.541, and that with the mean daily ranges an exceptionally high r=+.923.

On a bivariate basis, the first component shape and spread statistics' correlation was r=+.624, slightly less than San Diego's r=+.647.

4.3.1. – Combined First Component "Shape" and "Spread" (2-D Confidence Ellipsoid) Assessments

The Confidence ellipsoid, iterative plotting approach identified, at the .990 level, the year 1970 as having the most anomalous combined shape/spread configuration for San Francisco (See Figure 22), its point on the graph being well outside the ellipsoid boundary. Figure 23 shows the year's max/min temperature pattern, the configurations reflecting an above average January to March, followed by a predominantly cooler than average temperatures from April through October, punctuated over the latter by brief spikes of above average temperatures, including a relatively pronounced spell of these in late September.

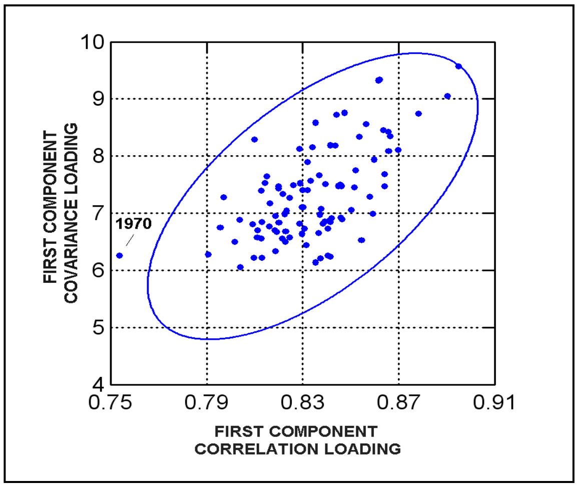

Figure 22 – 2-D Confidence Ellipsoid Plot for Downtown San Francisco First Component "Shape" and "Spread" Statistics

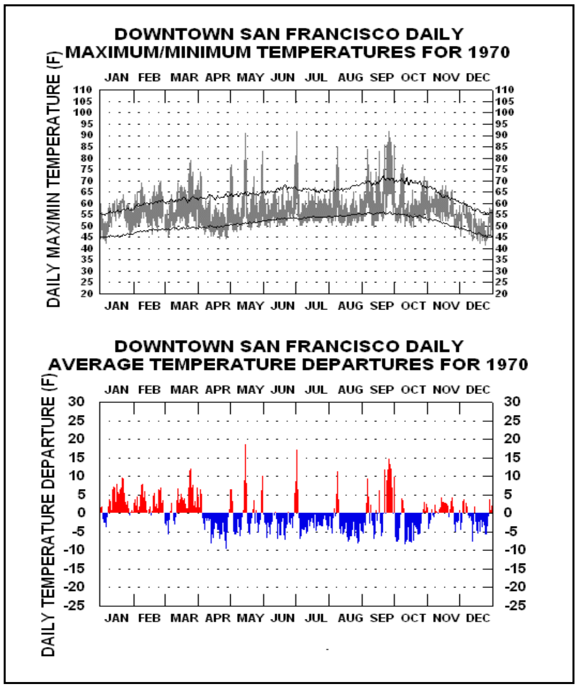

Figure 23 - - Daily Max/Min Temperature Pattern for Downtown San Francisco (1970)

Mean annual temperature for 1970 (56.9 F) was only slightly color than 1921-2016 climatology, the average daily range (11.312 F) lower than climatology (12.521 F), but ranking only as the 22nd lowest.

4.3.2. – Higher Order Component Extreme Patterns for Downtown San Francisco

Inspection of the PCA eigenvalue output and application of the confidence interval formulation at the .01 level of significance determined that no less than seven modes qualified as “original variables”, the 1.0 threshold value lying within each of their confidence interval bounds. Component two’s magnitude (1.124) qualified it outright, the remaining figures’ original values, in order of importance, 0.907, 0.836, 0.804, 0.790, and 0.759. Confidence ellipsoid treatment ascertained the year 1970, already identified as component one’s most prominent outlier, as the most extreme for the sixth component as well, not unlike the Los Angeles distinction for the year 1971, which was the most anomalous for both the first and second components. For brevity’s sake, daily max/min plots for the second through fourth outliers only (Figures 24 through 26, respectively) appear below.

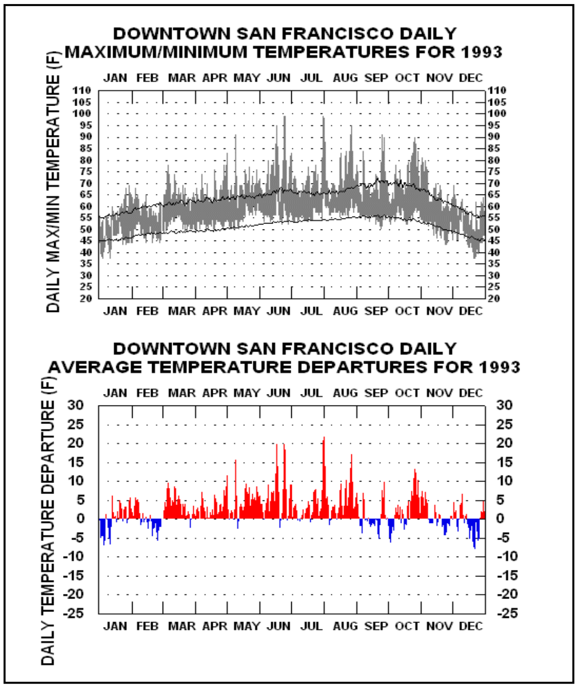

Figure 24 - - Daily Max/Min Temperature Pattern for Downtown San Francisco (1993) – Second Component most extreme “Spread”/”Shape” configuration

The year 1993 displayed consistently above normal temperatures (annual mean: 59.94 F the 5th highest on record) with large diurnal variability (mean daily range: 15.66 F, the 2nd highest, compared to 12.5 F for climatology). There was scarcely a day over March through August that was not above average, with frequent “spikes” of high daily maxima, quite uncharacteristic for San Francisco during those months.

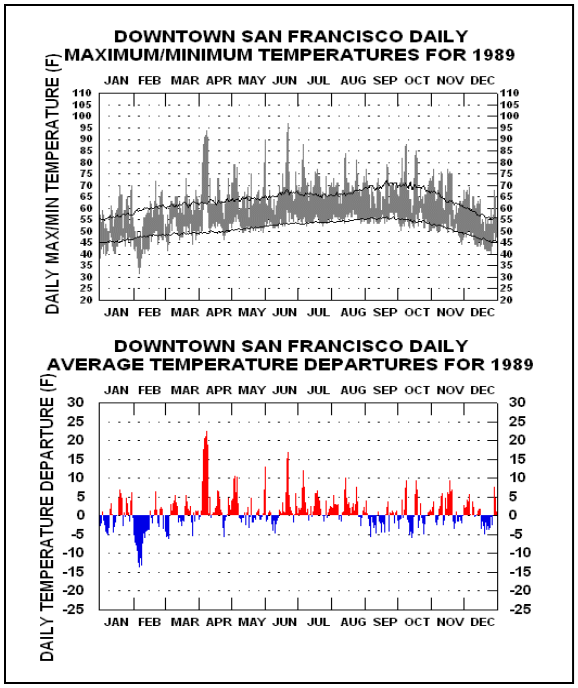

Figure 25 - - Daily Max/Min Temperature Pattern for Downtown San Francisco (1989) – Third Component most extreme “Spread”/”Shape” configuration

The year 1989’s most noticeable feature was the two spells of much below normal temperatures covering the first half of February, succeeded by a week-long spell of far above normal readings in early April. Overall, the year was modestly above average in annual mean temperature (mean: 58.3 F: climatology, 57.4 F). There was also relatively high diurnal variability (overall average daily range: 15.18 F, the 8th highest in the record).

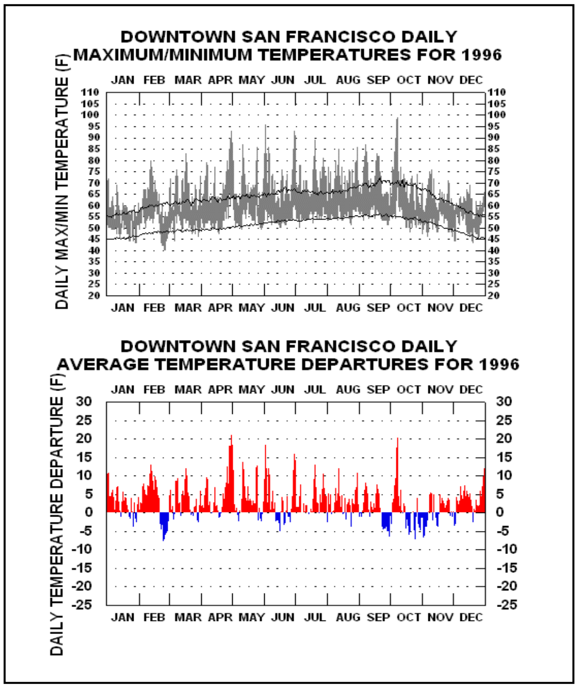

Figure 26 - - Daily Max/Min Temperature Pattern for Downtown San Francisco (1996) – Fourth Component most extreme “Spread”/”Shape” configuration

The year 1996 (see Figure 26) stands as the warmest year of the San Francisco history, the annual mean temperature: (60.28 F) nearly 3 F above climatology (57.4 F). Annual average daily range (15.51 F) was the 4th highest on record.

In summary, viewing Figures 23-26 as a group, it appears subjectively that Figure 23 is the most irregular, Figures 24 and 26 somewhat alike with the high daily ranges, but Figure 25 rather unremarkable except for the two short-term pronounced-in-anomaly spells.

4.4. – A few selected other stations (First Component results) only

The Los Angeles, San Diego, and San Francisco stations are similar in that they are close to the Pacific Ocean, the marine influences of which, predominant over much of the year, can lessen seasonal and diurnal temperature variability.

Thus, as a switch of focus, the remainder of the exploratory analysis explores results from three stations, the first just slightly inland from the Pacific coast (Fresno, in the California Central Valley), the second, in a highly continental more northerly portion of the U.S.: Minneapolis-St. Paul, MN, and the third an Atlantic coastal, subtropical station – Miami, Fl.

Fresno’s available digitized record was available back to 1931. Reflecting a higher amplitude in temperature variation through the year, first component correlation loadings (“shape”) results were higher than all of the three coastal California stations, the highest figure +.957, the lowest +.926; mean overall statistic for the 86 years was +.946. Linear correlation with the annual means was r= +0.064, that with the average annual daily ranges (r=+.342).

First component covariance loadings (“spread”) statistics were much higher than the previous three coastal stations, ranging from 16.614 to 20.045, the mean +18.387. Correlation with the annual means was actually negative (r=-0.331), and that with the mean daily ranges highly positive: r=+.754. The second eigenvalue’s magnitude was only .392, no indication of any chance of inclusion as an “original variable” through confidence interval consideration. On a bivariate basis, the first component shape and spread statistics’ correlation was r=+.459.

The 2-D confidence ellipsoid iteration methodology marked the year 1998 with the most irregular pattern, the daily max/min configuration chart appearing below as Figure 27. From the chart, the absolute contrasts in anomalies through the course of the year and the station’s climatological inclination to experience more variability through the year, both seasonal and diurnal, produces a rather striking anomaly impression. Annual mean temperature in Fresno for 1998 (62.4 F) was 1.4 F below average, the average daily range statistic (22.4 F) the lowest on record, some 3.6 F below climatology.

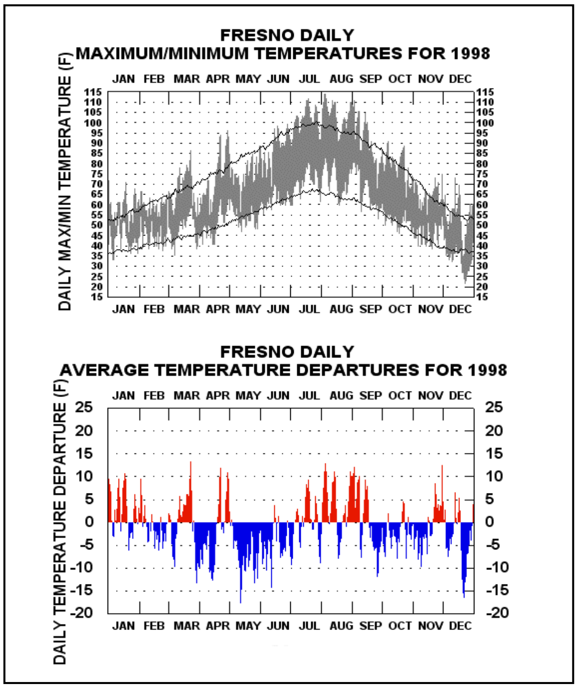

Figure 27 - - Daily Max/Min Temperature Pattern for Fresno, California (1998) – First Component most extreme “Spread”/”Shape” configuration

Moving inland to a highly continental location, the daily temperature record of the Minneapolis-St. Paul, MN area is next analyzed. Digitized observations are available back to 1873, comprising St. Paul recordings through 1890, Downtown Minneapolis into early 1938, and the International Airport thereafter. First component correlation loading results for the 144 years, ranged from +.897 to +.948 with the overall mean +.918. Corresponding covariance loadings varied from +18.959 to +27.262, the overall average +23.09; linear association between the two was r=+.644.

The confidence ellipsoid approach identified 1936 as the primary outlier year, and its daily max/min temperature pattern is depicted in Figure 28. Evident is the great seasonal contrast in temperatures, including protracted cold in January/February followed by the warmest July in history.

Again, the greater natural seasonal variability of this station (significantly greater than Fresno) helps makes for a rather striking display in an extreme case such as this one.

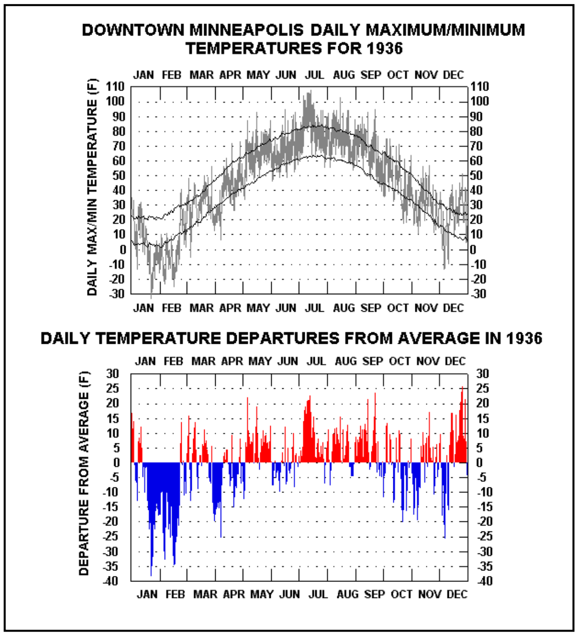

Figure 28 - - Daily Max/Min Temperature Pattern for Minneapolis-St. Paul, MN (1936) – First Component most extreme “Spread”/”Shape” configuration

Finally, moving southeast to Florida, the daily temperature records, by year, are considered for Miami International Airport, these available from 1949. First component correlation loading results for the 68 years ranged from +.829 to +.919 with the overall mean +.882. Corresponding covariance loadings varied from +7.501 to +10.953, the overall average +9.029; linear association between the two was r=+.385.

The most extreme pattern was identified via the confidence ellipsoid method as that for 2010 (see Figure 29). The graph shows a steady, uniform succession of daily temperatures a bit above average from April through October, flanked however by fluctuating spells of both below and above normal temperatures over January to March and September through December, the cold spells more pronounced. The first component covariance loading statistic for 2010 (10.953 F) was the highest of the 68-year history, the annual mean for the year (75.8 F) the coolest for Miami over the 1986-2016 period.

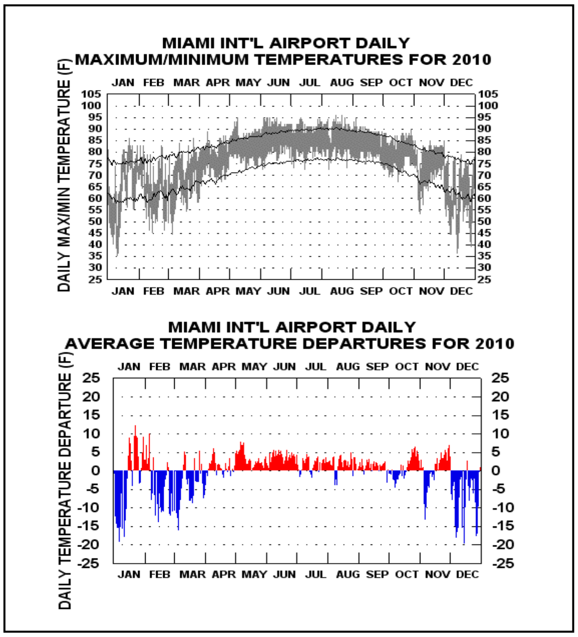

Figure 29 - - Daily Max/Min Temperature Pattern for Miami Int’l Airport (2010) – First Component most extreme “Spread”/”Shape” configuration

5. SUMMARY AND CONCLUSION

Utilizing non-rotated Linear PCA as an analytical tool, the purpose of the foregoing was to characterize and differentiate statistically, year-by-year calendar-day maximum/minimum temperature patterns, for six stations, in terms of two attributes, correlation loadings or "shape", and covariance loadings or "spread", respectively. Graphed on scatterplots, the yearly point representations were fitted to confidence ellipsoids at progressively higher levels of significance, until a single point only remained outside the ellipsoid boundary, this signifying the year with the most "anomalous" pattern for the particular component being considered. The extreme-most patterns were then plotted in floating-bar form. The requirement was that each component should have an eigenvalue of at least 1.0, either as an outright calculation, or with the 1.0 magnitude contained within a confidence interval for eigenvalues at the .01 level of significance.

This somewhat unconventional PCA approach, utilized for two previous similar type studies, was originally motivated after exploratory/experimental investigations involving daily maximum and minimum temperature data determined that unrotated PCA first component correlation and covariance loadings statistics could serve as a means of distinguishing year-to-year temperature patterns in terms of the above attributes. The PCA software's capacity to rapidly process rectangular matrix data made it an attractive option, as well the supplementary information provided in the form of output files and reports, including those on higher order components should they be of interest (as in the present investigation).

Subsequently it was discovered that these first component loadings metrics (by far the most prominent) were in reality perfect analogs of the more conventional correlation and covariance statistics, the correlations and correlation loadings equal in magnitude, as were the covariances vs. covariance loadings, the latter relationship holding true if the array of reference climatological statistics was expressed in standardized units. So aside from the software benefits described above and the capacity to examine higher components, the unrotated PCA method turned out to be an efficient if not slightly roundabout method to characterize daily max/min temperature patterns in terms of "shape" and "spread" at the climatological mean (or first component) level.

Since PCA is an advanced tool, perhaps not likely available or understood to a great depth by many, it should thus be said in conclusion that if extreme pattern identification was desired only at the first component level, a viable alternative option to a PCA would be a simple correlation and covariance analysis. All that would be required would be creation of an array of climatological daily maxima and minima statistics, standardized as a single unit, for the calendar period of interest, this to be matched up in a simple correlation/covariance exercise with the individual years' data for the period of record. Standardization of the daily max/min's, not necessarily required, would however express the covariance loadings statistics in reduced-to-common-scale units, allowing for more meaningful comparisons with results from other stations. As before, the 2-D confidence ellipsoid approach would be a suitable subsequent step.

The above said, there is of course no absolute, definitive answer as to what constitutes a most "anomalous" pattern of daily max/min temperatures, the method utilized in this exploratory analysis just one of perhaps many formalized ways of evaluation. The correlation and covariance metrics, however, in addition to being well-known statistical measures, were pertinent and useful here. The first component or overall comparions to climatological means' results carried considerably more interpretative weight, as the associated eigenvalue magnitudes and percents of variance explained were at least an order of magnitude greater than those of the lesser modes, capturing irregularities that characterized an annual pattern in its totality, in contrast with the others, which highlighted configuration irregularities of a more subperiod nature.

One, however, could refine this unrotated PCA approach by introducing orthogonal rotation options, such as Varimax. The Varimax option would 1.) enhance the component-by-component number of near-0 and near-1 correlation loadings and 2.) reduce inter-component contrasts in covariance loadings' absolute magnitudes, in effect drastically reducing the overwhelming influence of the first component, and as a result creating a new set of loadings' statistics utilizable in identifying "extreme-most" patterns; these selections would be more or less transparent to original eigenvalue magnitudes and/or percents of variance explained.

Floating-bars can take on many different forms, the type utilized here (narrow with no spaces between adjacent days) seemed to capture the patterns effectively, the precise readings of individual days' temperature magnitudes of secondary importance. The accompanying daily mean temperature departure graphs also seemed to serve as a useful complement.

6. REFERENCES

Fisk, C.J., 2004: "Objective Identification of Extreme-Most Anomalous Daily Max/Min Temperature Patterns using Principal Components Analysis", 17th Conference on Probability and Statistics in the Atmospheric Sciences, American Meteorological Society – Seattle, 2004. http:/ams.confex.com/ams/pdfpapers/69198.pdf

Fisk, C.J., 2007: "Identification of Intra-Month Daily Mean Temperature Modes Using Principal Components Analysis", 16th Conference on Applied Climatology, American Meteorological Society – San Antonio, TX https://ams.confex.com/ams/pdfpapers/119632.pdf

Fisk, C.J., 2012: "Objective Identification of Extreme-Most Midnight-to-Midnight Hourly Historical Temperature Patterns Utilizing Principal Components Analysis", 24th Conference on Climate Variability and Change - New Orleans, 2012 https://ams.confex.com/ams/92Annual/webprogram/Paper192749.html

Larson, R., and Warne, R., 2010: "*Estimating confidence intervals for eigenvalues in exploratory factor analysis*", Behavior Research Methods, 42 (3), pp 871-6.

Yarnal, B., 1993: *Synoptic Climatology in Environmental Analysis, A Primer*, Bellhaven Press, 195 pp.