QUALITY CONTROL FOR UNMANNED METEOROLOGICAL STATIONS IN MALAYSIAN METEOROLOGICAL DEPARTMENT

By

Wan Mohd. Nazri Wan Daud

Malaysian Meteorological Department, Jalan Sultan,

46667 Petaling Jaya, Selangor, Malaysia

Phone: +603-79678000, Fax: +603-79584840, wnazri@met.gov.my

ABSTRACT

The requirements for quality data is a prerequisite where meteorological data is concerned. While manned stations have trained observers to ensure data quality, data from unmanned stations needs stringent quality control.

In the MMD Integrated Meteorological Surface Observation System, these requirements are handled by an innovative rule set based quality control module which has the ability to accept data from both automatic as well “manual sources”. Therefore a single point of entry into the database allows for QC rules to be homogenous throughout the dataset whatever the temporal frequency of the dataset. Datasets are based on chunks of data with configurable temporal frequencies.

The system is effective as all data is transferred to a central data warehouse. The use of rule sets enables differing rules for differing purposes and gives a better view of actual weather phenomena with the inclusion of range, consistency, temporal and spatial rules.

The QC module has three stages where different functions are performed. The rule sets for First Stage Quality Checking (QC1) perform data validation with range, consistency and time. Second stage Quality Checking (QC2) performs interpolation based on rules with forward fill, constant fill and linear interpolation.

Datasets that pass both stages are ingested into the database, whereas data that fails is passed to the final stage which is manual QC (HQC). Here trained personnel make decisions on the quality and usefulness of the data.

1.0 INTRODUCTION

Despite its relatively small area, Malaysia has large variations in its local terrain and this in turn influences local airflow dynamics and weather patterns. All these areas are subject to different weather conditions during different times of the year ranging from severe thunderstorms during the inter-monsoon periods to dry and hazy conditions during the South-West monsoon and flooding during North-East monsoon season.

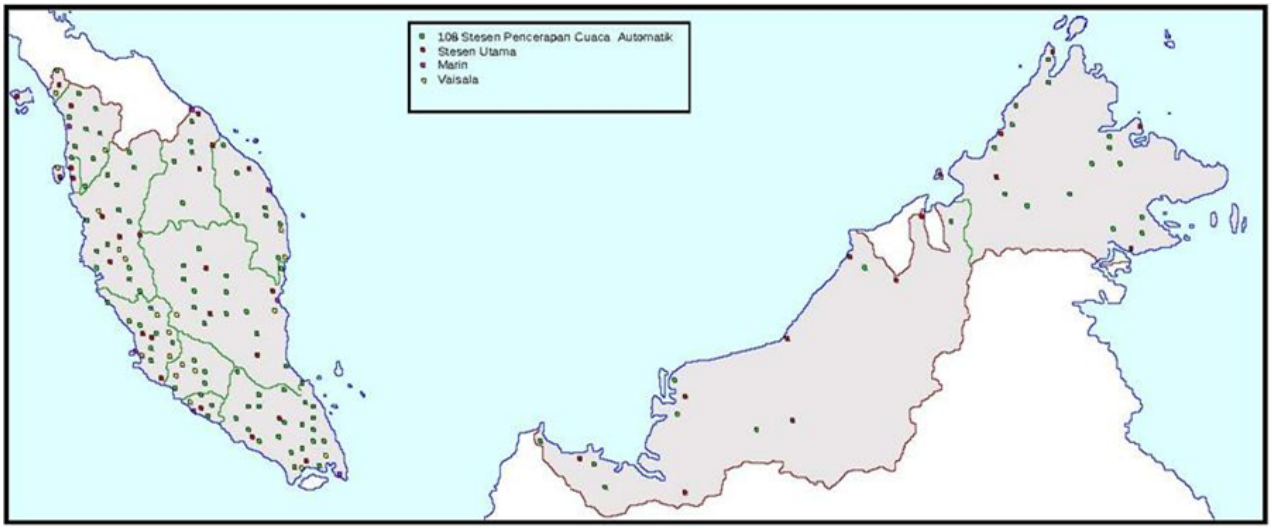
The Integrated Meteorological Surface Observation System which has been designed and built from scratch by the department itself, integrates data from the various surface meteorological systems to gather data in real-time using data streaming protocols, which enables the Malaysian Meteorological Department (MMD), to detect, monitor and predict more accurately meteorological phenomena affecting the country. The parameters measured by the real-time system are precipitation, wind speed and direction, temperature, relative humidity and atmospheric pressure.

The main function of the Integrated Meteorological Surface Observation System is to enable the MMD to:

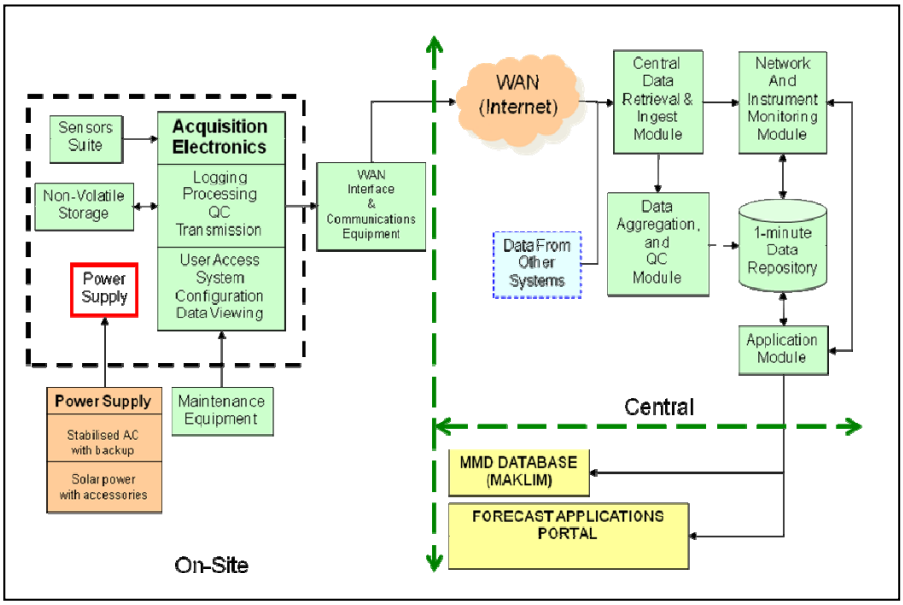
i) Provide detailed observation data from the network of stations, which can be processed for eventual forecasting and reporting of weather conditions

ii) Process and generate local meteorological messages such as climate hourly and climate daily for the Climatological database.

iii) Provide and display real-time streaming data, for the purposes of now-casting.



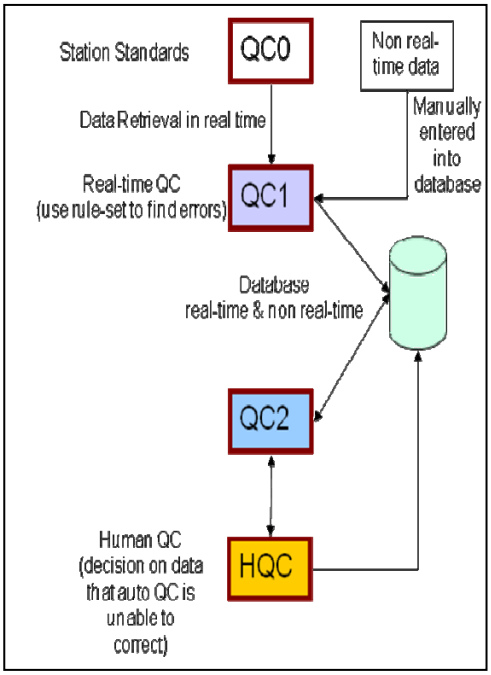
Malaysian Meteorological Department Surface Observations Network



System Layout and Data Flow of MMD Surface Observations Network

2.0 QUALITY CONTROL OF METEOROLOGICAL OBSERVATIONS DATA

While timely data is important, the requirements for quality data are also a prerequisite. These requirements are handled by an innovative rule-set based quality control module which has the ability to accept data from both automatic as well manual sources. Therefore a single point of entry into the database allows for QC rules to be homogenous throughout the dataset whatever the temporal frequency of the dataset. The use of rule-sets enables differing rules for differing purposes and gives a better view of actual weather phenomena with the inclusion of range, temporal and spatial rules.



Quality Control Flow Diagram of Meteorological Observations Data

2.1 First Stage Quality Checking (QC1)

The processing engine is centered on the concept of a chunk, which is a set of data for given product ID for a fixed time period, e.g. an hour. For a chunk to be eligible for processing all of its data must have been received in the Central Processing Facility or a specified timeout must have been exceeded, e.g. 12 hours or more. The administrator may also specify other related products that must be available before this product is considered eligible for processing. This is used for consistency checks or spatial checking, where the validation of a station data involves data from other stations as well.

Once a chunk is considered valid; its data is fetched and made available to the Data Quality Control engine (QC1). This performs a repeat of the checks which are performed at the station, namely:

i) Range Checking

The range check algorithm allows the specification of low and high limits for a given sensor parameter. Data that falls with the limits are flagged as “OK”. Data that falls outside of the limits are flagged as “NOT OK”. The value for these range checks shall be obtained from statistical limits for the particular station or area, or from established boundaries previously set by the Climatological Division. Due to the current trend of climate change, it is essential that some headroom has been included in the limits to cater for extreme values that occur right currently or would statistically occur in the future. These limits are also be periodically revised to ensure their relevancy. The values selected for limits shall also be influenced by

the sensor limitations, for example measurements beyond the capabilities of the sensor should not be allowed.

ii) Step Checking

Data for any given minute is compared with the value immediately preceding it. The difference between the two values must not exceed a given amount. If the difference is large than a given amount, the data is flagged as “NOT OK”. Like limits for range checks, the step value shall be obtained from statistical limits for the particular station or area, or from established boundaries previously set by the Climatological Division. Step limits can often be based on physical limits however, e.g. a 5 ºC jump in temperature in one minute is more likely to be a result of instrument failure or interference, rather than an actual weather phenomena, regardless of the where the station located.

iii) Temporal Checking

Data is expected at regular intervals. Where data is not present, it is considered to be missing and therefore defective.

iv) Spatial Checking

This check involves comparing the data from a set of stations and determining the median of the values. The checked value should not differ from the median value by a certain percentage. The median is used instead of the mean so as not to allow a defective station to affect the validation checks of other stations. When performing this form of spatial checking, it is important that the set of stations chosen possess comparable characteristics. This is not just dependent on the location of the station and elevation, but can be influenced by other physical factors (e.g. proximity to mountain range or the seaside) that may affect the prevailing weather conditions at the station. The suitability of station to form spatial check set with other stations needs to be determined on an element by element basis. For example, while a stations pressure reading may be comparable to another station, they may have different wind characteristics, and therefore should not be spatially checked for similarity in experience, by careful analysis of historical data.

v) Consistency Checking

This involves checking that a given value is consistent with other sensor parameters measured at the same time. The comparison is done on the basis of physical or climatological relationship between different elements being measure. Possible checks that can be determined here include:

- Difference between dew point temperature and air temperature should be within a given range.

- Rainfall only occurs when relative humidity is above a certain level.

vi) Spatial-Consistency Checking

This involves a combination of both the spatial and the consistency methods. Station weather parameters are tested against heuristic rules not just against another parameter from the same station, but based on the median of set of comparable stations.

As a result of these checks, certain readings within the “chunk” may be marked defective. If the number of reading is below a specified minimum, then all the data in the chunk is considered valid and is passed on for data aggregation. Where data does not pass the given minimum required, it is passed on to the next stage of QC.

2.2 Second Stage Quality Control (QC2)

When a failed chunk is detected, the chunk is passed to the next stage of quality control, known as QC2. This is where the system attempts to derive the missing values within the chunk. It does this through a process of interpolation. A number of interpolation algorithms are available:

i) Follow-through interpolation

This involves taking the last realistic value (if available), and declaring it as the missing value. This is only realistic for short bursts of missing data.

ii) Temporal Interpolation

This involves taking the mean of the last reading before the missing block and the first reading after the missing block, and declaring that as the intermediate value. This is useful where there is only one missing value but is also used where the missing block does not extend beyond logical temporal limits.

iii) Spatial Interpolation (median method)

This involves analyzing the value for data for comparable stations within the same spatial set, and assuming the median value of the set to be the missing one.

iv) Spatial Interpolation (weighted-distance method)

This also involves analyzing the value for the data for comparable station within the same spatial set. However, it derives the missing value from a weighted average of the data for comparable stations, with the weigh being based on the distance to each station.

The choice of algorithms to be used for each weather parameter is to be determined by the administrator after consultation with the technical committee of the department. Multiple algorithms may be selected, with the engine picking the

final algorithm based on the results returned. This intelligence is built into the system.

2.3 Manual Quality Control (HQC)

If the results of QC2 still have more missing data than is allowed for this product, then the system logs the chunks details into a “Manual QC” bin. It is then left for a human operator to review and manually edit.

2.4 QC Engine Architecture

The QC engine has been designed to be scalable and highly extensible. It can be expanded to cater for more products, and more importantly, additional rules. The module is entirely parameter driven, based on settings in a configuration file. This allows the administrator to specify, for each product:

i) The “chunk” size?

ii) What other dependent data is required? This can involve data from the previous or next chunk for the same product, or involve differing stations or sites (for consistency and spatial checks).

iii) The minimum number of valid readings for the “chunk” to be considered acceptable.

iv) The list of checks to be performed within QC1, together with their associated parameters.

v) The list of interpolation algorithms to be used within QC2, together with their associated parameters.

2.5 The User Interface

The rules for each parameter are specified from GUI interface. As many stations within similar environments have similar rule sets, a “template” based interface allows settings to be duplicated across similar stations. In operation, the QC engine scans the results of the Central Retrieval and Ingest Module, and determines which chunks are ready for processing, based on the dependent data.

For each chunk that is ready, the QC engine extracts the data and places it into a job file. The job file also contains the parameters for all the QC1 and QC2 algorithms from the Station Profile Database. The QC engine then executes the necessary QC algorithms, passing the job file to each QC algorithm. Each QC algorithm is written as a separate process, and can be run in parallel across either a standalone or a cluster based system. After each QC algorithm has done its processing, it updates the job file and returns control to the QC engine. The QC engine than determines if further algorithms need to be run on the job file, and continues assigning the job file to other algorithms if necessary. The process completes when the data in the chunk meets its

minimum number of valid readings to be considered “acceptable” and is ingested into the database, or is assigned to the Manual QC bin.

If the chunk is valid, it is written out to the Validated Data table, and further processing is triggered. Because the QC engine distributes the processing of QC algorithm out to the Application servers, the total QC processing capacity of the system can also increased in the future by adding additional Application Servers.

3.0 CONCLUSION

The quality checking module has three stages where different functions are performed. The rule sets for First Stage Quality Checking (QC1) perform data validation. The Second stage Quality Checking (QC2) performs data correction based on rules. Datasets that pass both stages are ingested into the database, whereas data that fails is passed to the final stage which is manual QC (HQC). Here trained personnel make decisions on the quality and usefulness of the data.

As the whole system uses data streaming over HTTP as the primary method of transporting data from sites to the Central Servers, it is imperative to have an excellent communication system. Where communications impede the transmission of data, the QC processes slow and degrade as the system waits for data.

The single point of entry for data has enabled a standard QC method to be applied to both automatic data and data which is manually keyed into the system. It has also enabled the department to redeploy staff from the previously laborious manual checking system to more technically challenging maintenance jobs.

On the flip side, the system was designed under the assumption that communications would easily enable data streaming. There are blind spots where throughput is not as assumed. Therefore there is a need for a review of communication methods with the addition of other protocols such as FTP of hourly data for sites that are in these blind spots.

The QC method used in the system has enabled the department to review many parts of the data acquisition system especially with regards to system design and averaging algorithms. This has empowered the staff in many fields such as system design, communications and electronics. It has also enabled the department to use the understanding of meteorological principles in designing these systems.