# EXTREME VALUE ANALYSIS OF K-MEANS SINGLE (K=1) CLUSTER STATISTICAL DISTANCES, A HEURISTIC APPROACH WITH LOS ANGELES AND SAN FRANCISCO HOURLY TEMPERATURE AND MONTHLY PRECIPITATION DATA

Charles J. Fisk *
Newbury Park, CA.

## 1. INTRODUCTION

K-Means clustering is a statistical method that partitions n observations into *k* clusters such that each individual case is assigned to the particular cluster in which its statistical distance to the centroid is a minimum. A K-means analysis will output, among other measures, each observation's statistical distance from its respective 'kth' cluster centroid, but such distances are only "local", relating to the particular cluster concerned, but what about consideration of overall statistical distances, those produced if the observations are not clustered (i.e., a "forced" k=1 treatment)? Departing from the usual objective of resolving and analyzing multiple clusters, the process of generating k=1 statistical distances can be fashioned heuristically into a "global" multivariate-type extreme-value analysis, taking advantage of the scale reduction standardizations that are a pre-processing step in a cluster analysis. Compared to a typical extreme-value analysis that makes use of one or two different physical variables (e.g., "copulas" for the latter), extreme-value results utilizing the scale-invariant statistical distances can identify atypical patterns comprising many variables. The "heuristic" description is meant to convey that the determination of models that best describe a given statistical distance variable's probabilistic character are done on a purely mechanical best-fit basis with no theoretical interpretations or assumptions made.

## 2. DATA AND PROCEDURES

To this end, an exploratory analysis is performed on two different climatological parameters for Los Angeles (KLAX) and San Francisco (KSFO) each 1.) daily 0000 LST-2400 LST hour-to-hour temperatures, by calendar month and 2.) July-June monthly precipitation totals (those for Downtown Los Angeles and San Francisco each). The former will be a 25-dimensional application, covering the period January 1948 through June 2023, the latter a 12 dimensional one, covering the July-June seasons 1876-77 to 2022-23 for Downtown Los Angeles, and 1849-50 thru 2022-23 for San Francisco Downtown. Results will include time-series graphs, probability density distribution fittings, and estimated return-periods. The Squared Euclidean method is chosen as the statistical distance metric after the standardizations..

---

*e-mail:cjfisk@att.net

A software package is employed that fits block-maximum type variables to more than 60 continuous probability density distributions with parameter counts ranging from one to six, and utilizing two goodness-of-fit techniques, the Kolmogorov\Smirnov and the Anderson-Darling, the models are ranked. The K-S is most sensitive to areas near the center of the distribution as opposed to the tails [1], hence it is somewhat less ideal to an extreme-value type analysis. However, the Anderson-Darling test, a modification of the K-S test, does give more weight to the tails, and in general it's considered a more sensitive test overall [2]. Consequently, the Anderson-Darling test's rank is chosen as the primary means of designating the best-fitting-model, assuming that the model's parameters are four or less. In most of these cases the K-S ranking is similarly high as well.

## 3. RESULTS

**3.1**. Los Angeles (LAX) Extreme-most Hour-to-Hour Temperature Patterns, by Calendar Month

As described above, the hourly temperature extreme value analysis utilized the block-maxima approach, each of the years' absolute largest statistical distance cases, month-by-month, assembled into data sets, preparatory to the curve fittings.

Figure 1 below is a tabular summary of the LAX calendar month results, comprising six columns. From left to right, Column 1 is the calendar month of interest, Column 2 the maximum statistical distance encountered for any one day for the calendar month, and Column 3 the actual date of the extreme-most distance. Column 4 identifies the best-fit model and Column 5 the approximate return period for the distance
Column 6 lists the sample size of days for the calendar month, followed in parenthesis by the mean statistical distance of all the daily observations, For example, January's extreme statistical distance, 4.848, was generated for 10 January 1949, an unusually cold day in the LAX vicinity with temperatures confined to the low 30's to mid 40's F. From the curve fitting, the Dagum 3P (i.e., Dagum- three-parameter) probability density distribution ranked highest, and analysis of its inverse cumulative distribution function at the 4.848 level enabled calculation of an estimated return period figure (Col. 5) of about 70 years. Figure 2 is the Dagum3P fitting histogram of the data, the small table inset within the body of the chart listing the model's parameter values. From Column 6, n=2356 total midnight to midnight January observations between 1948 and 2023

were processed, their mean statistical distance 0.444, only 9.1 % of the Column 2 extreme distance figure. Interestingly, the January extreme case being associated with colder than average temperatures for the day, was the only such calendar month of that distinction for LAX. All the other calendar month extreme-most cases related to warmer than average (offshore flow) episodes.

| (Col. 1) | (Col. 2) | (Col. 3) | (Col. 4) | (Col. 5) | (Col. 6) |
|---|---|---|---|---|---|
| MONTH | MAXIMUM | DAY/YEAR | BEST | RETURN | MEAN (N) |
| | K=1 | | FIT | PERIOD | |
| | DISTANCE | | MODEL | (years) | |
| | | | <= 4 PARAMETERS | | |
| JAN | 4.848 | 10 1949 | DAGUM 3P | ~70 | 0.444 (2356) |
| FEB | 4.676 | 28 2020 | GAMMA 3P | ~25 | 0.491 (2147) |
| MAR | 7.053 | 25 1988 | PEARSON 5 3P | ~110 | 0.365 (2356) |
| APR | 8.488 | 6 1989 | FATIGUE LIFE 3P | ~85 | 0.336 (2280) |
| MAY | 9.181 | 15 2014 | DAGUM 3P | ~55 | 0.300 (2356) |
| JUN | 8.925 | 16 1981 | DAGUM 3P | ~55 | 0.266 (2280) |
| JUL | 6.946 | 10 1959 | BURR (4P) | ~110 | 0.349 (2325) |
| AUG | 5.440 | 13 1994 | FATIGUE LIFE 3P | ~70 | 0.420 (2325) |
| SEP | 10.202 | 26 1963 | FATIGUE LIFE 3P | ~120 | 0.381 (2250) |
| OCT | 7.799 | 15 1961 | FATIGUE LIFE 3P | ~70 | 0.374 (2325) |
| NOV | 6.298 | 4 2010 | GAMMA 3P | ~95 | 0.442 (2250) |
| DEC | 4.969 | 30 1980 | FATIGUE LIFE 3P | ~85 | 0.414 (2325) |

**Figure 1:** Tabular Summary of LAX Extreme Value Distances, by Calendar Month, for the Hour-to-Hour Temperature Application

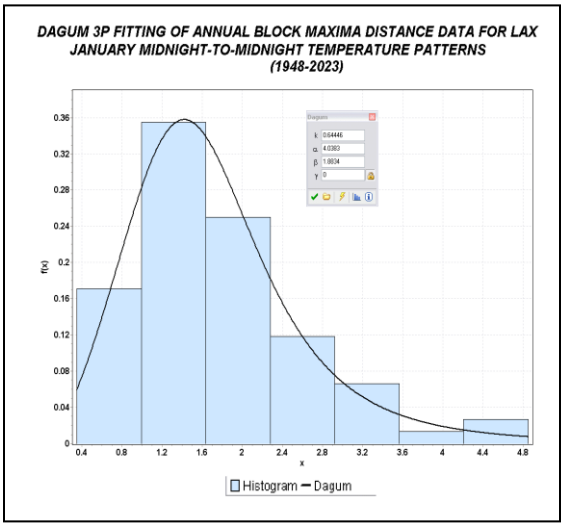


**Figure 2**: Histogram for Dagum 3P fitting of January Year-to-Year Block Maxima. Dagum 3P model parameter values contained in the table inset.

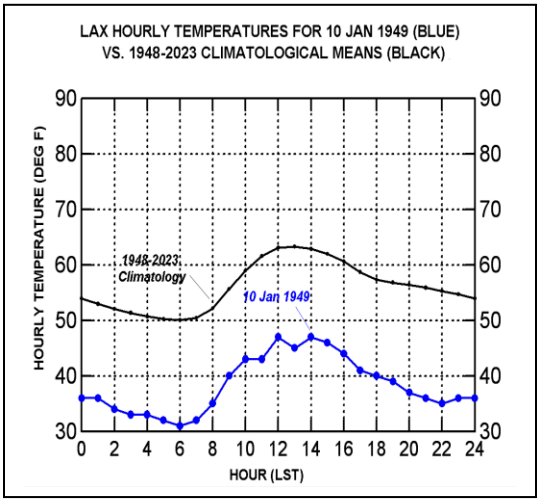


**Figure 3.** LAX Hourly Temperatures for 10 Jan 1949 vs. 1948-2023 Climatologiocal Means

Figure 3 above traces the actual hour-to-hour temperature progression (blue-trace) for 10 January 1949, compared to 1948-2023 climatological averages (black trace). In addition to being considerably lower than climatology, the amplitude is somewhat enhanced as well which also likely figured in the extreme statistical distance computation.

Since the statistical distance methodology utilized here involves reduction of each of the 25 hourly variables to a common scale with the same scaled standard deviations as well, Figure 4's chart of the standardized anomalies permits one-on-one hourly comparisons which identify those particular hours contributing most to the extreme distance magnitude. Early evening temperatures for 20 LST to 22 LST, inclusive, were each in excess of 4 standard deviations below average, not to mention the collection of near -3.5 magnitudes for the contiguous hours 00 LST to 08 LST. Thus, 10 January 1949's standing as the designated

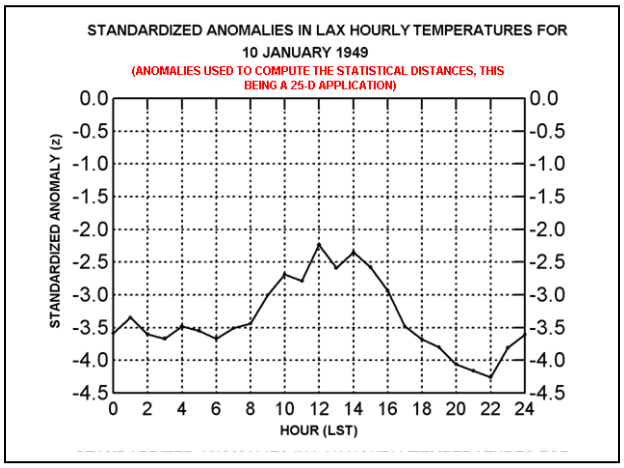


**Figure 4:** – Standardized Anomalies in LAX Hourly Temperatures for 10 January 1949

extreme-most January pattern covering 76 years' of history could be attributed most significantly to the nocturnal hours' anomalously cold temperatures.

Inspecting the other return-period statistics in Figure 1, most do not depart drastically from 75 or 76 years, the actual periods of record for the calendar months' data sets. Three do stand out to some extent, those for March and July with ~110 years each, and September ~120 years.

Figures 5 thru 7, respectively, show the actual hour-to-hour temperature progressions for these three (blue traces), compared to their respective 1948-2023 hourly climatological averages (black traces).

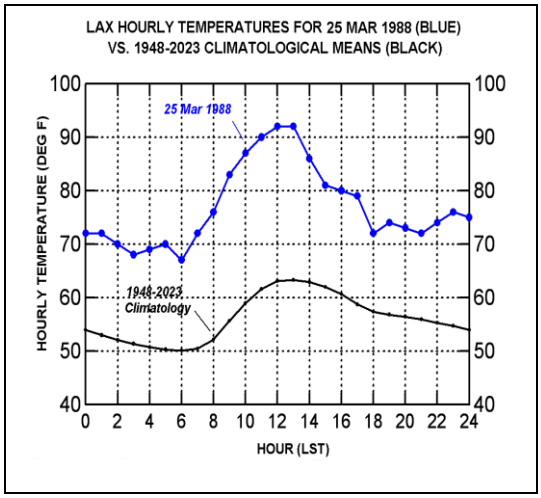


**Figure 5.** LAX Hourly Temperatures for 25 Mar 1988 vs. 1948-2023 Climatological Means

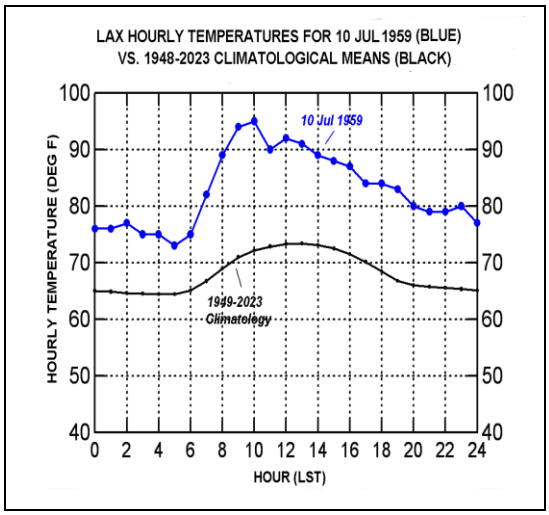


**Figure 6.** LAX Hourly Temperatures for 10 Jul 1959 vs. 1949-2023 Climatological Means

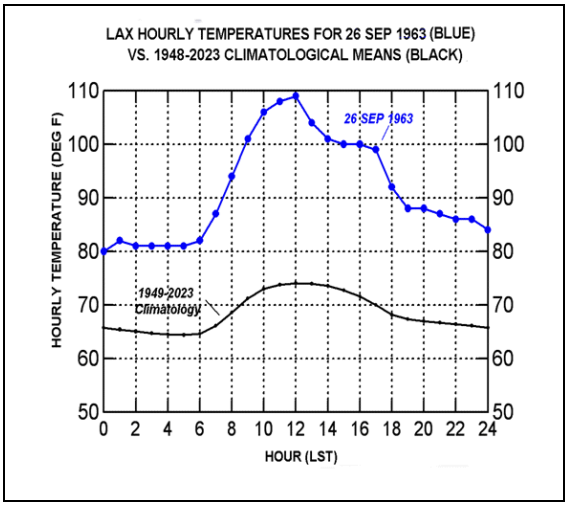


**Figure 7.** LAX Hourly Temperatures for 26 Sep 1963 vs. 1949-2023 Climatological Means

Figure 5 depicts an unusually late in the season extreme midnight-to-midnight Santa Ana episode for 25 March 1988, hourly temperatures reaching the low 90's F for the late morning into the early afternoon, nearly 30 F above average. The 91 F reading at 13 LST is 5.3 standard deviations above. Later in the day, past sunset, a slight upward trend set in from 18 LST (71 F observation) through midnight, the 75 F reading at 23 LST also 5.3 standard deviations above. These features were instrumental in producing the 7.053 maximum statistical distance figure for March.

Figure 6 shows the hourly temperature pattern for 10 July 1959, a day atypically devoid of the typical low stratus conditions that hinder diurnal temperature changes at this season, especially in nearshore area of LAX. Based on 1949-2023 climatology, July diurnal temperature ranges only average about 9 F, but on 10 July 1959 the range exceeded 20 F. Maximum temperature for the day (95 F) was reached early on at 10 LST (standardized anomaly: +5.6 z), several hours earlier than the norm, and the 83 F reading in the evening at 19 LST had a +5.3 z excess as well. The relatively low 6.946 extreme distance statistic is further indicative of the July climatological tendency for relatively small daily temperature ranges and uncommon occurrences of pronounced high readings,, which would otherwise "spike" the distance statistic upward.

Lastly, Figure 7 displays the diurnal pattern for 26 September 1963, an early season strong Santa Ana event, and one of the hottest days ever experienced midnight to midnight at LAX. At 9AM the temperature was already 106 F, and the 109 F reading at Noon was 6.4 standard deviations above the September norm for that hour (See Figure 8). The 10.202 distance statistic is noticeably higher than those for the March and July, but the return period only marginally so, indicative that similar but lesser in character far above normal warm

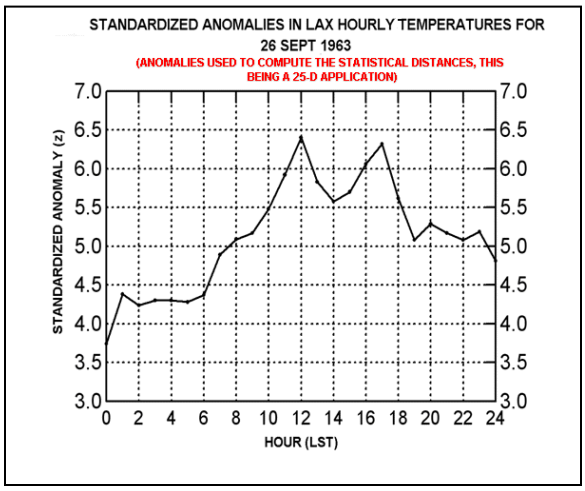


**Figure 8:** – Standardized Anomalies in LAX Hourly Temperatures for 26 September 1963

days are more common in September than March or July. The second ranking block maximum statistic for September, that for 1955's (7.314), exceeds both March's and July's top ranked figures (7.053 and 6.946, respectively).

It Is interesting to note overall, that eight of the twelve ranking best-fit results in Figure 1 were produced by just two 3-parameter models, the Fatigue Life (five best-fits) and Dagum 3P (three best fits). The remaining four were produced by the Gamma 3P (two best-fits), the Pearson 5 3P (one) and the Burr 4P (one). For what its worth, the Fatigue-Life Distribution, otherwise known as the Birnbaum-Saunders, was originally developed to model failure times due to cracks (3), the Dagum 3P (or Mielke Beta-Kappa) often used to model income and wealth distribution (4). The Burr (5) is also frequently used to model household income.

**3.2**. Downtown Los Angeles Extreme-most Month-to-Month Precipitation Patterns, by Season

Next, utilizing the same K=1 statistical distance approach, the extreme-most July to June month-to-month precipitation anomaly patterns for the 1877-78 thru 2022-23 period of record are identified and described.

This was more of a pure-pattern exploratory exercise, because in the reduction of the month-by-month data to a common scales, calendar months that normally receive very little or no measurable precipitation were given the same weight as those typically much wetter, this being counter-intuitive to some degree. This contrast was a minor factor with the hour-to-hour temperature data, as the variables were more alike, the analyses/interpretations still relatively "physical". Here they would be more abstract but possibly interesting in their own right.

| (Col. 1) SEASON | (Col. 2) MAXIMUM K=1 DISTANCE | (Col. 3) SEASON | (Col.4) BEST FIT MODEL <=4 PARAMETERS | (Col. 5) RETURN PERIOD (years) | (Col. 6) MEAN (N) |
|---|---|---|---|---|---|
| JULY-JUNE | 2.255 | 1883-84 | FRECHET (3P) | ~265 | 0.313 (146) |
| JULY-JUNE | 38.18 | 1883-84 | FATIGUE-LIFE 3P | ~110 | 14.74 (146) |
| JULY-JUNE | 3.21 | 2006-07 | FATIGUE-LIFE 3P | ~525 | 14.74 (146) |

**Figure 9:** Tabular Summary of Downtown Los Angeles Precipitation Statistics (1877-78 thru 2022-23 Period of Record)

Figure 9 is a tabular summary of the Downtown Los Angeles monthly precipitation pattern results. From Row 1, maximum extreme statistical distance, 2.255, in Column 2, was produced for the 1883-84 season, and the best ranked model, the Frechet 3-P (a bona-fide "Extreme-Value" one), produced an estimated return period of ~265 years (Col. 5). Mean statistical distance for the 146 cases (0.313 in Column 6), was less than one-seventh that of the absolute maximum statistic in Column 2.

The 1883-84 season's pronounced maximum distance statistic was attributable to the late season anomalously heavy rain falls of March, April, as well as June. That for June (1.39"), more than seven standard deviations above normal, obviously inflated the statistic (See Figure 10). The 146 seasonal distance statistics (See Figure 11) were positively correlated with the seasonal rainfall totals (r=+0.655).

.

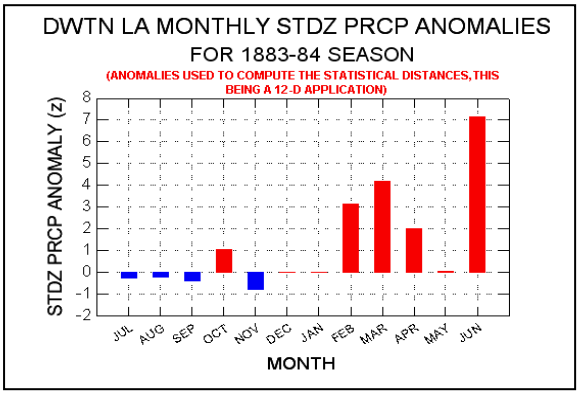


**Figure 10** – Downtown Los Angeles Month to Month Precipitation Anomaly Statistics for the 1883-84 Season

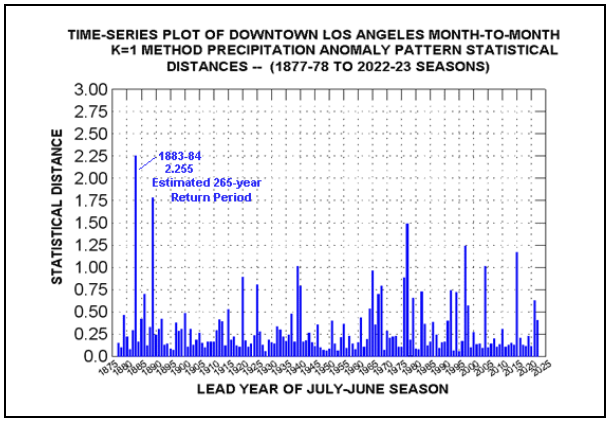


**Figure 11** – Time-Series plot of Downtown Los Angeles Month-to-Month Precipitation Anomaly Pattern Distance Statistics for 1877-78 thru 2022-23 Seasons

By way of comparison, the Downtown Los Angeles actual July-June seasonal totals for the 146 year history are fitted, the Fatigue-Life distribution again taking top ranking (See Figure 12 below), The 1883-84 season (38.18") being the wettest July-June season in history, the 2006-07 season (3.21") the driest, the Fatigue-Life inverse cdf produced a 110-year return period calculation to the former, a 525 year return period to the latter (See Column 5 in Figure 9).

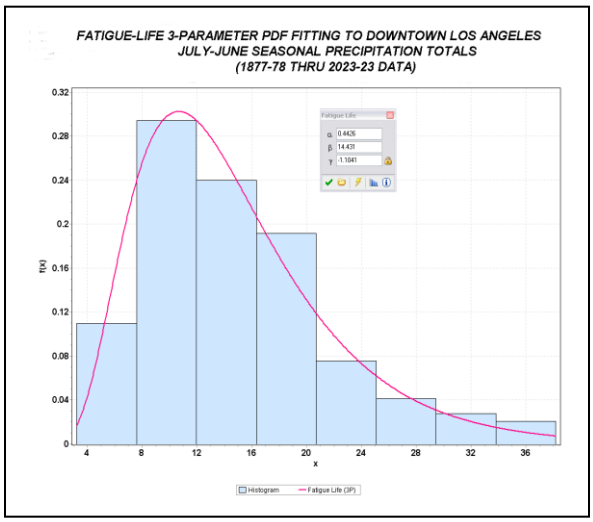


**Figure 12**. Histogram Fitting of Fatigue-Life 3P Distribution to Downtown Los Angeles July-June Seasonal Precipitation Totals (1877-78 thru 2022-23 Seasons)

**3.3**. San Francisco International Airport (SFO) Extreme-most Hour-to-Hour Temperature Patterns, by Calendar Month

Moving on to San Francisco International Airport (SFO), Figure 13 below is a tabular summary of its calendar month hourly temperature results (same format as LAX's in Figure 1). The 1948 to present SFO period-of-record was incomplete in that the hourly observations archives from 1969-1972 were in every third hour format, and therefore not usable for an hour-to-hour treatment like this one. The yearly period of record counts are thus reduced to either 71 (July-December) or 72 years (January-June).

For whatever reason, the selection of best fitting probability density distribution cases for SFO in Figure 13 was almost completely different than those for LAX, only the Pearson 5 3P appearing in each station's list (one case each). For SFO, the Log-Pearson 3 appears three-times (February, March, and April), all the others just once each.

Three of the twelve extreme cases were associated with cold days (January, February, and December) no doubt reflecting SFO's more northerly location, and somewhat unlike LAX, most of the calendar month return-period calculations were actually less than the actual periods of record (nine). The three exceptions

| (Col. 1) | (Col. 2) | (Col. 3) | Col. 4) | Col. 5) | (Col. 6) |
|---|---|---|---|---|---|
| MONTH | MAXIMUM | DAY/YEAR | BEST | RETURN | MEAN (N) |
| | K=1 | | FIT | PERIOD | |
| | DISTANCE | | MODEL | (years) | |
| | | | <= 4 PARAMETERS | | |
| JAN | 4.652 | 21 1962 | BURR (4P) | ~60 | 0.498 (2232) |
| FEB | 4.739 | 6 1989 | LOG-PEARSON 3 | ~60 | 0.459 (2304) |
| MAR | 5.169 | 16 2004 | LOG-PEARSON 3 | ~60 | 0.406 (2232) |
| APR | 7.241 | 30 2014 | LOG-PEARSON 3 | ~65 | 0.407 (2160) |
| MAY | 9.380 | 15 2008 | GEN. PARETO | ~70 | 0.383 (2232) |
| JUN | 9.270 | 14 2000 | INV. GAUSSIAN (3P) | ~50 | 0.376 (2160) |
| JUL | 6.704 | 31 1993 | LOG-NORMAL | ~70 | 0.289 (2199) |
| AUG | 7.326 | 1 1993 | LOG-LOGISTIC | ~40 | 0.358 (2201) |
| SEP | 8.529 | 2 2017 | GEN PARETO | ~215 | 0.442 (2130) |
| OCT | 6.456 | 12 2010 | PEARSON 5 (3P) | ~30 | 0.484 (2201) |
| NOV | 6.000 | 3 1950 | GEN. LOGISTIC | ~190 | 0.512 (2130) |
| DEC | 6.607 | 22 1990 | DAGUM (4P) | ~880 | 0.484 (2201) |

**Figure 13:** Tabular Summary of SFO Extreme Value Distances, by Calendar Month, for Hour-to-Hour Temperature Application

September, November, and December have exceptionally long estimated return periods and for that reason their patterns are selected for examination below.

Figures 14 thru 16 describe the extreme-most pattern experienced for 2 Sept 2017, including a best-fit histogram, one with the actual hourly temperature progressions, and one depicting the hourly standardized anomaly progressions.

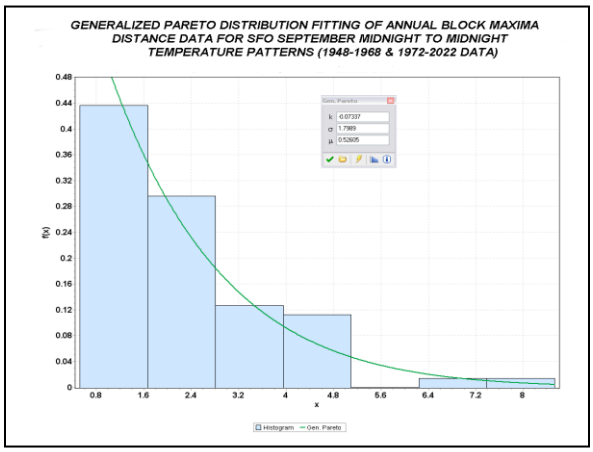


**Figure 14**: Histogram Fitting of Generalized Pareto Distribution to SFO Year-to-Year September Block Maxima Data

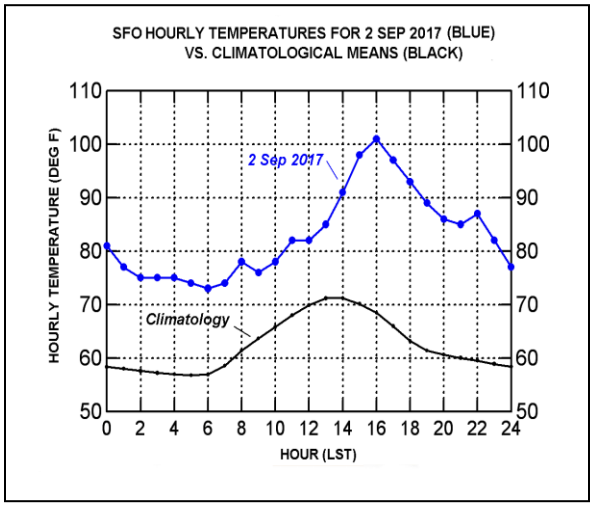


**Figure 15:** SFO Hourly Temperatures for 2 Sep 2017 vs Climatological Means

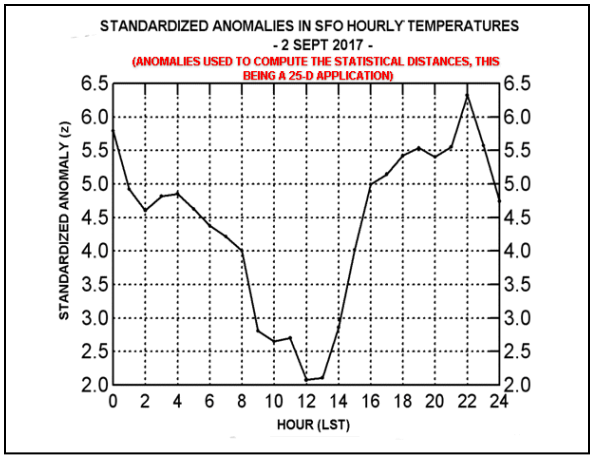


**Figure 16:** Standardized Anomalies in SFO Hourly Temperatures for 2 September 2017

Inspecting the charts, it appears that the high temperatures in general that day, the shift of the peak maxima to the late afternoon, and the extraordinarily high temperatures persisting into the early and late evening contributed to the extreme distance statistic. From Figure 15, readings were still only 81 F at 11 LST and 12 LST each, above normal to be sure, but not drastically so, but a subsequent offshore push brought a ~20 F rise over the next several hours, the 87 F reading (still) at 22 LST, more than 6 standard deviations above average (see Figure 16). Downtown San Francisco, 14 miles to the north, recorded 106 F this day, its all-time highest.

Next, Figures 17 thru 19 characterize the extreme pattern of 3 November 1950, one with an estimated return period of 190 years, a bit less than 2 September 2017's 215 year approximation.

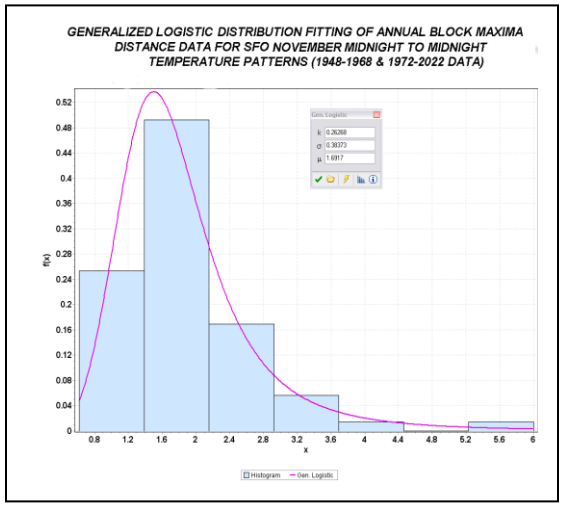


**Figure 17**: Histogram Fitting of Generalized Logistic Distribution to SFO Year-to-Year November Block Maxima Data

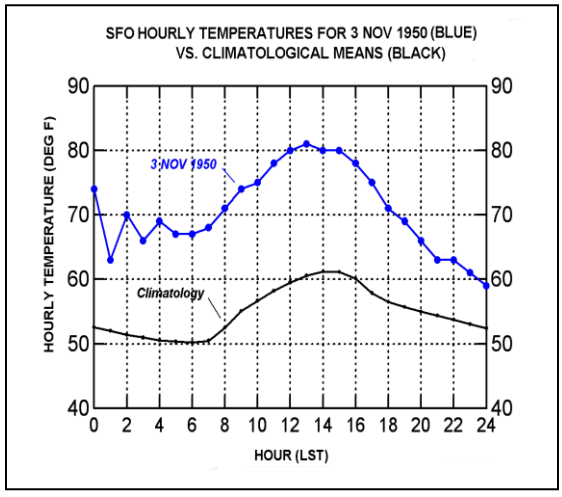


**Figure 18:** SFO Hourly Temperatures for 3 Nov 1950 vs Climatological Means

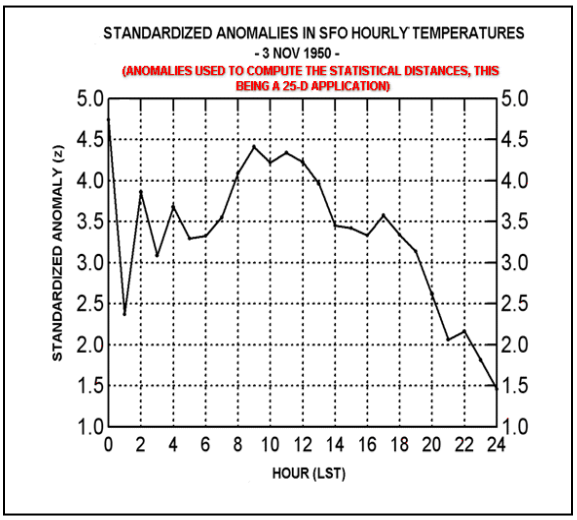


**Figure 19:** Standardized Anomalies in SFO Hourly Temperatures for 3 November 1950

From Figure 18, 3 Nov 1950 commenced the first few hours with an unusually warm, fluctuating offshore pattern. Midnight temperature was a balmy 74 F, more than 4.5 standard deviations above normal for that hour (See Figure 19 above), the preceding 23 LST reading on 2 November 75 F, the hourly reading prior to that 67 F. By 01 LST on the 3rd, though, the mercury had fallen to 63 F but oscillated back to 70 F at 02 LST. Midday temperatures reached 80 F, in excess of 4 standard deviations above, but hours later as the offshore flow episode abated, readings fell rapidly, to 60 F by midnight, the standardized anomaly for that hour only a moderately positive +1.5 z. On an absolute basis, the 6.00 extreme-most distance was not especially high, but inspection of the other November block maxima indicated no other statistics higher than 3.639. Thus, this somewhat odd configuration was sufficiently anomalous in 25-D space relative to the other November maxima to produce the marked return-period calculation.

Lastly, Figures 20 thru 22 illustrate the extreme pattern of 22 December 1990, one with the most pronounced estimated return period for either SFO or LAX: 880 years.

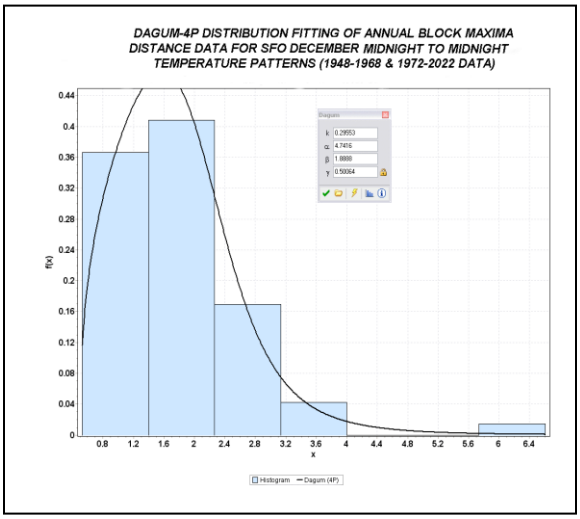


**Figure 20**: Histogram Fitting of Dagum-4P Distribution to SFO Year-to-Year December Block Maxima Data

From Figure 21 below, 22 Dec 1990's anomalous very cold character was due almost entirely to its far below normal hourly temperatures departures throughout, the sinusoidal diurnal pattern not markedly dissimilar from than climatology. The cold air mass's intense nature, of historic proportions, hindered temperature rises, the Noon reading, still 32 F was 4.4 standard deviations below normal, and the least negative standardized anomaly for the whole day was still -3.27 z at 22 LST. The 22 Dec 1990 absolute maximum 6.607 distance statistic was not especially high compared to some of the other months, but not unlike November, the second ranking December statistic was a considerably lower 3.484. Physically, this likely relates to the natural contrasts in anomalous temperature character between winter and the other seasons. Cold season temperature distributions at both LAX and SFO are typically skewed negatively, reflecting that most anomalous temperature spells on balance are cold ones. Yet, given LAX and SFO's locations next to the moderating ocean, the intensities and durations of these spells are probably constrained to some extent physically. The other calendar months of the year, with their higher sun angles and other factors have positively skewed distributions, and the physical constraints in the positive (warm) anomaly direction are likely less.

But In this regard, the intense steady intense cold of 22 Dec 1990 was quite remarkable, expressed in the extraordinary high return-period magnitude.

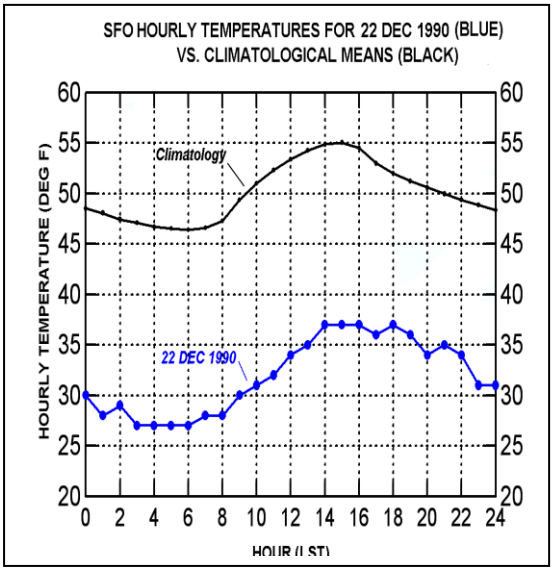


**Figure 21:** SFO Hourly Temperatures for 22 Dec 1990 vs. Climatological Means

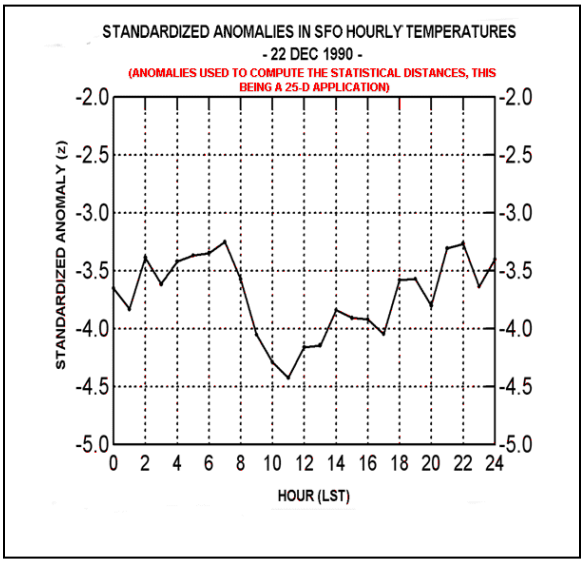


**Figure 22:** Standardized Anomalies in SFO Hourly Temperatures for 22 December 1990

**3.4**. Downtown San Francisco Extreme-most Month-to-Month Precipitation Patterns, by Season

As with the Downtown Los Angeles precipitation data, the same type exploration was performed with Downtown San Francisco's, the latter's period of record extending back to October 1849.

Figure 23 is the tabular summary of the Downtown San Francisco results along with some additional summary statistics discussed further below. From Row 1, the maximum extreme statistical distance, 2.304, in Column 1, was produced for the 1997-98 season, and the best ranked model, the Dagum 3-P, produced an estimated

| | (Col 1) MAXIMUM K=1 DISTANCE | (Col.2 ) DAY/YEAR | (Col. 3) BEST FIT MODEL <=4 PARAMETERS | (Col. 4) RETURN PERIOD (years) | (Col. 5) MEAN (N) |
|---|---|---|---|---|---|
| JULY-JUNE | 2.304 | 1997-98 | DAGUM | ~130 | 0.329 (174) |
| JULY-JUNE | 49.27 | 1861-62 | DAGUM | ~120 | |
| JULY-JUNE | 7.16 | 1976-77 | DAGUM | ~130 | |
| MAX MONTH | 24.36 | JAN 1862 | DAGUM (4-P) | ~850 | |

**Figure 23:** Tabular Summary of Downtown San Francisco Precip Distances (1849-50 thru 2022-23 Period of Record)

return period of ~130 years (Col. 3). Mean statistical distance for the 174 cases (0.329 in Column 5), was nearly the same as LAX's 0.313, and also about one-seventh the absolute maximum statistic of Column1's.

For San Francisco, the 1997-98 season's pronounced maximum distance statistic was attributable to the anomalously heavy rainfalls of January, February, and May (standardized anomalies each in excess of +2.0 z), and especially August, the modest 0.73" fall nevertheless producing a +6.5 z anomaly (See Figure 24). San Francisco's distance statistics were again positively correlated with the seasonal rainfall totals (r=+0.550).

Figure 25 is a plot of the distance statistics for the complete time-series.

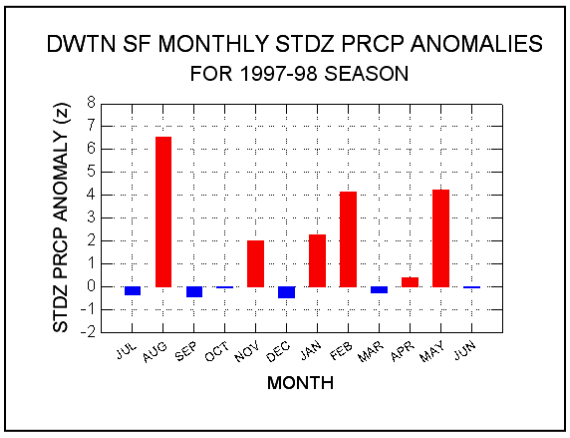


**Figure 24** – Downtown San Francisco Month to Month Precipitation Anomaly Statistics for the 1997-98 Season

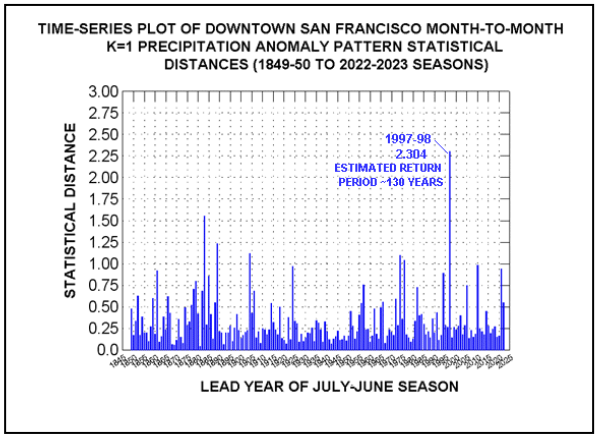


**Figure 25** – Time-Series plot of Downtown San Francisco Month-to-Month Precipitation Anomaly Pattern Statistics for 1849-50 thru 2022-23 Seasons

Once again, by way of comparison, the Downtown San Francisco actual July-June seasonal totals for the 174 year history are fitted, the Dagum 3P taking top ranking (See Figure 26 below). The 1861-62 season (49.27") being the wettest July-June season in history, the 1976-77 season (7.16) the driest, the Dagum 3P produced a 120-year return period calculation to the former, a 130 year return period to the latter.

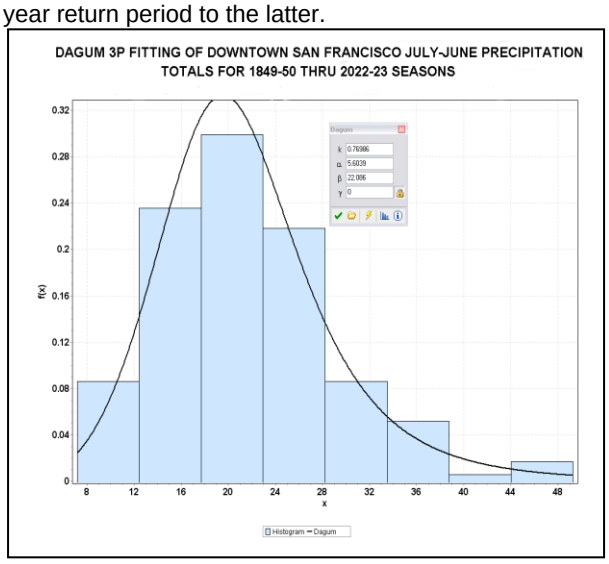


**Figure 26 -** Histogram for Dagum 3P fitting of Downtown San Francisco 1849-50 thru 2022-23 July-June Precipitation Totals

A prominent feature of the 1861-62 season's month-to month precipitation progression was the amazing 24.36" rainfall recorded for January, by far the wettest individual calendar month total ever recorded in the Downtown San Francisco history; this prompted an additional model fitting and return-period inquiry.

Assembling a data set of wettest individual calendar month statistics, season-by-season, the 174-year collection was fitted to the sets of models. Results again determined that the Dagum, this time the 4-.

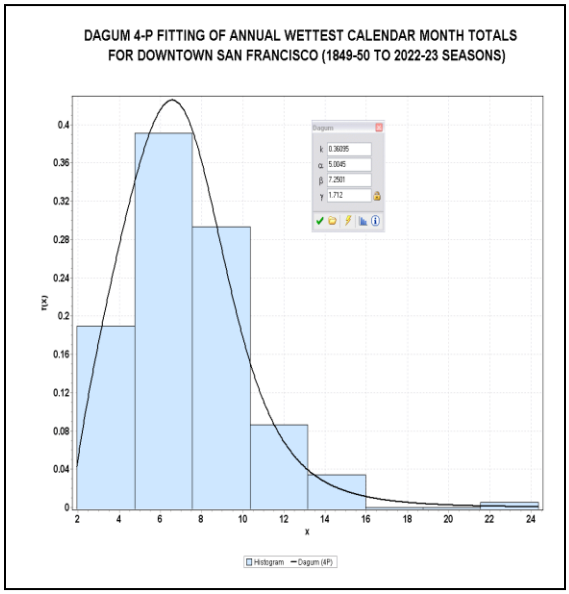


**Figure 27 -** Histogram for Dagum 4P fitting of Annual Wettest Calendar Month Precipitation Amounts for Downtown San Francisco July-June Rain Years

parameter model, ranked highest. Figure 27 above shows the histogram and best-fit curve, the return-period calculation yielding an ~850 year estimate.

## 4. SUMMARY AND CONCLUSION

In a heuristic-type "workaround" exploratory analysis, extrememost patterns of multivariate hourly temperature observations and monthly precipitation totals were resolved and characterized for Los Angeles and San Francisco utilizing curve-fittings and return-period calculations based on K-Means single-cluster statistical distance data.

For the 25-D hourly temperature data application, the approach seemed to be advantageous and precise, as the finer scale methodology not only considered hour-to-hour observation magnitudes, but the pattern progressions hour-to-hour. Observations with identical daily overall means but differing hour-to-hour patterns, for instance, would produce differing statistical distance and return-period calculations, atypical ones (i.e, non-first-harmonic or phase shifted-ones) yielding inflated distance magnitudes. The preliminary standardizations of the raw data insured that each of the 25 hourly variables' anomalies would contribute equally to a given day's statistical distance result and there would be no biases as to particular hours.

Another potential application, not demonstrated here, is the evaluation of heterogeneous variables, such as monthly precipitation amounts combined with precipitation-day counts, or any number of other variable combinations, correlated with one another, whose joint character at extreme multivariate levels is of interest.

In a heuristic treatment such as this, one major issue that would have be considered would be the choice of

the reduction scheme and distance metric. The data in this analysis, as described at the outset, were standardized by variable, with the squared Euclidean distance metric applied; the latter known to isolate out extreme values more effectively than the Euclidean method. Different reduction schemes and distance metrics would likely produce different model fittings and return-period calculations, so decisions would have to be made as the "best" ones to use, complete with caveats.

## 5. REFERENCES


Bousquet, Nicolas, and Bernardara, Pietro, Editors, 2021: Extreme Value Theory with Applications to Natural Hazard – From Statistical Theory to Industrial Practice. Springer, 481 pp.

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

(1) National Institute of Standards and Technology, *Anderson-Darling Test – NIST"* 1.3.5.14. Anderson-Darling Test (nist.gov)

(2) National Institute of Standards and Technology*,"Kolmogorov-Smirnov Goodness of Fit Test -NIST* 1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test (nist.gov)

(3) Wikipedia, *"Birnbaum-Saunders distribution"* Birnbaum–Saunders distribution - Wikipedia

(4) Wikipedia, *"Dagum distribution"* Dagum distribution - Wikipedia

(5) Wikipedia*, "Burr distribution"* Burr distribution - Wikipedia