Data quality assessment in hydrological information systems

Li Chao, Zhou Hui and Zhou Xiaofeng

ABSTRACT

The hydrological data fed to hydrological decision support systems might be untimely, incomplete, inconsistent or illogical due to network congestion, low performance of servers, instrument failures, human errors, etc. It is imperative to assess, monitor and even control the quality of hydrological data residing in or acquired from each link of a hydrological data supply chain. However, the traditional quality management of hydrological data has focused mainly on intrinsic quality problems, such as outlier detection, nullity interpolation, consistency, completeness, etc., and could not be used to assess the quality of application – that is, consumed data in the form of data supply chain and with a granularity of tasks. To achieve these objectives, we first present a methodology to derive quality dimensions from hydrological information system by questionnaire and show the cognitive differences in quality dimension importance, then analyze the correlations between the tasks, classify them into five categories and construct the quality assessment model with time limits in the data supply chain. Exploratory experiments suggest the assessment system can provide data quality (DQ) indicators to DQ assessors, and enable authorized consumers to monitor and even control the quality of data used in an application with a granularity of tasks.

Key words | data quality, data supply chain, hydrological information system, quality assessment, quality dimension, task

Li Chao (corresponding author)
Zhou Hui
Zhou Xiaofeng
College of Computer Science and Information Engineering,
Hohai University,
Nanjing 211100,
China
E-mail: *springxun@163.com*

Li Chao
Information Engineering College,
Hubei Minzu University,
Enshi 445000,
China

ABBREVIATIONS AND NOTATION

C	Connectors. It includes five types of connector: sequential structure C_{sq}, recursive structure C_{rc}, selective structure C_{sl}, concurrent structure C_{cc}, and composite structureC_{cs}
CS	Customer
DAA	Data analyst and administrator
DCD	Degree of cognitive difference on quality dimension contributions to an assessment result
DE	Degree of easiness in differentiating quality dimensions
DM	Decision maker
DT	Data set used in an application in the form of tasks under assessment
G	A quality graph, $G = <S_0, \text{TS}, \text{DT}, C, \text{RC}, R, S_e>$
M_{ir}	Mean score of the i-th quality dimension assessed by the r-th role
M_r	Mean score of quality dimensions to the r-th role
QDM_{jr}	Data quality has no limitations in time and the quality of data on the dimension j to the role r
QD_r	Data quality to the role r
W_{irk}	Score of the i-th quality dimension corresponding to the k-th person who belongs to the role r. Where $i \in \{1, 2, ..., I\}$, 'i' represents the i-th data quality dimension, 'I' represents the total number of quality dimensions under assessment; $r \in \{1, 2, 3\}$, the value of 'r' is 1, 2, or 3 which represents three roles, respectively; 'N' represents the sum of people that belong to the role r
W_{jr}	Weight of the j-th quality dimension assessed by role r, respectively
R	Roles of stakeholders. Where $R = \{U, \text{DAA}, \text{DM}\}$, every user must belong to one of those three roles

doi: 10.2166/hydro.2015.042

RC A finite set of associations. R_{CC} represents a set of mappings between connectors, R_{CT} represents a set of mappings between connectors and tasks, R_{TC} represents a set of mappings between tasks and connectors. In a quality graph, the correlation between each two nodes must belong to one of those three mapping types. This rule can be used to assert the validity of a quality graph

S_0 Initial state. Every task-quality graph has a unique initial state and in a task-quality graph it is tagged by the symbol ⓪

S_e End state. Every task-quality graph has a unique end state and it is tagged by the symbol ⓔ

TS A finite set of time-limited tasks in an application

INTRODUCTION

One of the most commonly asked questions about hydrological data and information services in hydrological decision support systems or information systems may be 'How good are the hydrological data and information services?' This generic question is asked or concerned about by almost all hydrological data users, such as consultants working on flood or water resources study, scientists undertaking hydrology research, decision makers (DMs) in government departments, the public, the media, and environmental managers. It is also a question for the hydrologists themselves. Hydrological information systems or decision support systems play a paramount role in flood forecasting and controlling, and the data fed to these systems often reside in or are acquired from some links in the hydrological data supply chain – through acquisition instruments, network nodes, branch data centers, data center, to control center, and even hydrological applications. The hydrological data fed to these systems might be untimely, incomplete, inconsistent and illogical for network congestions, and suffer from low performance of servers, instrument failures, human error, etc. The quality of data might be unstable and unreliable; it is imperative to assess, monitor, and even control the quality of hydrological data online.

Data quality (DQ), a multidimensional, complex and morphing concept, has evoked much interest in academia and industry in the last two decades (Karr 2006). Organizations have many different requirements related to DQ assessment (Molina *et al.* 2013; Woodall *et al.* 2013). The effective collection, representation, and application of organizational data are important to a firm because these activities facilitate business operations and business analytics (Motro *et al.* 2004; Ozmen-Ertekin & Ozbay 2012). Cappiello *et al.* (2003) argued that data of poor quality would sway the setting strategy and its execution in an organization, decrease the ability to derive value from data and reduce the competitiveness of the organization. It would also reduce the capability of decision-making, the trust shared among organizations, customer satisfaction, employee morale, etc. To enhance the quality of hydrological data and information services, the state council of China founded an organization in 2010, 'the leading organization for the first national water conservation census', to identify the basic situation of Chinese rivers and lakes, to master water resources development, utilization and protection of the status quo, to find out the requirements on water resources in the process of economic and social development, to understand the response capacity in water conversation, and then to develop a national infrastructure platform that would provide a reliable basis to the national economic and social development of water information support and security. The quality of census data is the basis and core of the census, and the attention to DQ becomes very important to all users from administrators to consumers in hydrological information systems.

Although the quality assessment of data and information services becomes imperative in hydrology, there are still some knots in hydrological DQ management. Now, let us suppose that some dangerous cases might be triggered by the upstream torrential rains along a river – where several sections of levees have failed, and a hydrologist engaging in flood forecasting and controlling wants to know the current circumstances of the upstream rivers and reservoirs – such as the water level and flow rate of rivers, the reserved storage capacity of reservoirs, the meteorological information of upstream areas, and others, and then use this information to make decisions with the aid of a hydrological decision support system deployed in the control center. The deployment diagram of the hydrological data and the control center are shown in Figure 1.

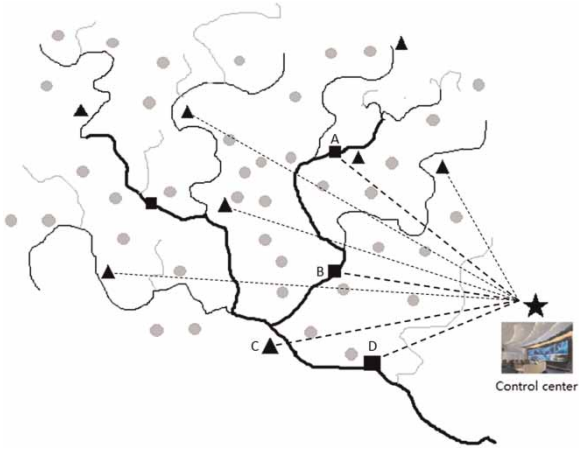

Figure 1 | A scenario of flood controlling.

In Figure 1, A, B, and D represent three reservoirs, C represents a hydrological monitoring station, and E represents a section of a river. The filled triangles, circles, squares, and star represent the hydrological monitoring stations, the rain-gauge monitoring stations, the reservoirs, and the control center, respectively. Every station can acquire hydrological data from lower layers, submit data to its database server and transmit data to higher layers on schedule. The control center can access the hydrological data from subordinate stations online, and use the data to achieve flood forecasting and controlling. However, the data fed to the system might reside in the branch data center(s), or could be acquired from acquisition instruments in time and then transmitted to the hydrological monitoring stations and the control center in a form of data supply chain. When the hydrologist monitors the circumstances of upstream rivers and reservoirs, they might confront DQ issues as follows:

1. Some information, such as the rainfall, water level, flood, and others, cannot be reported correctly for network failures, power failure, and other instrument failures. For example, the water level in the river at station C is monitored as 1,200 cm, but the highest water level in history was 300 cm and the possible maximal value set by hydrological experts is 500 cm. According to business rules or expert knowledge, the value can be defined as *abnormal*.
2. The required data on rainfall, water level, or flood cannot be transmitted stably for network failures. For example, the hydrological data controlled directly by stations A, B, or C cannot be transmitted or submitted to the control center, or the time consumed during transmission might be either longer or shorter in different circumstances and such time variations are significant. That is to say, the network service or data acquisition instrument is *unstable*.
3. The required data cannot be transmitted or (and) processed in real time for network congestions, network failures, low performance of servers, and other factors. For instance, if the hydrologist wants to monitor the current water level near to station C, but the information system can only provide the record 2 hours ago, then the value turns out to be *stale*.
4. There are many rain-gauge sensors deployed around station A and other areas to monitor the hydrological parameters, but there are significant differences between the values reported by the sensors adjacent to the same area. For example, if the values of rainfall monitored by two sensors deployed in the same place are 20 mm and 40 mm, respectively, then the two values are *inconsistent*.
5. The water level of a river adjacent to station C is acquired every half an hour in a specific station, but the number of records on the river may appear to be less than the expected number, i.e., the data may be *incomplete*.
6. Other issues.

In hydrological information systems or decision support systems, the hydrological data are often accessed in a form of Web Service, Remote Procedure Call, Application Interface, Work Flow, Cloud Computing, or others, and we abstract these forms of data access as tasks uniformly. Traditional quality assessment methodology on hydrological data has focused mainly on the intrinsic quality or the representation quality of data, and could not be used to assess the application-related DQ in the form of the data supply chain and with a granularity of tasks. To detect, monitor, and even control the quality of hydrological data, we present an application-related DQ assessment model which can depict the architecture of a hydrological application in a large granularity of tasks, and then an authorized consumer can monitor and even control the quality issues at an abstract level and with a large granularity of tasks. To achieve these objectives, we should detect the quality issues hidden in the hydrological data supply chain first, then analyze the correlations between the tasks embedded

in hydrological applications, and finally develop a compound model to achieve DQ assessment in the form of data supply chain.

The rest of this paper is organized as follows. The next section gives a brief research background of DQ and introduces the related work on hydrological DQ management. This is followed by a section that analyzes the hydrological data supply chain from data acquisition to data consumption in hydrological information systems or decision support systems, analyzes the related quality dimensions in each link of data supply chain, shows the procedure of DQ assurance and assessment, and finally, describes a methodology to derive quality dimensions and model the degree of the cognitive differences in quality dimension importance. A detailed process of building quality assessment models is provided in another section, followed by a section presenting experimental results and then the conclusions appear in the last section.

RESEARCH BACKGROUND

The notion of DQ has been widely discussed in the literature. Wang & Strong (1996) and Lee & Strong (2003) believed that data of high quality are 'getting the right data in the right place at the right time to complete the task at hand'. Madnick *et al.* (2009) held that DQ is a degree of inherent attributes meeting the requirements. As far as Zhu & Buchmann (2002) are concerned, DQ is the sum of the intrinsic attributes and the external attributes of the data. Some authors (Pitt *et al.* 1995; Strong *et al.* 1997) argued that DQ primarily refers to the degree of an information system to achieve a consistency of schema and instance, and how to achieve correctness, consistency, completeness, and minimalism. DQ reflects the degree to which specific application data are met (Rahm & Do 2000). The first step of developing a user-specific contextual DQ model is to understand the perceptions and requirements of users in a specific context (Sidi *et al.* 2012).

The work on DQ assessment ranges from the techniques that assess DQ to the design of data integration systems that integrate heterogeneous data sources with different degrees of quality, and is extended to warehouses (Zhu & Buchmann 2002), information systems (Strong *et al.* 1997; Rahm & Do 2000; Lee *et al.* 2002; Sidi *et al.* 2012), e-commerce (Aboelmeged 2000; Knight & Burn 2005; Peltier *et al.* 2013), e-learning systems (Alkhattabi *et al.* 2011), collaborative business processes (Falge *et al.* 2012), Web data (Caro *et al.* 2008; Sun *et al.* 2012; Yerva *et al.* 2012), genome annotation work (Huang *et al.* 2012), wireless sensor networks (Coen-Porisini & Sicari 2012), and hydrological DQ assurance (Dillon & Woithe 1988; Mosley & McKerchar 1989; Hudson *et al.* 1999; Burn & Hag Elnur 2002; Schlaeger *et al.* 2007; Li *et al.* 2012).

In hydrology, the research work on quality assurance can be traced back to Mosley & McKerchar (1989) and Hudson *et al.* (1999), who focused on the quality assurance in hydrologic measurements and hydrological data. Dillon & Woithe (1988) stated that the collection of hydrological data must be based on sound theoretical principles, well coordinated, and subjected to strict quality control and error estimation and correction techniques. Burn & Hag Elnur (2002) and Branisavljevic *et al.* (2011) emphasized the importance of DQ in hydrological data management and knowledge of the process for controlling DQ, and also mentioned how to detect the outlier(s) and interpolate the values based on the neighboring points in time and space. Meanwhile, they found that many inconsistencies exist in flow records and the DQ may also differ from year to year. WMO (2003) stated that hydrological observation data (obtained indirectly through a process of conversion) include abnormal values and observation errors caused by malfunctioning observation instruments or the analysis. This report also mentioned that the quality of hydrological data is related to the reliability, missing data, and abnormal values. Gourbesville (2009) stated that the technology evolution can provide quality data and new challenges to users. Krajewski *et al.* (2011) described how to manage and transform the data at different levels, and showed the techniques to extract metadata and manage hydrological data using the Hydro-NEXRAD software system. Abdullaev & Rakhmatullaev (2014) presented practical results on improving water management in Central Asia through the application of better data management tools at the operational level across diverse institutional settings. Schlaeger *et al.* (2007) said the quality assurance levels include data level, workflow level, system level, and IT level. As Watts *et al.* (2009) stated, some information may be highly relevant

to a task but not to others; meanwhile, a specific and user-related service is often directly related to a small percentage of the data set (Wang *et al.* 1995).

There has been a significant amount of work on the quality management of hydrological data or information services, but too much attention has been paid to the quality assurance in data level, and has focused mainly on general quality issues, such as detecting outliers, nullity interpolation, checks on completeness, gradient, and distance, and others. In hydrological information systems or decision support systems, the traditional DQ assessment methodology or strategies could not embody the influence of circumstances, such as the performance of network services, users' requirements in response time of a task, and others. In these systems, the quality of hydrological data or decisions are directly or indirectly related to one or more links of the hydrological data supply chain, and they are also affected by the preference and emphasis of users (different interested applications, time limits, etc.), the runtime environments of tasks (the capacity of software and hardware, memory, network bandwidth, energy, etc.), and others. Therefore, the quality assessment model should embody the differences in personal characteristics, contexts, task magnitudes, time-variances, and other interested and concerned aspects. To assess, monitor, and even control the quality issues residing in the data supply chain and with a granularity of tasks, we have done further research with the main contributions being as follows:

1. Analyze and detect the quality issues residing in the hydrological data supply chain – from acquisition instruments, network nodes, branch data centers, data center, control center to hydrological applications.
2. Illustrate the DQ assurance procedure that is applicable to hydrological information systems or decision support systems. The procedure can give potential assessors brief guidance on the framework of hydrological DQ management.
3. Analyze the correlations between the tasks accessing hydrological data, classify them into five categories and construct the related quality models with time limits. Meanwhile, we develop the quality assessment model on hydrological data application in the supply chain, and then the authorized consumers can use these models to monitor and even control the quality of hydrologic data, a form of data supply chain and with a granularity of tasks.

DERIVING QUALITY DIMENSIONS AND THE PROCEDURE OF DQ ASSURANCE IN HYDROLOGIC INFORMATION SYSTEMS

Some literature has described the criteria and measurements to improve data or information quality, but few studies have focused on quality dimension prioritization and quantitative quality assessment especially in hydrology. Meanwhile, many DQ definitions, models, and methodologies have been proposed to tackle the specific problems on DQ, but they may not be applicable to the domain of hydrology. To understand the processes of hydrological data production and achieve hydrological DQ management, we present the hydrological data supply chain first, then identify the quality issues, and next illustrate the procedure of DQ assurance and assessment; finally we show how to prioritize the quality dimensions and endow them with weights.

Data supply chain in hydrologic information systems

The data used in a hydrological information system are directly or indirectly related to each link of the hydrological data supply chain, such as data acquisition, transmission, transformation, validation and correction, storage, dissemination, representation (use), and other links. The hydrological data supply chain in a hydrological information system is shown in Figure 2.

In hydrological information systems or decision support systems, the DQ management recurs throughout the whole life cycle of data production and consumption, and the quality issues might spring from any link of hydrological data supply chain as follows:

1. The tools of hydrologic data acquisition. As shown in Figure 2, the phase of hydrologic data acquisition (phase 1) is to monitor the river, lake, reservoir, canal, ground water, and other hydrologic objectives and then to acquire related information (data) such as water level, discharge, velocity of flow, rainfall, snowfall,

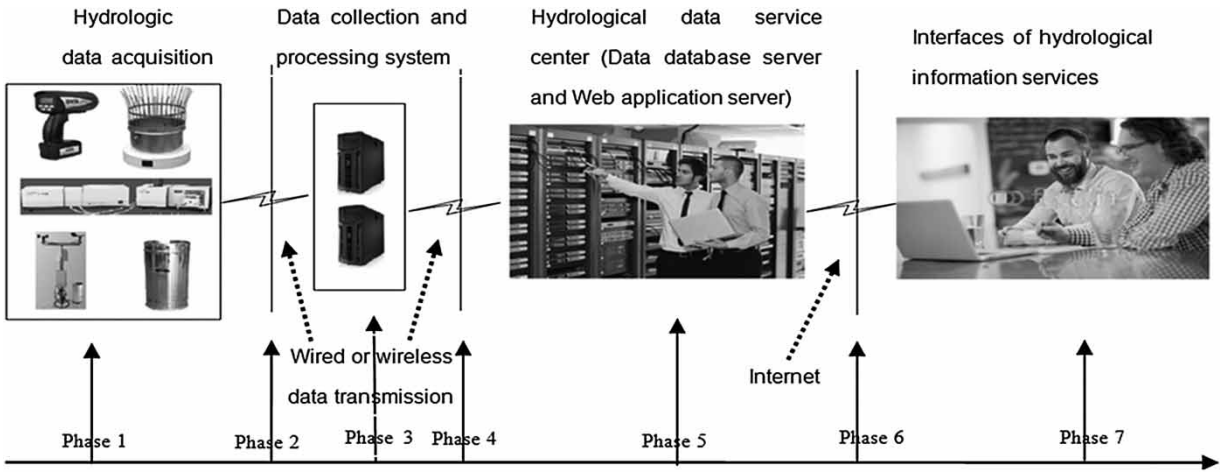

Figure 2 | The hydrological data supply chain.

sediment, ice, soil moisture, water quality parameters, and other information with the help of measuring instruments (such as water level monitor, evaporation meter, rainfall sensor.). In this phase, the DQ assessment mostly focuses on the quality dimensions (characteristics), such as accuracy (with low energy consumption), stability, sensitivity, operability, or others.

2. Data transmission. As shown in Figure 2, there are three phases on data transmission in hydrological data management (phase 2, phase 4, and phase 6). When the acquired hydrological data are transmitted from the acquisition instruments to the branch data center(s) – which can collect and process hydrological data, and transmit the data to the hydrological data center – database servers and application servers, or when the data (accessed in a form of Web Service, Remote Procedure Call, Application Interface, Work Flow, Cloud Computing, or Application Interfaces) in the data center are provided to consumers on a wired or wireless network, DQ problems (such as data loss, data falsification, data delays, and others) may occur for unstable and insecure network(s). In the three phases, the DQ assessment mostly focuses on the quality of service, such as transmission speed, error rate, response time, usability, throughput, reliability, and others.
3. Data collection and processing system. As shown in Figure 2, the data collection and processing system (corresponding to phase 3 in Figure 2) can carry out missions on data management. The data might be collected from acquisition instruments or (and) be transmitted by private or public network, Global Positioning System (GPS), or others media, and might temporarily or permanently reside in the branch data center(s) or the data center. In this phase, DQ assessment mostly focuses on quality dimensions such as accessibility, integrity, consistency, validity, timeliness.
4. Data service center (data database server and Web application server). The data service center is the center of data storage, data transformation, data mining, and data application (phase 5 in Figure 2). The main function of the data service center is to provide hydrological data (information) to consumers. It also provides interfaces to software developers in the form of Web Service, Remote Procedure Call, Application Interface, Work Flow, Cloud Computing, etc. The consumed data might have been uploaded from the branch data center(s) and be derived from the infrastructure database (including data on relief maps, established hydrologic stations and flood forecast stations in a river basin, adjacent telemetric stations, etc.), or be derived from the acquisition instruments in real time. In this phase, DQ assessment mostly focuses on quality dimensions such as integrity, consistency, validity, timeliness, redundancy, conciseness, conformity, and others.
5. Clients of hydrological information services. A client may be an administrator, a consumer, a software developer, or someone who wants to get a service from the hydrological information system or to make decisions relying on the decision support system. In this phase (phase 7 in

Figure 2), DQ assessment mostly focuses on quality dimensions such as consistency, validity, profitability, timeliness, accessibility, conciseness, interpretability, understandability, security.

The DQ management in hydrological information systems is directly or indirectly related to the above-mentioned links of the data supply chain, and the focuses of quality dimensions in each link are often different. Thus, in the next section, we will list and explain the related quality dimensions in each link of the hydrological data supply chain, then illustrate the DQ assurance procedure in hydrological information systems, and finally show how to prioritize and endow them with applicable weights.

Quality dimensions and correlations

An important basic task on DQ assessment in an information system is to identify the quality dimensions which describe the quality of data (Wand & Wang 1996; Wang 1998). Quality dimensions are the basis of DQ assessment, and different quality dimensions make different contributions to the results of a DQ assessment. As Wang & Strong (1996) described, the DQ dimensions are a set of DQ attributes which represent single aspects or constructs of DQ. The most frequently mentioned DQ dimensions in the literatures are accuracy, completeness, consistency, and timeliness (Berti-Equille 2007). Since Brodie's (1980) proposition, more than 200 quality dimensions have been collected to characterize DQ in the literature (Lim *et al.* 1999; Cappiello *et al.* 2003; Caro *et al.* 2008; Ge & Helfert 2008). Although numerous quality dimensions are revealed from different perspectives, it is a challenge to reach a consensus about which DQ dimensions should be taken into account in a quality assessment system. How to define a quality dimension precisely in a specific application? Which quality dimensions are more important and how to assign proper weights to them? Meanwhile, the quality dimensions are not independent of each other (Pipino *et al.* 2002). Those factors introduce complexity and inaccuracy into quality assessments process and results.

Making a correct decision is clearly dependent on the quality of data used (Berner *et al.* 2005), whereas, complete knowledge on the quality of data cannot be gained without knowing what the existing relationships are among quality dimensions (Barone *et al.* 2010). Some researchers (Kesh 1995; Karplus & Diederichs 2012) showed the major interactions between the various quality dimensions from a data modeling perspective, and offered the opinion that if users cannot understand the data model they will not be able to verify its accuracy and completeness. Motro *et al.* (2004) took quality dimensions, such as recentness, cost, accuracy, and availability as the features of knowledge about the performance of the source data, and combined the results with a flexible utility function to sort the results derived from inconsistent and multiple information sources. Wang & Strong (1996) and Lee & Strong (2003) classified DQ into contextual DQ, representational DQ, accessibility DQ and intrinsic DQ, and took second-order factors (such as accuracy, currency, completeness, timeliness, precision, reliability) as more specific quality attributes. Gorla *et al.* (2010) characterized information system quality and showed the correlations between system quality, information quality, and service quality and organizational impact at both a qualitative perspective and quantitative perspective. Data service quality is defined as the degree of discrepancy between customers' normative expectations for service and their perceptions of service performance, such as reliability, response time, assurance, and empathy (Pitt *et al.* 1995). System quality is the extent to which the system is technically sound (Seddon 1997), and is concerned with whether there are bugs in the system, the consistency of user interface, ease of use, quality of documentation, and even quality and maintainability of program codes.

Some information quality dimensions, such as accuracy, completeness, and consistency, cannot be measured without confirming the context. Some quality dimensions, such as accuracy, timeliness, friendliness, maintainability, relevance and believability, cannot be objectively measured and tend to vary with the usage context (Watts *et al.* 2009). Operations and metrics of quality dimensions would be changed with the context although the names and meanings of quality dimensions may stay the same (Stvilia *et al.* 2009). Those DQ dimensions are not independent of each other and typically only a subset of them are relevant in a specific application; the relevant quality dimensions and required levels of quality in an application are determined by the context and the subjective preferences of consumers. Some

quality dimensions in hydrology are defined and shown in Table 1.

There are many quality problems such as incompleteness, invalidity, and other intrinsic quality dimensions present in a hydrological data center. For instance, reported data sets – recording some parameters on a prominent river in China and reported by a specific station – refer to the parameters of daily average stages (HY_DZ_C), daily average discharges (HY_DQ_C), and daily average water temperature (HY_DWT_C). These abbreviations are encoded by the Ministry of Water Resources of the People's Republic of China, and they also represent the table names or the data types on the corresponding data. We have checked the data set and discovered many quality problems, such as value invalidation, value nullity, loss of records. The data set has been checked by the software and then judged by hydrologists, and the statistical results are shown in Table 2.

Table 1 | Some DQ dimensions in hydrological information systems

Quality dimensions	Explanation
Accessibility	The extent to which hydrological data or information services are available or retrieved easily and quickly anywhere and anytime. It is related to the quality dimensions of reliability, performability, etc.
Accuracy	The extent to which data values agree with the corresponding correct or genuine values. For example, in a specific station, the real value of rainfall with probability 0.99 is contained in a confidence interval $[40-40, 40+40]$, a reported value of the rainfall is 100 mm in an observation station, and then value may not be accurate with a high probability. It is related to cost, correctness, etc.
Appropriateness	The extent to which data are appropriate for the tasks or applications.
Completeness	The extent to which data are not missing or can be used in applications.
Consistency	The extent to which data values are not contradictory with facts.
Conciseness	The extent to which data are compactly represented.
Correctness	The extent to which data values are correct in range and logic verification.
Cost	An amount paid or required in accessing data, by means of cost-effectiveness.
Conformity	The extent to which data are not coded and stored according to uniform format; it often refers to completeness.
Integrity	The extent of the absence of alteration between two instances or two updates of a data record or object.
Interpretability	The extent to which data conform to the technical ability of consumers. It is related to maintenance, conformity, and other quality dimensions.
Maintenance	The extent to which data are well documented and manipulated.
Profitability	The state or condition of yielding a financial profit or gain. It is often measured by price to earnings ratio.
Redundancy	The extent to which hydrological data can be deleted but do not affect the implementation of functions.
Reliability	The extent to which hydrological data are regarded as true and credible.
Reusability	The extent to which hydrological data can be reused in other projects. For example, if the records on an entity are stored with different formats and do not remain consistent with the standards of a national authority organization, then the reusability is low.
Security	The extent to which access to hydrological data are restricted appropriately to its users and applications. For example, if the confidential data stored in the data center can be accessed, even modified by anyone, then the degree of security is low.
Timeliness	The extent to which hydrological data are sufficiently up-to-date for the task. For example, if a hydrological expert wants to know the current water level of a specific station, but the information system can only provide a record 24 hours ago, then the value is stale. It relates to the quality dimensions of accessibility and reliability.
Understandability	The extent to which data are easily comprehended. For example, the public wants to know the water level in the next 3 days, but the information provided by the hydrological information system is presented with jargon.
Validity	The extent to which data values and data records are valid; this requires expert knowledge, and can, for example, be judged by a hydrologist.

Table 2 | Statistical results on value invalidation, value nullity, and loss of records

	Number of null, invalid, and existing records		
Years	**HY_DZ_C**	**HY_DQ_C**	**HY_DWT_C**
1980–1984	30–21–1,746	38–12–883	22–7–1,469
1985–1989	190–76–1,462	127–37–1,528	91–52–1,254
1990–1994	76–22–1,362	71 –19–1,643	75–32–1,643
1995–1999	290–37–1,515	45–22–1,419	36–22–1,465
2000–2004	75–32–1,643	43–37–1,723	40–37–1,528
2005–2009	6–22–1,761	14–32–1,692	15–8–1,747
2010	12–31–1,158	16–15–1,265	13–22–1,257

In Table 2, we list the number of null values, invalid values, and lost records residing in the data sets of HY_DZ_C, HY_DQ_C, and HY_DWT_C. From Table 2, we can see quality problems exist in the data center, such as value invalidation, value nullity, and loss of records. For example, from 2005 to 2009, the number of records on HY_DZ_C should be 1,825 if none are missing. Nevertheless, the number of records stored in the data center is 1,761, that is to say, some of the records on HY_DZ_C are missing or were never reported. Therefore, it is imperative to discover the quality problems, and solve them and even improve the quality of data in the data center.

The data used in hydrological information systems or decision support systems may refer to several links of hydrological data supply chain, and the quality of data and information services is related to many factors and not all factors are equally important. There are many opinions on quality assessment from different perspectives, but there is agreement over the idea that the definitions and prioritizations of quality dimensions are crucial to gain an objective, application-specific, multi-user satisfied and up-to-date quality assessment result. Thus, the next section will present the prioritization of DQ dimensions and how to endow them with weights.

DQ assurance and assessment procedures in hydrological information systems

Hydrological DQ management recurs throughout the whole life cycle of data production and consumption, and quality problems reside in each link of the hydrological data supply chain. DQ is influenced by the adopted technologies, consumers, rules and regulations, application context, and other factors, and the quality assurance procedure can be implemented at different levels – data level, workflow level, system level, and IT level. In our research, we mainly assess the quality of data and information service in a granularity of tasks based on our previous work (Liu *et al.* 2011; Li *et al.* 2012). The procedure of DQ assurance and assessment is shown as follows:

Step 1: Analyze the correlations between tasks or modules of a hydrologic information system, and depict the architecture of the information system using the five basic structures which are stated in the section 'Quality assessment model on hydrological applications'. For example, the hydrologist (mentioned in the introductory section) wants to get the current water level of the river C and the water level and reserved storage capacity of the reservoirs A, B, and D. These data can be acquired in the form of tasks, and these tasks can be implemented independently and in a parallel manner. In this situation, the architecture can be depicted by the concurrent structure (stated in the section 'Quality assessment model on hydrological applications'). Meanwhile, if the hydrologist wants to forecast the future water level, both the current information and the existing records on the water level should be accessed, and then the previous tasks will include two subtasks – one is to obtain the current water level that might be acquired by acquisition instruments immediately, and the other is to access the records that might be stored in the branch data center. According to the requirements of an application, a quality assessor can depict the architecture of the application and develop the corresponding quality view diagram.

Step 2: Identify or select the related quality dimensions from personal perspectives and specific circumstances. For example, if the hydrologist finds that some records on the water level of river C are not complete, or the information on reservoir B cannot be submitted to the control center in time, they might want to know the completeness of the records and the accessibility of information on reservoir B. Meanwhile, the hydrologist might confront other quality issues, and then might implement a quality assessment on relevant quality dimensions.

Step 3: Set proper values for quality parameters. For example, the hydrologist can set the expected response time for specific tasks and then get the quality of experience – the definition of quality of experience is defined by de Carvalho Costa & Furtado (2011). In other words, if the time spent in implementing the task is over the threshold limit, then the returned results might be considered invalid or stale – the task might not be implemented according to some strategies.

Step 4: Carry out the user-related and task-related hydrologist data or information service quality assessment on specific quality dimensions with a granularity of tasks. The control center can develop the required programs or employ other existing programs to achieve the quality assessment on specific quality dimensions (such as stability, currency, consistency, completeness, etc.). Meanwhile, the assessment system should provide interfaces to integrate other programs. If an end-user wants to know the general quality of application-related data, the assessment system should fuse each quality dimension assessment results into a compound result – in this case, the weights of quality dimensions should be set by the end-user. Finally, the assessment system presents the assessment result(s) to the end-user in a form of quality view diagram or other type of display.

Step 5: Discern and identify the links with low quality data and information service in a data supply chain, and then improve the DQ by employing advanced instruments, excellent hydrological information system architecture, and other measures if necessary.

Prioritization of DQ dimensions and endowing them with weights

To assess the DQ of an application, the first thing is to classify users into different roles, then ascertain related quality dimensions in a quality assessment system, and finally, balance their weights. In this assessment system, the users are classified into three classes (roles): they are the customer (CS), data analyst and administrator (DAA), and DM.

This section first presents some criteria to prioritize quality dimensions and show how to endow them with weights. To prioritize the DQ dimensions, let M_{ir} represent the mean score of the i-th quality dimension assessed by the r-th role, as

$$M_{ir} = \frac{\sum_{k=1}^{N} W_{irk}}{N} \tag{1}$$

where W_{irk} represents the score of the i-th quality dimension corresponding to the k-th person who belongs to the role r. Where $i \in \{1, 2, \ldots, I\}$, '$i$' represents the i-th DQ dimension, 'I' represents the total number of quality dimensions under assessment; $r \in \{1, 2, 3\}$, the value of 'r' is 1, 2, or 3 which represents three roles, respectively; 'N' represents the sum of people that belong to the role r. For example, W_{133} represents the score of quality dimension accessibility assessed by a decision maker no. 3 (DMNO.3). If a quality dimension is more important than others, then it will be assigned with a higher weight. Let M_r represent the mean score of quality dimensions and W_{jr} represent the weight of the j-th quality dimension assessed by role r respectively, as follows:

$$M_r = \frac{\sum_{I=1}^{I} M_{ir}}{I} \tag{2}$$

$$W_{jr} = \frac{M_{jr}}{\sum_{I=1}^{I} M_{ir}} \tag{3}$$

Without loss of generality, we assume that the perceived scores of quality dimensions are stated in 11 levels, which are integers between 0 and 10, with 10 corresponding to the highest priority and 0 corresponding to the lowest. If we assume the DQ has no limitations in time and the quality of data on the dimension j to the role r is represented by 'QDM_{jr}', then we can get the DQ 'QD_r', as shown in Equation (4)

$$\mathrm{QD}_r = \sum_{i=1}^{I} \mathrm{QDM}_{jr} * W_{jr} \tag{4}$$

where W_{jr} represents the weight of the j-th quality dimension as shown in Equation (3), the constant I represents the total quality dimension number that stakeholders are concerned with except the time-limitation factors.

In the process of analyzing our questionnaires, we find that the users that belong to a specific role may have an obvious cognitive difference in quality dimension contributions (we call it the degree of cognitive difference

(DCD)). To detect the cognitive difference on quality dimension contributions to an assessment result, we demonstrate DCD as

$$\mathrm{DCD}_{ir} = \frac{\sum_{k=1}^{N} (W_{irk} - M_{ir})^2}{N} \tag{5}$$

If the value of DCD_{ir} is bigger than DCD_{jr}, the cognitive difference on quality dimension contributions is more obvious to the role r. That is, a specific quality dimension would make different points in different people's views although they belong to the same role.

In addition to the cognitive difference, there is another troublesome concern that differentiates quality dimensions from the case that all the contributions (weights) are almost the same for a specific role (called the degree of easiness among differentiating quality dimensions, DE). This phenomenon often springs from the individual differences in education levels, professions, degrees of technical proficiency, focus points, and so on. To detect the degree of difficulty in differentiating quality dimensions to the role r, DE can be calculated as

$$\mathrm{DE}_r = \frac{\sum_{i=1}^{M} (M_{ir} - M_r)^2}{M} \tag{6}$$

It is more easy to choose the relevant quality dimensions to the role r if the DE_r owns a bigger value.

QUALITY ASSESSMENT MODEL ON HYDROLOGICAL APPLICATIONS

The level of hydrological DQ or information service quality is not only related to the inherent quality of data, but also associated with the task accomplishment (in a hydrological information system or a decision support system, the accomplishment of data access and information service is reliant on the execution of SQL, Web Service, Remote Procedure Call, Application Interface, Work Flow, Cloud Computing, etc., which we regard as tasks in our quality assessment system), the runtime environment, the stakeholders, and others factors. In a flood controlling and drought resisting application or real-time water quality monitoring, the performance of a task is directly related to the quality of data used, the limitations of time (such as 'the earliest start time', 'the latest finish time', 'the response time', 'the expected response time', and 'the expected latest finish time'), the runtime environments of tasks, and others. Here, we first define these concepts:

- The earliest start time. It represents the earliest start time of all the tasks in an application, and its value is based on the start time of the first implemented task.
- The latest finish time. It represents the latest finish time of all the tasks in an application and its value is based on the finish time of the latest implemented task.
- The response time. It represents the time spent on a task and is also related to the network transmission time, the data access time and the data processing time by the application program.
- The expected response time. It represents the expected time spent on a task, and it is often used to assess the quality of experience. The expected response time set by different assessors might be different.
- The expected latest finish time. It represents the expected time of the latest finished task in an application. The expected latest finish time set by different assessors might be different.

In a time-limited application in hydrology, an applicable and flexible quality assessment model should embody the differences in personal characteristics, application contexts, task magnitudes, time-variances, and other related aspects. In this research, we mainly focus on the influences of factors such as the quality of data sources, the time limitations of tasks, and the roles of users on DQ assessment results. To evaluate the DQ in a time-limited application, we present a quality assessment model which consists of six components. The model can be represented with the form of a task-quality graph

$$G = <S_0, \mathrm{TS}, \mathrm{DT}, C, \mathrm{RC}, R, S_e> \tag{7}$$

where S_0: Initial state. Every task-quality graph has a unique initial state and in a task-quality graph it is tagged by the symbol ⓪.

TS: A finite set of time-limited tasks in an application. $\mathrm{TS} = \bigcup_{i=1}^{N} \mathrm{TS}_i$, $\mathrm{TS}_i = \langle T_i, \beta_i, \delta_i, t_i, \mathrm{QST}_i \rangle$. T_i represents the i-th task, QST_i represents the intrinsic quality of task i without time-limitation, β_i represents the earliest start time, δ_i represents the latest finish time of task i, and t_i represents the response time of the task i. In this quality assessment system, we suppose the response time of a task can be obtained in advance. Of course, it is not difficult to adjust this system to achieve quality assessment in the case that the response time is dependent on the scenario of task execution.

DT: The data set used in an application in the form of tasks under assessment.

C: A finite set of connectors. Where $C = \{C_{\mathrm{sq}}, C_{\mathrm{rc}}, C_{\mathrm{sl}}, C_{\mathrm{cc}}, C_{\mathrm{cs}}\}$, it includes five types of connector, these being sequential structure C_{sq}, recursive structure C_{rc}, selective structure C_{sl}, concurrent structure C_{cc}, and composite structure C_{cs}, and they are tagged by the symbols (sq), (rc), (sl), (cc) and (cs), respectively, in a graph. (?) represents one non-deterministic state which belongs to one of the above five types. The time spent on these five types of connectors will be ignored in quality assessment. In this assessment system, time spent on network transmission is ignored.

RC: A finite set of associations. R_{CC} represents a set of mappings between connectors, R_{CT} represents a set of mappings between connectors and tasks, R_{TC} represents a set of mappings between tasks and connectors. In a quality graph, the correlation between each two nodes must belong to one of those three mapping types. This rule can be used to assert the validity of a quality graph.

R: Roles of stakeholders. Where $R = \{U, \mathrm{DAA}, \mathrm{DM}\}$, every user must belong to one of those three roles.

S_e: The end state. Every task-quality graph has a unique end state and it is tagged by the symbol (e).

There are five types of connectors and homologous quality assessment models, that is, the correlations between tasks can be classified into five basic categories (they are single task structure, sequential structure, recursive structure, selective structure, and concurrent structure – we will illustrate them in the following sections, and then one can construct more complex structures based on these five basic structures), and the pertinent quality assessment models to assess the DQ in time-limited applications as follows. First, let QST_{ir} represent the quality of data used in a single task to the role r without considering the limitations in time, then we can get the assessment results by using the aforementioned Equation (4).

Single task structures

Let QS_{ir} stand for the quality of a single time-limited task i to the role r, and let β represent the start time and t_i represent the response time (spent on task i); the model is shown in Figure 3.

Then, the quality of a single time-limited task is modeled as follows:

$$\mathrm{QS}_{ir} = B(t_i + \beta \le \delta_i) * \mathrm{QST}_{\mathrm{ir}} \tag{8}$$

where $B(t_i + \beta \le \delta)$ represents a Boolean function used to assert whether the close time $t_i + \beta$ is surpassed by the threshold δ. If the Boolean value of function B is false, the QS_{ir} will be directly assigned with 0.

Let Q_r stand for the quality of an application in the form of tasks, then the quality of corresponding application as shown in Figure 3 is as follows (the application only refers to a single task to the role r here)

$$Q_r = \mathrm{QS}_{ir} \tag{9}$$

Sequential structures

Let Q_r stand for the quality of an application which refers to two sequential time-limited tasks; the corresponding structure is shown in Figure 4.

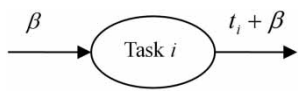

Figure 3 | Single task structures.

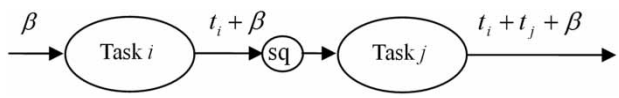

Figure 4 | Sequential structures.

Then, the Q_r can be derived from Equation (10) as follows:

$$Q_r = \mathrm{Min}(\mathrm{QST}_{ir}, \mathrm{QST}_{jr}) * \mathrm{B}(\mathrm{t_i} + \beta \le \delta_\mathrm{i}) * \mathrm{B}(\mathrm{t_i} + \mathrm{t_j} + \beta \le \delta_\mathrm{j}) = \mathrm{Min}(\mathrm{QS_{ir}}, \mathrm{QS_{jr}}) \quad (10)$$

where QST_{ir} and QST_{jr} represent the quality of single task i and task j to the same role r, respectively. In the function of Min, $\mathrm{QST_{jr}}$ represents the quality of data used in a single task to the role r without time limit, and $\mathrm{QST}_{jr} = (\sum_{k=1}^{I} \mathrm{QDM}_{rk} * \mathrm{W_{kr}}) * \mathrm{B}(\mathrm{t_i} + \mathrm{t_j} + \beta \le \delta_\mathrm{j})$. The function Min $(\mathrm{QST}_{ir}, \mathrm{QST}_{jr})$ is to get the lesser one between QST_{ir} and QST_{jr}. Note that the start time of task j is $t_i + \beta$. Generally, the data source used in the two tasks is different, i.e., if the data source is identical, this structure will become the following recursive structure.

Recursive structure

Let Q_r stand for the quality of a time-limited application which refers to a recursive time-limited task to the role r; the corresponding structure is shown in Figure 5.

Then, the Q_r can be derived from Equation (11) as follows:

$$Q_r = \mathrm{QST}_{ir} * \prod_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{B}(\mathrm{j} * \mathrm{t_i} + \beta \le \delta_\mathrm{j}) = \mathrm{QS_{ir}} * \prod_{\mathrm{j}=2}^{\mathrm{n}} \mathrm{B}(\mathrm{j} * \mathrm{t_i} + \beta \le \delta_\mathrm{j}) \quad (11)$$

where QST_{ir} represents the quality of single time-limited task i to the role r, $B(n * \mathrm{t_i} + \beta \le \delta_\mathrm{i})$ represents a Boolean function used to assert whether the close time, referring to the start time β and the response time $n * \mathrm{t_i}$ (spent on task i after n-1 times recurrence, and every close time of the anterior n-2 times recurrence does not surpass the threshold δ_i), does not surpass the threshold δ_i.

Selective structure

Let Q_r stand for the quality of an application which refers to two selective time-limited tasks to the role r; the corresponding structure is shown in Figure 6.

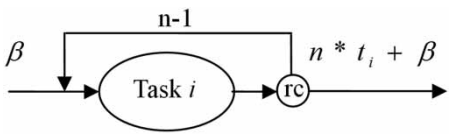

Figure 5 | Recursive structures.

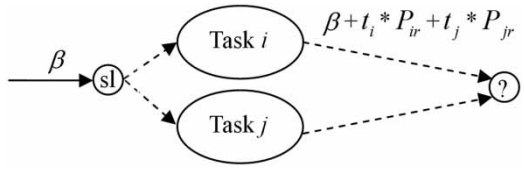

Figure 6 | Selective structures.

Then, the Q_r can be derived from Equation (12) as follows:

$$Q_r = \mathrm{QST}_{ir} * \mathrm{B}(\mathrm{t_i} + \beta \le \delta_\mathrm{i}) * \mathrm{P_{ir}} + \mathrm{QST_{jr}} * \mathrm{B}(\mathrm{t_j} + \beta \le \delta_\mathrm{j}) * \mathrm{P_{jr}} \quad (12)$$

where QST_{ir} and QST_{jr} represent the quality of single task i and task j, respectively, to the role r, $B(t_i + \beta \le \delta_j)$ represents a Boolean function used to assert whether the close time, referring to the start time β and the response time (Max (t_i, t_j), spent on task i or task j), surpasses the threshold δ_i. P_{ir}, and P_{jr} represent the probability of task i and task j carried out, respectively.

Concurrent structure

Let Q_r stand for the quality of an application which refers to two concurrent time-limited tasks to the role r; the corresponding structure is shown in Figure 7.

Then, the Q_r can be derived from Equation (13) as follows:

$$Q_r = \frac{\mathrm{QST}_{ir} * \mathrm{B}(\mathrm{t_i} + \beta \le \delta_\mathrm{i}) + \mathrm{QST_{jr}} * \mathrm{B}(\mathrm{t_j} + \beta \le \delta_\mathrm{j})}{N} \quad (13)$$

where QST_{ir} and QST_{jr} also represent the quality of single task i and task j to the role r, respectively, the function B is defined as mentioned before, and the quality of concurrent tasks equals the mean value of task quality. 'N' represents the number of concurrent tasks.

In this section, five basic quality structures are listed. One can extend them to other complex structures. The

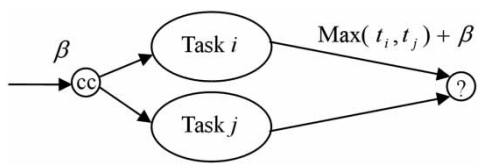

Figure 7 | Concurrent structures.

main algorithms of quality assessment in the form of pseudo codes based on the aforementioned models are shown in Appendix A (available online at http://www.iwaponline.com/jh/017/042.pdf).

STATISTICAL AND EXPERIMENTAL RESULTS

DQ dimensions and weights in hydrological information systems

The scores of quality dimensions

In our survey, 300 effective respondents can be used to prioritize the quality dimensions, and of these, 200, 60, and 40 stakeholders, respectively, belong to the role CS, DAA, and DM, respectively. Over half the participants ($n = 210$) have a hydrological background and work in a hydrological institute department, and over two-thirds of the participants have had more than 5 years of hydrological information system experience. The survey results are closely related to the respondents' own interests, so the percentage of effective questionnaires is close to 100%. To obtain the importance of each quality dimension to each role, we calculated the mean score of a quality dimension assessing the role according to Equation (1). The statistical results are shown in Table 3.

In Table 3, *D1* represents the first quality dimension that appeared in Table 1, that is, accessibility; *D2* represents the second quality dimension that appeared in Table 1, that is, accuracy; *D3–D20* correspond to the third quality dimension to the 20th quality dimension that appeared in Table 1, respectively. CS, DAA, and DM represent the three roles that are CS, DAA, and DM, respectively. From Table 3, we can find that the quality dimensions of cost, understandability, and validity are more important than the others to the role CS; the quality dimension of maintenance is more important than others to the role DAA; and the quality dimension of accuracy is more important than others to the role DM. The survey is mandatory, not voluntary, and it is executed in official institutions from top to bottom. However, there are some flaws in our survey, such as a single nation without reflecting the diversity and the instability of quality dimension importance to a role. According to the results, we can distinguish more important quality

Table 3 | DQ dimensions and weights in hydrology

Quality dimensions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
CS	9.60	7.05	5.30	6.90	7.00	8.85	8.55	9.80	3.90	2.05	9.25	3.00	9.70	5.70	9.20	5.35	8.05	9.70	9.80	9.80
DAA	9.80	9.80	8.40	9.45	9.30	8.50	9.60	4.95	9.70	9.85	5.80	9.45	3.50	9.85	8.60	8.00	8.00	8.25	5.00	9.50
DM	9.50	9.85	7.00	9.25	9.20	7.95	9.30	6.75	6.85	9.00	4.80	3.00	9.35	5.00	9.35	4.00	9.00	9.65	9.30	9.50

dimensions from listed dimensions for a role and get the weight of each quality dimension (as shown in Equation (3)).

From Table 3, we can find out which quality dimensions are more important than others to a specific role. However, the weight of each quality dimension is not clear, so based on the statistical data, we calculate the weight of each quality dimension to each role according to Equations (2) and (3), and get the preliminary weights of each quality dimension for the three roles.

The preliminary weights of dimensions for the three roles are shown in Table 4. They are called preliminary weights because of the correlations that exist between some dimensions. Proper weights can be assigned by experts or adjusted by a feedback system, and this additional work will be tested in our future work. We assume the qualities of data on the 20 dimensions are 0.95, 0.90, 0.99, 0.75, 0.90, 0.90, 0.98, 0.40, 0.80, 0.90, 0.95, 0.95, 0.30, 0.90, 0.98, 0.95, 0.98, 0.80, 0.50, and 0.95, respectively, and then we can get the quality assessment results, which are 0.81, 0.87, and 0.83 for the three roles, respectively. It is obvious that the quality assessment results are closely related to the roles. So, in the process of quality assessment, it is imperative to take the differences in roles into consideration for a quality assessment system.

The differences in quality dimension preferences for each role

To project the preference and emphasis of a user on quality dimensions in an application, we prioritize the quality dimensions for each role by the mean score, and show the top ten quality dimensions, as shown in Figure 8.

From Figure 8, we can find the differences in quality dimension preferences for each role. For example, all roles (namely all users) regard correctness, validity, and accessibility as more important dimensions in a quality assessment system.

The cognitive differences on quality dimension weights for each role

According to the collected data from the questionnaire on each quality dimension, Equation (5) and the calculated weights of DQ dimensions for each role, we get the cognitive

Table 4 | Weights of quality dimensions for the three roles

	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}	W_{17}	W_{18}	W_{19}	W_{20}
CS	0.065	0.047	0.036	0.046	0.047	0.060	0.058	0.066	0.026	0.014	0.062	0.020	0.065	0.038	0.062	0.036	0.054	0.065	0.066	0.066
DAA	0.059	0.059	0.051	0.057	0.056	0.051	0.058	0.030	0.059	0.060	0.035	0.057	0.021	0.060	0.052	0.048	0.048	0.050	0.030	0.057
DM	0.060	0.063	0.044	0.059	0.058	0.050	0.059	0.043	0.043	0.057	0.030	0.019	0.059	0.032	0.059	0.025	0.057	0.061	0.059	0.060

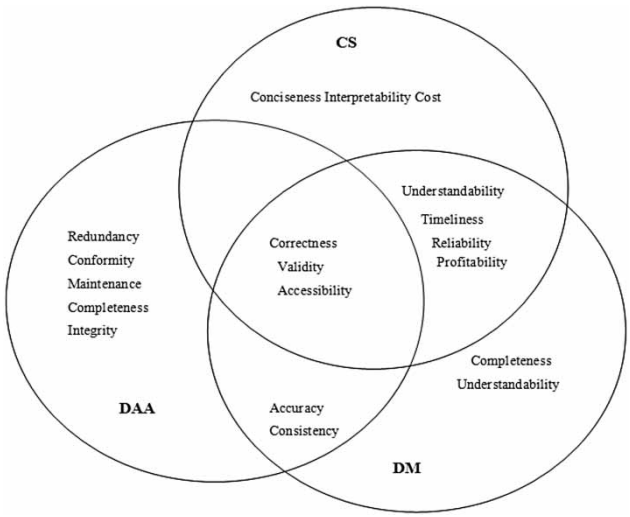

Figure 8 | The top 10 quality dimensions for the three roles.

differences on those quality dimensions for each role, and the results are as shown in Figure 9.

From Figure 9, we can find the line for CS varies with greater fluctuation than the others. According to Equation (5) and relevant data (the collected data from the questionnaire on each quality dimension and the calculated weights of DQ dimensions for each role), the cognitive differences on quality dimensions for role CS (the results of DCDs on three roles are 1.597, 0.234, and 0.292, respectively) are sharper than the other two roles, typically represented by the quality dimensions of maintenance and reusability (in other words, the cognitive differences for CS on maintenance and reusability are more volatile than the others, see Equation (5)).

The difficulty degree of choosing relevant quality dimensions

To make clear the individual differences in education levels, professions, degrees of technical proficiency, and evaluate the degree of easiness in differentiating quality dimensions to the role r, based on the above results and the aforementioned Equation (5), we get the degree of easiness in differentiating quality dimensions, as shown in Figure 10.

According to Equation (6), we get the degree of easiness among differentiating quality dimensions and the results are 1.579, 0.292, and 0.234, respectively, corresponding to the roles of CS, DM, and DAA. The results indicate that differences between quality dimensions for the people belonging to the role CS are more apparent than the other roles, that is, a CS can have a clear understanding of which quality dimensions are more important to them. In contrast, it is hard to identify which quality dimensions are more important to a DM and an administrator.

Tasks deployment diagrams on applications and quality assessment

The hydrologist engaging in flood forecasting and controlling wants to know the current circumstances of the upstream rivers and reservoirs, and the required information (data) can be acquired or accessed in the form of Web Service, Work Flow, Cloud Computing, and others. The tasks on accessing information and quality assurance refer to the following:

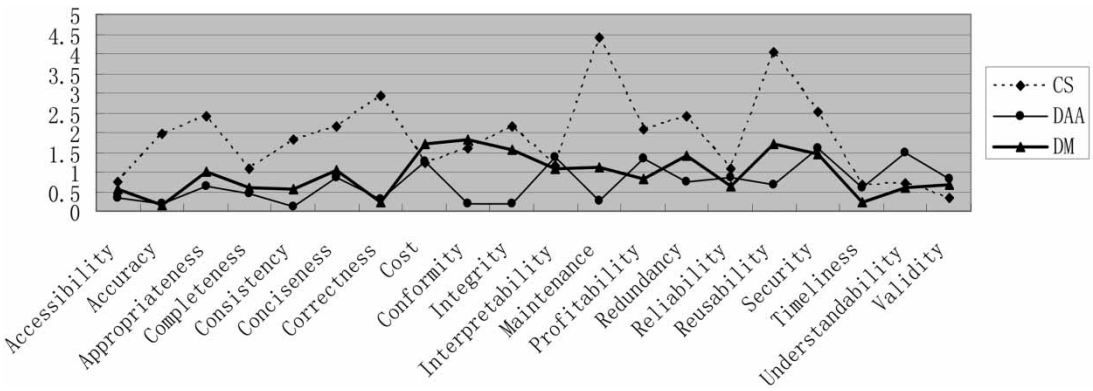

Figure 9 | The cognitive differences between quality dimensions for the three roles.

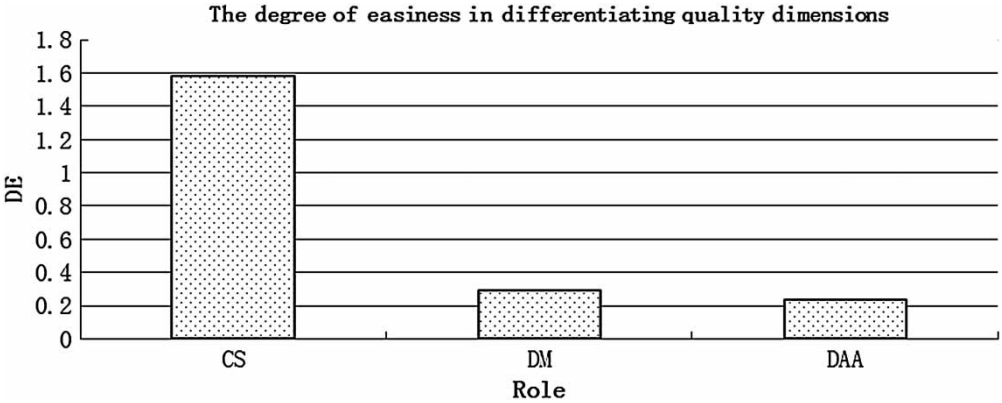

Figure 10 | The degree of easiness in differentiating quality dimensions.

(a) The acquisition instruments deployed around river C receive and implement the commands coming from the control center, then acquire the water level and submit it to station C. In this procedure, the data might not be acquired on instrument failure or submitted in time for network faults. We can assess and monitor the quality on the dimensions of accessibility, timeliness, and completeness, and then abstract the DQ related to this link with task 1 as shown in Figure 11.

(b) Transform and integrate the acquired data, and get a fused data set according to the domain standard. In this procedure, the data might not be processed and should be submitted to the control center in time. We can assess and monitor the quality on the dimensions of timeliness, completeness, and consistency, and then abstract the DQ related to this link with task 2 as shown in Figure 11. Obviously, task 1 and task 2 constitute a sequential structure as shown in the section 'Quality assessment model on hydrological applications'.

(c) To forecast the flood, the hydrologist also wants to get the records on water level. However, the data on water level are stored in two different databases – one stores the up-to-date records and the other stores the historic records. The task should be implemented twice, and then we can abstract the DQ related to this link with task 3 as shown in Figure 11.

(d) Similarly, the task on acquiring the water level of reservoirs A and B, the task on calculating the reserved storage capacity of reservoirs A, B, and D, and the task on forecasting the possible discharge based on the computational models and their DQ are abstracted and depicted as task 4, task 5, and task 6, respectively.

(e) Get the meteorological information of upstream areas and assess its quality on the computational models. This information can be acquired in the form of Web Services, or submitted manually. These tasks and the corresponding quality view diagrams are abstracted and depicted as task 7 and task 8, respectively, as

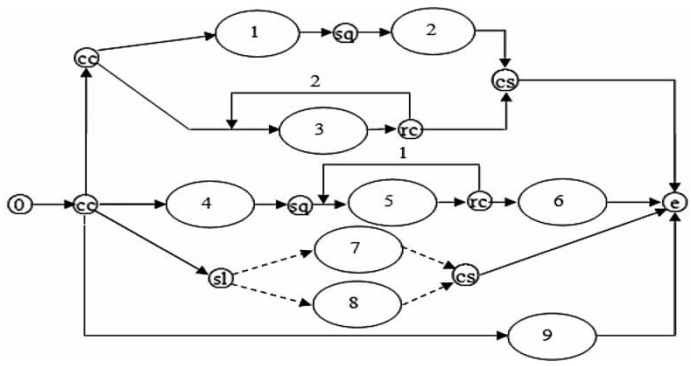

Figure 11 | Tasks' deployment in a specific application.

shown in Figure 11. To forecast and control the flood, the hydrologist might acquire other information, and the task and the corresponding DQ are abstracted and depicted as task 9.

The architecture and the corresponding quality diagram of required data used in flood forecasting and controlling can be abstracted and depicted as shown in Figure 11. The architecture belongs to concurrent structure, task 1 and task 2 constitute a sequential structure, and task 3 constitutes a recursive structure itself. Meanwhile, task 4, task 5, and task 6 constitute a sequential structure; task 5 constitutes a recursive structure, and task 7 and task 8 constitute a selective structure.

According to the aforementioned models, we can get a midterm assessment result of Figure 11 shown in the application as follows:

$$Q_r = \frac{1}{4}*\{\frac{1}{2}*\left[\mathrm{Min}(\mathrm{QST}_{1r}, \mathrm{QST}_{2r}) + \mathrm{QST}_{3r}*\prod_{j=2}^{3}\mathrm{B}(j*t_i+\beta\le\delta_j)\right] + \mathrm{Min}\left[\mathrm{QST}_{4r}, \mathrm{QST}_{5r}*\prod_{j=2}^{2}\mathrm{B}(j*t_i+\beta\le\delta_j), \mathrm{QST}_{6r}\right] + \left[\mathrm{QST}_{7r}*P_{7r} + \mathrm{QST}_{8r}*P_{8r}\right] + \mathrm{QST}_{9r}\} \quad (14)$$

As mentioned in the 'Introduction', in this quality assessment system, we mainly focus on the impact of time limitations (referred to as the earliest start time, the latest finish time, and the response time) of tasks and the roles of users on DQ assessment results. There are many factors affecting the assessment results, but we only show the influences of several time limits on the quality of experience.

The influences of the expected response time and the expected latest finish time on quality assessment results

The response time of a task represents the time difference between the response time and the request time of the task. If a task includes several requests (subtasks), the time difference will become the time difference between the last response time and the first request time of the requests. The time difference of a request is related to the time in submitting the request, the time in carrying out the request, the time in transmitting and network delay, and others. In a distributed multisource information system, the task (service) with short response time and low variability in service response time will become a candidate with high probability. We assume that the expected response time (units: s) of a task can be calculated and resides in the form of an XML file. In our research, we do not show how to set the time and the corresponding XML file (used to save the configuration on activities and time limitations temporarily or permanently).

We assume the expected response time (units: s) of each subtask will be carried out in the final form of Web Service, Remote Procedure Call, Application Interface, Work Flow, Cloud Computing, or other forms. The response time of a task can be monitored in real time with the help of tool SOAPUI (a world-leading open source functional testing tool for Application Programming Interface (API) testing) as the following matrix (Equation (15)) shows. The matrix gives five different groups of the expected response time to the nine tasks

$$\begin{pmatrix} T1 \\ T2 \\ T3 \\ T4 \\ T5 \end{pmatrix} = \begin{pmatrix} 0.3 & 0.3 & 0.2 & 0.1 & 0.3 & 0.2 & 0.5 & 0.7 & 0.6 \\ 0.3 & 0.3 & 0.3 & 0.1 & 0.3 & 0.2 & 0.5 & 0.7 & 0.6 \\ 0.3 & 0.3 & 0.2 & 0.1 & 0.3 & 0.4 & 0.5 & 0.7 & 0.6 \\ 0.3 & 0.4 & 0.2 & 0.1 & 0.3 & 0.2 & 0.5 & 0.7 & 0.6 \\ 0.3 & 0.4 & 0.2 & 0.1 & 0.3 & 0.4 & 0.5 & 0.7 & 0.6 \end{pmatrix} \quad (15)$$

Meanwhile, we assume the expected latest finish time (namely thresholds) of each task is as the next matrix shows. In this matrix, every row represents the thresholds of a group of tasks. Each value of the next row are smaller than or equal to the previous corresponding ones and the threshold of task i should be not smaller than the response time; otherwise, the task is meaningless in an application

$$\begin{pmatrix} \mathrm{TH}_1 \\ \mathrm{TH}_2 \\ \mathrm{TH}_3 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.8 & 0.9 & 0.3 & 1.0 & 1.2 & 0.7 & 0.9 & 1.0 \\ 0.3 & 0.6 & 0.7 & 0.1 & 0.8 & 1.0 & 0.6 & 0.7 & 0.8 \\ 0.3 & 0.5 & 0.6 & 0.1 & 0.7 & 0.9 & 0.6 & 0.7 & 0.6 \end{pmatrix} \quad (16)$$

In this situation, we want to show the influences of the response time and the latest finish time on assessment results. So, we assume $\beta_i = 0$ and $\mathrm{QST}_{ir} = 1$ for the sake of simplicity. That is to say, the earliest start time of every task is equal to 0 and each task quality without time limitations (namely QST_{ir} defined in the section 'Quality

assessment model on hydrological applications') is equal to 1 for all roles, i.e., here we assume the intrinsic quality of each task is identical to every user (may belong to different roles).

Let us take the case (the expected response time is listed as T4 and the expected latest finish time is listed as TH_2) as an example. The influences of the response time and the latest finish time on the performance are as shown in Figure 12.

As shown in Figure 12, for the limitation in the expected response time and the expected latest finish time in the application, task 2 and task 6 will not be executed (in a quality diagram, if a task will not be executed, it will be noted with cross ticks, and the affected tasks before this task will be filled with oblique lines). If the execution condition could not be checked before executing task 2 and task 6, the executions of task 1, task 4, and task 5 would be futile. Let us suppose the execution burdens of all tasks are equal to 1, with the help of this assessment system, the futile cost on executing task 1, task 4, and task 5 can be checked out and be avoided, and the efficiency of executing these tasks then becomes high, from 55.6% or 45.4% to 100.0% – if task 2 and task 6 are not executed in the beginning but tasks 1, 4, and 5 are executed, the efficiency is 55.6%; if tasks 2 and 6 are executed and terminated for the expected finish time, the efficiency will be 45.4%; if the assessment system can check out the tasks that cannot satisfy the execution conditions, the tasks and the front 'futile tasks' will be not executed, and the improved efficiency will then be up to 100.0%. An assessment system (often embedded into a specific application) can recognize the tasks which cannot be satisfied in the applied conditions and note these situations to users or other calling applications.

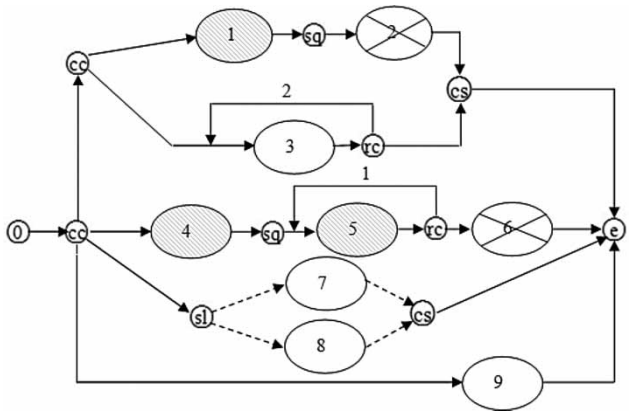

Figure 12 | The influences of the expected time on the performance.

According to the quality assessment models and in combination with the limitations shown in Equations (15) and (16), we can get the assessment results as shown in Table 5. From these results, we can find that there are tiny differences in the response time between two groups of tasks, but obvious changes may be elicited in the quality assessment results.

Under the given latest finish time shown in Equation (16), we can find when the response time of one or more tasks changes a little, the assessment result may change dramatically. For example, when the latest finish time is as shown in the second row in Equation (16) and the response time of the sixth task is changed from 0.2 to 0.4 s, the quality assessment result of the application is changed from 1.000 to 0.750. From the assessment results shown in Table 5, we can see the quality assessment results are greatly influenced by the limitation in response time.

The influences of the response time, the latest finish time, and the stakeholder roles on quality assessment results

Without loss of generality, we assume $\beta_i = 0$ and non-time-limited qualities of data source used for the nine tasks are represented by vector Q, $Q = \langle 1, 0.9, 0.7, 0.9, 0.6, 0.8, 1, 0.8, 0.5 \rangle$. The response time and the latest finish time are still the same as shown in Equations (15) and (16). That is to say, the earliest start time of every task is also equal to 0 (namely $\beta_i = 0$) and non-time-limited quality of each task i can be derived from the models as shown in the section 'Quality assessment model on hydrological applications'. According to the quality assessment models, we can get the assessment results as shown in Table 6.

Table 5 | Quality assessment results

	T1	T2	T3	T4	T5
TH1	1.000	1.000	1.000	1.000	1.000
TH2	1.000	0.875	0.750	0.875	0.625
TH3	0.875	0.750	0.625	0.875	0.625

Table 6 | Quality assessment results

$r=1$	T1	T2	T3	T4	T5
TH1	0.700	0.700	0.700	0.700	0.700
TH2	0.700	0.613	0.550	0.588	0.438
TH3	0.588	0.500	0.438	0.588	0.438

From Table 6, we can see the quality assessment results are greatly influenced by the limitation in the response time and the latest finish time of tasks.

From the experimental results of quality assessment, we find the quality assessment results of a time-limited application are greatly influenced by the data source, the preference and emphasis of users, the limitation in the earliest start time, the response time, and the latest finish time of tasks, and other factors. In fact, they are also influenced by the capacity of software and hardware, such as CPU, memory, network bandwidth, energy, etc.

CONCLUSIONS

In this research, the aim was to get an objective, application-specific, multi-user satisfied, and up-to-date quality assessment result. To do this, we first analyzed and detected the quality issues residing in the hydrological data supply chain – from acquisition instruments, network nodes, branch data centers, data center, control center to hydrological applications. Then, we illustrated the DQ assurance procedure that is applicable to hydrological information systems or decision support systems. Finally, we analyzed the correlations between the tasks accessing hydrological data, classified them into five categories and constructed the related quality models with time limits. Meanwhile, we developed the quality assessment model on hydrological data application in the supply chain, which means the authorized consumers can use these models to monitor and even control the quality of hydrologic data, a form of data supply chain and with a granularity of tasks. Exploratory experiments suggest the assessment system can provide DQ indicators to DQ assessors, and enable authorized consumers to monitor and even control the quality of data used in an application with a granularity of tasks.

There remains much important work to do in our future research. First, we have demonstrated how to assess the DQ in hydrological information systems, but we have not provided methods to improve the DQ of data in each process (data acquisition, transmission, transformation, integration, storage, application, or others) of data management. In the next research report, we will elaborate upon the details. Second, how to achieve an objective and subjective quality assessment with high efficiency and low energy cost in a resource-limited application has not been discussed. Third, in our further research, we will embody the characteristics of personality to get a more objective and personalized quality assessment result.

ACKNOWLEDGEMENTS

We thank the editor and two anonymous reviewers for giving us their very valuable and constructive comments and recommendations after reviewing our submission. We wish to acknowledge the help given by the experts and users of hydrological institutions; they gave us advice and data source to test. We gratefully acknowledge the financial support of the National Natural Science Foundation of China (No. 61040006) and the Education Department of Hubei Province (No. B2014245).

REFERENCES

Abdullaev, I. & Rakhmatullaev, S. 2014 Data management for integrated water resources management in Central Asia. *Journal of Hydroinformatics* **16** (6), 1425–1440.

Aboelmeged, M. 2000 A soft system perspective on information quality in electronic commerce. In: *Proceedings of the Fifth Conference on Information Quality*. MIT, Cambridge, MA, USA, Vol. 3 (1), pp. 318–319.

Alkhattabi, M., Neagu, D. & Cullen, A. 2011 Assessing information quality of e-learning systems: a web mining approach. *Computers in Human Behavior* **27** (2), 862–873.

Barone, D., Stella, F. & Batini, C. 2010 Dependency discovery in data quality. In: *Advanced Information Systems Engineering:* [Proceedings of the] *22nd International Conference, CAiSE 2010, Hammamet, Tunisia, June 2010* (Pernici, B., ed.). Springer, Berlin, Heidelberg, pp. 53–67.

Berner, E. S., Kasiraman, R. K., Yu, F., Ray, M. N. & Houston, T. K. 2005 Data quality in the outpatient setting: impact on clinical. Decision support systems. In: *AMIA Annual*

Symposium Proceedings. American Medical Informatics Association, University of Alabama, Birmingham, AL, USA, pp. 41–47.

Berti-Equille, L. 2007 Data quality awareness: a case study for cost optimal association rule mining. *Knowledge and Information Systems* **11** (2), 191–215.

Branisavljevic, N., Kapelan, Z. & Prodanovic, D. 2011 Improved real-time data anomaly detection using context classification. *Journal of Hydroinformatics* **13** (3), 307–323.

Brodie, M. L. 1980 Data quality in information systems. *Information Management* **6** (3), 245–258.

Burn, D. H. & Hag Elnur, M. A. 2002 Detection of hydrologic trends and variability. *Journal of Hydrology* **255** (1), 107–122.

Cappiello, C., Francalanci, C. & Pernici, B. 2003 Time-related factors of data quality in multichannel information systems. *Journal of Management Information Systems* **20** (3), 71–92.

Caro, A., Calero, C., Caballero, I. & Piattini, M. 2008 A proposal for a set of attributes relevant for Web portal data quality. *Software Quality Journal* **16** (4), 513–542.

Coen-Porisini, A. & Sicari, S. 2012 Improving data quality using a cross layer protocol in wireless sensor networks. *Computer Networks* **56** (17), 3655–3665.

de Carvalho Costa, R. L. & Furtado, P. 2011 Quality of experience in distributed databases. *Distributed and Parallel Databases* **29** (5–6), 361–396.

Dillon, P. J. & Woithe, S. D. 1988 *Hydrological Data Acquisition Systems*. NSIDC, Washington, USA, Patent No. 4,720,799. 1998, pp. 1–13.

Falge, C., Otto, B. & Osterle, H. 2012 Data quality requirements of collaborative business processes. In: *Proceedings of the 45th Hawaii International Conference on System Science*. Maui, HI, IEEE Computer Society, Vol. 1 (5), 4316–4325.

Ge, M. & Helfert, M. 2008 Data and information quality assessment in information manufacturing systems. *Business Information Systems* **8**, 380–389.

Gorla, N., Somers, T. M. & Wong, B. 2010 Organizational impact of system quality, information quality, and service quality. *Journal of Strategic Information Systems* **19** (3), 207–228.

Gourbesville, P. 2009 Data and hydroinformatics: new possibilities and challenges. *Journal of Hydroinformatics* **11** (3–4), 330–343.

Huang, H., Stvilia, B., Jörgensen, C. & Bass, H. W. 2012 Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology* **63** (1), 195–207.

Hudson, H. R., McMillan, D. A. & Pearson, C. P. 1999 Quality assurance in hydrological measurement. *Hydrological Sciences Journal* **44** (5), 825–834.

Karplus, P. A. & Diederichs, K. 2012 Linking crystallographic model and data quality. *Science* **336** (6084), 1030–1033.

Karr, A. F. 2006 Exploratory data mining and data cleaning. *Journal of the American Statistical Association* **101** (473), 399–402.

Kesh, S. 1995 Evaluating the quality of entity relationship models. *Information and Software Technology* **37** (12), 681–689.

Knight, S. A. & Burn, J. M. 2005 Developing a framework for assessing information quality on the World Wide Web. *Informing Science: International Journal of an Emerging Transdiscipline* **8** (5), 159–172.

Krajewski, W., Kruger, A., Smith, J., Lawrence, R., Gunyon, C., Goska, R., Seo, B.-C., Domaszczynski, P., Baeck, M. L., Ramamurthy, M. K., Weber, J., Bradley, A. A., DelGreco, S. A. & Steiner, M. 2011 Towards better utilization of NEXRAD data in hydrology: an overview of Hydro-NEXRAD. *Journal of Hydroinformatics* **13** (2), 255–266.

Lee, Y. W. & Strong, D. M. 2003 Knowing-why about data processes and data quality. *Journal of Management Information Systems* **20** (3), 13–39.

Lee, Y. W., Strong, D. M., Kahn, B. K. & Wang, R. Y. 2002 AIMQ: a methodology for information quality assessment. *Information & Management* **40** (2), 133–146.

Li, C., Zhou, H. & Zhou, X. 2012 Design and implement of hydrological data quality assessment system based on business rules. In: *IET International Conference on Information Science and Control Engineering 2012*. IET, Shenzhen, China, Vol. 3 (1), pp. 925–929.

Lim, K. K., Ahmed, P. K. & Zairi, M. 1999 Managing for quality through knowledge management. *Total Quality Management* **10** (4–5), 615–621.

Liu, S. J., Zhou, H. H. & Li, C. 2011 Design of remote engine room monitoring system based on niche stack TCP/IP. *Lecture Notes in Electrical Engineering* **1** (2), 417–425.

Madnick, S. E., Wang, R. Y., Lee, Y. W. & Zhu, H. 2009 Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)* **1** (1), 1–22.

Molina, R., Unsworth, K., Hodkiewicz, M. & Adriasola, E. 2013 Are managerial pressure, technological control and intrinsic motivation effective in improving data quality? *Reliability Engineering & System Safety* **119**, 26–34.

Mosley, M. P. & McKerchar, A. 1989 Quality assurance programme for hydrometric data in New Zealand. *Hydrological Sciences Journal* **34** (2), 185–202.

Motro, A., Anokhin, P. & Acar, A. C. 2004 Utility-based resolution of data inconsistencies. In: *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*. ACM, Paris, France, Vol. 5 (2), pp. 35–43.

Ozmen-Ertekin, D. & Ozbay, K. 2012 Dynamic data maintenance for quality data, quality research. *International Journal of Information Management* **32** (3), 282–293.

Peltier, J. W., Zahay, D. & Lehmann, D. R. 2013 Organizational learning and CRM success: a model for linking organizational practices, customer data quality, and performance. *Journal of Interactive Marketing* **27** (1), 1–13.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. 2002 Data quality assessment. *Communications of the ACM*. **45** (4), 211–218.

Pitt, L. F., Watson, R. T. & Kavan, C. B. 1995 Service quality: a measure of information systems effectiveness. *MIS Quarterly* **19** (2), 173–187.

Rahm, E. & Do, H. H. 2000 Data cleaning: problems and current approaches. *IEEE Data Engineering Bulletin* **23** (4), 3–13.

Schlaeger, F., Natschke, M. & Witham, D. 2007 *Quality Assurance for Hydrometric Network Data as a Basis for Integrated River Basin Management*. IAHS-AISH Publication, Wallingford, UK, pp. 327–337.

Seddon, P. B. 1997 A respecification and extension of the DeLone and McLean model of IS success. *Information Systems Research* **8** (3), 240–253.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H. & Mustapha, A. 2012 Data quality: A survey of data quality dimensions. In: *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*. IEEE Computer Society, Kuala Lumpur, Malaysia, pp. 300–304.

Strong, D. M., Lee, Y. W. & Wang, R. Y. 1997 Data quality in context. *Communications of the ACM* **40** (5), 103–110.

Stvilia, B., Mon, L. & Yi, Y. J. 2009 A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology* **60** (9), 1781–1791.

Sun, M., Dou, H., Li, Q. & Yan, Z. 2012 Quality estimation of Deep Web Data Sources for data fusion. *Procedia Engineering* **29**, 2347–2354.

Wand, Y. & Wang, R. Y. 1996 Anchoring data quality dimensions in ontological foundations. *Communications of the ACM* **39** (11), 86–95.

Wang, R. Y. 1998 A product perspective on total data quality management. *Communications of the ACM* **41** (2), 58–65.

Wang, R. Y. & Strong, D. M. 1996 Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems* **12** (4), 5–33.

Wang, R. Y., Reddy, M. P. & Kon, H. B. 1995 Toward quality data: an attribute-based approach. *Decision Support Systems* **13** (3), 349–372.

Watts, S., Shankaranarayanan, G. & Even, A. 2009 Data quality assessment in context: a cognitive perspective. *Decision Support Systems* **48** (1), 202–211.

WMO 2003 *Hydrological data management present state and trends*. Report of the World Meteorological Organization. World Meteorological Organization, Geneva, Switzerland.

Woodall, P., Borek, A. & Parlikad, A. K. 2013 Data quality assessment: the hybrid approach. *Information & Management* **50** (7), 369–382.

Yerva, S. R., Miklós, Z. & Aberer, K. 2012 Quality-aware similarity assessment for entity matching in Web data. *Information Systems* **37** (4), 336–351.

Zhu, Y. & Buchmann, A. 2002 Evaluating and selecting web sources as external information resources of a data warehouse. In: *Proceedings of the Third International Conference on Web Information Systems Engineering*. IEEE Computer Society, Singapore, Vol. 2 (1), pp. 149–160.

First received 21 March 2014; accepted in revised form 21 January 2015. Available online 27 March 2015