Journal of Water Resources Research 水资源研究, 2025, 14(5), 451-457
Published Online October 2025 in Hans. https://www.hanspub.org/journal/jwrr
https://doi.org/10.12677/jwrr.2025.145049

Hans汉斯

# 基于大数据多模式互验的水资源计量数据异常检测方法

雷四华[1]，高若凡[2]，单心怡[2]，许　怡[1]，吴　巍[1]，张擎天[2]

[1]南京水利科学研究院，水灾害防御全国重点实验室，江苏 南京
[2]河海大学计算机与软件学院，江苏 南京

收稿日期：2025年7月9日；录用日期：2025年10月13日；发布日期：2025年10月30日

## 摘　要

**水资源计量过程环节多，异常数据难于避免，开展数据异常检测是水资源计量管理重要内容。模型算法是快速自动发现水资源计量异常数据的核心技术，大数据挖掘技术已有较多算法在数据异常检测中应用，对模型算法有效利用是准确发现异常数据的工作基础。本异常检测方法是基于统计、分类、预测等3种大数据模型算法进行数据异常检测，研究完整设计了计算处理流程，并应用软件技术完成系统开发，实现计算参数自定义，采用消息控制多模式计算，有效实现交互验证处理；经实测数据应用检验，实验分析表明，该检测方法能有效提升异常发现准确率、提高修正的便利性。**

## 关键词

**水资源计量数据，数据异常检测，大数据方法，多模式互验**

# A Method for Abnormal Detection in Water Resource Measurement Data Based on Big Data Multimode Mutual Verification

**Sihua Lei[1], Ruofan Gao[2], Xinyi Shan[2], Yi Xu[1], Wei Wu[1], Qingtian Zhang[2]**

[1]The National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing Jiangsu
[2]College of Computer Science and Software Engineering, Hohei University, Nanjing Jiangsu

Received: July 9, 2025; accepted: October 13, 2025; published: October 30, 2025

## Abstract

**The process of water resource measurement involves multiple stages, and abnormal data is difficult to**

作者简介：雷四华，江西南昌人，硕士，正高，研究方向：水文水资源、水利信息化。Email: shlei@nhri.cn

文章引用: 雷四华, 高若凡, 单心怡, 许怡, 吴巍, 张擎天. 基于大数据多模式互验的水资源计量数据异常检测方法[J]. 水资源研究, 2025, 14(5): 451-457. DOI: 10.12677/jwrr.2025.145049

**avoid. Conducting data anomaly detection is an important part of water resource measurement management. Model algorithms are the core technology for quickly and automatically discovering abnormal data. Big data mining technology has been widely applied in data anomaly detection, and effective utilization of model algorithms is the foundation for accurately discovering abnormal data. Based on three big data model algorithms, including statistics, classification, and prediction, research was conducted, and a calculation processing flow was designed. Software technology was applied to complete system development, achieving custom calculation parameters and using message control for multi-mode calculation and effectively achieving interactive verification processing; Applying measured data for application calculations, experimental analysis shows that this detection method can effectively improve the accuracy of anomaly detection and the level of intelligence for correction.**

## Keywords

**Water Resource Measurement Data, Data Anomaly, Big Data Methods, Multimode Mutual Verification**

Copyright © 2025 by author(s) and Wuhan University & Bureau of Hydrology, Changjiang Water Resources Commission.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

水资源计量是指在管道或明渠上安装满足技术标准要求的专用水量计量设施，并按规定将计量数据采集传输至指定位置存储。随着信息技术发展及管理需求，水资源计量普及范围、准确性要求不断提高。因数据采集、传输存储、数据交换等设备设施安装及运行故障或人为干预影响，水资源计量数据出现突变、缺失等异常现象仍无法避免。开展数据异常检测，及时发现异常，为异常原因发现、故障排除、修正数据提供技术支撑，一直是水资源计量技术研究重点内容[1] [2]。

水资源管理中基于大数据方法已有较广泛应用，应用领域包括水资源公共事件的监测和预警、异常数据检测、水资源管理决策与政策评估[3]，包括 ADWICE 聚类[4]、ARIMA 模型、3$\sigma$ 准则[5]、支持向量机和深度神经网络[6]等算法[7]，应用前景广泛[8]。

水资源计量数据具有时间序列特征，同时与计量监测对象生产经营活动密切相关，经多年建设，水资源计量监测数据已规模化存储[9] [10]。随着水资源计量数据的多样化、复杂化、海量化，独立的异常检测模型已经无法满足数据异常检测准确性需求。本文从水资源计量数据的特性出发，采用大数据机器学习技术，研发了多模型算法处理技术，并建立了可靠的水资源计量数据异常检测分析系统。

## 2. 模型算法及互验处理

### 2.1. 模型算法

水资源计量是按规定时段间隔定时读取取用水量的过程，其数据产品是以计量时间为标记的系列数据，与系列众多数据具有明显差异的称为突变异常数据；另外，按水资源计量管理要求应报而未报或系列中存在数据缺失，称为缺失异常。针对水资源计量数据异常特点，本研究分别采用基于统计、分类、预测等 3 种模型算法对突变异常检测。采用时间点比对是检测缺失异常的常用方法。

1) 基于统计标准差法

标准差法是计量统计及异常检测的常用方法[11] [12]，计算各待检测计量值和已知序列计量值均值之间的差值，得到残差序列，设定残差序列的阈值用于判断残差是否异常，进而判断计量值是否为异常值，因该阈值可

动态调整，故又称为动态阈值方法，即 K-sigma 检测法。如下计算公式所示，标准差法方法简练、计算过程清晰，具有运算效率高的优势。

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}} \tag{2}$$

$$\left|Y - \bar{X}\right| > k\sigma \tag{3}$$

式中：$X_i$ 为已知序列 $X_1, X_2, \cdots, X_n$，$\bar{X}$ 为均值，$\sigma$ 为标准差，$k$ 为调整系数，$Y$ 为待检测计量值，$\left|Y - \bar{X}\right|$ 为残差系列。作为判断异常残差范围，由 $k$ 乘以标准差确定，故 $k$ 值对该范围具有较大影响。

2) 基于分类孤立森林算法

孤立森林(Isolated forest)是基于无监督机器学习的异常检测算法，通过隔离数据中的离群值识别异常，作为大数据方法近年在异常检测中得到深入研究应用[13]。本研究基于决策树构建孤立森林，孤立森林由多颗树组成，实现时，每棵树构成是从给定的计量数据序列中随机抽取一部分数据，确定该部分数据最大值和最小值，然后在最大值和最小值间随机选择一个分割值，形成二叉树，来隔离离群值。这种随机划分会使异常数据点在树中生成的路径更短，从而将它们和其他数据分开。

3) 基于预测 ARIMA 算法

ARIMA 模型的全称是差分自回归滑动平均模型(Autoregressive Integrated Moving Average Model)也记作 ARIMA(p,d,q)，该方法常用于对系列数据预测[14]，通过对预测值与计量值对比可判断计量是否异常。ARIMA 模型的方程式为：

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{p}\phi_i L^i\right)\varepsilon_t \tag{4}$$

模型参数 $L$ 是滞后算子(Lag operator)，参数 $p$ 是代表时间序列数据自身的滞后性，参数 $d$ 是代表时间序列数据变成平稳性数据需要差分的次数，参数 $q$ 代表模型中采用的预测误差的滞后数。

### 2.2. 互验处理

水资源计量数据特征与计量监测管理对象密切相关，各管理对象所在区域、所属行业、工作时段等影响其取用水过程，本研究采用不同时段、不同行业、不同区域等作为模型输入，应用不同模型、不同输入进行检测可认为是一套计算参数。

基于统计、分类、预测等 3 种算法分别构建计算模型，各模型对应不同计算参数，每个模型与计算参数组合形成一种计算模式。当同一序列数据使用不同计算模式检测，计算得出的异常值不尽相同，对各异常值序列使用三维图形分析，可实现互验，确定检测结果的准确性。

## 3. 计算处理流程及软件系统实现

### 3.1. 计算处理流程

模型算法为检测处理核心内容，各模型算法封装为独立模块，计算处理流程主要有待检数据预处理、模型计算处理、异常值标记、异常值预修正、互验处理等环节，计算处理流程如图 1，其中模型训练仅适用于 ARIMA 算法。各环节计算处理内容如下。

1) 待检数据预处理：按选择的时间段、水资源计量对象监测范围等从数据库中读取待检测数据，形成数据

系列矩阵，检测是否存在数据缺失，对数据缺失时间点给出标记。

2) 模型计算处理：获取待检测数据的数据特征，使用相应的模型对训练数据集进行训练、参数求解，按模型算法检测数据，检测发现异常值。本研究利用消息控制及数据库表实现了多个模式并行计算。

3) 异常值标记：对异常数据进行标记，包括检测算法标识、异常分类等，分类类型包括突变大、突变小、突变零、持续不变、负值等[15]。

4) 异常值预修正：若需要对异常数据进行自动修正，则读取修正规则，在保持原序列数据不变情况下，按照预设修正条件计算得出修正值，并另行存储。

5) 互验处理：以水资源计量监测对象取用水特征为划分，确定互验数据范围，包括取用水周期、行业类别、企业规模等，对各模型检测结果进行对比处理，计算各异常值差别，统计异常特征，提供检测结果数据图形可视化展示，进一步确定异常检测结果准确性，对确定的结果写入应用数据库表。

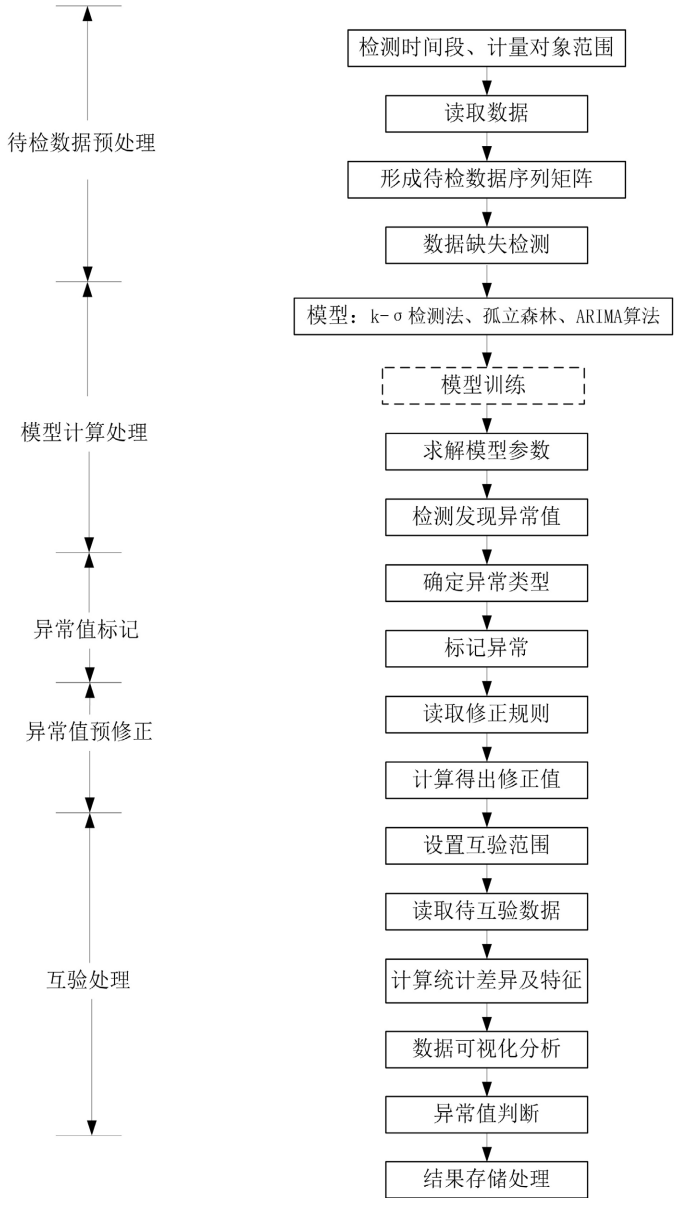


**图 1.** 水资源计量数据异常检测计算处理流程

### 3.2. 软件系统实现

本研究以水资源计量管理现有数据库为基础[16] [17]，扩展异常检测数据库表，采用通用跨平台语言开发建立异常检测软件系统。异常检测数据库表组成如表 1 所示。

**表 1.** 异常检测数据库表组成

| 序号 | 数据库表名称 | 主要字段 | 说明 |
|---|---|---|---|
| 1 | 异常检测结果 | 监测点代码、时间、水量、是否异常、异常类型、修正值、ARIMA 模型差值、检测模型标识、检测时间戳 | 监测点代码、时间、水量等由检测数据序列直接摘录，其余字段内容为模型检测结果及相应标识 |
| 2 | 检测任务控制表 | 行政代码、行业代码、方差起始时间、方差结束时间、是否包括节假日、阈值、异常检测开始时间、异常检测结束时间、连续个数、是否自动修正、空值插值补充、按区域还是行业检测、任务状态、任务创建时间、任务更新时间、任务执行间隔时间、任务类型 | 对检测任务给定的检测范围、参数等进行记录 |

软件系统采用开源跨平台 JAVA 语言开发，系统框架为 springboot，集成 mybatis；应用 python 中的机器学习库 sklearn 引入 ARIMA 模型，按照水资源计量数据序列及检测数据库表结构对模型进行修改定制，并在 JAVA 调用 python 执行该代码模块；采用 RabbitMQ 消息控制技术实现多模式并行计算；应用 Layui、ECharts 等组件实现图表展示。

## 4. 实验及结果

### 4.1. 实验数据

本实验数据来源于水资源计量数据的取用水计量自动监测数据，按照取用水计量自动监测传输要求，各计量监测点每日需报送 24 个整点小时水量、1 个日水量，共 25 个数据，分别存入小时水量数据库表、日水量数据库表[15]。

实验小时水量数据是在收集的某省自 2021 年 1 月 1 日 0 时至 2022 年 11 月 11 日 0 时的小时水量数据基础上，筛选出数据量最多的站点所对应时段，确定时段为 2021 年 1 月 1 日 1 时至 2021 年 8 月 5 日 11 时，对应计量监测点 1527 个，各监测点应报数据为 5219 条。实验日水量数据时段为自 2021 年 1 月 1 日至 2022 年 10 月 31 日，各监测点应报送日水量数据为 668 条。为了检验模型计算及软件系统处理的准确性与效率，考虑数据可比性及代表性，按照小时水量、日水量数据报送完整性，分为 5 个等级，数据完整性方面监测点分布如表 2。

**表 2.** 数据完整性小时水量监测点分布

| 序号 | 完整性等级 | 小时水量数据分布 | | 日水量数据分布 | |
|---|---|---|---|---|---|
| | | 数据量范围 | 监测点数 | 数据量范围 | 监测点数 |
| 1 | 差 | [1, 2000) | 603 | [1, 200) | 1157 |
| 2 | 较差 | [2000, 3000) | 277 | [200, 300) | 281 |
| 3 | 一般 | [3000, 3500) | 192 | [300, 450) | 542 |
| 4 | 较好 | [3500, 4000) | 249 | [450, 550) | 571 |
| 5 | 良好 | [4000, 4500) | 183 | [550, 650) | 2339 |
| 6 | 优 | >=4500 | 23 | >=650 | 21 |

### 4.2. 结果分析

为了检验模型算法，在模型研发过程使用实验数据进行了多方式大量测试，为了便于图形描述，本文选取

单监测点部分时段检测结果分析模型算法互验处理性能。本文选取的测站编码为 XX04310001002，在完整性方面小时水量、日水量分别处于较好、良好范围。分析时段为 2021 年 2 月 1 日至 2021 年 3 月 31 日，该时段小时水量实测数据共 974 个，缺失 442 个；日水量实测数据共有 59 个，无缺失。

ARIMA 模型检测对小时水量检测出 8 个异常后，因数据序列存在缺失而中止运行；对日水量共检测出 17 个异常值，包括突变大 3、突变小 14 个；孤立森林模型对小时水量共检测出 89 个异常值，均标记为孤立异常，未区分异常类别；对日水量共检测出 7 个异常值；K-sigma 算法对小时水量共检测出 848 个异常值，分别有持续不变 309 个、连续波动 100 个、数据缺失 419 个、突变大 1 个、突变零 19 个；对日水量共检测出 13 个异常值，均为持续不变异常。检测结果图形过程如图 2 所示。

在小时水量检测中，各模型算法均检测到发生在 2021 年 2 月 27 日 18 时的异常大值(38,860,656 $m^3$)，因该值远大于其他值，为了图形表达效果，图 2 已剔除该值，K-sigma 算法能对异常数据进一步清晰分类。

从日水量检测结果分析，ARIMA 模型与 K-sigma 算法具有良好的互验作用，孤立森林模型检测结果不宜采纳；在突变大的检测结果中，因 K-sigma 算法基于标准差倍数，各数据变幅均在其设定范围内，未有突变大异常。

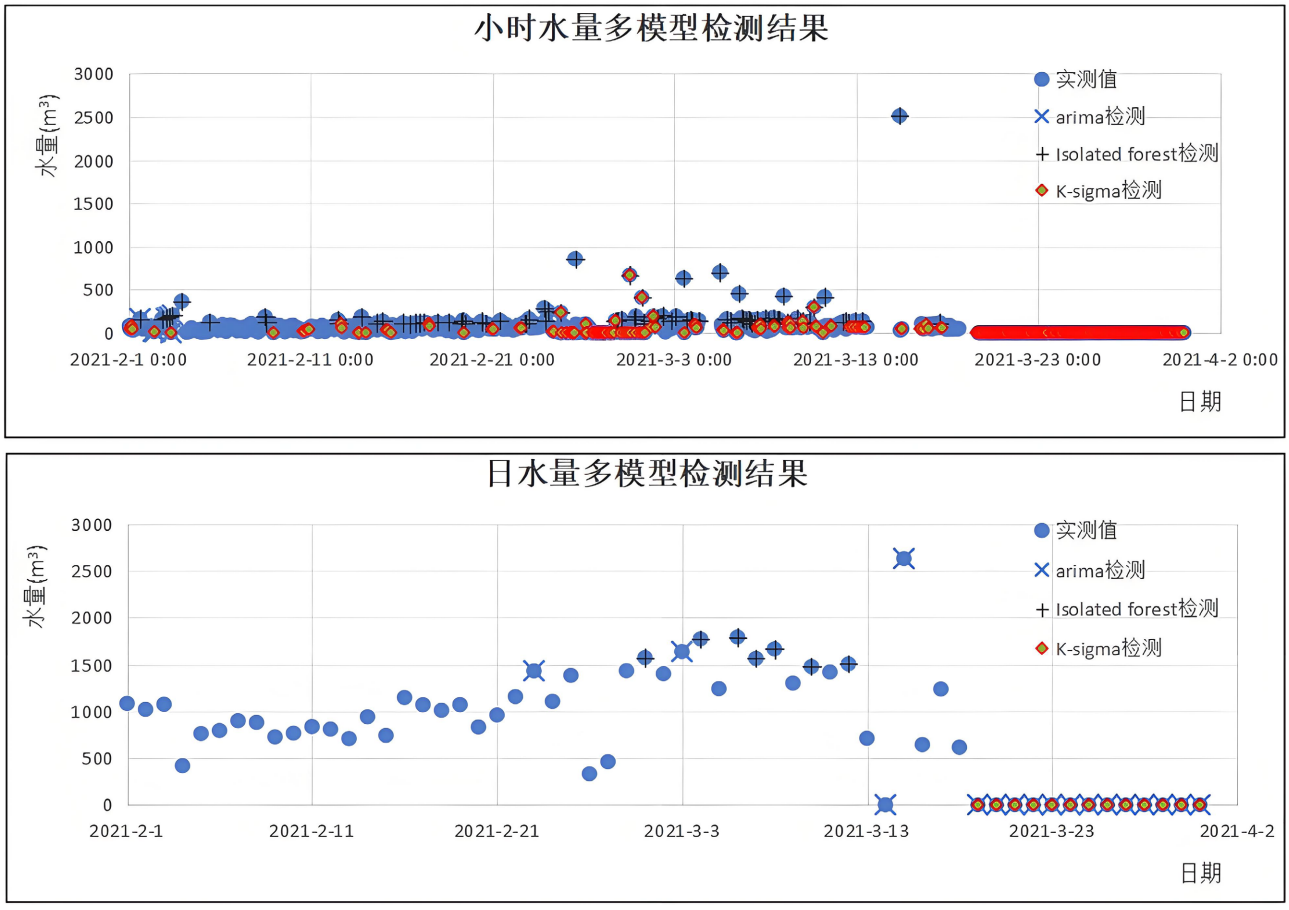


**图 2.** 各模型对实验数据检测结果

## 5. 主要结论

本研究利用基于统计(K-sigma)、分类(Isolated forest)、预测(ARIMA)等 3 种模型算法开展水资源计量数据异常检测，是大数据挖掘算法综合利用方式，通过多模式互验能直观反映检测结果，改进了判断便利性，相对单

模式核对，有效提高了异常检测准确性。主要结论有以下几点。

1) 基于统计检测法相对其他 2 种算法具有运行占用资源少、速度快，程序依赖关系少，开发实现便利等优势；基于预测的 ARIMA 模型算法运行的资源需求大、耗时长，对数据序列完整性要求高。

2) 基于统计检测法需预设 $k$ 值，检测结果受 $k$ 值影响大；采用机器学习的基于分类、预测的检测算法，具有较好的自适应性，ARIMA 模型算法还具有自动给出预测值的优势。

3) 采用多模式互验计算处理，具有同时多情景计算优势，对水资源计量对象复杂性具有较好的适用能力，为实际应用提供了多种组合方式选择。

## 基金项目

国家重点研发计划课题 2023YFC3006501；国家自然科学基金项目 42075191；南京水科院基本科研业务费项目 Y524008。

## 参考文献

[1] 熊明, 梅军亚, 杜耀东, 吴琼. 水资源监测数据的质量控制[J]. 人民长江, 2018, 49(9): 41-46.

[2] 刘怀利, 徐浩. 水资源取水计量数据精确采集方法探讨[J]. 水利信息化, 2014(5): 46-50.

[3] 刘予伟, 刘东润, 陈献耘. 大数据在水资源管理中的应用展望[J]. 水资源研究, 2015, 4(5): 470-476.

[4] RACITI, M., CUCURULL-JUAN, J. and NADJM-TEHRANI, S. Anomaly detection in water management systems. In Critical Infrastructure Protection. Berlin: Springer, 2012: 98-119. https://doi.org/10.1007/978-3-642-28920-0_6

[5] 赵和松, 王圆圆, 孙爱民. 一种基于 ARIMA 模型与 3$\sigma$ 准则的取水异常检测方法[J]. 水利信息化, 2022(1): 35-41.

[6] INOUE, J., YAMAGATA, Y., CHEN, Y., et al. Anomaly detection for a water treatment system using unsupervised machine learning. In 2017 IEEE international conference on data mining workshops (ICDMW). Washington DC: IEEE Computer Society, 2017: 1058-1065. https://doi.org/10.1109/icdmw.2017.149

[7] 巫朝星. 基于孤立森林模型的企业用水异常检测研究[J]. 企业科技与发展, 2019(11): 61-64.

[8] 赵东升, 姜蒙, 彭婷婷. 大数据时代计量管理质量提升路径研究[J]. 中文科技期刊数据库(全文版)工程技术, 2021(12): 3.

[9] 李达, 邢智慧. 水资源监测网络研究[J]. 水资源研究, 2009, 30(3): 9-10.

[10] 雷四华, 吴永祥, 王高旭. 国家水资源监控能力建设项目数据库设计[J]. 水资源研究, 2020, 9(4): 386-393.

[11] 于善奇. 计量抽样检验的标准差法[J]. 中国质量与标准导报, 2022(2): 76-77+80.

[12] 穆宝胜, 刘欣, 朱文艳. 基于 n 个标准差法和箱线图法识别变形监测中异常值的应用探究[J]. 南通职业大学学报, 2023, 37(2): 100-104.

[13] 魏泰, 贺少雄, 胡子武, 等. 基于改进孤立森林算法的风电机组异常数据清洗[J]. 科学技术与工程, 2024, 24(9): 3691-3699.

[14] 丁红翔. 基于降雨-蒸发量的靛坑河流量时间序列预测研究[J]. 水利科技与经济, 2023, 29(11): 112-114.

[15] 国家市场监督管理总局. JJF2210-2025, 取水计量数据质量控制技术规范[S]. 北京: 质检出版社, 2025.

[16] 雷四华, 吴永祥, 毛学文, 等. SZY301-2018, 国家水资源监控能力建设标准——基础数据库表结构与标识符[S]. 北京: 国家水资源监控能力建设项目办, 2018.

[17] 吴永祥, 雷四华, 毛学文, 等. SZY302-2018, 国家水资源监控能力建设标准——监测数据库表结构与标识符[S]. 北京: 国家水资源监控能力建设项目办, 2018.