

基于渐进式GAN增强的农村供水水质异常预警方法及应用

张亚晴¹, 瞿暄², 余轶鹏¹, 王婷², 陈兴宇², 刘丽²

(1.合肥市市政设计研究总院有限公司, 安徽 合肥 230041; 2.合肥工业大学 土木与水利工程学院, 安徽 合肥 230009)

摘要:【目的】构建农村供水水质小样本数据增强与异常预警模型, 提升突发性污染识别精度, 为农村供水安全监测提供技术路径。【方法】以安徽省两家农村供水企业 pH、浊度时序数据为研究对象, 提出三阶段渐进式 GAN 数据增强策略, 结合 CNN-LSTM-DA 模型实现水质预测与自适应异常预警, 对比 RNN 生成、均值插值法验证方法有效性。【结果】①三阶段渐进式 GAN 能够较好保持原始水质序列的趋势特征、波动结构和突变信息, 生成数据在分布一致性和时序保真度方面优于 RNN 生成方法和均值插值法。②GAN 增强数据有效缓解了农村供水水质监测样本不足的问题, 使模型对 pH 值和浊度峰值、周期波动及极端变化的刻画能力明显提升。③在不同水质波动场景下, GAN 增强均能改善模型预测稳定性, 其中对平稳序列的提升更为显著, 对高波动浊度序列仍具有一定适应性。④CNN-LSTM-DA 模型基于预测残差构建动态阈值, 可在连续监测过程中识别潜在水质异常, 降低固定阈值预警对人工经验的依赖。【结论】三阶段渐进式 GAN 数据增强与 CNN-LSTM-DA 融合方法可有效缓解小样本约束, 提升水质预警模型的精度与泛化能力, 可为监测数据稀缺地区的水质风险识别、预警阈值动态设定和农村供水安全管理提供参考。

关键词:深度学习; 农村供水水质; 数据增强; 异常预警; 生成对抗网络

中图分类号:TU991.31

文献标识码:A

文章编号:

农村饮用水安全直接关系居民健康与生态环境的可持续发展^[1]。世界卫生组织数据显示, 全球约 20 亿人口使用的饮用水源存在不同程度污染, 其中发展中国家农村地区由于监测设施不足与管理能力有限, 潜在水质异常波动呈上升趋势^[2]。突发性水质污染不仅威胁公共健康, 还可能对区域生态系统造成持续性影响。因此, 构建高效、可靠的水质异常预警方法已成为农村供水安全保障的关键技术问题^[3-4]。

现有水质预警方法大致经历了由经验模型向数据驱动模型的发展过程^[5]。早期方法多依赖固定阈值或线性模型, 如水质指数法和多元回归模型^[6], 在水质变化平稳条件下具有一定适用性。然而, 实际水质数据往往呈现非线性、非平稳及多尺度特征^[7], 突发污染事件常伴随分布突变或极端波动, 使得传统方法易出现漏报或误报。随后, 控制图及机器学习方法被引入水质异常检测

^[8-10], 在一定程度上提升了模型的灵活性与预测能力, 但仍依赖人工特征构建, 且在小样本条件下泛化性能受限^[11]。

近年来, 深度学习方法为水质时序建模提供了新的思路。长短期记忆网络 (Long Short-Term Memory, LSTM) 能够融合历史数据与未来趋势, 显著提高时间序列预测的稳定性和精度^[12]; 双向 LSTM 可同时捕获正向与反向时间依赖, 增强对复杂地下水水质变化规律的识别能力^[13]。卷积神经网络 (Convolutional Neural Network, CNN) 可自动提取局部特征, 提升对水质微小波动和异常模式的敏感性^[14-15]。Baek S S 等^[16]提出的 CNN-LSTM 融合模型, 通过 CNN 提取局部波动特征, 再利用 LSTM 模拟总氮、总磷等关键指标动态变化, 实现河流综合水质的全面监控。然而, 这类方法通常依赖大规模训练数据, 而农村供水场景普遍存在监测数据稀缺、异常样本不足的问题

收稿日期/Received: 2026-04-29

修回日期/Revised: 2026-06-22

基金项目:安徽省住房城乡建设科学技术计划项目 (2025-YF034); 国家自然科学基金项目 (51479046); 安徽省自然科学基金项目 (2505085ME116)

第一作者:张亚晴(1991—), 女, 高级工程师, 硕士, 主要从事给排水设计工作方面的研究。E-mail:1594159343@qq.com。

通信作者:刘丽(1972—), 女, 教授, 硕导, 博士, 从事城市管网安全理论与技术方面的研究。E-mail:lliu@hfut.edu.cn。

题，导致模型易过拟合且难以识别低频但高风险的突发污染事件^[17]。因此，如何在小样本条件下提升模型泛化能力与异常感知能力，成为当前研究的关键瓶颈。

针对上述问题，本研究提出一种融合生成对抗网络（Generative Adversarial Network, GAN）数据增强与 CNN-LSTM-DA 的水质异常预警方法。首先，设计面向小样本时序数据的三阶段渐进式 GAN 增强策略，通过对生成过程的分阶段约束，保证扩充数据在统计特征与时间动态上的一致性，从而缓解传统 GAN 在时序任务中的模式坍塌与趋势失真问题。其次，构建 CNN-LSTM-DA 预测预警模型，结合 CNN 的局部特征提取能力与 LSTM 的时序建模优势，并引入基于预测残差的差值分析（Difference analysis, DA）模块，实现异常阈值的自适应设定，降低人为经验依赖。在此基础上，基于两家农村供水企业的实际监测数据开展实验，对比不同数据增强策略对模型性能的影响，并通过连续监测验证方法的有效性。

1 模型与方法

1.1 生成对抗网络

在对 A、B 企业取水水质指标进行预测与预警建模时，由于原始数据量有限，难以充分反映水质复杂的变化规律，导致模型在训练过程中无

法学习到足够丰富的样本信息，从而限制了泛化能力与预测精度。为解决这一问题，本研究引入数据增强技术，通过生成与原始数据分布相似的新样本扩充数据集规模，提高模型的训练效率，使其能够更准确地捕捉水质指标的内在联系和变化趋势，为构建高性能预测与预警模型提供数据支撑。

本研究采用生成对抗网络（GAN）进行水质指标数据增强。GAN 由生成器（Generator）和判别器（Discriminator）组成，其优化目标函数为：

$$\min_G \max_D V(D, G) = E_{x \sim P(x)} [\log D(x)] + E_{z \sim P(z)} [\log (1 - D(G(z)))] \quad (1)$$

式中： $P(x)$ 为真实样本 x 的数据分布； $P(z)$ 为输入噪声变量 z 的先验分布； $G(z)$ 为生成器根据噪声 z 生成的样本； $D(x)$ 为判别器对真实样本 x 的输出概率； $D(G(z))$ 为判别器对生成样本 $G(z)$ 的输出概率。

GAN 核心在于生成器与判别器的对抗训练机制，其模型结构如图 1 所示。生成器通过随机噪声生成逼真样本以迷惑判别器，判别器则不断提高对真实与生成样本的区分能力，二者在训练过程中不断优化，直至达到平衡状态，此时生成器能够生成与真实数据分布高度一致的样本，从而实现数据增强目标。

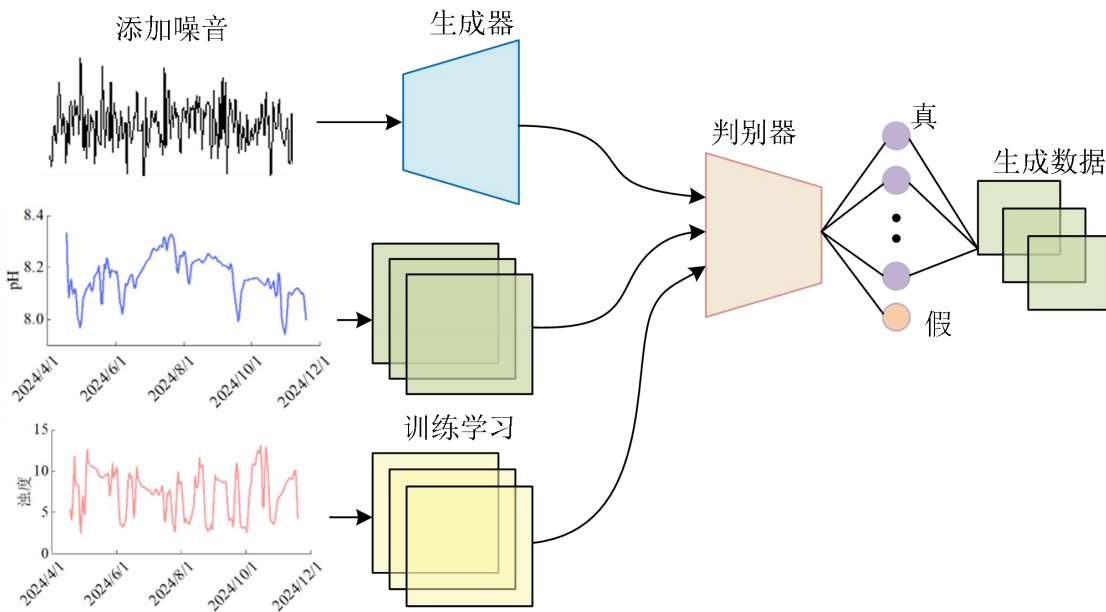


图 1 GAN 模型结构

Fig.1 Structure of the GAN model

1.2 LSTM 模块

长短期记忆网络（LSTM）是循环神经网络

（Recurrent Neural Network, RNN）的改进模型，旨在解决 RNN 在长序列训练中存在的梯度消失

或梯度爆炸问题^[18]。LSTM 通过遗忘门、输入门和输出门选择性地保留或更新信息，实现跨时间步长的信息传递，提高对短期和长期依赖的建模能力。其计算公式如下：

遗忘门：

$$f_t = \text{sigmoid}(w_f \cdot [h_{t-1}, X_t] + b_f) \quad (2)$$

输入门：

$$i_t = \text{sigmoid}(w_i \cdot [h_{t-1}, X_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(w_c \cdot [h_{t-1}, X_t] + b_c) \quad (4)$$

输出门与细胞状态更新：

$$o_t = \text{sigmoid}(w_o \cdot [h_{t-1}, X_t] + b_o) \quad (5)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (6)$$

模型输出：

$$h_t = o_t \cdot \tanh(C_t) \quad (7)$$

式中： t 为当前时间步； X_t 表示当前输入； h_{t-1} 为上一时间步长输出； C_t 与 C_{t-1} 分别为当前及上一时间步的细胞状态； $\tilde{C}_t$ 为 t 时刻候选细胞状态； f_t 、 i_t 和 o_t 分别为遗忘门、输入门和输出门； w_f 、 w_i 、 w_c 、 w_o 为相应权重矩阵； b_f 、 b_i 、 b_c 、 b_o 为相应偏置； h_t 为当前隐藏状态，即模型单元输出。

1.3 CNN 模块

卷积神经网络（CNN）能够自动提取局部特征，本研究通过多层卷积层与池化层提取水质指标的空间特征^[19]。卷积层通过不同大小的卷积核对输入数据进行滑动卷积，池化层采用最大或平均池化进行下采样，降低数据维度的同时保留关键特征，有效缓解过拟合^[20]。多轮卷积与池化操作可充分挖掘水质数据的复杂特征，为 LSTM 提

供丰富输入。

1.4 差值分析模块

差值分析（Difference analysis, DA）模块用于捕捉预测值与实际值之间的偏差，从而反映水质数据的异常波动。其计算公式为：

$$\Delta x_i = p_i - x_i, \quad i=1, 2, \dots, N-1 \quad (8)$$

式中： Δx_i 为第 i 个时间点预测值与实际值的差值； p_i 为第 i 个时间点的预测值； x_i 表示第 i 个时间点的实际值。

基于预测残差的动态阈值采用滑动窗口置信区间法计算，公式如下：

$$\Delta x_{\text{threshold}} = \mu_{\Delta x} \pm z_{\alpha/2} \cdot \sigma_{\Delta x}, \quad i=1, 2, \dots, N-1 \quad (9)$$

式中： $\mu_{\Delta x}$ 为滑动窗口内预测残差的均值； $z_{\alpha/2}$ 为标准正态分布的分位数； $\sigma_{\Delta x}$ 为滑动窗口内预测残差的标准差； α 为置信水平，取 0.05。

1.5 CNN-LSTM-DA 模型

本研究构建的 CNN-LSTM-DA 模型结合了局部特征提取、时间序列预测与异常检测三大功能。首先，CNN 提取水质指标的局部波动与周期性特征；随后，LSTM 对每个时间步进行预测，强化数据的时序依赖和峰值特征；最后，DA 模块计算预测差值，并设定动态阈值捕捉异常波动，实现高精度水质预警模型，结构图如图 2。该模型结构设计旨在充分利用小样本时序数据，提升预测精度与异常检测能力，为农村供水水质管理提供技术支撑。

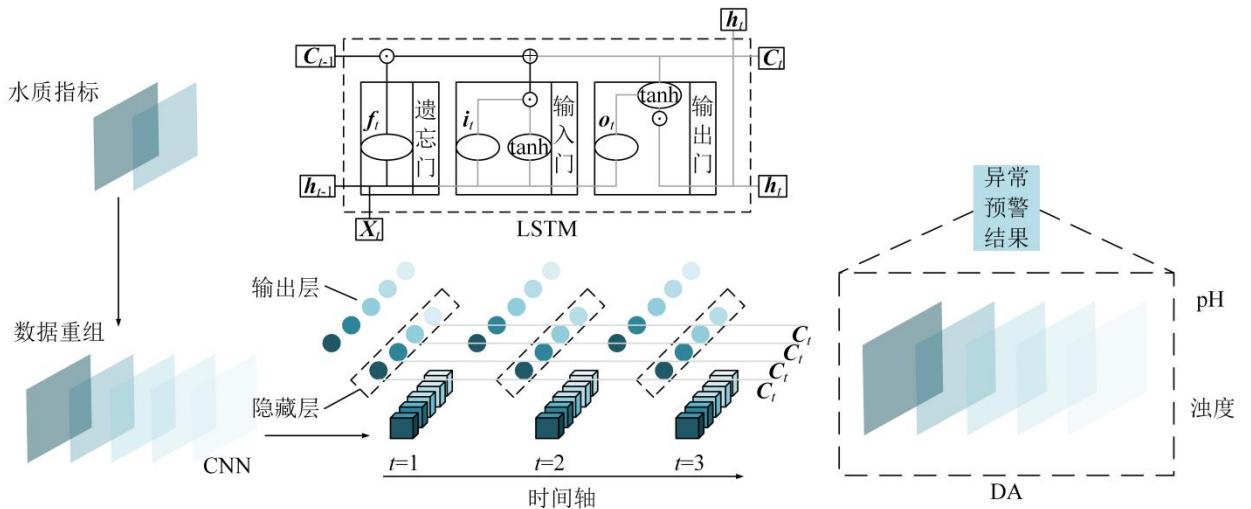


图 2 CNN-LSTM-DA 模型结构示意图

Fig.2 Schematic diagram of the structure of the CNN-LSTM-DA model

1.6 评价指标

为全面评估数据增强效果及模型预测与预警

性能，本研究选取以下指标：均方误差（Mean Squared Error, MSE）、均方根误差（Root Mean

Squared Error, RMSE)、平均绝对误差 (Mean Absolute Error, MAE) 以及平均绝对百分比误差 (Mean Absolute Percentage Error, MAPE)。这些指标在水质预测领域应用广泛, 能够有效量化模型拟合精度和预测能力。其计算公式如下:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{\text{pred}} - y_{\text{test}})^2 \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{\text{pred}} - y_{\text{test}})^2} \quad (11)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{\text{pred}} - y_{\text{test}}| \quad (12)$$

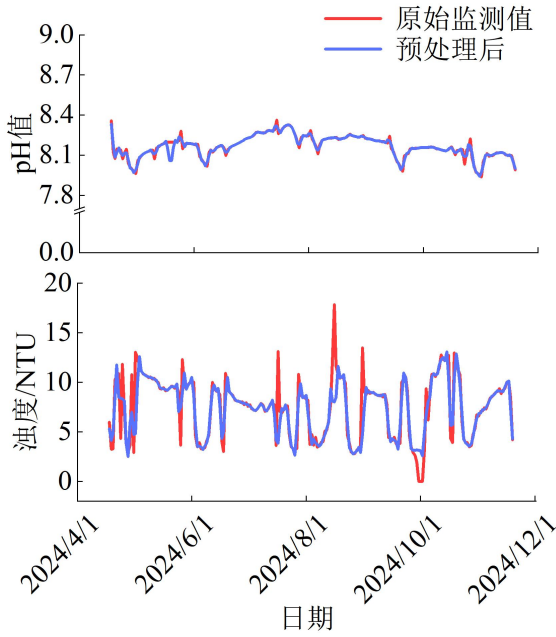
$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_{\text{pred}} - y_{\text{test}}}{y_{\text{test}}} \right| \times 100\% \quad (13)$$

式中: y_{pred} 为模型预测值; y_{test} 实测值; n 为样本总数。MSE、RMSE、MAE 与 MAPE 值越低, 表明模型预测精度越高, 对水质变化的刻画能力越强。

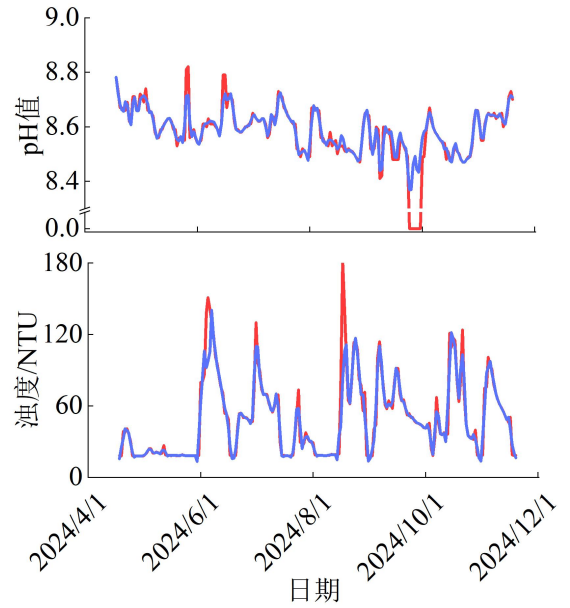
2 实例应用

本研究以安徽省两家供水企业(以下简称 A、B 企业)为对象, 分析其取水水质指标的时间序列特征。数据采集时间覆盖 2024 年 4 月 18 日至 11 月 19 日, 共计 216 天, 每天记录一次。每家企业的监测指标包括 pH 值和浊度, 共获得 216 条有效观测数据。

原始数据经检查无缺失值, 部分离群点采用线性插值法进行去噪处理, 预处理后 A 企业的 pH 均值为 8.148, 浊度均值为 6.817; B 企业的 pH 均值为 8.568, 浊度均值为 52.138。对比分析显示, A 企业的 pH 值整体较为稳定且略低于 B 企业, 同时浊度显著低于 B 企业, 表明两企业水质在总体水平和波动幅度上存在明显差异。两家企业的 pH 与浊度指标的分布特征及变化趋势, 为后续预测模型的设计、数据增强策略选择及模型性能评估提供了重要参考依据。原始监测值及预处理结果见图 3。



(a) A 企业



(b) B 企业

图 3 A、B 企业取水水质监测值及预处理结果

Fig.3 Water quality monitoring values and preprocessing results for raw water intake at enterprises A and B

3 结果分析

3.1 GAN 数据增强效果评估

为确保增强数据在统计分布与时间序列结构上与原始数据保持一致, 本研究提出一种三阶段渐进式数据增强策略。各阶段设置如下: 阶段一以原始数据训练基础 GAN, 训练轮次 300 轮, 生

成器与判别器训练比例 1:1, 目标是学习水质数据的整体趋势与基本统计特征; 阶段二以阶段一生成的数据集微调模型, 训练轮次 300 轮, 训练比例调整为 3:1, 目标是捕捉局部波动与周期性特征; 最后在阶段二的基础上最终优化, 训练轮次 400 轮, 训练比例调整为 5:1, 加入峰值特征损失, 目标是强化极端值与突变点生成能力。该

策略基于 GAN 生成器与判别器的对抗学习机制，在逐步扩展样本规模的同时，约束生成数据在趋势、波动与周期性特征上的一致性，并与 RNN

生成方法及均值插值法进行对比分析。预处理后全部 216 条原始数据经三阶段增强，水质指标样本量从 216 条扩展至 1728 条，见图 4。

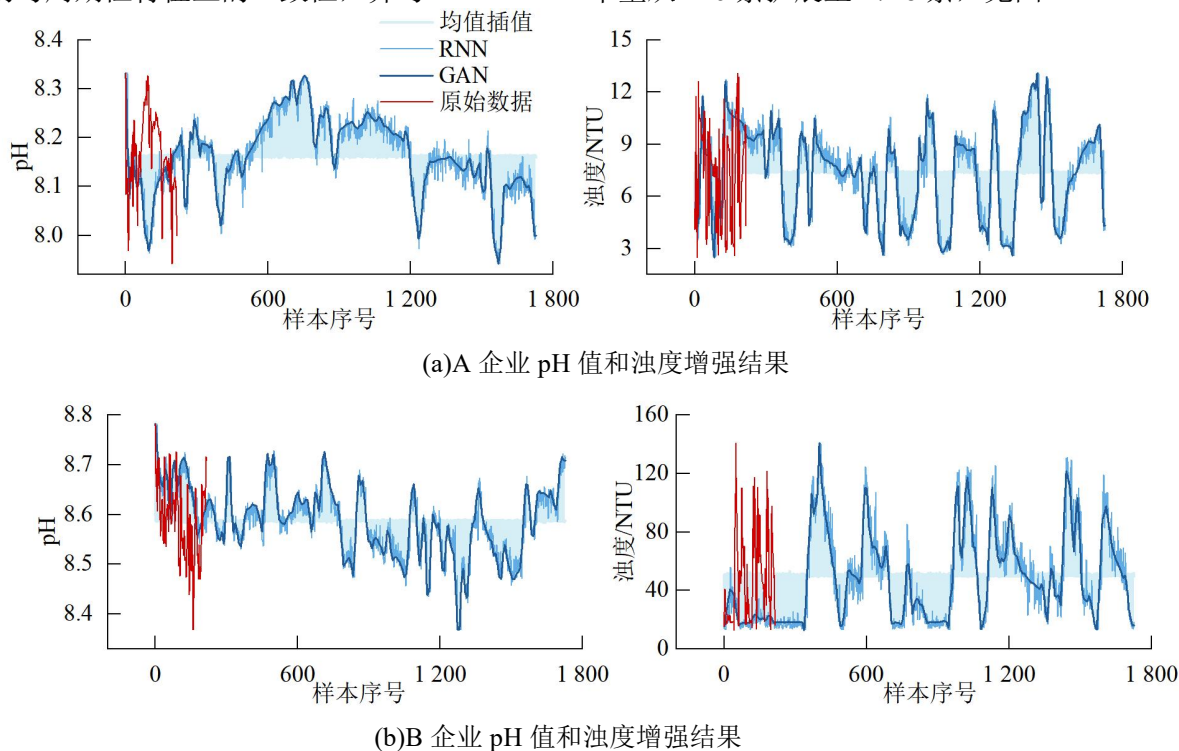


图 4 A、B 企业不同方法的数据增强结果

Fig.4 Data augmentation results obtained using different methods for Enterprises A and B

A 企业结果表明，均值插值法由于仅依赖统计均值，破坏原始序列趋势（如 pH 上升过程被削弱），难以反映真实动态；RNN 虽可保持整体趋势，但局部波动明显；相比之下，GAN 生成数据不仅保持趋势连续性，还能有效刻画非线性波动结构，在细节保真度与多样性方面表现最佳。B 企业结果进一步验证该结论，尤其在浊度指标上，GAN 能够同时保留周期性突变特征。

为进一步说明阶段一生成数据的可靠性，本研究补充分析了基础 GAN 生成数据与原始监测数据之间的量化误差。结果显示，阶段一 GAN 在 A 企业 pH 指标上的 MAE 和 MSE 分别为 0.008 92 和 0.000 18，在浊度指标上的 MAE 和 MSE 分别为 0.277 72 和 0.128 52；在 B 企业 pH 指标上的 MAE 和 MSE 分别为 0.010 75 和 0.000 181，在浊度指标上的 MAE 和 MSE 分别为 4.058 81 和 29.671。进一步与 RNN 模型对比可知，阶段一 GAN 在 A 企业 pH、A 企业浊度、B 企业 pH 和 B 企业浊度上的 MAE 分别降低 15.3%、42.0%、14.5%和 1.9%，MSE 分别降低 10.0%、76.3%、38.6%和 24.0%。上述结果表明，阶段一基础 GAN 已能够较好地拟合原始水质数据的整体变化趋势和基本统计特征，生成数据与真实数据之间保持

了较高的一致性。因此，将阶段一生成结果作为后续第二阶段和第三阶段继续训练与微调的数据基础具有一定可靠性。

数据增强的有效性不仅取决于样本量，更关键的是生成数据与原始数据特征的契合度。为量化增强效果，本研究计算了各阶段生成数据的误差指标。结果显示，GAN 在所有指标上显著优于 RNN 和均值插值法，并且其优势随增强阶段逐步增强，结果如图 5 所示。以 A 企业为例，第一阶段 pH 指标 MAE 为 0.008 92，比 RNN 降低 15.3%，MSE 为 0.000 18，降低 10%；浊度的 MAE 为 0.277 72，比 RNN 下降 42%，MSE 降幅高达 76.3%（从 0.541 47 降至 0.128 52）。这表明 GAN 的对抗训练能够有效捕捉浊度数据中的非线性突变，使生成数据高度贴近真实分布。相比之下，均值插值法误差显著偏高（pH 的 MAE 为 0.062 61，浊度的 MAE 为 2.223 93），凸显线性方法在非线性时序数据增强中的局限性。在 B 企业中，尽管原始数据波动性更大导致误差绝对值高于 A 企业（如浊度的 MAE 达到 4.058 81），GAN 仍保持明显优势。pH 的 MSE 较 RNN 降低 38.6%，浊度的 MSE 降低 24%，在抑制极端异常值和提升整体稳定性方面效果突出。综合三阶段增强结果，

GAN 在两家企业的关键指标（MAE、MSE）上相较 RNN 平均降低约 50%，相较均值插值法降低超过 90%，充分验证了其在复杂水质时序数据

增强中的优越性和适用性。均值插值法则仅适合对数据真实性要求较低的场景。

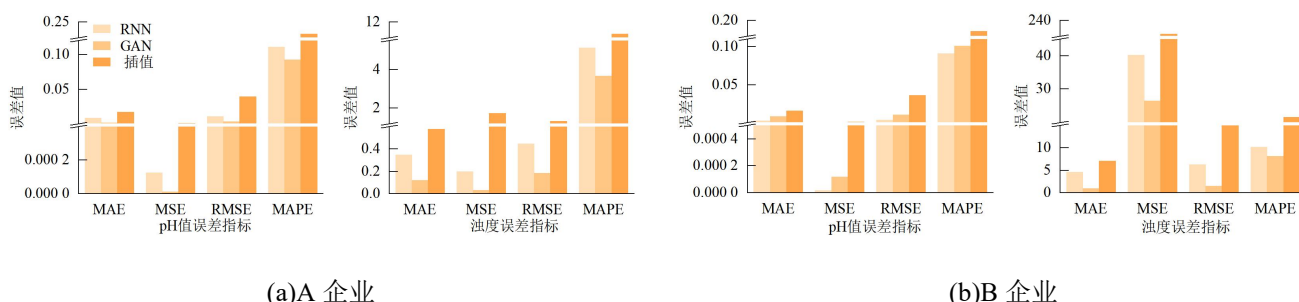


图 5 不同增强模型输出与原始监测数据的误差指标对比

Fig.5 Comparison of error indicators between outputs of different enhancement models and original monitoring data

3.2 预测模型性能分析

本研究采用 CNN-LSTM 模型对原始数据及经 RNN 和 GAN 增强后的数据进行预测，通过比较不同数据源的预测性能，可进一步评估数据增强效果。在模型训练中，迭代次数设置为 1 000 次，按照 7：2：1 划分训练集、验证集和测试集，

训练集均方误差（MSE）随迭代变化趋势如图 6 所示。结果表明，三类数据均能稳定收敛，但 GAN 增强数据表现最优。其初始误差最低、最终误差最小且收敛速度最快。这归因于 GAN 生成数据在扩充样本规模的同时保持高质量分布，使模型更易学习并快速达到收敛阈值。

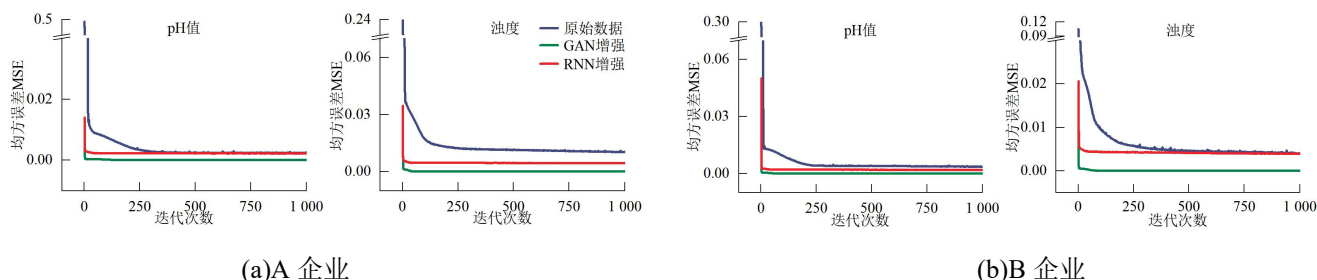


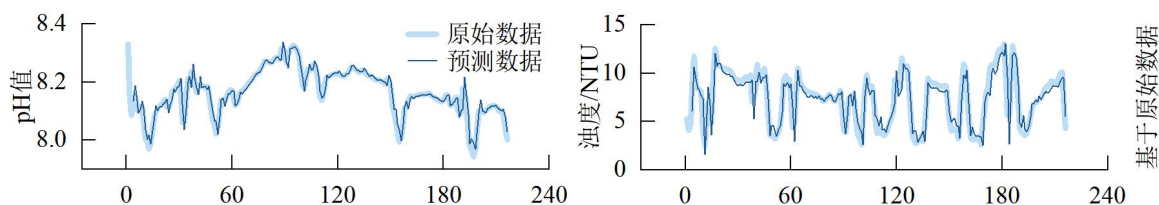
图 6 训练集在不同增强方法下的预测误差对比

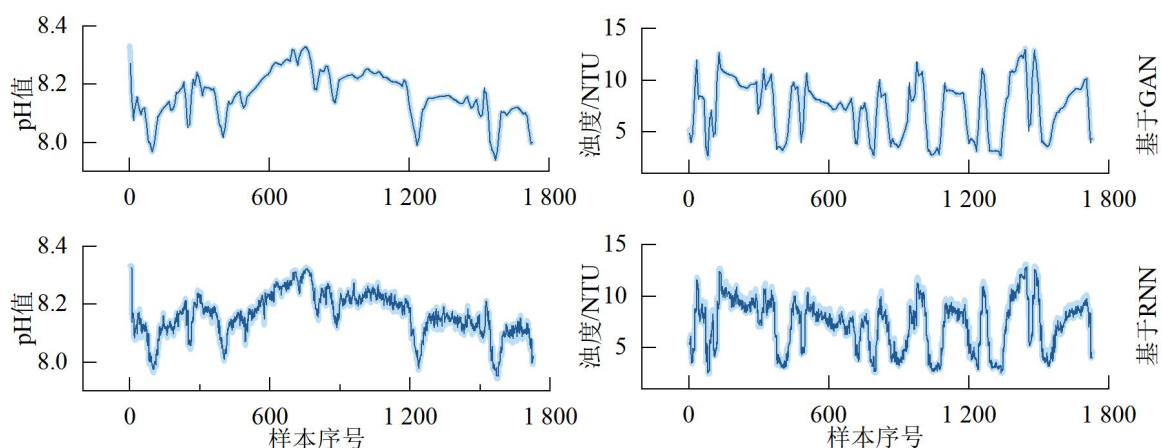
Fig.6 Training set prediction error comparison across different enhancement methods

农村供水系统的稳定运行依赖水质指标的准确预测。为评估增强方法对模型连续预测能力的影响，分别采用原始数据、RNN 增强数据及 GAN 增强数据进行连续预测（图 7）。对于 A 企业，原始数据训练的模型在峰值处预测效果不佳，尤其是浊度指标偏差明显。这主要源于原始样本量有限（216 条），模型在小样本条件下易过度拟合局部噪声而非全局趋势，导致对周期波峰及极端值表征不足。GAN 增强数据显著改善了峰值预测精度，同时保留了原始序列的动态依赖结构，周期性变化捕捉更合理。相比之下，RNN 增强数据虽有所改善，但局部震荡仍存在，限制了峰值

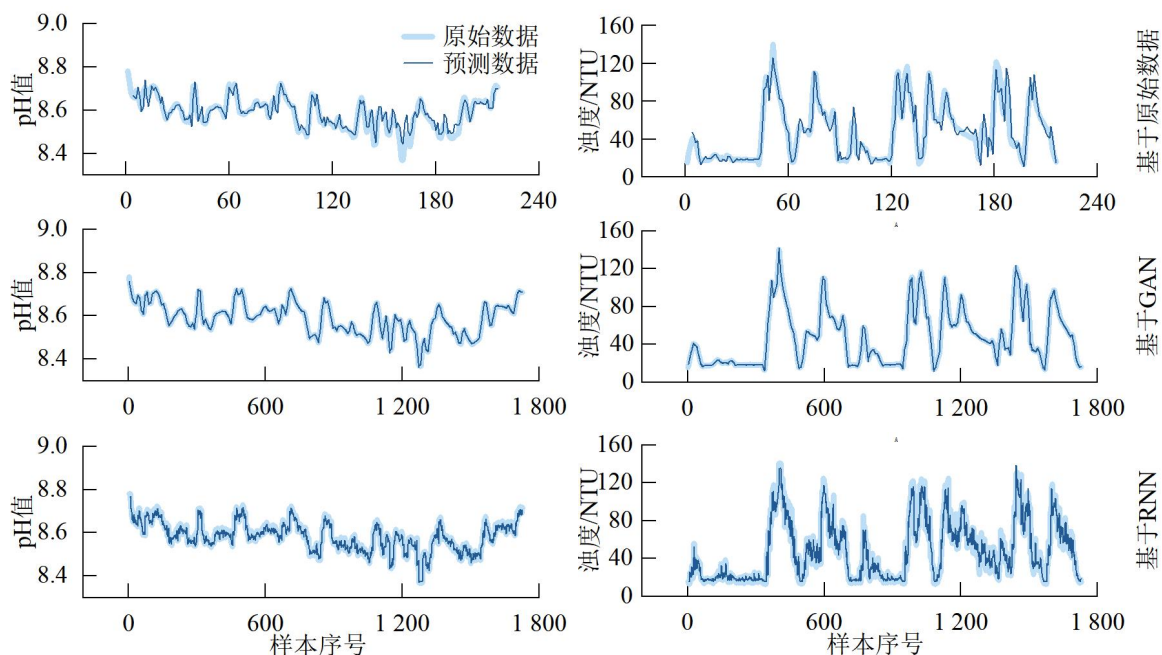
拟合质量与整体时序一致性。B 企业结果呈现类似趋势，但差异更为显著。原始数据在 pH 和浊度峰值处预测性能不足，反映数据量对复杂水质变化建模能力的限制。RNN 增强在 B 企业表现进一步下降，尤其在浊度指标上。分析认为，RMSE 序列波动剧烈、脉冲性强（方差约为 A 企业 140 倍），RNN 自回归生成模型在训练时可能将高幅噪声“记忆”下来，导致新样本中混入非物理极端

波动，从而降低连续预测准确性。相比之下，GAN 增强数据仍保持较高预测精度，进一步验证其对高复杂度时序的适应性。





(a)A 企业不同数据增强方法下 pH 和浊度预测结果



(b)B 企业不同数据增强方法下 pH 和浊度预测结果

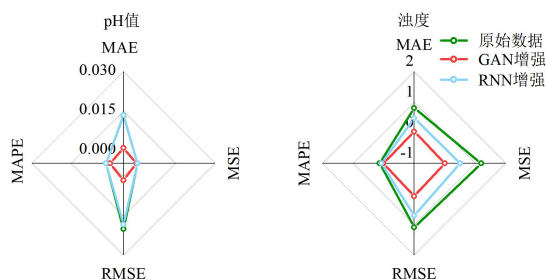
图 7 基于不同增强方法的水质指标连续预测结果对比

Fig.7 Comparison of continuous prediction results for water quality indicators based on different enhancement methods

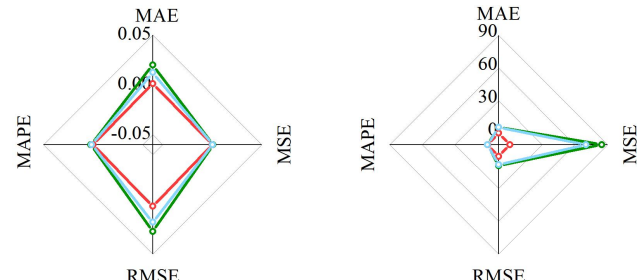
综上，GAN 数据增强不仅扩充样本量，更保留了对预测关键的非线性依赖和尖峰信息，显著缓解了原始数据不足对模型性能的制约。需要指出的是，这种优势受数据复杂度影响，在低复杂度序列（A 企业）中提升明显，而在高度混沌序列（B 企业浊度）中，MAE 降幅从 91%收窄至约

50%，说明该方法在复杂高波动场景下仍需结合数据特征进一步优化。

为进一步评估数据增强对模型泛化能力的影响，将训练好的 CNN-LSTM 模型分别应用于 A、B 企业的独立测试集（未参与任何增强或训练），比较不同增强方法下的预测误差（图 8）。



(a)A 企业



(b)B 企业

图8 测试集水质指标连续预测误差在不同增强方法下的对比

Fig.8 Test set water quality indicator continuous prediction error comparison across different enhancement methods

在A企业pH预测中, GAN增强效果最佳, $MAE=9.30 \times 10^{-4}$, $MSE=2.06 \times 10^{-6}$ 。相比原始数据 ($MAE=1.335 \times 10^{-2}$, $MSE=3.33 \times 10^{-4}$), MAE和MSE分别降低93.0%和99.4%, MAPE由16.38%降至1.14%。这一结果表明, GAN生成数据不仅高度保留了原始分布, 其对抗训练过程还起到隐式正则化作用, 有效抑制了孤立噪声。相比之下, RNN增强数据的pH预测结果 ($MAE=0.013\ 23$, $MSE=4.02 \times 10^{-4}$) 与原始数据基本持平甚至略差, 原因可能在于: 标准RNN/LSTM作为生成模型存在梯度消失或爆炸问题, 且自回归生成方式易累积误差。浊度预测中, 差异更为显著。原始数据预测 $MAE=0.476$, GAN增强后降至 $0.041\ 79$ (降幅91.2%), MAPE从7.14%降至0.646%, 表明GAN能精准捕捉浊度的非线性波动模式(包括突发上升或下降)。RNN增强数据反而使MAE升至 $0.806\ 98$ (较原始增加69.5%), MSE明显恶化。这说明, 即便生成数据的单点统计均值接近原始数据, 低质量生成数据仍可能误导模型学习虚假的时序依赖, 导致测试集泛化误差增大。此外, 原始浊度数据预测的MSE为 $0.502\ 74$, GAN增强后进一步降至 $0.005\ 38$, 凸显了增强数据对模型鲁棒性的提升作用。

在B企业pH预测中, GAN增强后的MAE和MAPE分别为 $0.001\ 27$ 和 $0.014\ 8\%$, 较原始数据 ($MAE=0.020\ 06$, $MAPE=0.234\%$) 下降约93.7%。浊度预测中, 原始数据 $MAE=6.056\ 59$, $MSE=84.65$; GAN增强后MAE降至 $0.540\ 19$ (降幅91.1%), MSE降至 $0.540\ 01$ (降幅99.4%)。需注意, MSE的降幅巨大与原始数据方差极大及平方量纲有关。更合理的解读是, GAN增强使测试集RMSE从 0.74 降至 0.73 (降幅约92.0%), 该指标更直观反映预测精度改善。RNN增强在B企业表现出明显的场景依赖性。对平稳的pH指标, RNN增强后的MAE ($0.012\ 49$) 略优于原始数据 ($0.020\ 06$), 表明短期记忆能力在平稳序列中可发挥正向作用; 但对高波动性浊度指标, RNN增强后的RMSE虽由原始数据的 $9.200\ 61$ 降至 $8.385\ 18$, 改善幅度相对有限, 且MAPE由

12.881%升至13.744%, 表明其对复杂非线性序列的建模能力仍存在一定局限。

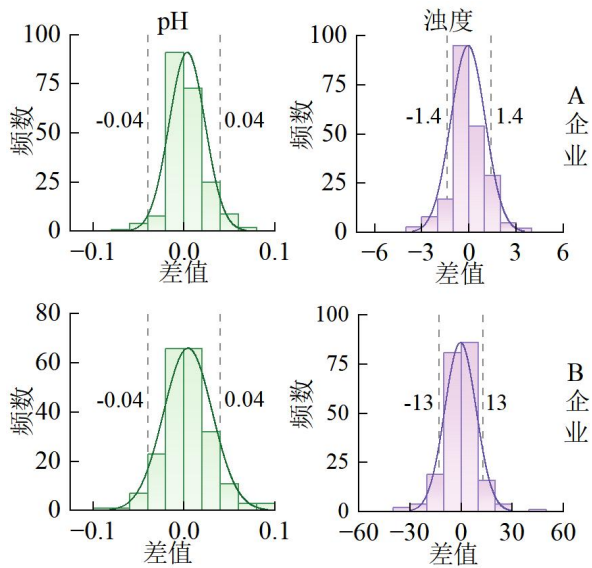
综上, GAN增强在两家企业及不同水质指标的实验中均展现出更优性能, 尤其在高波动性场景(如B企业浊度)下, 其对模型泛化能力的提升具有不可替代性。RNN增强效果高度依赖数据特性, 在平稳序列中可部分替代原始数据, 但对复杂非线性特征的建模能力明显不足。

3.3 水质异常预警

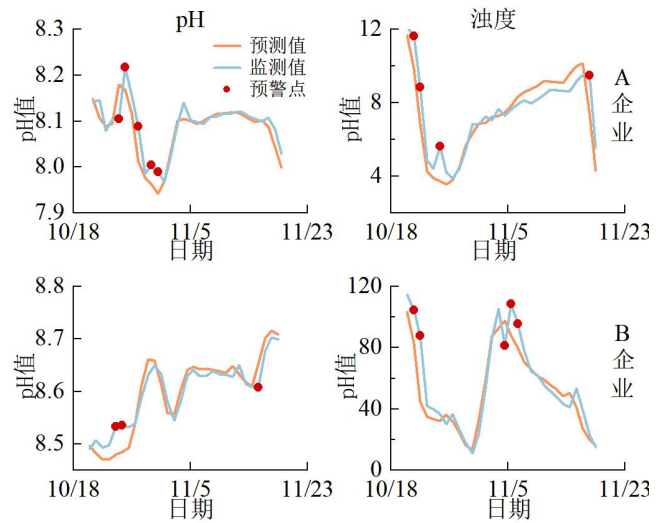
前期研究表明, 基于GAN的数据增强能显著提升水质预测模型性能, 其核心在于GAN可依据原始数据分布生成大量逼真且多样化的样本, 有效扩充数据集规模与丰富度。本研究在此基础上进一步创新, 将GAN增强后的水质数据与CNN-LSTM-DA模型结合, 通过DA模块对预测残差进行拟合, 并自适应设定阈值, 实现水质异常波动的动态预警。

针对预测值与实际监测值的差异进行系统分析, 残差分布与预警识别结果如图9所示。以A企业为例, pH指标预测差值 ≤ 0.01 的时间点占总体的83.1%, 浊度预测差值 ≤ 0.5 的时间点占83.7%, 表明模型在绝大多数时间内具有高精度。B企业呈现类似趋势, pH差值 ≤ 0.03 的时间点占92.0%, 浊度预测差值 ≤ 5 的时间点占90.2%, 同样表现出优异的预测性能。从差值分布看, pH与浊度的预测误差均呈现近似正态分布, 集中于零附近, 波动规律可控, 说明模型长期预测稳定可靠。基于95%置信区间, 分别设定A、B企业异常预警阈值: A企业pH为 ± 0.04 、浊度为 ± 1.4 ; B企业pH为 ± 0.04 、浊度为 ± 13 。当预测值与实测值差值超出阈值时, 判定该时刻水质指标存在异常波动。对2024年10月21日—11月19日共30天监测数据开展预警分析。A企业pH触发5次预警, 集中于10月25日—11月16日, 偏差较小, 与水源地季节性波动相关; 浊度触发4次预警, 实测值显著高于预测值, 提示该时段存在潜在异常。B企业pH相对稳定, 仅触发3次预警; 浊度异常突出, 10月22日、11月6日偏差分别达26.5%、23.2%, 提示水质可能存在异常。

综上, GAN增强数据结合CNN-LSTM-DA模型在两家企业水质指标的连续预测与异常预警中均表现出高精度、强鲁棒性与良好可靠性, 为小样本条件下水质监控与实时预警提供可行高效的技术方案。



(a)A、B 企业水质监测数据与预测值的差值分析结果



(b)A、B 企业水质异常预警分析结果

图9 A、B 企业水质预测残差分布及预警分析结果

Fig.9 Residual distribution and early-warning analysis results for enterprises A and B

4 结论

本研究针对农村供水水质监测样本不足及固定阈值适应性较差的问题,提出 GAN 数据增强与 CNN-LSTM-DA 融合方法,主要结论如下:

1) GAN 生成的数据在统计特征、趋势和周期性上高度接近原始水质数据,各阶段增强结果表明,GAN 的 MAE 和 MSE 总体低于 RNN 和均值插值法,验证了其在小样本水质时序增强中的优势。

2) 基于 GAN 增强数据的 CNN-LSTM-DA 模型在 A 企业 pH 预测 MAE 下降约 93%,浊度 MAPE 下降约 91%;B 企业浊度预测 RMSE 下降约 92.0%,显示对峰值及周期变化的捕捉能力明显增强。

3) 通过基于预测残差的差值分析,并按 95% 置信区间设定阈值,模型在 30 天监测期内识别多次水质潜在异常波动,为农村供水系统早期预警提供可行方案。

4) 当前研究基于有限指标和短期数据,尚缺乏跨年度和跨区域验证。未来可扩展更多水质指标及环境因素,引入迁移学习提升泛化能力,并结合可解释性人工智能技术增强预警结果透明度。

参考文献

[1] Bopp C,Nicolas-Artero C,Blanco E,et al.Infrastructure conditions and service quality in rural drinking water systems: a cluster analysis of community-based organizations in Chile[J].Journal of Water Resources

Planning and Management,2024,150(6):11.

- [2] Kayembe-Mulumba B,N'gattia A K,Belizaire M R D.One health,many gaps:rethinking epidemic intelligence in resource-limited settings to prepare for the global threat of disease X[J].Microorganisms,2025,13(11):2615.
- [3] Nong Xizhi,Zeng Jun,Chen Lihua,et al.A novel water quality risk assessment framework for reservoir water bodies coupling key parameter selection and dynamic warning threshold determination[J].Scientific Reports,2025,15:14377.
- [4] 黄羽,袁文博,郭丹阳,等.水资源配置工程突发性水污染事故预警体系构建研究[J].环境污染与防治,2022,44(8):1115-1120.[Huang Yu,Yuan Wenbo,Guo Danyang,et al.Study on construction of early warning system for sudden water pollution accidents in water resources allocation projects[J].Environmental Pollution & Control,2022,44(8):1115-1120.]
- [5] Zhang Chi,Nong Xizhi,Behzadian K,et al.A new framework for water quality forecasting coupling causal inference,time-frequency analysis and uncertainty quantification[J].Journal of Environmental Management,2024,350:119613.
- [6] Rajae T,Boroumand A.Forecasting of chlorophyll-a concentrations in South San Francisco Bay using five different models[J].Applied Ocean Research,2015,53: 208-217.
- [7] Wang Xianhe,Qiao Mu,Li Ying,et al.Deep-learning-based water quality monitoring and early warning methods:a case study of ammonia nitrogen prediction in

- rivers[J].Electronics,2023,12(22):4645.
- [8] Wang Lu,Wang Yueru,Li Xin.An evaluation method of marine water quality based on cloud-model chart representation theory and its role in tourism of coastal cities[M]//Information Technology and Agricultural Engineering.Berlin:Springer,2012:337-344.
- [9] Essamlali I,Nhaila H,El Khaili M.Advances in machine learning and IoT for water quality monitoring:a comprehensive review[J].Heliyon,2024,10(4):e27920.
- [10] Ahmadianfar I,Shirvani-Hosseini S,Samadi-Koucheksaraee A,et al.Surface water sodium (Na⁺) concentration prediction using hybrid weighted exponential regression model with gradient-based optimization[J].Environmental Science and Pollution Research,2022,29:53456-53481.
- [11] 冯婧涵,钱龙霞,唐加山,等.数据稀缺条件下水文频率分布模拟及应用[J].北京师范大学学报(自然科学版),2023,59(1):72-84.[Feng Jinghan,Qian Longxia,Tang Jiashan,et al.Simulation and application of hydrological frequency distribution under data scarcity[J].Journal of Beijing Normal University (Natural Science Edition),2023,59(1):72-84.]
- [12] Wang Yuanyuan,Zhou Jian,Chen Kejia,et al.Water quality prediction method based on LSTM neural network[C]//2017 12th International Conference on Intelligent Systems and Knowledge Engineering. Nanjing:IEEE,2017:1-5.
- [13] Alfwzan W F,Selim M M,Althobaiti S,et al.Application of Bi-LSTM method for groundwater quality assessment through water quality indices[J].Journal of Water Process Engineering,2023,53:103889.
- [14] Zhu Guocheng,Lin Jialin,Fang Haiquan,et al.A flocculation tensor to monitor water quality using a deep learning model[J].Environmental Chemistry Letters,2022,20:3405-3414.
- [15] He Fei,Lai Qiuying,Ma Jie,et al.Design and application of an early warning and emergency response system in the transboundary area of the Taihu Lake Basin[J].International Journal of Environmental Research and Public Health,2023,20(2):1340.
- [16] Baek S S,Pyo J,Chun J.Prediction of water level and water quality using a CNN-LSTM combined deep learning approach[J].Water,2020,12:3399.
- [17] 唐文忠,黄鑫,张洪,等.饮用水安全保障技术挑战与对策建议[J].科技导报,2025,43(11):34-44.[Tang Wenzhong,Huang Xin,Zhang Hong,et al.Technical challenges and countermeasure suggestions for drinking water safety guarantee[J].Science & Technology Review,2025,43(11):34-44.]
- [18] Schmidhuber J,Hochreiter S.Long short-term memory[J].Neural Computation,1997,9(8):1735-1780.
- [19] 杨璐霞,王梦冉,林兴亮,等.基于 CNN-LSTM 的工业出水水质预测模型[J].净水技术,2024,43(5):143-149.[Yang Luxia,Wang Mengran,Lin Xingliang,et al.Prediction model of industrial effluent water quality based on CNN-LSTM[J].Water Purification Technology,2024,43(5):143-149.]
- [20] Wu Xijuan,Zhang Qiang,Wen Fei,et al.A water quality prediction model based on multi-task deep learning:a case study of the Yellow River,China[J].Water,2022,14(21):3408.

Rural Water Supply Water Quality Anomaly Early Warning Method Based on Progressive GAN Enhancement and Its Application

Zhang Yaqing¹, Qu Xuan², Yu Yipeng¹, Wang Ting², Chen Xingyu², Liu Li²

(1. Hefei Municipal Design and Research General Institute Company Limited, Hefei 230041, China; 2. School of Civil and Hydraulic Engineering, Hefei University of Technology, Hefei 230009, China)

Abstract: 【Objective】 To construct a small-sample data augmentation and anomaly early warning model for rural water supply water quality, improve the identification accuracy of sudden pollution, and provide a technical approach for the safety monitoring of rural water supply. **【Methods】** Taking the time-series data of pH and turbidity from two rural water supply enterprises in Anhui Province as the research objects, a three-stage progressive GAN data augmentation strategy was proposed. Combined with the CNN-LSTM-DA model, water quality prediction and adaptive anomaly early warning were realized. The effectiveness of the proposed method was verified by comparing with RNN generation and mean

interpolation methods. **【 Results 】** ① The three-stage progressive GAN effectively preserved the trends, fluctuation patterns, and abrupt-change information of the original water quality series. The generated data showed better distributional consistency and temporal fidelity than those produced by RNN-based generation and mean interpolation. ② GAN-augmented data alleviated the shortage of water quality monitoring samples and improved the model's ability to capture peaks, periodic fluctuations, and extreme variations in pH and turbidity series. ③ GAN augmentation improved prediction stability under different fluctuation conditions, with greater improvement for relatively stable series while maintaining adaptability to highly variable turbidity series. ④ The CNN-LSTM-DA model established warning thresholds based on prediction residuals, enabling the identification of potential water quality anomalies during continuous monitoring and reducing reliance on empirically defined fixed thresholds. **【 Conclusions 】** The proposed method can alleviate small-sample constraints and improve the prediction accuracy and generalization performance of the water quality warning model. It provides a reference for water quality risk identification, residual-based warning threshold determination, and rural water supply safety management in regions with limited monitoring data.

Keywords: Deep learning; Rural water supply quality; Data augmentation; Anomaly early warning; Generative adversarial network