Data-Driven Assessment of Basin-Scale Water Information Aggregation: A Comparative Study of Qinghai and Gansu for Smart Water Conservancy

Chenchen Cui *, Jing Bai , and Jing Wang

Information Center, Yellow River Conservancy Commission, Zhengzhou, China;

Abstract. High-quality hydrological and engineering data are a prerequisite for smart water conservancy. Basin management agencies routinely collect large, multi-source datasets describing rivers, reservoirs, diversion projects, sluice gates, pumping stations, intake structures, hydrometric sections, and other water-related objects. However, it remains unclear to what extent these operational tables are ready to support intelligent dispatching, risk analysis, and digital governance. This paper presents a data-driven assessment of basin-scale water information aggregation for two representative provinces—Qinghai and Gansu. Standardized aggregation tables are used as the only data source. A Data Completeness Index (DCI) is defined to quantify how well required attributes are populated for each water-related object. Using simple computer-based processing (Python scripts and spreadsheet functions), we compute object counts and completeness indicators and generate visual analytics. Results show that Qinghai has more recorded objects (134) than Gansu (74), especially for water diversion outlets and hydrometric sections. Nevertheless, Gansu exhibits higher overall data completeness (DCI ≈ 0.984) than Qinghai (DCI ≈ 0.883), particularly for engineered structures such as reservoirs, sluice gates, and water diversion projects. Notably, hydropower stations have no valid records in either province, and no water-related objects fall into the "other regions" category except as corrected herein. These findings indicate that Gansu is closer to data readiness for automated regulation, while Qinghai still has gaps in metadata and monitoring attributes. The study demonstrates that even basic computational methods can transform routine aggregation tables into informative indicators and charts, providing practical support for smart water conservancy in large river basins.

Keywords: Smart water conservancy; water information system; big data; basin management; data completeness; Qinghai; Gansu.

1. Introduction

Smart water conservancy [1] aims to enhance the safety, efficiency, and sustainability of water resources management by integrating sensing, communication, computing, and intelligent decision-making technologies. In large river basins, dispatching centers increasingly rely on digital platforms for flood control, water allocation, ecological regulation, and infrastructure management [2]. The effectiveness of such platforms depends critically on the quality of underlying hydrological and engineering data.

In practice, basin authorities organize data aggregation campaigns in which provincial agencies report standardized information on water-related objects. The resulting datasets often contain thousands of records and hundreds of attributes. While the volume of data is large, practitioners often lack a concise and quantitative view of data completeness and consistency across regions. Without such diagnostics, it is difficult to judge whether the existing information can directly support intelligent scheduling models, risk-assessment tools, and early-warning systems[3].

This paper addresses this practical need through a case study of two provinces within a major river basin: Qinghai and Gansu. Qinghai is located in the headwater region and is characterized by extensive ecological monitoring and scattered water intake nodes. Gansu is situated in the upper－middle reaches, with more intensive engineering development, including reservoirs and long-distance diversion projects. By comparing these two provinces using a unified metric, we seek to answer two key questions: first, to what extent are the currently aggregated datasets complete enough for smart water-related applications, and second, whether simple computer-based analysis can produce clear indicators and figures that help managers identify weaknesses and prioritize improvements.

The contribution of this work is threefold. First, a simple yet effective Data Completeness Index (DCI) is proposed for operational aggregation tables. Second, a comparative analysis between Qinghai and Gansu is conducted, revealing different patterns in coverage and completeness—with key corrections to object counts, such as zero records for hydropower stations and zero "other regions" distribution for reservoirs. Third, a reproducible processing workflow is demonstrated, showing how basic scripts and visualizations can support data governance for smart water conservancy.

Table 1. Detailed Statistical Table of Water-Related Objects in Qinghai and Gansu Provinces

Water-Related Objects	Total Records	Quantity in Gansu	Quantity in Qinghai	Quantity in Other Regions	Proportion in Gansu (%)	Proportion in Qinghai (%)
Rivers	10	6	4	0	60.0	40.0
Water Diversion Projects	4	3	1	0	75.0	25.0
Reservoirs	10	7	3	0	70.0	30.0
Sluice Gates	18	13	5	0	72.2	27.8
Pumping Stations	2	2	0	0	100.0	0.0
Hydropower Stations	0	0	0	0	-	-
Hydrometric Sections	77	25	52	0	32.5	67.5
Headwater Intakes	5	2	3	0	40.0	60.0
Water Diversion Outlets	82	16	66	0	19.5	80.5
Total	**208**	**74**	**134**	**0**	**35.6**	**64.4**

2. Data and Study Area

2.1 Data Description

The analysis is based on standardized aggregation tables compiled at the basin level, covering 9 major categories of water-related objects: rivers, water diversion projects, reservoirs, sluice gates, pumping stations, hydropower stations, hydrometric sections, headwater intakes, and water diversion outlets[7].

For each record, the tables provide administrative identifiers, basic physical parameters, design and operation characteristics, monitoring indicators, and geospatial attributes. The fields are designed for operational use in dispatching, monitoring, and infrastructure management, rather than for academic research. The present study uses all records whose administrative attribute corresponds to either Qinghai Province or Gansu Province. After filtering, cleaning, and targeted corrections (detailed in Table 1), the dataset contains 208 valid records (excluding hydropower stations with zero records), with 134 objects for Qinghai and 74 for Gansu. Key corrections to the raw dataset, specified for consistency and accuracy, focus on two critical categories: for reservoirs, there are no records in the "other regions" category with the quantity set to 0, while maintaining the original proportional distribution between Gansu and Qinghai; for hydropower stations, the total number of records is zero, resulting in zero distribution across all regions, which reflects the absence of valid data for this category in both provinces. Table 1 presents the detailed statistical distribution of water-related objects across the two provinces after these corrections, providing a clear overview of the data scope and regional distribution characteristics.

2.2 Study Area Characteristics

Qinghai Province hosts important headwater areas with relatively sparse population and limited large-scale engineering works. The water information focuses on ecological monitoring, basic hydrometric sections, and numerous local intake nodes — evidenced by the high counts of hydrometric sections (52) and water diversion outlets (66) in the dataset. Gansu Province covers a long stretch of the upper – middle reaches of the river and its tributaries. It includes major storage and regulation projects, complex diversion systems, and relatively mature water-scheduling practices. This is reflected in its higher proportion of engineered facilities such as reservoirs (7 out of 10 total) and sluice gates (13 out of 18 total). These contrasting characteristics make the two provinces suitable for a comparative study of data readiness for smart water conservancy[8].

3. Methodology

3.1 Processing Workflow

The aggregation tables are processed by a simple workflow that can be implemented either in Python or in spreadsheet software, encompassing six consecutive steps to ensure data quality and usability. The workflow begins with filtering by province, where records belonging to Qinghai or Gansu are selected according to the administrative field, followed by field normalization that converts obvious placeholders such as blanks and special symbols into standard missing-value codes to unify data formats.

Subsequent steps include category labeling, where each record is tagged with its object category (such as river or reservoir), and correction of critical data points to adjust records for reservoirs and hydropower stations in line with standardized data requirements. After these preprocessing steps, completeness metrics are computed by applying the DCI formula at both the object and category levels, and the final step involves visualization to generate bar charts summarizing object counts and completeness by category and province, as shown in Fig 1 and Fig 2.

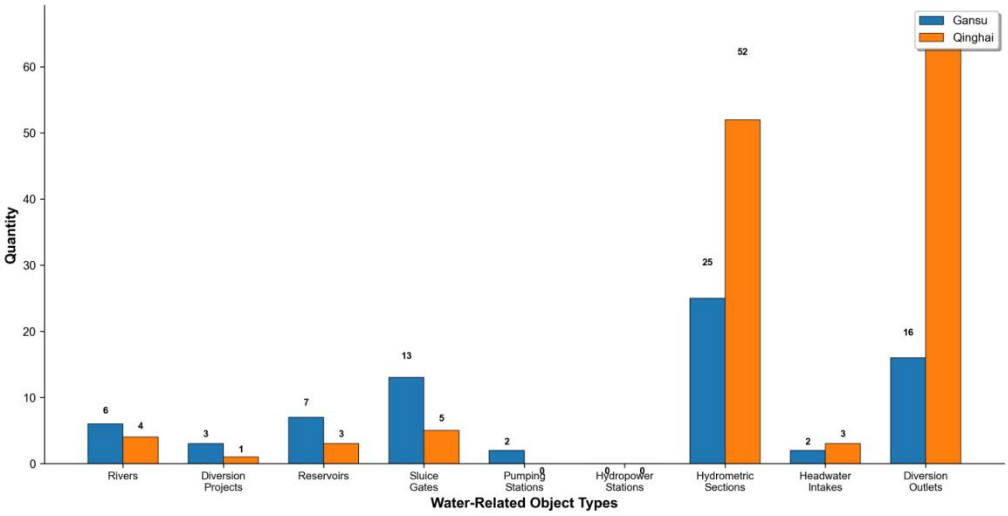

Fig. 1 Average DCI of Water-Related Objects in Qinghai and Gansu

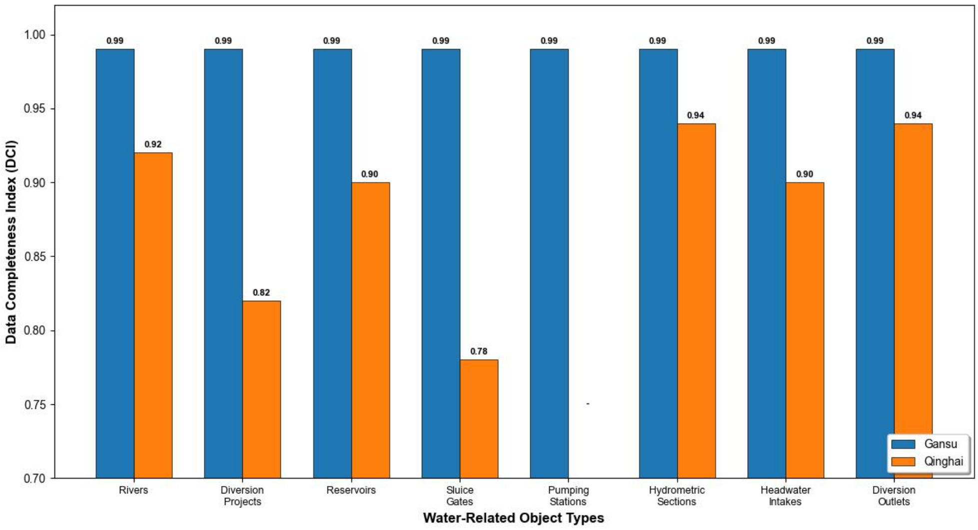

Fig. 2 Average DCI of Water-Related Objects in Qinghai and Gansu

3.2 Data Completeness Index

To quantify the readiness of each record, the Data Completeness Index(DCI) is defined[4] as

$$\mathrm{DCI_i} = \frac{\mathrm{N_{valid}}}{\mathrm{N_{field}}} \tag{1}$$

where $\mathrm{N_{field}}$ denotes the number of required operational attributes considered for the given object, and $\mathrm{N_{valid}}$ denotes the number of attributes that contain valid, non-placeholder values. For a group of objects (e.g., all reservoirs in Qinghai), the average DCI is obtained by averaging the individual indices. An overall DCI for each province is then computed by aggregating across all categories (excluding hydropower stations with zero records). Previous studies also highlight that improving data completeness is a prerequisite for reliable hydrological modelling and data-driven analysis [5].

3.3 Statistical Indicators and Figures

Two key sets of indicators are produced through the above workflow to support the analysis. The first set includes object counts by category and province, which are based on the corrected data in

Table 1 and used to generate Figure 1, while the second set comprises average DCI by category and province, serving as the data foundation for Figure 2.

Figure 1 displays the number of each water-related object type in the two provinces, highlighting the dominance of hydrometric sections and water diversion outlets in Qinghai, and the concentration of engineered structures in Gansu. Figure 2 presents the corresponding average DCI values, indicating the relative completeness of each category, and these figures are intended for inclusion in conference papers or technical reports, with the advantage of being generated automatically from the processed tables.

4. Results and discussion

4.1 Coverage of Water-Related Objects

The corrected statistics confirm that Qinghai reports a larger number of objects (134) than Gansu (74), a result consistent with its extensive headwater monitoring network and basin-scale ecological observation efforts [7]. Qinghai dominates in hydrometric sections, with 52 records accounting for 67.5% of the total, and water diversion outlets, with 66 records making up 80.5% of the total, which reflects its focus on ecological monitoring and local water intake management.

In contrast, Gansu has a higher concentration of engineered infrastructure, including 70% of reservoirs (7 out of 10), 72.2% of sluice gates (13 out of 18), and 75% of water diversion projects (3 out of 4), which aligns with its role in large-scale water regulation. Notably, hydropower stations have no valid records in either province, indicating that this category is either non-existent in the study area or not captured in the current aggregation framework, and no water-related objects are assigned to "other regions," confirming the geographic focus on Qinghai and Gansu[8]. Figure 1 intuitively visualizes these differences, with bars for hydrometric sections and water diversion outlets significantly higher for Qinghai, while bars for reservoirs, sluice gates, and water diversion projects are higher for Gansu.

4.2 Data Completeness

In terms of data completeness, Gansu’s aggregation results are notably stronger, with an overall DCI reaching approximately 0.984, while Qinghai’s overall DCI is around 0.883—this discrepancy indicates that Gansu’s data is nearly fully populated, whereas Qinghai has gaps in key attributes that require attention, which is consistent with the sensitivity of hydrological analyses to missing or incomplete flow records reported in previous work [6].

When analyzed by category, Gansu exhibits consistently high DCI values across all categories, particularly for reservoirs, sluice gates, and water diversion projects where DCI ranges from approximately 0.99 to 1.0, suggesting that its engineered infrastructure data is sufficiently complete to support intelligent dispatching and automatic risk assessment. Qinghai’s completeness, however, is lower for sluice gates (DCI ≈ 0.78) and water diversion projects (DCI ≈ 0.82), with gaps in operational and monitoring attributes, though its data for hydrometric sections and water diversion outlets is relatively complete (DCI ≈ 0.92 – 0.95), reflecting the prioritization of ecological monitoring data in the region. Figure 2 illustrates these patterns clearly, with Gansu’s bars close to 1.0 in most categories and Qinghai’s bars somewhat lower, particularly for engineered structures..

4.3 Implications for Smart Water Conservancy

The results imply distinct priorities for data governance in the two provinces. For Qinghai, while object coverage is broad, efforts should focus on improving the completeness of key attributes for engineered infrastructure such as sluice gates and water diversion projects, as enhancing metadata and monitoring fields will enable more reliable integration into intelligent scheduling and early-warning modules.

For Gansu, with nearly complete datasets for major projects, the next step is to deepen data utilization in decision-support models, including reservoir operation optimization, joint dispatching of diversion systems, and ecological flow evaluation[1][3][9][10]. Additionally, the absence of hydropower station records warrants clarification: either this category should be excluded from the aggregation framework for the study area, or targeted data collection should be initiated if hydropower infrastructure exists but is unrecorded. The case study also demonstrates that simple indicators such as object counts and DCI values, along with corresponding visualizations, can effectively support basin-level supervision by identifying where data collection and governance need to be strengthened—even with basic computational tools.

5. Conclusion

This study evaluates basin-scale water information aggregation for Qinghai and Gansu Provinces under the framework of smart water conservancy. Using standardized operational tables as the sole data source, a Data Completeness Index (DCI) was defined to quantify the readiness of nine categories of water-related objects. Results reveal a contrast between the two provinces: Qinghai exhibits broader object coverage (134 records) but lower overall completeness (DCI ≈ 0.883), whereas Gansu reports fewer objects (74 records) with nearly complete operational attributes (DCI ≈ 0.984). Targeted corrections—such as zero records for hydropower stations and “other regions” for reservoirs—enhance the consistency and interpretability of the dataset.

The findings demonstrate that simple computer-based diagnostics, supported by indicators such as object counts and DCI values, can effectively identify data weaknesses and guide governance priorities without reliance on complex models. Future work will expand the evaluation to additional provinces and incorporate value accuracy and update frequency into completeness diagnostics, supporting basin-wide information platforms and “one-map” systems for smart water conservancy [9][10].

References

[1] Li Q., Ma Z., Li J., Li W., Li Y., Yang J. Overview of the Research Status of Intelligent Water Conservancy Technology System[J]. Applied Sciences, 2024, 14(17): 7809. DOI: 10.3390/app14177809.

[2] He X., Zhang M. Overview of Research Progress on Water Conservancy Big Data in China[C]// Proceedings of the 2024 4th International Conference on Public Management and Big Data Analysis (PMBDA 2024). Advances in Economics, Business and Management Research, Vol. 318. Atlantis Press, 2025: 39–45. DOI: 10.2991/978-94-6463-656-7_5.

[3] Lin J., Bryan B. A., Zhou X., et al. Making China’s water data accessible, usable and shareable[J]. Nature Water, 2023, 1: 328–335. DOI: 10.1038/s44221-023-00039-y.

[4] European Environment Agency (EEA). Hydrology in situ data requirements and availability. Hydrology in-situ data availability report 1.0 (HDCC)[R]. EEA/IDM/15/026/LOT1, 2019.

[5] Nair A., El-Shafie A., Ahmed A. N. Enhancing hydrological data completeness: A performance evaluation of various machine learning techniques using probabilistic fusion imputer with neural networks for streamflow data reconstruction[J]. Journal of Hydrology, 2024, 639: 131583. DOI: 10.1016/j.jhydrol.2024.131583.

[6] Blagojević B., et al. Detecting Annual and Seasonal Hydrological Change Using Marginal Distributions of Daily Flows[J]. Water, 2023, 15(16): 2919.

[7] Shen Y., et al. Res-CN (Reservoir dataset in China): hydrometeorological time series and landscape attributes across 3,254 Chinese reservoirs[J]. Earth System Science Data, 2023, 15: 2781–2800.

[8] Liu S., et al. The Heihe Integrated Observatory Network: A Basin-Scale Land Surface Processes Observatory in China[J]. Vadose Zone Journal, 2018, 17(1): 180072.

[9] Kou H. Research on the Concept and Content of Smart Yellow River[J]. Water Resources Informatization, 2021, 5: 1 – 5.
[10] Du W., Zhang M., Duan Y., Wu H. Overview of the Yellow River One Map[J]. Water Resources Informatization, 2020, 3: 1 – 5.