LADSS Version 2.0 - Climate Data "Cleaning" Process

"Cleaning" data refers to the completion of the data set and the eradication of NULL data. There are two methods for cleaning the climate data:

- Clean missing records
- Clean incomplete records

Clean missing records searches the database for missing records and NULL records. NULL records refers to days that have NULL values in *all* of the following columns:

- min_temp
- max_temp
- precipitation
- solar_radiation
- sunshine_hours

Firstly, the database is queried for the non-NULL records. The results are processed to determine missing blocks of data.

These blocks of missing/NULL records are processed together in order to reduce the day-to-day errors of this process. The aim is to replace a block of missing/NULL data with a complete block of data from another year rather than various days from different years. These blocks must not exceed 35 days during the growing season (day 75 – 273).

For each block of missing/NULL records the following strategy is used to create/complete the records:

Criteria: search all years for same location for records between days X and Y
Result: the count of non-null records by year.
Order: sort the results by the number of non-null records (most first).
Tiebreak: the absolute difference between the search year and replacement year (smallest first).

Such a query might yield the following results:

YEAR	COUNT(*)	YEAR DIFF
1981	31	1
1983	31	3
1976	31	4
1984	31	4
1974	31	6
1982	20	2
1977	20	3
1975	20	5
1985	15	5
1973	13	7

Note that the first column order is the number of complete (non-null) records. Since all years that are displayed have full record sets the 'year diff' tiebreak is needed.

However, if no full records existed for any year we modify of the query to retrieve all partial record sets.

Values that are created using this process are flagged to indicate this – ie. the columns MEASURED_MIN_TEMP, MEASURED_MAX_TEMP, MEASURED_SOLAR_RADIATION, MEASURED_PRECIPITATION and MEASURED_SUNSHINE_HOURS are set to zero.

Clean incomplete records searches the database for records that contain NULL values for some but not all of the following:

- min_temp
- max_temp
- precipitation
- solar_radiation
- sunshine_hours

For each record, we search the database for records for the same location where day of year is +/- 10 the current day.

This returns a set of data similar to the following:

Day	*Precip*	*Min T*	*Max T*	*SR*
253	0.0	9.2	19.9	10.9
254	1.1	7.1	18.5	13.1
255	2.3	7.9	20.0	14.2
…	…	…	…	…
272	4.0	6.4	13.5	6.8

From this set of data we wish to select a record that best represents the record we are cleaning. We do this by taking the absolute difference of each of the searched data sets from the known values in the record we are cleaning. We then rank the search records for each criterion. This produces the following, including a total:

Day	*Precip*	*Min T*	*Max T*	*SR*	*Total*
259	1	7	2	1	11
254	2	3	1	5	11
255	3	4	8	10	25
…	…	…	…	…	…
272	20	10	15	2	47

Days 259 and 254 give the best score using this technique. We tiebreak them by selecting the record with the lowest *maximum individual score*. In this example, this would be 254.

When this still results in a tie, the record with day of year closest to the day of year for the record we are cleaning wins the tie.

When no replacement record is found the average for the day of year is used to fill the missing data.

The modified record is flagged to indicate cleaned data.