

Box and Whisker Diagrams

Box and whisker diagrams are often used to display statistically analyzed data. The traditional box and whisker diagram displays the range (maximum to minimum) of the data, the median, and the 1st and 3rd quartile about the median. The second quartile is also the median. A review of quartiles is provided below. An alternate form of the box and whisker diagram is to show the mean and 1 standard deviation about the mean. This latter form of the box and whisker diagram is easier to compute. This document will describe the two variations of the box and whisker diagram.

The traditional version of the box and whisker diagram:

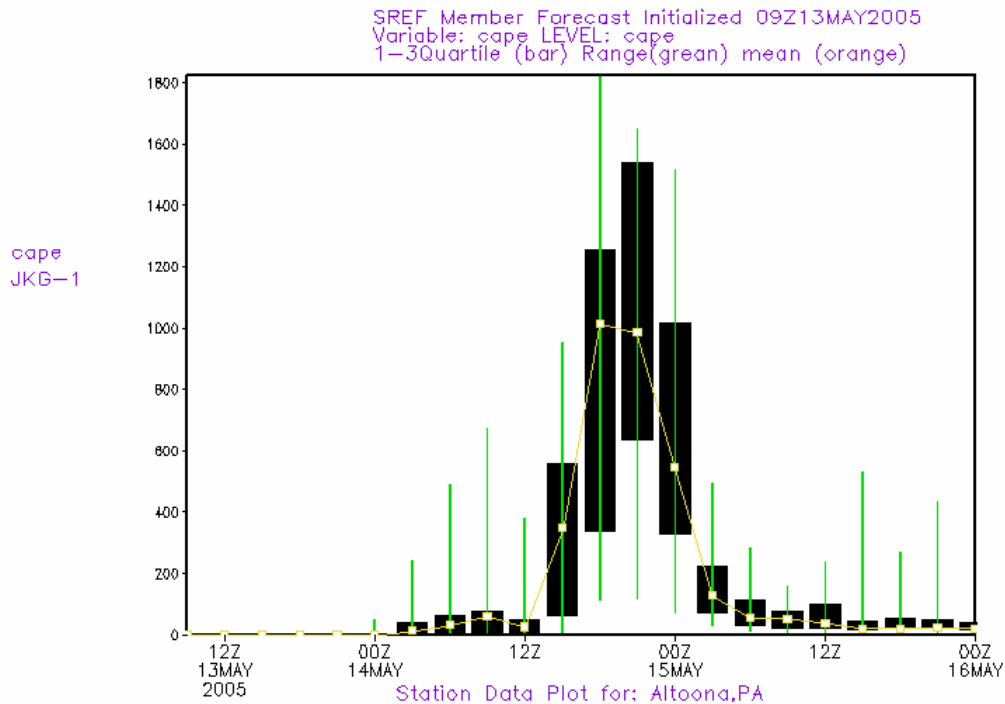


Figure 1 Box and whisker diagram of CAPE for Altoona, Pennsylvania from SREF forecasts issued at 0900 UTC 13 May 2005 for the time period shown in the x-axis. Median is the do and light line. The 1st and 3rd quartiles are defined by the boxes and the range is provided by the green lines.

The traditional *box and whisker diagram* as stated above shows the complete range of the data. The minimum value represents the 0th quartile. The 1st quartile represents the lower 25% of the members in the dataset, the 2nd quartile represents 50% of the data set, and the 3rd quartile 75% of the dataset members. The maximum value is the top limit of the 4th and final quartile. Thus by plotting the median and the 1st and 3rd quartile, the box portion shows the range of the most “middle” 50% of the members with the median being the mid-point. Often, the mean is close to the mean. One variation of the box and whisker

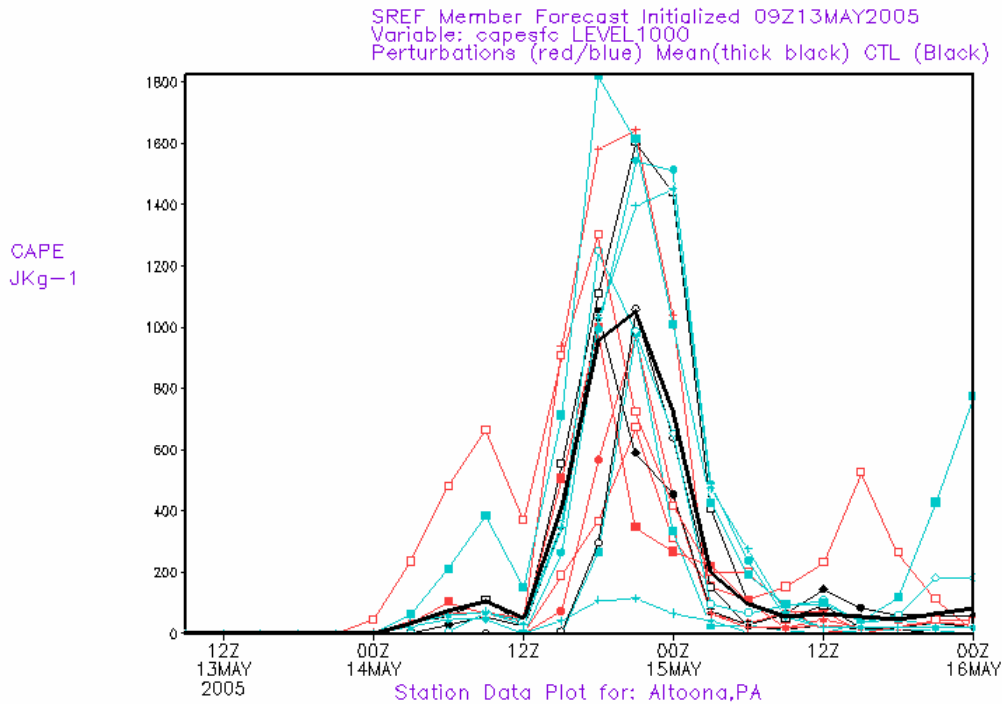


Figure 2 As in Figure 1 except traditional plume diagram with mean in thick black.

diagram is to show the mean and the median. This provides a first approximation of the skewness of the data.

The whiskers of the box and whisker diagram provide the full extent of the range of the data. The length of the whiskers from the end of the box and from the median provides a measure of how large the difference is between the median and outlier values are. This distance from the box to the end of the whisker represents the range of 25% of the entire data sample.

Computing the traditional box and whisker diagram requires a sorting algorithm. The each data point must be put into an array. Once all the data is collected, the data must be sorted from lowest to highest. In a data set with **16 elements** the quartiles are compute as **16/4**. The top end of the 1st, 2nd, and 3rd quartiles are 4, 8, and 12 respectively. The value at array position 8 is also known as the **median value**. These data arrays also offer a means to compute the mode of the data by determining how many elements contain the same value. An example of computing the quartiles is [provided below](#).

Figure 1 is an example of a traditional box and whisker diagram. The data show the convectively available potential energy (CAPE) for Altoona, Pennsylvania from forecasts of the NCEP SREF initialized at 0900 UTC 13 May 2005. These data have unique limit in that the CAPE value can never be lower than 0. Few members show any significant CAPE until after 1200 UTC on the 14th. Note at 0000 UTC 14 May, only the whisker of the top quartile is visible suggesting most members forecast the CAPE to be 0 Jkg⁻¹. The

median CAPE peaks around 1800 UTC around 1000 Jkg^{-1} . The boxes show that there is a clustering in the 3rd quartile close to the median. There are a few strong outliers with CAPE values of at least one forecast of 2000 Jkg^{-1} . At least 50% of the members fall between 400 and 1200 Jkg^{-1} at 2100 UTC. The median is lower at 2100 UTC though there are more members with higher values in the 3rd quartile than at 1800 UTC where it peaks near 1500 Jkg^{-1} and overall the 1st and 3rd quartiles show higher CAPE than the previous forecast.

Figure 2, a traditional plume plot shows that the mean peaks at 2100 UTC. Comparing Figures 1 and 2 provides some insight on how to use the box and whisker diagrams.

On the eyewall.met.psu.edu/plumes/ website, all traditional box and whisker diagram diagrams show are identified with labels showing the 1st and 3rd quartiles and the variable and file names contain the name BAR in them. We hope to introduce a choice menu allowing the user to select between plumes and box-and-whisker diagrams in the near future.

Modified version of the box and whisker diagram:

Modified versions of the box and whisker diagram typically do not contain the median and the quartiles though they do show the range of the data. The short rationale for this is that it is easier to compute the maximum, minimum, mean, and standard deviation of the

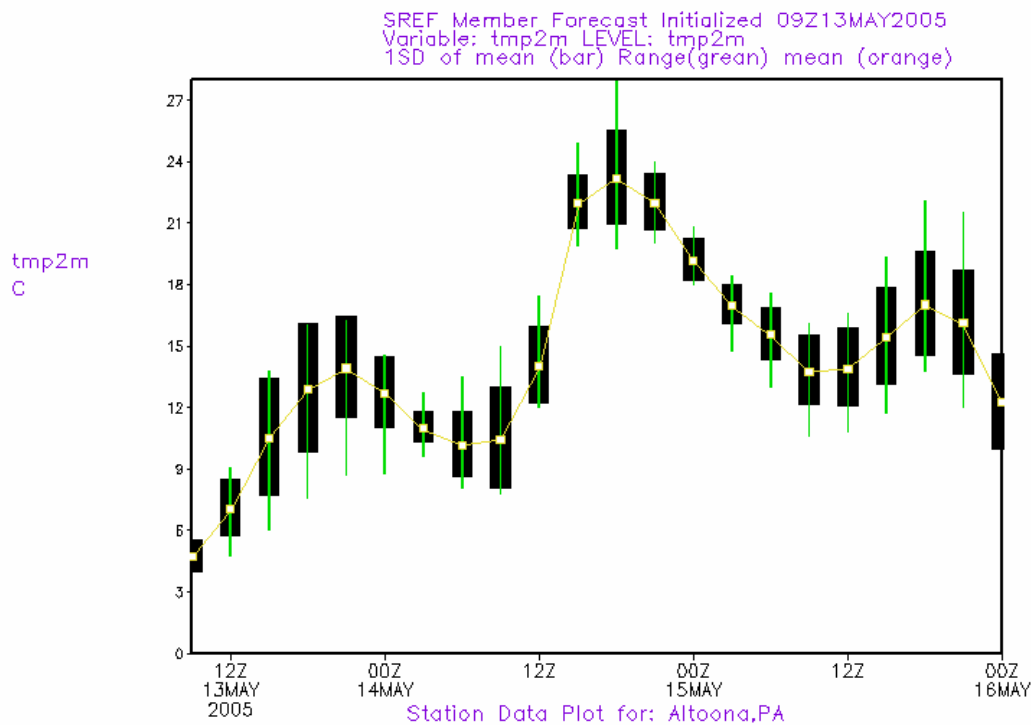


Figure 3 As in Figure 1 except for 2m temperatures (C) and boxes are standard deviations about the mean. The mean is plotted as a yellow dot connected by a yellow line.

data then it is to bin the data to compute the other variables. When the data has a probability density function (PDF) similar to that of a normal distribution, this method of data display is rather effective and efficient. The diagram normally includes the range, the mean, and value one standard deviation about the mean. These diagrams clearly show the location of 66% of the values by the range of the box. Recall in the traditional box and whisker diagram the bars show 50% of the data about the median.

This method of display fails to show if the data does not have a near normal PDF. Highly skewed and bimodal data are more difficult to discern using this data display method. The median and the traditional box-and-whisker diagram are often more representative when the data is bimodal.

Figure 3 is an example of the standard deviation plot. This example shows 2m temperatures (C) for Altoona from forecasts initialized at 0900 UTC 13 May 2005. Note that the bars are of equal length on either side of the mean. Unlike the data in Figure 1, the spacing is even either side of the mean but can be skewed either side of the median. These boxes represent data within 66% of the mean while on the 50% of the members lie either side of the median in the boxes in Figure 1.

Summary:

The traditional box and whisker diagram displays the median, and the 1st and 3rd quartiles about the median, which is of course, the top of the second quartile. The range is depicted by the lower and upper limits of the 1st and 4th quartiles respectively. ***The bars are often not equidistant about the median like the standard deviation is about the mean in the modified box and whisker plots.*** The traditional version has the advantage when showing skewed or bimodal data. The modified version, using the standard deviation or spread about the mean, performs well when the data approximate a normal distribution.

Both box and whisker diagrams are quick methods of summarizing large amounts of data. When the number of ensembles grows sufficiently large, traditional spaghetti plumes will become difficult to use, thus requiring the use of diagrams that quickly summarize the data.

Quartiles:

Computing quartiles:

Quartile bounds = $N/4$

If there are 12 data points the bounds are 3,6,9 and 12. The data in bin 6 is the median and the data in bin 12 is the maximum. The data is assumed to be sorted from low to high as shown as shown for 12 ensemble members CAPE forecasts:

1	2	3	4	5	6	7	8	9	10	11	12
80	200	400	550	800	1000	1050	1100	1200	1400	1600	2000

For these data the bars would start at 400 and span to 1300. The median is 1000 and the range is 0 to 2000. The calculations are simple with an even number of bins. With bins not divisible by 4, an average or weighted average between the bins must be taken: $QB = N/4$ and we have 18 members then $QB = 4.5$. Thus the first quartile is at 4.5 so the average of bins 4 and 5 are used. The median would be bin 9 ($2 * 4.5$), and the 3rd quartile would be $3 * 4.5 = 13.5$ or the average of bins 13 and 14.

The above was a clean example as it was easily divisible by 4. If there is a residual, say there were 15 members, the break points would be in increments of 3.75. Thus the quartiles would be 3.75, 7.50, 11.25, and 15.00 respectively. Thus the median would be computed using the 7th and 8th elements of the array while the 1st quartile would be a weighted average of the 3rd and 4th elements ($.25 * \text{element } 3 + .75 * \text{element } 4$). The reverse weights would hold for the 3rd quartile.