RainQC Job Manager Daily/Monthly Report Anomaly Reports

Version 1 / August 3, 2023

As part of the daily and monthly reports that the RainQC Job Manager produces, there is a listing of the anomalies identified. An anomaly is a precipitation reading for a target station that the RainQC data quality system believes is either too low or too high, based on the corresponding daily precipitation readings at one or more neighboring stations.

These anomalies are reported into the TAHMO system using a set of flags. A flag value of 2 indicates an anomalous daily precipitation value. A flag value of 1 indicates a nominal value and a value of 0 indicates that a score could not be computed, usually due to insufficient data at the target and/or its neighbors. For brevity, only flag=2 anomalies are presented in the daily/monthly reports.

RainQC currently utilizes a linear regression model for each station that has been trained on historical daily precipitation data for the target station and its neighboring station(s). These neighboring stations have been selected due to their geographical proximity, e.g., they are within roughly 100km or less of the target station. These regression models are used to produce a score for each target station.

Let's look at an example of the daily report for a single station:

```
TA00003 2023-07-02 | score:   317.164 (thresh:  220.434) -- 'pr' t:  0.867 mm n: (8.759 mm, 26 km), (15.349 mm,
37 km), (6.922 mm, 51 km), (40.803 mm, 51 km)
```

If we think of these report lines as having columns, we have the following:

Station Name	Scoring Date	Score	Model Threshold	Daily Precipitation at Target Station	Daily Precipitation at Neighbors and Distance from Target for each Neighbor
TA00003	2023-07-02	317.164	220.434	0.867 mm	(8.759 mm, 26 km), (15.349 mm, 37 km), (6.922 mm, 51 km), (40.803 mm, 51 km)

The target station is the listed first, TA00003. Following the station name is the date for which the score is being computed, July 2, 2023, in ISO format.

Next, the prediction score itself is given, 317.164. Next to the score is a station-specific threshold that was computed when the model was trained. If the computed daily score is greater than the threshold -- 220.434 -- the precipitation at the target station is deemed to be anomalous. One can think that the higher the score is compared to the threshold, the more anomalous the target precipitation. The computed score will often exceed the threshold when daily precipitation readings vary significantly from the model's representation of historical precipitation values.

After the score and threshold, the daily precipitation reading of the target station is given (0.867mm), along with the readings associated with the station's neighbors and their distances from the target station. Daily precipitation values are millimeters and distances are kilometers. We can see in our example that there are four neighboring stations that are 26km, 37km, 51km, and 51km away. The distances themselves are geodesic distances that do not take into account the elevation of the stations. Each neighbor is presented a precipitation reading and a distance from the target within parentheses. Neighbors are always listed in the same order, from nearest to farthest.

In simple terms, there are two signals that can be used to help analyze each reported anomaly.

The first is how much higher the score is versus the threshold. The higher the score, the greater the predicted difference between that day's daily precipitation measured at the target and what the model believes the target's precipitation should be based on the readings at its neighbors.

The second signal is the precipitation readings themselves. One can look at the daily precipitation at the target and make an assessment of how it compares to the daily readings at the neighboring stations. The closer the neighbors are to the target, the more we might think that there should be some correspondence between the target station and the neighbors.

The more close neighbors that we have, the greater the expected correspondence. However, keep in mind that while neighboring stations may be within roughly 100km of the target station, there may be intervening geographical features, changes in geography, etc. For instance, there may be a mountain range between two stations, or one station might be in a desert area 100km inland of a coastal station. The strength, shape, direction, and velocity of the weather systems themselves of course will often be a factor in the readings of the target versus its neighbors.

Additionally, one might also reason about the precipitation values for each station individually. If there is a repeated value day over day, that can be a sign of a problem. This repeated value may not be the target itself, but one or more of its neighbors.

Similarly, if there are some extremely high values, that can also be a sign of an issue. For example, TA00370's neighboring station registered a daily precipitation value of 523.241 mm on July 16, 2023. That's over half a meter of rainfall. Did that really happen?

Finally, keep in mind that currently RainQC does not look across more than a single day to assess whether or not a station’s precipitation readings are anomalous. RainQC also does not have awareness of the TAHMO ticketing system, e.g., a human may know there is already an open trouble ticket for a particular station, but the RainQC system doesn’t.

Thus it is up to the human analyst to use all available information to assess a station’s status and sensor health, using RainQC’s anomaly scores as one tool in their assessment toolbox.