first-report-rvs

September 12, 2022

1 How likely is the difference between AWS and LCD sensors to diverge signficantly from the mean?

This notebook prepares the raw data from the 2018 sensor reference data set and uses it to construct the following random variable to analyzse how the difference between station control measurements evolve through time.

1.1 Theory and definition of random variable

Let (x_i, y_i) represent measurements from a given AWS and LCD station pair at a given point in time $i \in I$ where I is the set of all observation pairs with complete data. Let $\overline{xy_i}$ and σ_{xy_i} be the mean difference and standard deviation of this difference between aws and lcd measurements for all pairs of observations (x_i, y_i).

Define the discreet Random Variable V where $V =$ (the number of observations that exceed a given number of standard deviations for a given time period). This can be written as $g_k(x_i, y_i) \; \forall \; (x_i, y_i \in (X, Y) \; \forall \; k \in K$ where K is the set of all time groupings and k is the set of observations i within each time grouping. This be written as the following for an arbitrary number of standard deviations c set as the cut off limit for success.

$$g_k(x_i, x_j) = 1 \; if \; \frac{x_i - y_i - \overline{\mathrm{xy_i}}}{\sigma_{xy_i}} > c$$

$$g_k(x_i, x_j) = 0 \; otherwise$$

Therefore the random variable can be written as:

$$V_k = \sum^{i} g(x_i, y_i) \; \forall \; i \in \; k$$

Concretely, there are readings every 10 minutes, therefore each hour of each day can be treated as a period k with 6 readings $i = (1, 2, 3, 4, 5, 6)$ in each period and $range(V_k) = [0, 6]$. Since the study period runs from May 15th - September 15th any given hour $k \in K = [0, 23]$ has around 120 trials.

If radiation drives deviations from the mean we should see the CDF for this random variable fluctuate according to the k chosen (e.g. if $k = 13$ we expect a lot of radiation and thus a lot of

devation from the mean, while if $k = 00$ we expect few signficant deviations fromm the mean. This notebooke implements this random variable test.

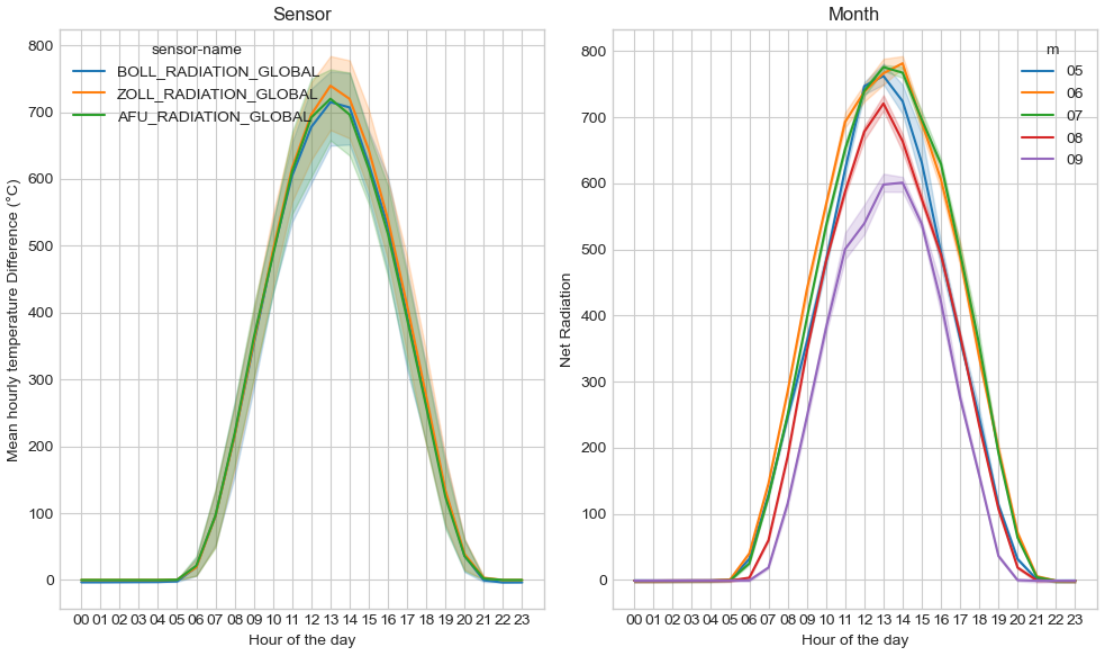
As will be shown graphically, time of day has a clear correlation with deviations from the mean, while day of the year seems to have little impact.

Interestingly the flucatuations are markedly different at each sensor site, with the Zollikofen sensors all fluctuating in a similar fashion. This test will be repeated with the 2019, 2020, and 2021 data to confirm the stability of the distribution of this random variable year on year.

1.2 General Mean Difference Results

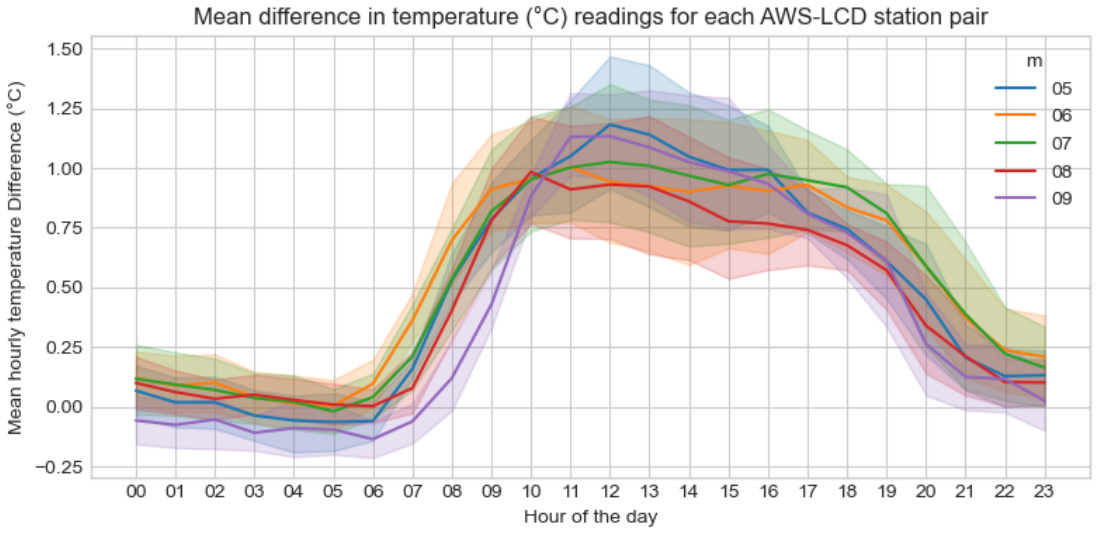
1.2.1 Radiation

The radiation shown below for all aws stations is the most likely culprit. However, the smooth curve is not observed for all figures.



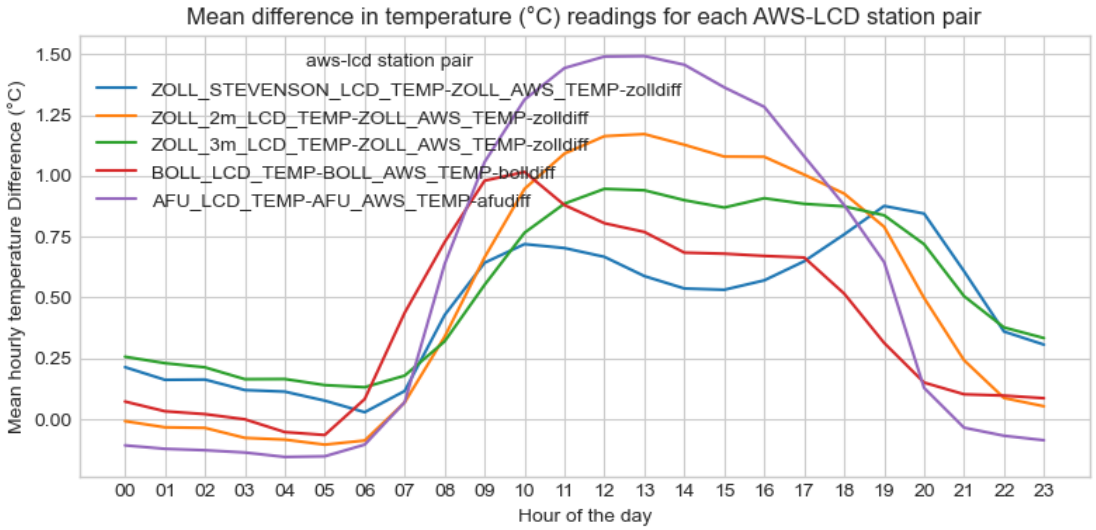
1.2.2 Grouped by hours and by month

The hours show a clear evolution throught the day that is relatively unchanged when grouped by month.



1.2.3 Grouped by hours and by station

Overall, the trajectories are similar across stations, but there are marked differences in their respective daily evolutions.



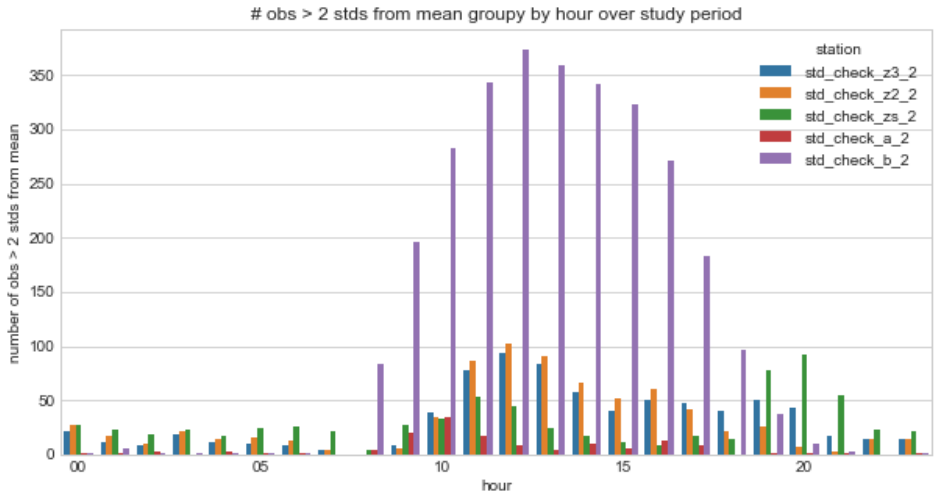
1.3 Random Variable

1.3.1 Hours

The random variable reflect this evolution as well with two strong divergences, namely the results for Bollwerk and AFU. AFU has very few significant deviations from the mean (successes) across the day, while Bollwerk has large number of successes during the peak hours (50% !). The below graph is of the *total number* of successes for each hour across time with

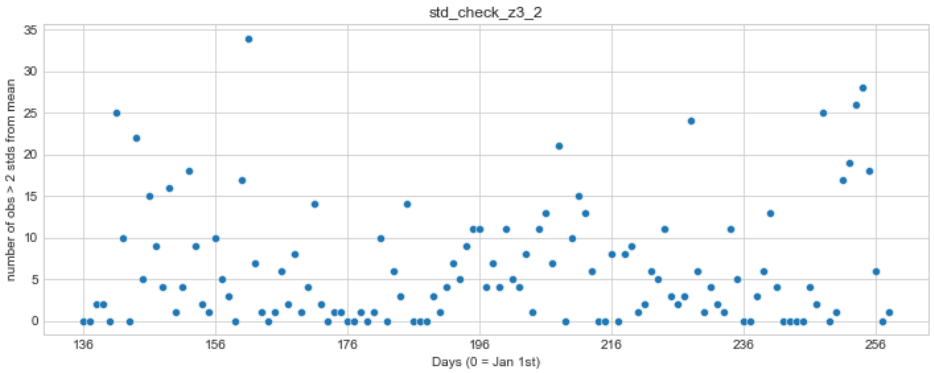
$$range(X_k) = [0, 720*]$$

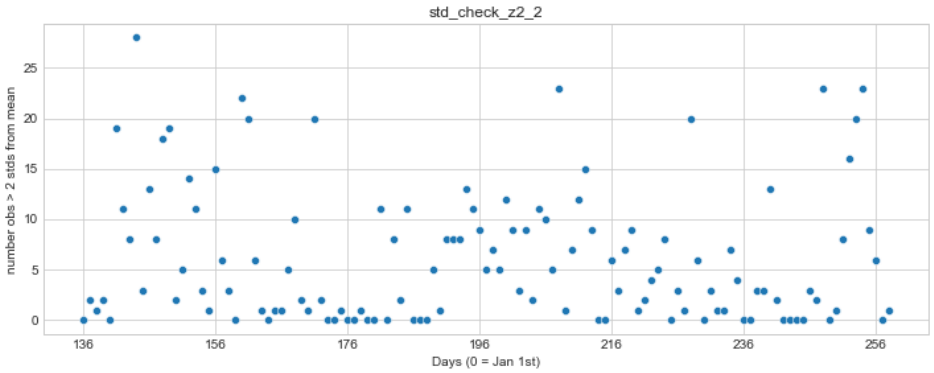
(note that the actual number of observations is variable and depends on the exact hour due to observations being removed during sensor readings).

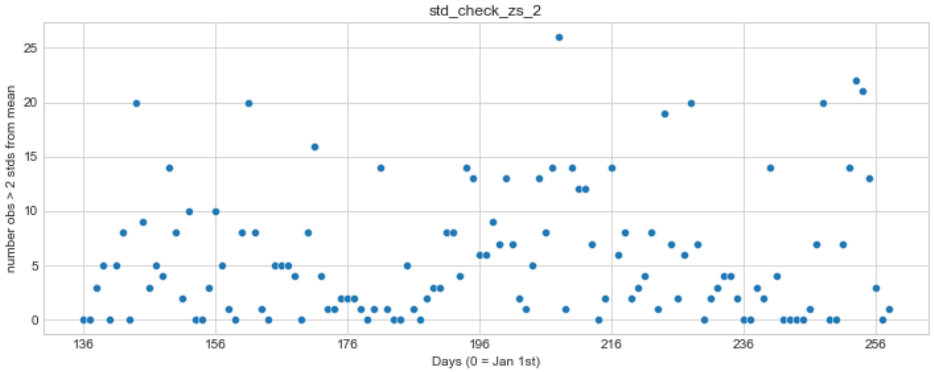


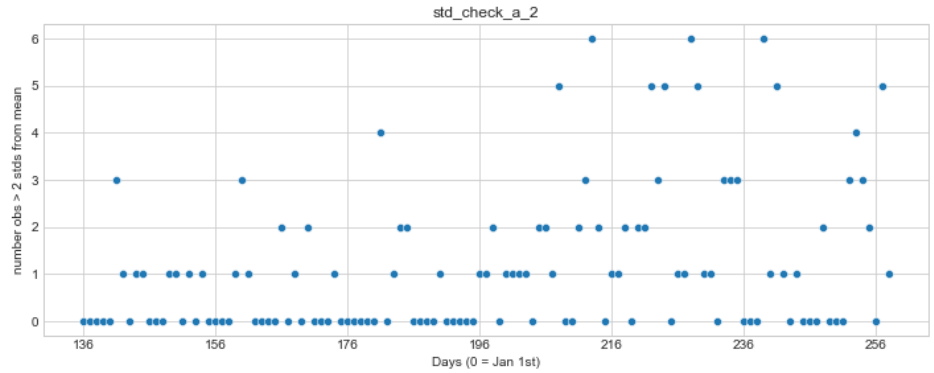
1.3.2 Days

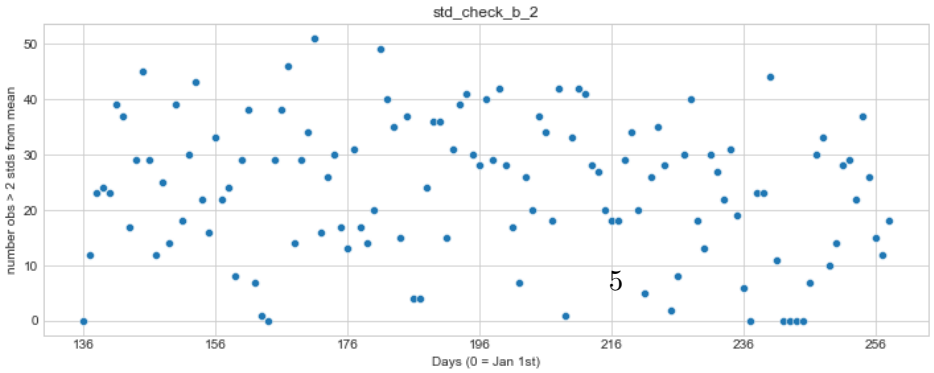
As the below charts show for the number of sucesses per day there is not a clear overall trend as the year progresses from May 15th to September 15th (except perhaps for AFU..)


std_check_z3_2
number of obs > 2 stds from mean
Days (0 = Jan 1st)


std_check_z2_2
number obs > 2 stds from mean
Days (0 = Jan 1st)

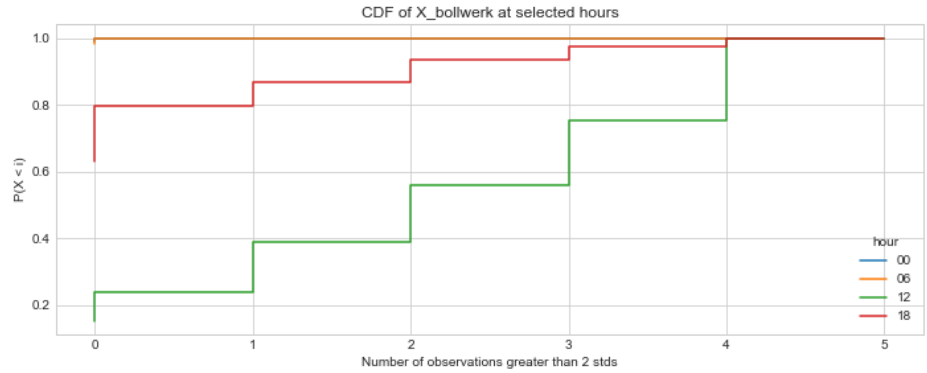

std_check_zs_2
number obs > 2 stds from mean
Days (0 = Jan 1st)

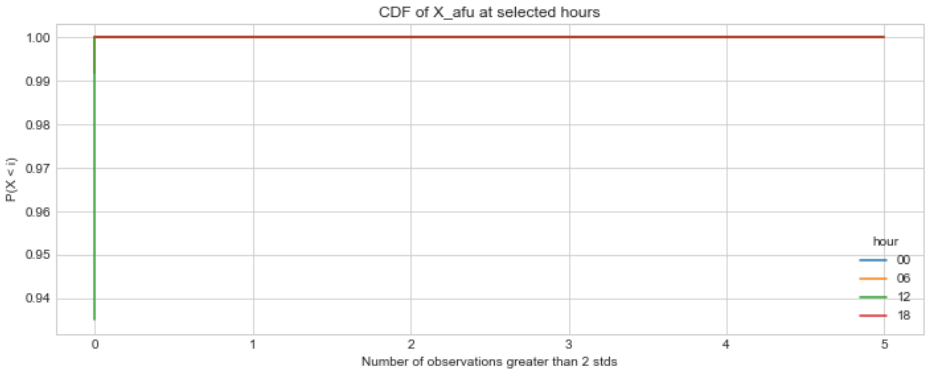

std_check_a_2
number obs > 2 stds from mean
Days (0 = Jan 1st)

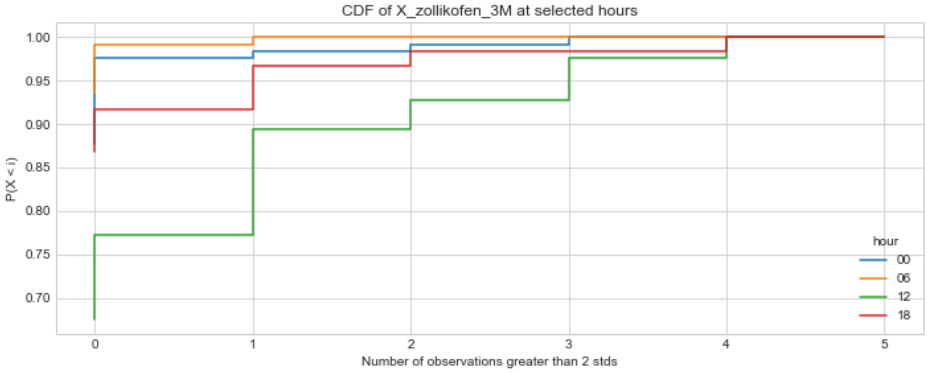

std_check_b_2
number obs > 2 stds from mean
Days (0 = Jan 1st)

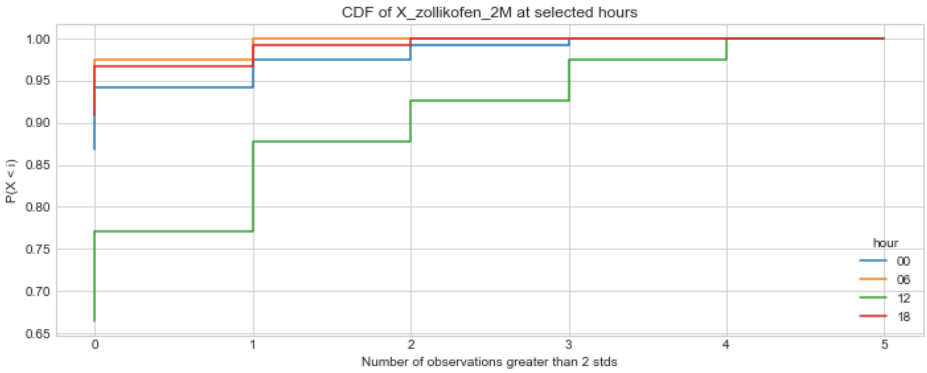
1.3.3 CDF for select hours of the day

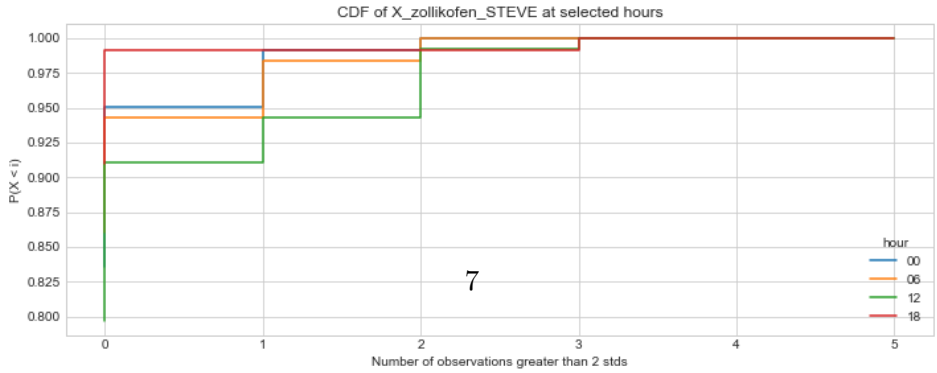
This graph of the CDF for select hours of the day based on the sample data summarizes the above charts. The lack of successes in AFU is striking. The station divergences are not expected as they do not correlate with the largest aws/lcd sensor differences.


CDF of X_bollwerk at selected hours
P(X < i)
Number of observations greater than 2 stds
hour
00
06
12
18


CDF of X_afu at selected hours
P(X < i)
Number of observations greater than 2 stds
hour
00
06
12
18


CDF of X_zollikofen_3M at selected hours
P(X < i)
Number of observations greater than 2 stds
hour
00
06
12
18


CDF of X_zollikofen_2M at selected hours
P(X < i)
Number of observations greater than 2 stds
hour
00
06
12
18


CDF of X_zollikofen_STEVE at selected hours
P(X < i)
Number of observations greater than 2 stds
hour
00
06
12
18

1.4 Quick Discussion

- The unexpected results could come from dividing by the variance. AFU has a higher highs and lower lows than other stations, and thus dividing by the variance underestimates the absolute changes in AFU values relative to the other stations.
- Selecting the std variation cut off will change results significantly, this has not been explored.
- Apply to 2019, 2020,2021 data to see if the station distinctive trends remain.

1.5 improvements

- code relatively inneficient in reading in data
- apply data corrections prior to any analysis in seperate notebook and then load that data directly.
- change std cut off values (1.5X mean?)
- include other data
- correct x axis cdf (range is 0-6 not 0 - 5)

[ ]: