PAPER *Special Section on Data Engineering and Information Management*

Designing a Framework for Data Quality Validation of Meteorological Data System

Wen-Lung TSAI[†a)], ***Member and*** **Yung-Chun CHAN**[†], ***Nonmember***

SUMMARY In the current era of data science, data quality has a significant and critical impact on business operations. This is no different for the meteorological data encountered in the field of meteorology. However, the conventional methods of meteorological data quality control mainly focus on error detection and null-value detection; that is, they only consider the results of the data output but ignore the quality problems that may also arise in the workflow. To rectify this issue, this paper proposes the Total Meteorological Data Quality (TMDQ) framework based on the Total Quality Management (TQM) perspective, especially considering the systematic nature of data warehousing and process focus needs. In practical applications, this paper uses the proposed framework as the basis for the development of a system to help meteorological observers improve and maintain the quality of meteorological data in a timely and efficient manner. To verify the feasibility of the proposed framework and demonstrate its capabilities and usage, it was implemented in the Tamsui Meteorological Observatory (TMO) in Taiwan. The four quality dimension indicators established through the proposed framework will help meteorological observers grasp the various characteristics of meteorological data from different aspects. The application and research limitations of the proposed framework are discussed and possible directions for future research are presented.

key words: *data quality, meteorological data, meteorological observatory, total quality management, data warehouse*

1. Introduction

Data quality plays an important role in data engineering and information systems [1]–[3]. This is especially true in the field of meteorology, which relies heavily on data sources. Meteorological observation data is not only the critical basis for weather forecasting operations, but is also an important basis for the establishment of climate data. In addition, long-term cumulative observations are often cited in data engineering and information management studies. Jeffrey et al. [4] pointed out that the meteorological data cited by most users are from long-term records of government agencies. These users believe that the data are completely accurate or have never been aware of errors in the data, leading to research findings or operational decisions that may be flawed. It can be seen that the quality of meteorological observation data not only affects the accuracy of meteorological forecasting, but any research or management decisions based on such data.

In the field of meteorology, data quality is an important research issue. Data quality studies explore how existing technology can be used to generate reliable datasets, which in turn can provide the right meteorological data for use by weather forecasting centers or external organizations [6]–[10]. Given the importance of meteorological data quality and the scope of its influence, many countries and regions have established meteorological observatories to obtain more complete meteorological data and have transmitted these large amounts of data to central meteorological stations using information technology processes. They have also simultaneously developed various types of validation checks to ensure that the meteorological datasets are reliable and accurate.

Many meteorological data quality validation methods currently exist, such as range validation, change rate validation, and static data validation methods. These methods are based primarily on physical characteristics and focus on validating the result data. However, in this study, we found that these methods are still inadequate. Consequently, we focused on comprehensive data quality problems based on total quality management (TQM) and information technology (IT). In the TQM perspective, the generation and presentation of meteorological data is process-oriented and should not only be results-based [11], [12]. For example, when the data are collected from a sensor, stored in the database, and manually corrected, do they suffer interference from other factors, which in turn affects data quality. From the IT perspective, the process of generating and using meteorological data is a form of IT presentation. Specifically, a data warehousing system processes large amounts of data and, therefore, there is a set of relevant data quality assurance requirements. Given these cross-disciplinary considerations, what can be done from a more comprehensive perspective when validating meteorological data? To answer this question and consequently solve this problem, in this study, the existing quality assurance methods in the field of meteorology and the concept of TQM were utilized as the foundation for integrating the IT and process focus perspectives. Consequently, the contribution in this paper is a proposed TMDQ framework that assists key personnel to fully ascertain meteorological data quality.

Regarding the application of data quality, Fu and Easton [13] used the four data quality dimensions—accuracy, consistency, completeness, and timeless understanding—in their design of the UK rail industry. They illustrated the process for deriving a data quality schema embedded in a data model. Karkouch et al. [14] categorized data quality into four dimensions for Internet of Things (IoT) data lifecycle: intrinsic, contextual, representational, and accessibil-

Manuscript received June 20, 2018.
Manuscript revised November 3, 2018.
Manuscript publicized January 22, 2019.

[†] The authors are with OIT, New Taipei 22061, Taiwan.
a) E-mail: wltsai@mail.oit.edu.tw (Corresponding author)

DOI: 10.1587/transinf.2018DAP0021

ity. Three of those dimensions (intrinsic, contextual, and representational) are similar to accuracy, completeness, and timeliness. Jones-Farmer et al. [15] used the control chart method, a statistical monitoring approach, to measure aircraft maintenance data quality.

The system proposed in this paper employs the IT object concept and is established based on the four data quality dimensions: accuracy, consistency, completeness, and timeliness. In addition, whereas the dimensions of meteorological data quality in previous studies [11], [12] stressed the accuracy, completeness, and consistency of the result-based historical data, in addition to the quality of the historical data, this study also focused on real time—specifically, the timeless quality dimension. Consequently, this paper defines the relevant operational metrics based on the practical needs observed during the implementation of the system. Further, in this study, Unified Modeling Language (UML) was applied to design and develop a system that not only verifies the feasibility of the proposed framework but also automates monitoring of the large amounts of meteorological data involved and improves its efficiency. Temperature data from 2005 to 2010 at the TMO were applied to implement system functions and usages and three labels—specifically, correct, suspicious, and missing—were applied to check the meteorological data.

The remainder of this paper is organized as follows. Section 2 gives a literature review that includes studies from the fields of data warehousing and data quality along with an exploration and review of existing methods of data validation in the field of meteorology. Section 3 outlines the research method and design and gives an overview of the TMDQ framework. Section 4 describes the implementation of the system. Section 5 discusses the experiments conducted and analyzes the results obtained. Section 6 summarizes the paper, discusses the application and research limitations, and presents possible directions for future research.

2. Literature Review

2.1 Data Warehousing

The concept of data warehousing originated in the 1960s. With advancements in technology and substantial increases in the amount of information in enterprise information systems, related data warehousing applications have attracted increasing attention in recent years. Most companies use data warehousing as the core of their database and then build various analytical applications, such as financial analysis systems, customer relationship management systems, and decision support systems on top of it [17]. In the field of meteorology, the practice of storing meteorological data in data warehouses by the meteorological centers of many nations has gradually become more commonplace [18], [19]. Data warehousing is a common and necessary step in business data analysis. According to the definition given in Janssen et al. [20], data warehousing is thematically oriented and integrated, and considers time-correlated datasets that are used to assist in the management decision-making process.

The applications of data warehousing are extensive, but the fundamental reason for establishing a data warehouse is to provide users with higher quality data to assist their decision-making. Therefore, the quality of data provided by data warehouses is a key success factor. In the relevant literature on data warehousing, "quality management" is often mentioned. In quality management, the question of whether the information provided by the system is of a sufficient quality is critical. Therefore, this paper also discusses data quality and total quality management.

2.2 Data Quality

For a considerable period, researchers have widely used a variety of properties or facets to measure data quality synergies; these are the more objective assessment methods [2], [3], [21]. In this paper, data quality in data warehouses is discussed from the total quality management perspective. Several scholars have proposed process management as a critical component [21]–[23]. Therefore, when assessing the output quality of systems, it is necessary to start with the process flow.

Ballou and Pazer [16] studied data models and process quality models and revealed that data quality can be approached from many angles. They used four dimensions to assess inadequacies in the data quality of information systems—specifically, accuracy, completeness, consistency, and timeliness. Chen et al. [1] also pointed out that in order to measure the correctness of information systems, especially those used in data warehousing, it is important to use the above four dimensions to evaluate the data separately. Similarly, in this study, among the many aspects of quality, the four dimensions—accuracy, completeness, consistency, and timeliness—were deemed to be more in line with the TQM perspective for measuring data quality in a data warehousing scenario. Consequently, these four quality dimensions are used in this paper as the standard for measuring data quality and evaluating problems arising from data quality.

2.3 Existing Quality Management Methods in the Field of Meteorology

In the field of meteorology, data quality control is the most critical part of quality management. Thus, the main purpose of null data detection, error detection, and error correction is to ensure that the data are highly accurate before being provided to users [24]. Table 1 classifies the validation checks and definitions of quality control procedures for meteorological data based on existing research.

Combining the four quality dimensions, the concept of process focus, and the existing meteorological data quality assurance methods mentioned above, we found that although existing meteorological data validation checks can identify suspicious data, they are too narrow and can only indicate whether the data have passed the validation checks;

they cannot adequately express the meaning behind the data. In addition, after comparing the quality of information required from the IT viewpoint and that of the existing meteorological data, we found that the existing validation checks for meteorological data are still inadequate. The existing meteorological data quality checks tend to use a variety of validation checks to determine whether the observed values meet certain meteorological characteristics or are aligned with certain meteorological relationships; that is, they only consider the results of data output. However, such validation methods ignore comprehensive data quality problems, which may arise from operational processes. For example, it is impossible to assess whether observers can apply manual corrections in a timely manner. Therefore, from the IT and TQM perspectives, data quality validation must also consider the impact the process flow has on data quality in addition to assessing data quality on the basis of results. As such, this paper integrates this concept and establishes a meteorological data quality framework, which is discussed below.

Table 1 Validation checks and definitions

Validation checks		Definition	Source
Range checks	Sensor-based range test	Find the observed values that exceed the limit, which is based on the allowable range of the measurement instrument	[25, 26]
	Climate-based range test	Define a reasonable range for each meteorological element and find observations that exceed said range	[5][9] [24-26]
Temporal checks	Rate of change test	Check whether the difference between the observed values before and after the validations is within a reasonable range	[5][9] [24-26]
	Step test	Check whether the difference in observations over a period is within a reasonable range	[5] [25-26]
	Persistence test	Check whether the observed value remains unchanged beyond a reasonable time interval	[5][9] [24-26]
Spatial checks		Check whether the difference in observations over a period is within a reasonable range	[5] [26]
Consistency checks		Find unreasonable variable relationships between relevant meteorological elements using the relationship between physical and meteorological values	[9] [24-26]
Validation checks		Definition	Source
Missing data checks		Check if the observation value is a null value	[5] [24, 25]

3. Research Design and Methods

3.1 Research Framework

TMDQ uses the works of [1] and [16] as a theoretical foundation for the use of the four data quality definitions as high-level collections of performance indicators: accuracy, consistency, completeness, and timeliness. As shown in Fig. 1, if the framework based on the TMDQ is presented as an IT concept and visualized using UML, it can be seen that data quality includes four dimensions in a so-called aggregation association, represented by the diamond shape (has-a relationships). This paper assesses the meteorological data quality using these four dimensions from different perspectives.

If the basic framework of TMDQ is presented as the objects concepts of IT and UML, it is possible to determine that data quality possesses the four dimensions of (has-a), i.e., the so-called aggregation association, which is represented by the hollow diamond shapes in Fig. 1. Therefore, this paper, in assessing data quality, measures meteorological data through these four dimensions from different perspectives.

The TMDQ framework can be further expanded as represented in Fig. 2 by integrating existing meteorological data quality assurance methods on the basis of the four quality dimensions mentioned above. Data quality has four dimensions, with each dimension inheriting different meteorological validation rules, such as temporal checks, spatial checks, and consistency checks. Some of the validation rules can be subdivided into different tests.

In this paper, the temperature data at TMO are used as an example. The range check method is subdivided into sensor-based range test and climate-based test, and the temporal check method is subdivided into rate of change test, step test, and persistence test. The details of these tests are as follows.

- Sensor-based range test: This test is defined by the

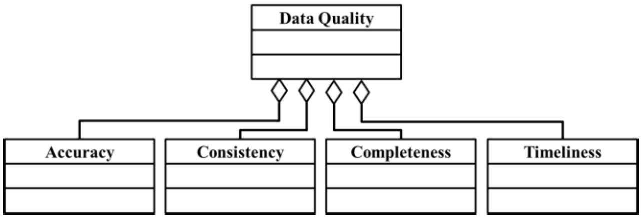


Fig. 1 TMDQ framework

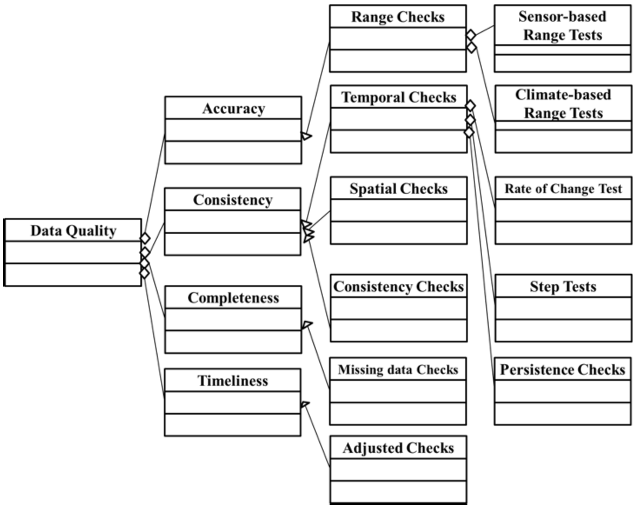


Fig. 2 TMDQ quality dimensions and meteorological validation checks

measurement range of the temperature sensors at TMO. The default range is defined between −30°C and 80°C. The temperature data are treated as being in error if the value is outside the measurement range.

- Claim-based range test: This paper substitutes the record of the maximum temperature (38.8°C) and minimum (15.4°C) in summer, and the maximum (30.5°C) and minimum (2.3°C) in winter from 1943 to 2009 at TMO into Eqs. (1) and (2), provided by Fiebrich et al. [26]. The maximum and minimum temperature are also applied as the upper and lower thresholds for each season. After calculating the upper and lower thresholds in each season, any measured temperature that is outside of the following ranges is considered suspicious.

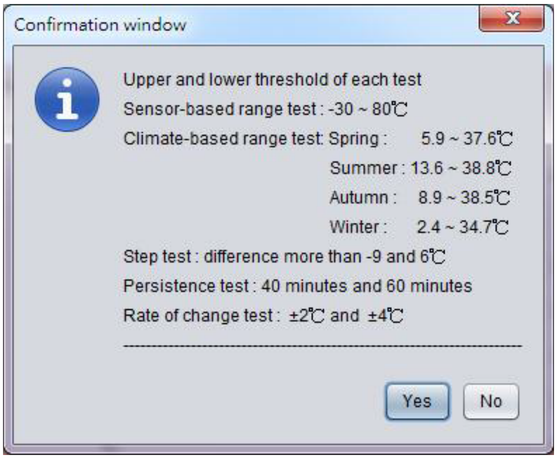
Spring:	5.9–37.6°C
Summer:	13.6–38.8°C
Autumn:	8.9–38.5°C
Winter:	2.4–34.7°C

$$T_{\max}(d) = T_{\max COL}D + (T_{\max HOT} - T_{\max COLD})\cos\{0.5\pi(d - 183)/183\} \quad (1)$$

$$T_{\min}(d) = T_{\min COLD} + (T_{\min HOT} - T_{\min COLD})\cos\{0.5\pi(d - 183)/183\} \quad (2)$$

[Variable Description]
d: day ordinal of a year, among 1 and 366
$T_{\max}(d)$: upper temperature on the d^{th} day.
$T_{\min}(d)$: lower temperature on the d^{th} day
$T_{\max}$HOT: maximum temperature in summer from historical data
$T_{\min}$HOT: minimum temperature in summer from historical data
$T_{\max COLD}$: maximum temperature in winter from historical data
$T_{\min COLD}$: minimum temperature in winter from historical data

- Step test: According to the Fiebrich et al. [26] and the experts in the Central Weather Bureau (CWB) in Taiwan, if the temperature increases by more than 6°C or decreases by more than 9°C, the measured data can be considered suspicious.
- Persistence test: According to the guidelines on quality control procedures in the World Meteorological Organization [24] and the experts at CWB in Taiwan, if the temperature is persistent with no change for 40 minutes, the data in the interval can be considered suspicious. If the temperature is persistent with no change for 60 minutes, the measurement is considered extremely suspicious.
- Rate of change test: According to Reek et al. [9] and the experts at CWB, if the temperature varies by 2°C or more in one minute, the data can be considered as suspicious. Further, if the temperature varies by 4°C or more within one minute, the data would be considered extremely suspicious.


Confirmation window

Upper and lower threshold of each test
Sensor-based range test : -30 ~ 80℃
Climate-based range test: Spring : 5.9 ~ 37.6℃
Summer : 13.6 ~ 38.8℃
Autumn : 8.9 ~ 38.5℃
Winter : 2.4 ~ 34.7℃
Step test : difference more than -9 and 6℃
Persistence test : 40 minutes and 60 minutes
Rate of change test : ±2℃ and ±4℃

Fig. 3 Default value range settings

Considering the above tests, in this study, the default value range for the proposed system was configured as depicted in Fig. 3.

In the process of collecting information from information systems, data errors are often introduced because of instrument failures, false records owing to human intervention, clerical errors, and input errors. For this reason, this paper defines accuracy as “the observed value falling within a reasonable predefined range.” The so-called reasonable range is defined by the upper and lower limits of the existing range-validation rule. The equation is as follows:

$$\textit{Accuracy} = 1 - \frac{\text{Number of data points that failed range checks}}{\text{Total number of data points tested}} \quad (3)$$

Generating meteorological observations is a continuous process. Therefore, the problem of missing data throughout the observation process must be minimized and factors such as sensor failures or problems during the transmission process, which can cause data interruptions, avoided. Therefore, this paper defines completeness as “the absence of null values in the dataset." The equation is as follows:

$$Completeness = 1 - \frac{\text{Number of missing data points}}{\text{Total number of data points tested}} \tag{4}$$

In the field of meteorology, most changes in meteorological elements have limitations defined by relationships between time, space, and other factors. Therefore, this paper defines consistency as "the observed value conforming to relationships between time, space, and other variables." The equation is as follows:

$$Consistency = 1 - \frac{\begin{array}{c}\text{Number of data points that failed temporal checks +}\\ \text{Number of data points that did not pass the spatial checks +}\\ \text{Number of data points that did not pass the consistency checks}\end{array}}{\text{Total number of data points tested}} \tag{5}$$

During system operation, the generation or presentation of data is a process. The process of data generation often affects the quality of the data finally presented. Therefore, this paper defines timeliness as "received (or transmitted) observations reflecting the latest state of the dataset." The equation is as follows:

$$Timeliness = 1 - \frac{\text{Number of data points not corrected for each receive/transmit period}}{\text{Total number of data points tested}} \tag{6}$$

This paper limits the quantified values of the dimensions between zero and one. The higher the value is, the more accurate, complete, consistent, and timely is the system.

3.2 Process and Functional Architecture

Meteorological data are typically collected through various meteorological sensors from various stations. The data collected by various sensors are called real-time data. The substantial meteorological data gathered by observatories are transmitted to the central station for storage and subsequent use. In other words, meteorological data are mainly stored in various observatories and the central station. These data stored in various observatories and the central station are called historical data. In general, the stored meteorological data (i.e., historical data) are used or applied monthly.

The data from observatories include information about various meteorological elements in the area, whereas the data in the central station are an aggregation of the data stored by the various observatories. This paper uses UML to explain the TMDQ, thereby presenting its processes and functions.

3.2.1 TMDQ: Discussion of System Functions

The main function of the use case diagram in UML is to describe system requirements and model system functions and applications. Therefore, this paper uses such a diagram to explain the functions of the TMDQ framework, as shown in Fig. 4. It can be seen from the figure that the actors include human observers and system administrators. The oval shapes represent use cases, which describe actions and functions that the validation system can perform. Here, the source of data, real-time data validation, historical data validation, the statistical quality of historical data, quality label generation, observed value correction, and quality criteria revision have been selected.

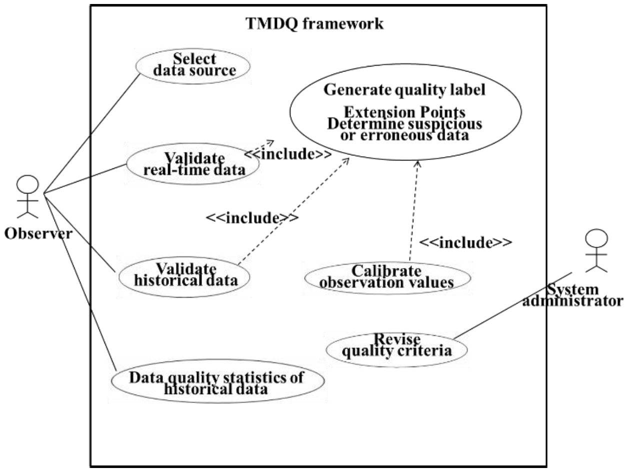


Fig. 4 Use case diagram of the TMDQ framework

The following is a functional description of each use case:

- Select data source: select the source of the meteorological data to be validated
- Validate real-time data: validate the observations received every period
- Validate historical data: validate historical meteorological data
- Generate quality label: each observation value is tagged with the corresponding label (i.e., correct, suspicious, or error) based on the results of validation against the quality dimensions
- Correct observation value: observers perform numerical corrections directly on suspicious or erroneous data through the system interface
- Data quality statistics of historical data: the number of correct, suspicious, and missing historical data
- Revise quality criteria: revise the content, boundaries, execution time, and other properties of the quality criteria in response to climate changes

Note that the observer and the system administrator are in charge of different system functions. However, owing to human resource allocation at TMO, the same person performed the system role of the observer and system administrator. That is, the observer at TMO was the system administrator.

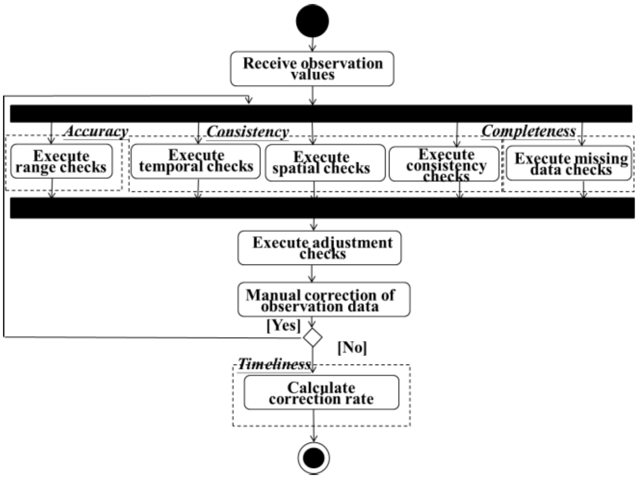

Receive observation values
Accuracy
Consistency
Completeness
Execute range checks
Execute temporal checks
Execute spatial checks
Execute consistency checks
Execute missing data checks
Execute adjustment checks
Manual correction of observation data
[Yes]
[No]
Timeliness
Calculate correction rate

Fig. 5 Detailed activity diagram for validating real-time data in the TMDQ framework

3.2.2 TMDQ: Discussion of System Processes

The purpose of the activity diagram in the UML is to model the execution process of a predefined information system process workflow. Therefore, this subsection analyzes the process further using the activity diagram, shown in Fig. 5. In general, the system receives meteorological observation data every predetermined period. In order to assess the quality of the observed values, the system, in parallel, quantifies the three main quality dimensions—namely, accuracy, consistency, and completeness. According to the definitions in the previous section, these three dimensions are primarily used to determine whether the received values conform to the meteorological, physical, and logical relationships. At the same time, different validation results are stored depending on the implementation method. Therefore, after quantifying the three major quality dimensions, the system generates quality labels based on the results. Through this step, the system can first filter out suspicious data, and the observer can directly perform data correction on it. If the observer has corrected the observed value, the system recalibrates the quality values and generates new quality labels using the previous process. The purpose of this step is to avoid the influence of human input errors on data quality during manual corrections. Finally, the system periodically quantifies the timeliness dimension by calculating the correction ratio to ensure that the observer has been correcting suspicious or misleading data in a timely manner.

In summary, the quality assessment process for meteorological data can be divided into two major steps. The first step is generation of quality labels according to the accuracy, consistency, and completeness dimensions in order to screen out suspicious or erroneous data. The purpose of this portion is to validate meteorological data. The second step is to have observers correct incorrect values based on the data labels generated in the first step. This step also ensures that manual corrections are performed in a timely manner by quantifying the timeliness dimension at the end.

4. System Implementation and Presentation

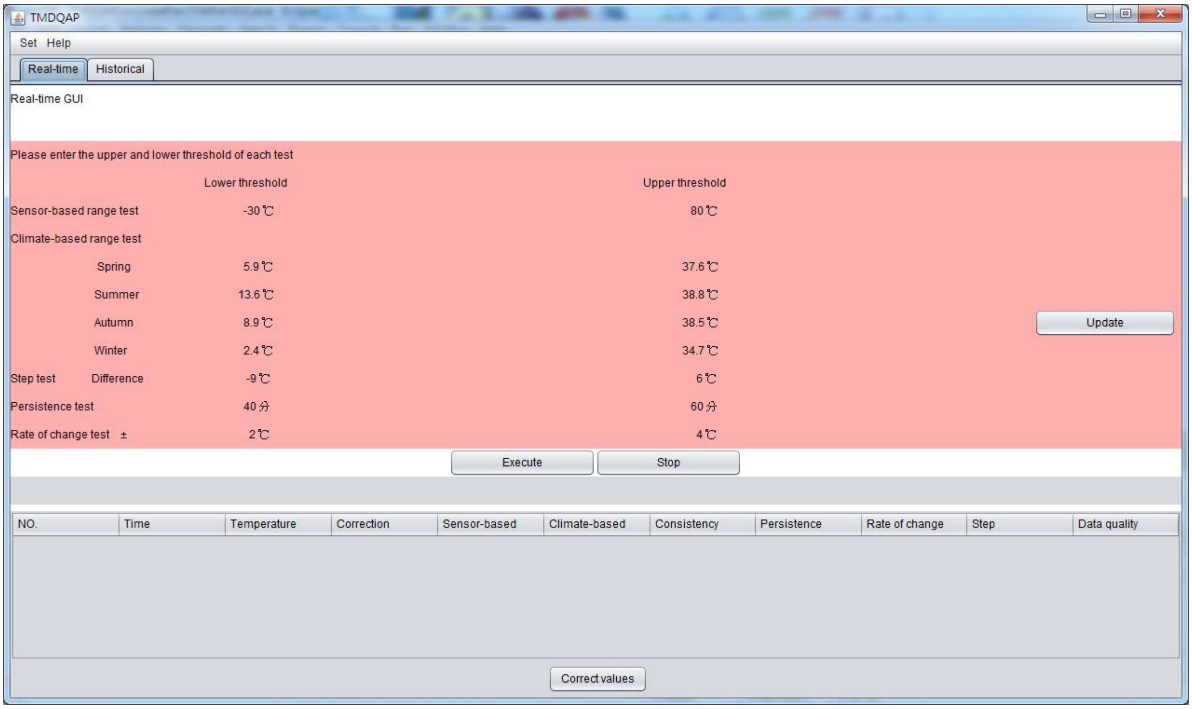
In this study, a system (called TMDQAP) was in the Java programming language to automate data processing and quantify quality aspects through built-in formulas to assist observers in determining meteorological data quality. A Microsoft SQL Server 2008 R2 database was used to store data. For data sourcing, minute-by-minute temperature data from TMO from 2005 to 2010 were used. A total of 2,779,205 data points were used as test objects. Owing to limitations in data acquisition and to avoid having an excessive amount of content, the scope of implementation was restricted to the TMO and tests were conducted only on the minute-by-minute temperature data.

Prior to the implementation of the TMDQAP, system administrators could select the source of the data to be verified and confirm the validation checks. After the initial settings have been selected, the data can be validated. The meteorological data validated by the TMDQAP can be divided into two main types: real-time data and historical data. The system interface for validating real-time data, as shown in Fig. 6, is divided into two main parts: the boundary-value display area (upper portion of Fig. 6) and the suspicious or null data display area (lower portion of Fig. 6). The system interface for validating historical data, as shown in Fig. 6, is divided into three main parts: the time-selection area for data to be validated (upper portion of Fig. 7), validation-check boundary-value display area (middle portion of Fig. 7), and suspicious or null data display area (lower portion of Fig. 7).

In the time-selection area for data validation, a drop-down menu is used to provide users with a quick way to select the time interval of the data they want to validate. The validation-check boundary values pane provides observers with a view of the currently used validation checks and allows changes to be made. The values currently shown in the system interface are those used by this paper based on the literature reviewed and are combined with boundary values derived from historical data and the local climate characteristic boundaries of the TMO. Finally, the suspicious or null data panes display individual information points determined by the system to be suspicious or containing null elements.

4.1 Validating Real-Time Data

When observers click on the “execute” button to validate real-time data, the system pops up a confirmation window to allow them to confirm the limits of each validation check. If the boundary values are correctly set, the system begins to perform data quality validation and quantify accuracy, consistency, and completeness. Therefore, when observation data are received, the system performs validation based on the methods corresponding to the three major dimensions and stores these results in the database. After quantifying the three quality dimensions, the system combines the re-



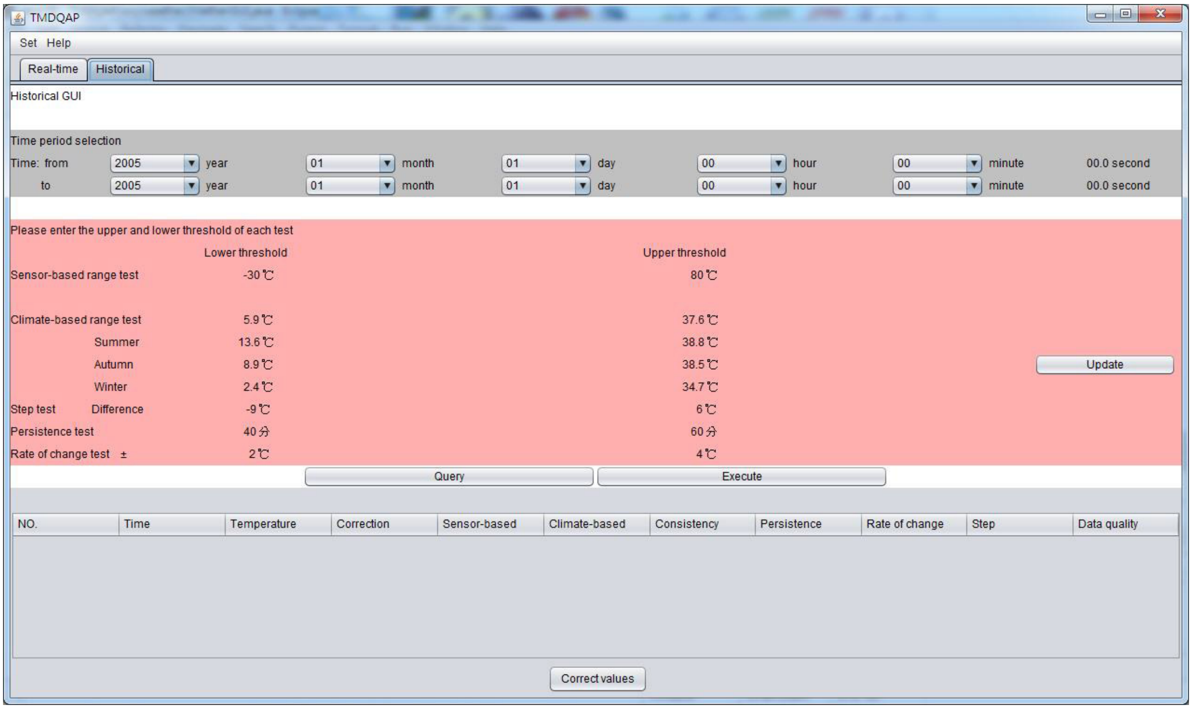
Real-time GUI

Please enter the upper and lower threshold of each test

		Lower threshold	Upper threshold
Sensor-based range test		-30℃	80℃
Climate-based range test			
	Spring	5.9℃	37.6℃
	Summer	13.6℃	38.8℃
	Autumn	8.9℃	38.5℃
	Winter	2.4℃	34.7℃
Step test	Difference	-9℃	6℃
Persistence test		40分	60分
Rate of change test	±	2℃	4℃

NO.	Time	Temperature	Correction	Sensor-based	Climate-based	Consistency	Persistence	Rate of change	Step	Data quality

Fig. 6 System interface: real-time data validation



Historical GUI

Time period selection

Time: from	2005	year	01	month	01	day	00	hour	00	minute	00.0 second
to	2005	year	01	month	01	day	00	hour	00	minute	00.0 second

Please enter the upper and lower threshold of each test

		Lower threshold	Upper threshold
Sensor-based range test		-30℃	80℃
Climate-based range test		5.9℃	37.6℃
	Summer	13.6℃	38.8℃
	Autumn	8.9℃	38.5℃
	Winter	2.4℃	34.7℃
Step test	Difference	-9℃	6℃
Persistence test		40分	60分
Rate of change test	±	2℃	4℃

NO.	Time	Temperature	Correction	Sensor-based	Climate-based	Consistency	Persistence	Rate of change	Step	Data quality

Fig. 7 System interface: historical data validation

sults and generates a quality label for the minute-by-minute temperature data. When the system determines that the data is suspicious or may be missing certain elements, it presents these data points individually in the form of bars below the execution display.

As such, observers can use this interface to perform numerical corrections directly on suspicious or erroneous data. Finally, the system calculates a correction rate (a type of adjustment check) every 60 min to ensure that observers have been making data corrections in a timely manner each hour. When the preset 60 min period elapses, the system automatically stores the four quality dimension indicators and the list of suspicious or erroneous data in a .txt file. Using this output file, the user can clearly see the validation time of the file, the results of the four quality indicators, the original suspicious temperature data generated, the validation results before manual correction, and the missing values. The file represents the data quality report of a specific timeframe and serves as proof of quality assurance. The user can use the quantified quality results and the list of suspicious data points to determine whether to check and correct data that fall below a certain standard.

4.2 Validating Historical Data

When the observer wants to validate historical data, he/she must first select the period to be validated through the dropdown menu. Similarly, after the observer clicks on the "execute" button, the system pops up a confirmation window to allow them to confirm the period and the limits of each validation check. If it is determined that the set values are correct, the system starts data validation according to the settings.

The difference between real-time data validating and historical data is in the system interface, because validating historical data involves a much larger amount of data. Thus, when presenting the validation results, only the statistics pertaining to each data quality tag are shown. If the observer wishes to go further and see which data points are in question, he/she can click on the "list" button to view detailed information, including the validated time interval, temperature, and other values of the suspicious data points, the results of each validation, and the null data. Similarly, the list can be stored as a file in a location designated by the observer and serve as a reference for subsequent data inquiries or as proof of data assurance.

Moreover, if the observer did not click the "update" button after validating historical data, he/she can click the "query" button to retrieve the past validation results after querying the relevant time interval. The system then counts the number of correct, suspicious, and null values, and provides a detailed list of these values based on the selected interval so that the observers can update them.

5. Analysis and Discussion

5.1 Reliability Analysis

To evaluate the reliability of the system, the following tests were performed: First, 129,100 correct data points were selected from the dataset of TMO for 2010, out of which 500 data points were randomly selected for conversion into 262 null values and 238 error values as test data and imported into the system for data validation, as shown in Table 2. The results of the data validation performed by the system are shown in Table 3.

Table 2 Original data before validation

Data type	Number
Correct data	129,100
Error data	238
NULL	262
Total	129,600

Table 3 Validation results

Data type	Number
Correct data	129,069
Suspicious data	269
NULL	262
Total	129,600

Based on the above results, we determined the system to be 100% accurate in detecting missing values (262/262 = 100%). However, the system found more instances of suspicious data than there actually were. After cross-checking, it was found that the suspicious data identified by the system included the original 238 erroneous data points and 31 misjudged data points. The reason for this is that after the system discovers a data error, the next data point is affected by the error detected in the previous data point and is unable to pass the step test and the rate of change test, consequently being misjudged by the system.

In addition, we adopted the method of statistical hypotheses and error to determine the reliability of the proposed system. A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. The dataset used in this study is the historical temperature data for 2010 at TMO. Type I error is the rejection of a true null hypothesis (also known as a "false positive" finding); that is, errors in the test results can occur even when the system is able to filter out the error data. This is one of the specific reasons for rechecking for correct, suspicious, and missing data in the meteorological data used in this study.

From the test results presented above, it can be seen that although the system was able to filter out all the error data, it also committed Type I errors. Thus, it is clear that validation of meteorological data differs from the theoretical research of the general sciences. In this case, committing a Type II error causes a more considerable and irreparable impact compared to a Type I error. Therefore, the guiding principle behind the design is that "it is better to convict wrongly than to let a criminal escape," which means that it is better to wrongly tag a data point as suspicious rather than allow a suspicious data point to slip through. Although there are instances of Type I errors in the current system, the misjudgments are allowed, provided that the number is below a certain threshold.

5.2 Expected Benefits

In this study, a semi-structured interview format was primarily used to conduct in-depth interviews with CWB observers and system administrators to predict the benefits that the system will bring. The technology acceptance model was used as the basis for the interviews. Two variables, cognitive usefulness and cognitive ease of use, were used to predict the users' acceptance of the system and then infer the values and benefits the system will bring to users [27].

Based on the results of the interviews, it was discovered that if the respondents' job is heavily reliant on information systems, then it is easier to establish cognitive usefulness. At present, the process of validating meteorological data requires a system to assist the relevant work units in validating the data. Therefore, all respondents responded positively to the system in terms of cognitive usefulness. In terms of cognitive ease of use, the results of the interviews show that the respondents found it easy to operate the system and that the interface design conforms to the requirements of user-friendly interfaces; in other words, it was easy to perform system operations.

Therefore, based on the above discussion, it can be concluded that positive cognitive usefulness and cognitive ease of use of the system can produce a positive attitude and willingness to use in the target users. According to the inferences made based on the technology acceptance model, we expect that the system will bring a certain degree of benefit and is confident that said system can improve the quality of meteorological data validation. However, it should be noted that in the future, the system must cover validation of all meteorological data elements in order to bring about tangible benefits.

5.3 Research Limitations

The quality validation system developed in this study is aimed at meteorological data in the field of meteorology and carries out data quality validations. At the same time, the data validation performed as part of this study was limited to temperature data, and so the system cannot be readily applied to the validation of other meteorological data elements. The boundary values of this system were set according to the temperature characteristics of the TMO. The boundary values of the validation checks should be adjusted based on the locations of different stations. If the system is to be used to validate the temperature data of other locations, the weather conditions and characteristics of the target location must be obtained in advance and then used to modify the upper and lower limit values of the validation checks of the system.

6. Conclusion

The TMDQ framework proposed in this paper is based on the quality requirements of total quality management and data warehousing. It helps functional units handling meteorological data validation to examine the quality of meteorological data according to four quality dimensions. Previously, in the field of meteorology, only the results of data output were considered when evaluating quality. Compared to conventional methods, the proposed TMDQ framework not only provides observers with a view of data quality from four quality dimensions, it also spots quality issues that may be caused by manual corrections, thereby taking a more comprehensive approach to validating meteorological data. The meteorological data quality validation system developed in this study can assist observers by improving their work efficiency and helping them grasp and monitor meteorological data quality more effectively.

However, the system currently only tags data according to three quality labels: correct, suspicious, and missing. In subsequent studies, the "suspicious" data quality label can be further divided into different degrees, such as "slightly suspicious," "moderately suspicious," and "definitely incorrect." By so doing, the observer can start performing manual corrections on the more obviously wrong data points while the system is sorting data according to the degree of suspiciousness. In addition, the system currently performs data validation only on minute-by-minute temperature data. It is hoped that the methods presented in this paper will be used to validate many types of meteorological data rather than just one type (temperature data). Therefore, in the future, the four quality dimensions of the TMDQ framework will be further developed for application with different meteorological elements, thereby enabling comprehensive meteorological data validation.

References

[1] C.Y. Chen, Y.L. Chi, and P. Wolfe, "An object-oriented quality framework with optimization models for managing data quality in data warehouse applications," Int. J Oper. Res. 2, no.2, pp.1–81, 2005.

[2] Y.W. Lee, D.M. Strong, B.K. Kahn, and R.Y. Wang, "AIMQ: A methodology for information quality assessment," Inf. Manage., vol.40, no.2, pp.133–146, 2002.

[3] R.Y. Wang and D.M. Strong, "Beyond accuracy: What data quality means to data consumers," J. Manage. Inf. Syst., vol.12, no.4, pp.5–33, 1996.

[4] S.J. Jeffrey, J.O. Carter, K.B. Moodie, and A.R. Beswick, "Using spatial interpolation to construct a comprehensive archive of Australian climate data," Environ. Modell. Softw., vol.16, no.4, pp.309–330, 2001.

[5] S. Feng, Q. Hu, and W. Qian, "Quality control of daily meteorological data in China, 1951–2000: A new dataset," Intl. J. Climatol., vol.24, no.7, pp.853–870, 2004.

[6] J.F. González-Rouco, J.L. Jiménez, V. Quesada, and F. Valero, "Quality control and homogeneity of precipitation data in the southwest of Europe," J. Clim., vol.14, no.5, pp.964–978, 2001.

[7] A.M.G.K. Tank, J.B. Wijngaard, G.P. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. Müller-Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A.F.V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J.A. López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L.V. Alexander, and P. Petrovic, "Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment," Int. J. Climatol., vol.22, no.12, pp.1441–1453, 2002.

[8] M.J. Manton, P.M. Della-Marta, M.R. Haylock, K.J. Hennessy, N. Nicholls, L.E. Chambers, D.A. Collins, G. Daw, A. Finet, D. Gunawan, K. Inape, H. Isobe, T.S. Kestin, P. Lefale, C.H. Leyu, T. Lwin, L. Maitrepierre, N. Ouprasitwong, C.M. Page, J. Pahalad, N.

Plummer, M.J. Salinger, R. Suppiah, V.L. Tran, B. Trewin, I. Tibig, and D. Yee, “Trends in extreme daily rainfall and temperature in Southeast Asia and the South Pacific: 1961–1998,” Int. J. Climatol., vol.21, no.3, pp.269–284, 2001.

[9] T. Reek, S.R. Doty, and T.W. Owen, “A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network,” Bull. Amer. Meteor. Soc., vol.73, no.6, pp.753–762, 1992.

[10] L.A. Vincent, X. Zhang, B.R. Bonsal, and W.D. Hogg, “Homogenization of daily temperatures over Canada,” J. Clim., vol.15, no.11, pp.1322–1334, 2002.

[11] M. Vakili, S.R. Sabbagh-Yazdi, S. Khosrojerdi, and K. Kalhor, “Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data,” J. Clean. Prod., vol.141, pp.1275–1285, 2017.

[12] C. Leauthaud, B. Cappelaere, J. Demarty, F. Guichard, C. Velluet, L. Kergoat, T. Vischel, M. Grippa, M. Mouhaimouni, I. Bouzou Moussa, I. Mainassara, and B. Sultan, “A 60-year reconstructed high-resolution local meteorological data set in Central Sahel (1950–2009): Evaluation, analysis and application to land surface modelling,” Int. J. Climatol., vol.37, no.5, pp.2699–2718, 2017.

[13] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, “Data quality in internet of things: A state-of-the-art survey,” J. Netw. Comput. Appl., vol.73, pp.57–81, 2016.

[14] Q. Fu and J.M. Easton, “Understanding data quality: Ensuring data quality by design in the rail industry,” 2017 IEEE International Conference on Big Data (Big Data), pp.3792–3799, IEEE, 2017.

[15] L.A. Jones-Farmer, J.D. Ezell, and B.T. Hazen, “Applying control chart methods to enhance data quality,” Technometrics, vol.56, no.1, pp.29–41, 2014.

[16] D.P. Ballou and H.L. Pazer, “Modeling data and process quality in multi-input, multi-output information systems,” Manage. Sci., vol.31, no.2, pp.150–162, 1985.

[17] V.C. Storey and I.-Y. Song, “Big data technologies and Management: What conceptual modeling can do,” Data Knowl. Eng., vol.108, pp.50–67, 2017.

[18] H. Jinghua, Y. Mei, L. Xiaowei, and S. Xinna, “The design and implementation of MDSS based on data warehouse,” 2010 International Conference on Computing, Control and Industrial Engineering (CCIE), vol.1, pp.42–45, IEEE, 2010.

[19] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer, “Visplause: Visual data quality assessment of many time series using plausibility checks,” IEEE Trans. Vis. Comput. Graph., vol.23, no.1, pp.641–650, 2017.

[20] M. Janssen, H. van der. Voort, and A. Wahyudi, “Factors influencing big data decision-making quality,” J. Bus. Res., vol.70, pp.338–345, 2017.

[21] I. Sila and M. Ebrahimpour, “Critical linkages among TQM factors and business results,” Int. J. Oper. Prod. Manage., vol.25, no.11, pp.1123–1155, 2005.

[22] J. José Tarí, “Components of successful total quality management,” The TQM magazine, vol.17, no.2, pp.182–194, 2005.

[23] D.T. Hoang, B. Igel, and T. Laosirihongthong, “The impact of total quality management on innovation: Findings from a developing country,” Int. J. Qual. Reliab. Manage., vol.23, no.9, pp.1092–1117, 2006.

[24] I. Zahumenský, Guidelines on quality control procedures for data from automatic weather stations, World Meteorological Organization, Switzerland, 2004.

[25] D.Y. Graybeal, A.T. DeGaetano, and K.L. Eggleston, “Complex quality assurance of historical hourly surface airways meteorological data,” J. Atmospheric Ocean. Technol., vol.21, no.8, pp.1156–1169, 2004.

[26] C.A. Fiebrich, C.R. Morgan, A.G. McCombs, P.K. Hall, and R.A. McPherson, “Quality assurance procedures for mesoscale meteorological data,” J. Atmospheric Ocean. Technol., vol.27, no.10, pp.1565–1582, 2010.

[27] F.D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” MIS Quarterly, vol.13, no.3, pp.319–340, 1989.

Wen-Lung Tsai is an Assistant Professor in the Department of Information Management, Oriental Institute of Technology (OIT), Taiwan R.O.C. He received his Ph.D. degree in Information Management at National Central University (NCU) in 2015. His current research and teaching interests include software engineering, project management, data quality, and information institution.

Yung-Chun Chan is a Research Assistant in the Department of Information Management, Oriental Institute of Technology (OIT), Taiwan R.O.C. Her current research interests include software engineering, project management, cloud service, and business intelligence.