杨靖仁, 王超, 雷晓辉, 等. 南水北调东线江苏段典型泵站运行效率模拟模型[J]. 南水北调与水利科技(中英文), 2024, 22(2): 388-398.
YANG J R, WANG C, LEI X H, et al. Simulation model for operational efficiency of typical pumping stations in the Jiangsu section of the Eastern Route of the South-to-North Water Transfers Project[J]. South-to-North Water Transfers and Water Science & Technology, 2024, 22(2): 388-398. (in Chinese)

南水北调东线江苏段典型泵站运行效率模拟模型

杨靖仁[1,2], 王超[2], 雷晓辉[1,2], 何中政[3]

（1. 河北工程大学水利水电学院, 河北 邯郸 056002; 2. 中国水利水电科学研究院, 北京 100038;
3. 南昌大学工程建设学院,江西 南昌 330031）

摘要: 泵站机组运行受多种因素影响, 导致泵站运行理论效率与实际效率误差较大。针对泵站机组运行效率精准模拟难题, 运用基于高价多项式回归、回归树、多元线性回归、向量机回归、高斯过程回归、神经网络的 10 个回归算法, 建立泵站机组效率模拟模型并开展对比分析, 优选出有效的泵站运行效率模拟建模方法。讨论分析采用“上下游水位+流量”代替传统“扬程+流量”开展泵站运行模拟的效果。以南水北调东线邳州站和遂宁二站共 8 台机组的历史数据开展实例分析, 相关实验结果表明: 在所有方法中, 高斯过程回归(Gaussian process regression, GPR)模型在均方根误差(E_{RMS})、平均绝对误差(E_{MA})、均方误差(E_{MS})、决定系数(R^2)和最大个体误差(E_{MI})指标上综合表现最佳, R^2逼近 0.95; 使用站上、站下水位代替传统的扬程对模型进行训练, 所有模型的综合评价指标整体有所改善。综合来看, 使用 GPR 模型并使用上游、下游水位代替扬程进行模拟效率表现最好, 以邳州站 4 号机为例, 可将模拟效率的 E_{MA} 和 E_{MI} 分别从 16.49% 和 20.40% 减少至 0.41% 和 2.30%, 研究成果具有一定实际意义, 可为我国调水工程泵站经济运行提供有力支撑。

关键词: 机器学习; 深度学习; 高斯过程回归; 泵站效率模拟; 南水北调东线
中图分类号: TV675 **文献标志码:** A **DOI:** 10.13476/j.cnki.nsbdqk.2024.0040

随着我国调水工程的规划、建设和运行, 调水工程研究进入从规划建设转向优化调度运行的关键转型期, 泵站经济运行相关研究越来越受到关注[1-2]。泵站效率特性曲线是泵站系统的基本特征之一, 在泵站优化运行过程中具有非常重要的作用, 通常采用实测扬程-流量数据表征, 是运行效率计算的基本依据[3]。

泵站效率误差是指在泵站长期运行中, 多种外界因素导致泵站的实际效率与设计效率之间的差异。泵站的设计效率是指在设计时预计的理论效率, 而实际效率是指在实际运行中泵站能够实现的效率。泵站机组效率的计算误差受众多因素的相互作用与影响, 表现出复杂的非线性特性, 具有显著的非平稳和模糊性等复杂特征[4]。效率误差通常由以下因素引起: 设计误差, 设计效率通常基于理论分析和计算得出, 而实际运行中可能会存在与设计不符的情况, 例如设计流量或设计压力与实际需求不符合等; 机械损失, 泵站内部的机械部件存在摩擦和损耗; 流体摩擦, 泵站内部的管道和阀门等部件会导致流体摩擦; 操作误差, 操作人员的技能水平和操作方式等因素也可能导致泵站效率的误差; 维护不当, 泵站的维护不当可能会导致泵站内部部件的损坏或降低, 从而影响泵站的效率[5]。

南水北调东线一期工程的邳州泵站位于江苏省邳州市八路镇刘集村徐洪河与房亭河交汇处, 是东线工程的第六个梯级。如图 1 所示, 邳州泵站 4 号机组在不同工况下的实际运行效率与理论计算效率偏差巨大, 2023年 3 月 13 日—2023 年 3 月 30 日泵站模拟效率平均绝对误差和最大误差分别为 16.47% 和 20.40%。泵站效率模拟误差使得根据优化获得的调度或控制方案偏离实际最优状态, 研究模拟精度高和稳定性好的效率模拟模型将有望解

收稿日期: 2023-09-28 **修回日期:** 2024-03-04 **网络出版时间:** 2024-03-26
网络出版地址: https://link.cnki.net/urlid/13.1430.TV.20240322.1447.004
基金项目: 国家重点研发计划项目(2022YFC3204603; 2023YFC3209402); 国家自然科学基金项目(52209024); 江西省自然科学基金项目(20224BAB204075; 20212BAB214065); 河北省自然科学青年基金(E2021402039)
作者简介: 杨靖仁(2000—), 男, 河南漯河人, 主要从事水资源综合利用及保护研究。E-mail: yangjingren@live.com
通信作者: 何中政(1992—), 男, 湖北大冶人, 讲师, 博士, 主要从事复杂水资源系统建模及其优化调控研究。E-mail: he_zz@ncu.edu.cn

决这一难题[6]。近年来，学者们[7-9]将多种常见的机器学习方法引入水利领域的预报模拟中，如多项式回归法（polynomial regression, PR）[10]、多元线性回归（multivariate regression, MLR）[11]、高斯过程回归（Gaussian process regression, GPR）[12]、回归树（decision regression tree, DRT）[13]、支持向量回归（support vector regression, SVR）[14]和神经网络（neural network, NN）[14-17]等。这些方法在相关领域的研究和应用均取得了较好的效果，也有望在泵站效率模拟方面发挥作用，人工智能技术开展泵站运行效率模拟也将成为研究热点[18-19]。为此，实验采用人工智能技术，以南水北调东线一期工程典型泵站为研究对象，开展泵站效率模拟相关研究。

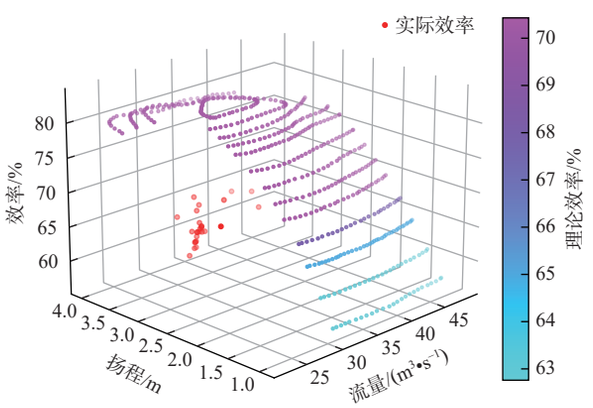


图 1　邳州泵站机组实际运行效率与理论效率对比

Fig. 1　Comparison of actual operating efficiency and theoretical efficiency of units in Pizhou pumping station

1　泵站机组能效特性模拟方法

1.1　数据预处理

由于采集端设备的老化或自然界因素水位流量波动等导致测量出现异常数据，为了避免这些异常值对效率矫正产生负面影响，需要先对测量数据进行数据预处理[20]。传统的数据预处理方法有处理缺失值、删除重复项、处理异常值、数据归一化、真值转化等[21]。

1.1.1　对齐时间

采集端中水位、流量和效率等数据的时间信息并不是一一对应的，因此需要将同一时间窗口内的数据相关联起来。使用基于 Python 3.9 版本 Pandas 工具库中的 Merge Asof 方法对数据表进行左连接，以时间为索引，找到每个数据表中最接近的时间戳，将两者数据关联成一个表，最大时间相差容忍度为 10 min。最后再以时间为索引对关联后的表排序。

1.1.2　划分数据窗口

采用改进的移动平均法（exponential moving average）的思想将数据聚类处理[22]。使用 Pandas 工具库的 Groupby 方法对对齐后的数据进行分组，以 n 条数据为 1 组将数据划分，将数据分为 m 组，并记录每组数据的特征信息，包括开始时间 T_0、结束时间 T_n、时间间隔Δt、切尾均值 A_m、标准差 S_m。

$$A_m=\frac{\sum_{i=1}^{n}x_i-x_{\max}-x_{\min}}{n-2} \tag{1}$$

式中：$x_{\max}$ 为和 $x_{\min}$ 分别为该组数据的极大值和极小值。

1.1.3　处理异常数据

在时间相近的情况下，流量和水位波动不会特别明显。设每组数据允许最大标准差为 $S_{\max}$，对分组后的数据进行筛选，找出 S_m 大于设定参数 $S_{\max}$ 的组，说明该组数据中存在突变数据，需要对其中的异常数据进行处理。

令该组数据$\Delta t=T_n-T_0$，设最大容忍时间间隔为 $T_{\max}$，当Δt 大于设定参数 $T_{\max}$ 时，说明此组为非连续时间序列数据，数据波动是正常情况，反之则说明相邻时间内存在数据突变的异常值。接着令该组中的各个数据与该组的切尾均值 A_m 逐个做差并取绝对值得到$\Delta a_0, \Delta a_1, \cdots, \Delta a_n$，设最大允许误差为 $E_{\max}$，若Δa 大于设定参数 $E_{\max}$，需对其进行删除或同化。整个数据预处理的流程见图 2。

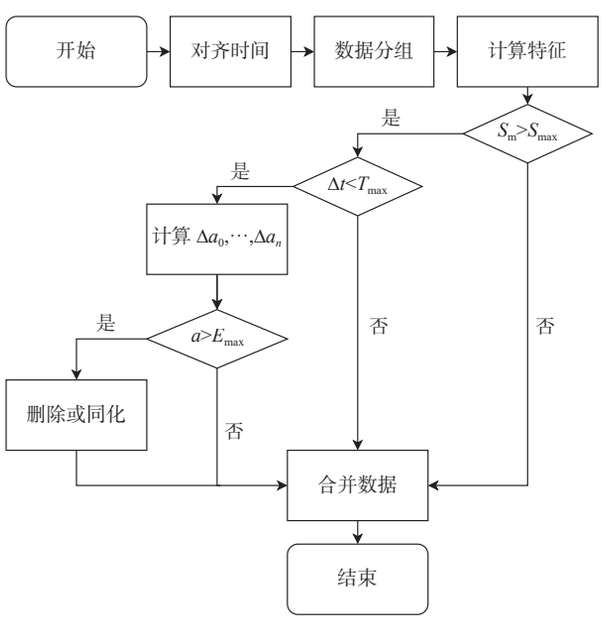


图 2　数据预处理流程

Fig. 2　Flow chart of data preprocessing

1.2 影响因子的选择

一般情况下,传统泵站效率计算由流量-扬程-转角拟合完成[3],其中扬程直接由站上水位减去站下水位得到,考虑到扬程计算的水头损失、泵站管道内摩擦以及液体密度随温度的变化和重力加速度在不同时刻的差异等多种误差影响因素,传统方法缺少对中间过程可能存在的误差分析,得到的效率误差较大。

为了研究使用站上水位、站下水位作为两个独立的影响因子代替扬程对泵站机组效率模拟的效果,如表 1 所示,实验使用站上水位、站下水位、单位机组流量及机组转角 4 个影响因子作为特征输入对 10 种模型进行训练,并与使用传统扬程、流量和转角 3 个影响因子作为特征输入来对多种不同的回归模型进行训练的效果进行对比分析。

表 1 影响因子选择

Tab. 1 Parameter setting in polynomial model

方法	特征数量	影响因子
传统方法	3	扬程,流量,机组转角
本研究实验方法	4	流量,机组转角,站上水位,站下水位

1.3 模拟效率的方法

泵站机组的运行效率模拟通常以回归问题的形式呈现。为了比较研究泵站效率模拟方法,选择 PR 回归,并将其与机器学习和深度学习领域最具潜力的 5 类回归方法(MLR、GPR、DRT、SVR、NN)[4]进行对比。

1.3.1 多项式回归

由于任一函数都可以用多项式逼近,因此 PR 有着广泛应用。高阶 PR 模型是线性回归模型的一种,在传统水力学问题中,泵站机组效率通常使用二阶或三阶多项式回归进行模拟[10]。其原理如下:

$$X_i = x_1, x_2, \cdots, x_i, \quad \cdots, \quad x_n \tag{2}$$

式中:x 代表输入的实测值,X_i 代表有 n 个输入特征的向量。

那么 n 元 p 阶多项式回归方程 $f(x)$ 为:

$$f(x) = \mathrm{a}_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 + a_4{x_1}^2 + a_5{x_2}^2 + \cdots + a_m{x_n}^p \tag{3}$$

式中:$a_0, a_1, \cdots, a_m$ 为多项式系数。实验使用二阶多项式回归模型(2nd PR)和三阶多项式回归模型(3rd PR)作为代表传统效率模拟方法的对照组,与实验所研究的机器学习和深度学习回归模型做对比参照。

1.3.2 回归树

DRT 是用树模型做回归问题,具有很高的复杂度和高度的非线性关系。DRT 通过将输入空间划分为多个矩形区域来构建树结构,每个矩形区域对应于树的一个叶节点,每个叶节点中包含一个预测值,用于表示该区域内样本的平均值或其他统计量。通过选择最佳的特征和切分点,根据数据的特征值将样本逐步分配到适当的叶节点。构建回归树的过程通常采用递归的方式,从树的根节点开始,根据选择的特征和切分点将数据划分到子节点。然后对每个子节点递归地重复这个划分过程,直到满足预定义的停止条件[13]。

1.3.3 支持向量机回归

SVR 是一种用于回归问题的机器学习算法,与传统的回归算法不同,它使用支持向量机(support vector machine, SVM)的思想来解决回归问题。本研究使用的线性支持向量回归(linear support vector regression, LSVR)是 SVR 的一种变体[14],它试图找到一条直线(或者超平面),使得尽可能多的数据点都落在这条直线上,并且离这条直线的距离最小,这个距离被称为边界。LSVR 的优点是能够在高维空间中处理大量数据,同时具有较好的泛化能力。不同于 SVM, LSVR 的目标是最小化预测值与真实值之间的差距,而不是将数据点分为两个类别。为此,LSVR 使用一个损失函数来惩罚那些预测误差较大的数据点。

1.3.4 多元线性回归

MLR 已在水利等相关领域得到广泛应用[11]。已知样本($x_0, x, x_2, \cdots, x_n, y$)的实际值为 y。设$\hat{y}$为预测值,那么多元线性回归方程为:

$$\hat{y} = b + \sum_{i=1}^{n} \omega_i x_i \tag{4}$$

式中:$\boldsymbol{x}=[1, x_0, x_1, x_2, \cdots, x_n]^{\mathrm{T}}$ 代表样本值;$\boldsymbol{\omega}=[b, \omega_0, \omega_1, \omega_2, \cdots, \omega_n]^{\mathrm{T}}$ 为每个样本对应的权重。

1.3.5 高斯过程回归

GPR 是基于贝叶斯理论和统计学习理论发展起来的一种全新机器学习方法[23]。对于实际工程问题考虑输出受噪声影响,使用如下模型:

$$y = f(x) + \varepsilon \tag{5}$$

式中:y 是受到噪声污染的观测值;f 为函数值;x 为输入向量。假设噪声 ε 服从均值为 0、方差为σ_n^2,以得到观测值 y 的先验分布为

$$y \sim N(0, k(\vec{X}, \vec{X}) + \sigma_n^2 I_n) \tag{6}$$

观测值 y 和预测值 $f(x_*)$ 的联合先验分布为

$$\begin{bmatrix} y \\ f(x_*) \end{bmatrix} \sim N\left(0, \begin{bmatrix} \boldsymbol{K}\left(\vec{X},\vec{X}\right)+\sigma_n^2 I_n & \boldsymbol{K}\left(\vec{X},x_*\right) \\ \boldsymbol{K}\left(x_*,\vec{X}\right) & \boldsymbol{k}(x_*,x_*) \end{bmatrix}\right) \quad (7)$$

式中：x_*为输入特征向量；$\boldsymbol{K}\left(\vec{X},\vec{X}\right)$是 $n\times n$ 阶的 GP（对称正定）协方差矩阵；$\boldsymbol{K}\left(\vec{X},x_*\right)$是输入集 $\boldsymbol{X}$ 与测试输入向量 x_*之间的 $n\times 1$ 阶协方差矩阵；$\boldsymbol{K}\left(x_*,\vec{X}\right)$同理可知，$\boldsymbol{k}(x_*x_*)$ 为 x_*自身的协方差；$\boldsymbol{I}_n$ 是 n 维单位矩阵。

由此，计算出的预测值 $f(x_*)$ 的后验分布为

$$f(x_*)\left|\vec{X},y,x_* \sim N\right|\left[\bar{f}(x_*),cov(f(x_*))\right] \quad (8)$$

式中：预测均值 $\bar{f}(x_*)$即是观测值 y 的预测值。

1.3.6　神经网络

人工神经网络（artificial neural network, ANN）是一种模仿生物神经网络结构和功能的计算模型，用于机器学习和模式识别任务[24]。深度神经网络（deep neural network, DNN）是一种 ANN 的扩展，具有多个隐藏层的结构[25]。与传统的浅层神经网络相比，DNN 具有更多的层和更复杂的结构，可以更好地处理复杂的特征和学习更高级的表示。

2　南水北调东线江苏段典型泵站运行效率模拟建模

2.1　实例基本情况

实例分析对象为南水北调东线徐洪河段邳州站和睢宁二站的 8 台机组，其中：邳州站是南水北调东线一期工程的第 6 级抽水泵站，主要任务是与泗洪站、睢宁二站一起，通过徐洪河输水线向骆马湖输水 100 m³/s，与中运河共同满足向骆马湖调水 275 m³/s 的目标，并结合房亭河以北地区的排涝；睢宁二站位于徐州市睢宁县沙集镇境内的徐洪河输水线上，与睢宁一站及运河线上的刘老涧泵站枢纽共同组成南水北调东线工程的第五个梯级，主要任务是与睢宁一站共同实现向骆马湖调水 100 m³/s 的目标，与运河线共同满足向骆马湖调水 275 m³/s 的目标。泵站安装 4 台套立式混流泵，具体机组配置见表 2。

表 2　泵站机组配置信息

Tab. 2　Information on the configuration of the pumping station unit

泵站	机组数量/台	单机流量/($m^3\cdot s^{-1}$)	单机功率/kW	总流量/($m^3\cdot s^{-1}$)	总功率/kW	设计站下水位/m	设计站上水位/m	设计扬程/m
邳州站	4	33.4	1 950	100.0	7 800	20.1	23.2	3.1
睢宁二站	4	20.0	3 000	80.0	12 000	13.3	21.6	8.3

采用 10 种模型对邳州站的 4 台机组和遂宁二站的 4 台机组经过数据预处理后的历史工况数据进行训练和测试，邳州站的 1 号、2 号、3 号机组的数据集为 2022 年 5 月 20 日至 2023 年 3 月 30 日的历史工况，邳州站 4 号机组的数据集为 2022 年 12 月 3 日至 2023 年 3 月 16 日的历史工况，睢宁二站的 4 台机组均采用 2022 年 12 月 3 日至 2023 年 3 月 16 日的历史工况作为数据集。实验选取了均方根误差（root mean squared error, E_{RMS}）、平均绝对误差（mean absolute error, E_{MA}）、均方误差（mean squared error, E_{MS}）、决定系数（R-square, R^2）、最大个体误差（max individual error, E_{MI}）评价指标来对各个模型的结果进行评估，其中 E_{MI} 为模型对每个历史数据进行模拟的结果与真实值的绝对误差中最大的值，可以评价对模型对个体模拟的最差效果。

2.2　训练方法

研究主要使用 Python 和 MATLAB 回归学习器构建模型。对于使用 python 构建的模型，借助 scikit-learn 机器学习库中的 train_test_split 方法从机组历史工况数据集中抽取 20% 的数据作为测试集，剩余的作为训练集。为避免深度学习模型记忆数据的顺序，并且提高模型泛化能力，实验使用 shuffle 方法打乱数据顺序。对于使用 MATLAB 回归学习器构建的模型，同样使用 20% 的数据作为测试集，并采用 k 倍交叉验证交叉验证的方式训练模型。根据经验，k 为 0.5 时被认为可以在过度拟合和准确性之间取得适当的平衡[6]。最后采用常用的决定系数 R^2作为拟合优度标准。各类模型所用的工具或模块见表 3。

表 3　6 类模型的实现方法

Tab. 3　Implementation methods of the six types of models

模型	软件/语言	工具/类
MLR	MATLAB2023b	回归学习器
DTR	MATLAB2023b	回归学习器
SVM	MATLAB2023b	回归学习器
GPR	MATLAB2023b	回归学习器
NN	Python 3.9	Keras
PR	Python 3.9	Scikit-learn

对于 MLR 和 SVM 模型均使用 MATLAB 回归学习器的默认参数配置,DTR 使用以往研究中被证明效果较好的参数设置,而 PR、GPR 和 NN 模型的具体配置有所不同,接下来说明其他模型的详细参数设置。

2.2.1　多项式模型参数设置

传统泵站效率模拟通常使用高阶多项式回归方法,为了更全面地考虑不同阶数对传统模拟方法效果的影响,实验采用二次多项式和三次多项式这两种不同基函数构建传统泵站效率模拟方法的模型。

表 4 中样本 Q 为 t 时刻泵站流量,H 为 t 时刻的扬程,R 为 t 时刻单位机组转角。本模型使用 Scipy 库中的 Polynomial Features 方法生成 n 阶特征矩阵模型,使用 Fit Transform 方法将模型转换为输入数据矩阵,使用最小二乘法线性回归模型(least square linear regression, LSLR)进行回归模拟[26]。

表 4　多项式模型中的参数设置

Tab. 4　Parameter setting in polynomial model

多项式	参数设置		
	Degree	特征因子	模型
三元二次	2	Q, H, R	LSLR
三元三次	3	Q, H, R	LSLR

2.2.2　选择 GPR 的核函数

GPR 可以选用不同的协方差函数[27],即核函数。不同的核函数作用下回归特征各不相同,常用的核函数有径向基、二次有理、马顿和指数核函数等[24]。实验选用二次有理、径向基和指数核函数进行回归预测。各个核函数的公式如下:

二次有理核函数(rational quadratic kernel, RQ),本文称为 GRP(RQ)模型:

$$\kappa(r)=\left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha},\alpha,l>0 \tag{9}$$

式中:$r=\|X_1-X_2\|$,α, l 为超参数。二次有理核函数可以理解为是无穷个径向基核函数的线性叠加,当 $\alpha\to\infty$ 时,二次有理核函数等价于 σ=l 的径向基核函数。与径向基核函数相比,更为省时,作用域广,但是对参数十分敏感。

径向基函数核也被称为平方指数(squared exponential, SE)或高斯核,本文称为 GRP(SE)模型:

$$\kappa(r)=\exp\left(-\frac{\|X_1-X_2\|^2}{2\sigma^2}\right) \tag{10}$$

式中:$r=\|X_1-X_2\|$,σ 为超参数,表征了学习样本间的相似度。径向基核函数也被称为高斯核,常被应用于 SVM 和 GPR 等各类机器学习算法中,参数简单,但在处理样本数量大或特征多时效果一般。

指数核函数(Exponential kernel, E),本文称为 GRP(E)模型:

$$\kappa(r)=\exp\left(-\frac{r}{l}\right) \tag{11}$$

式中:$r=\|X_1-X_2\|$,l 为超参数,指数核函数是马顿核在 v=0.5 的特殊形式,这样的变动减少对参数的依赖性降低,使得模型的训练学习难度相当于马顿核模型大大降低。

2.2.3　神经网络模型参数设置

实验使用两种神经网络对效率数据进行训练:一种是单层 ANN,其中包含输入层、一个具有 25 个神经元使用 ReLU 激活函数的隐藏层、输出层;另一种是 DNN,其中包含输入层、5 个隐含层、输出层。其中,隐含层的参数设置见表 5[28]。

表 5　神经网络模型参数设置

Tab. 5　Neural network model parameter setting

模型	参数设置		
	Layer	Activation	Dense
ANN	1	ReLU	25
DNN	1	ReLU	16
	2	Sigmoid	32
	3	Sigmoid	64
	4	Sigmoid	32
	5	ReLU	16

3　结果分析

3.1　扬程-流量-转角模拟计算效率实例分析

为评价预测结果的综合表现,首先使用传统的扬程-流量与转角进行效率模拟。图 3 至图 7 分别给出了 10 种模型对 8 台机组效率模拟结果的测试数据集的 E_{RMS}、E_{MA}、R^2、E_{MS} 和 E_{MI} 5 种指标,其中 E_{RMS}、E_{MA}、E_{MS} 和 E_{MI} 值越接近 0 表示模拟误差越小,精度越高,与泵站的真实效率越接近。决定系数 R^2越接近 1 说明回归模型的拟合效果越好。

由图 3 至图 7 的 8 个机组平均指标可知,GPR(RQ)、GPR(SE) 和 GPR(E) 3 种核函数的 GPR 模型对邳州 4 台机组和睢宁二站 4 台机效率模拟综合表现最佳。从 E_{MA} 指标来看,3 种 GPR 模型约为 0.34~0.36, ANN、DNN 和 MLR 略大于 0.5, 其余模型表现较差。从 R^2指标来看,除 DRT 和 SVM 模型在 0.7~0.8 外,其他模型均在 0.9 以上。从 MIE 指标来看,传统多项式(2nd PR 和 3rd PR)表现最差,其

他模型约在 5 以内，3 种 GPR 模型约在 3.2~3.5，表现较好。综合 E_{RMS}、E_{MA}、R^2、E_{MS} 和 E_{MI} 5 项指标来看，在众多传统模型和机器学习等方法中，GPR 模型整体测试的效果最优。

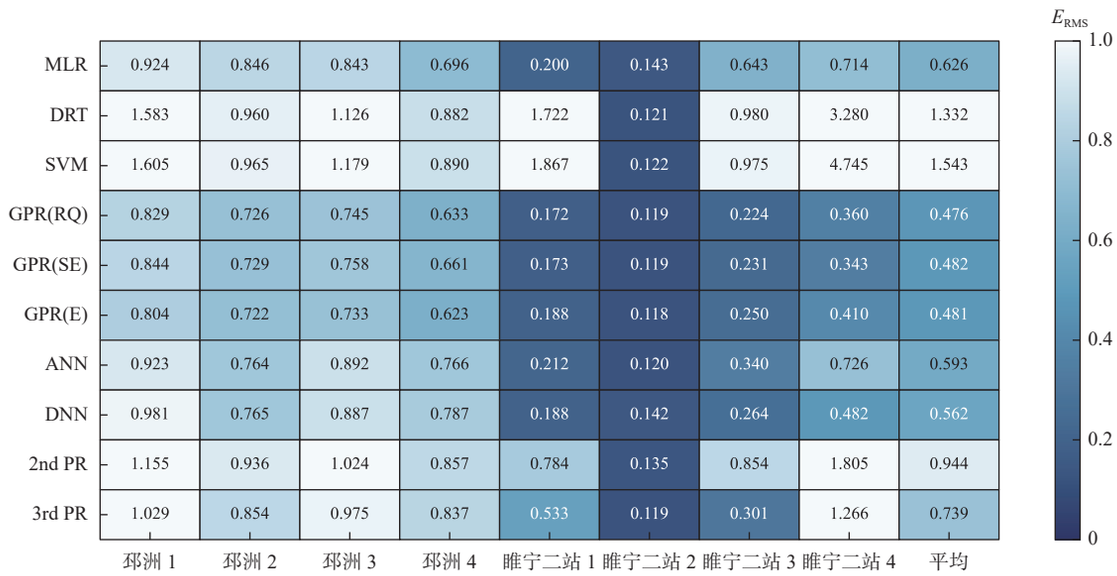


图 3　10 种模型测试集的 E_{RMS} 指标

Fig. 3　E_{RMS} index of ten model test sets

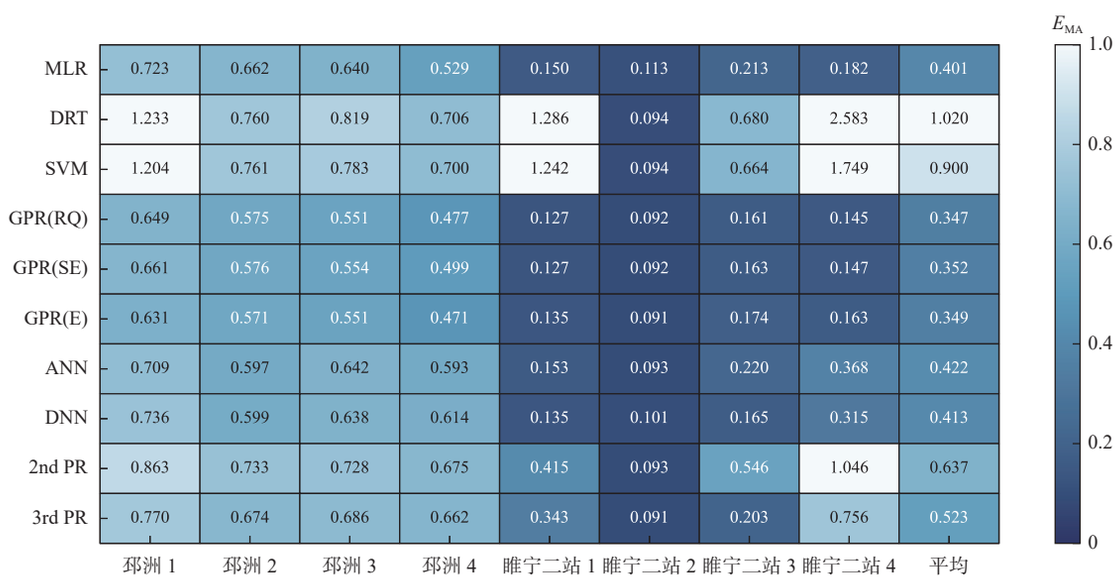


图 4　种模型测试集的 E_{MA} 指标

Fig. 4　E_{MA} index of ten model test sets

3.2　站上水位、站下水位与扬程参与模型训练对比分析

进一步分析站上水位、站下水位与传统扬程作为特征输入的效率模拟效果，研究以邳州站第 4 号机组为例，使用站上水位、站下水位、流量和转角 4 个特征输入，对比传统的以扬程、流量、转角作为特征输入，分别用 10 种模型对邳州泵站 4 号机组的历史数据进行实验分析。

图 8 反映了 1.2 节中两种影响因子在泵站效率模拟过程中与实际值之间的差异，通过各种模型对邳州站 4 号机组的不同工况拟合得到的效率，与对应工况下的实际效率对比得到的误差。由图 8 可知，当使用站上水位和站下水位代替扬程对邳州站第 4 号机组的历史数据进行训练后，各模型测试集的 R^2、E_{RMS}、E_{MS} 和 E_{MA} 结果指标均有不同程度的提升。由图 8(c)可知，对于最大误差 E_{MI} 的结果指标，仅 DTR 和 3 种 GPR 模型在使用站上水位和站下水位作为特征输入后比使用扬程模拟的效果有所提升。由此可见，使用此方法训练模型时，GPR 在邳州 4 号机组效率模拟的所有指标均比之前更好。

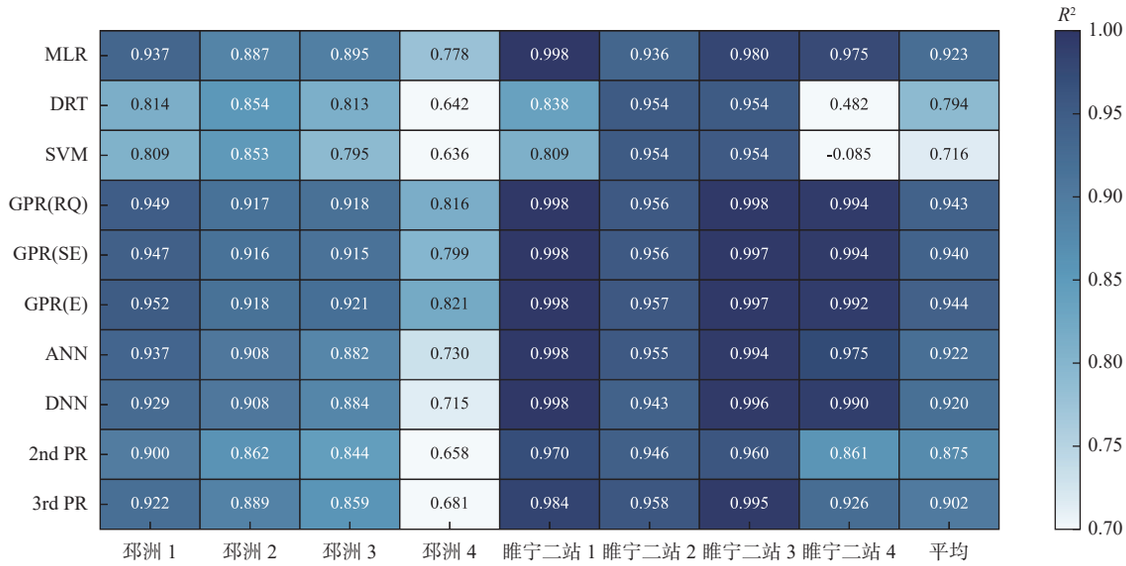


图 5　10 种模型测试集的 R^2 指标

Fig. 5　R^2 index of ten model test sets

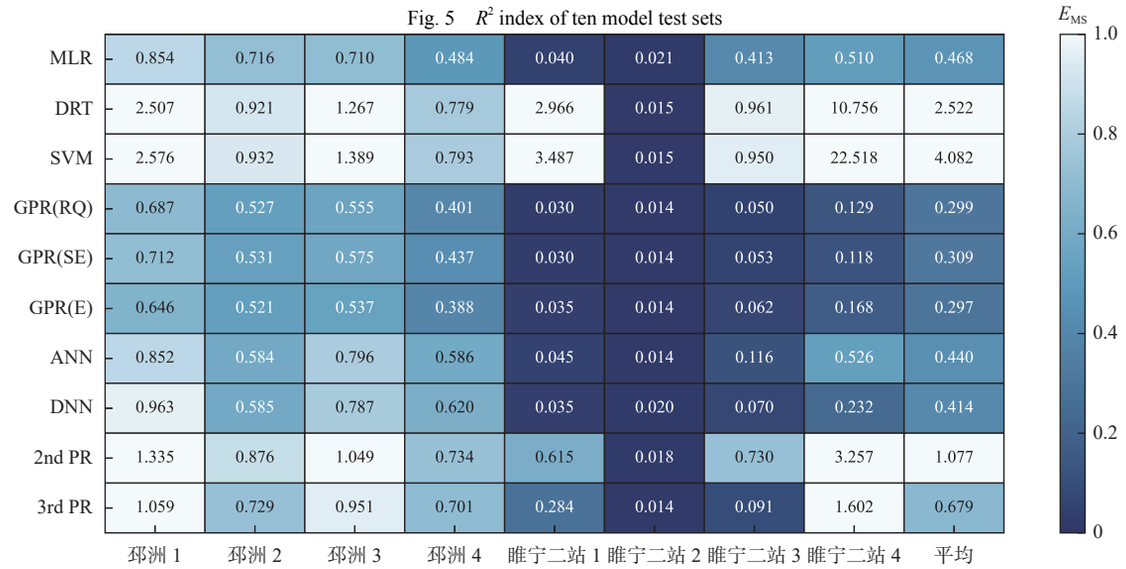


图 6　种模型测试集的 E_{MS} 指标

Fig. 6　E_{MS} index of ten model test sets

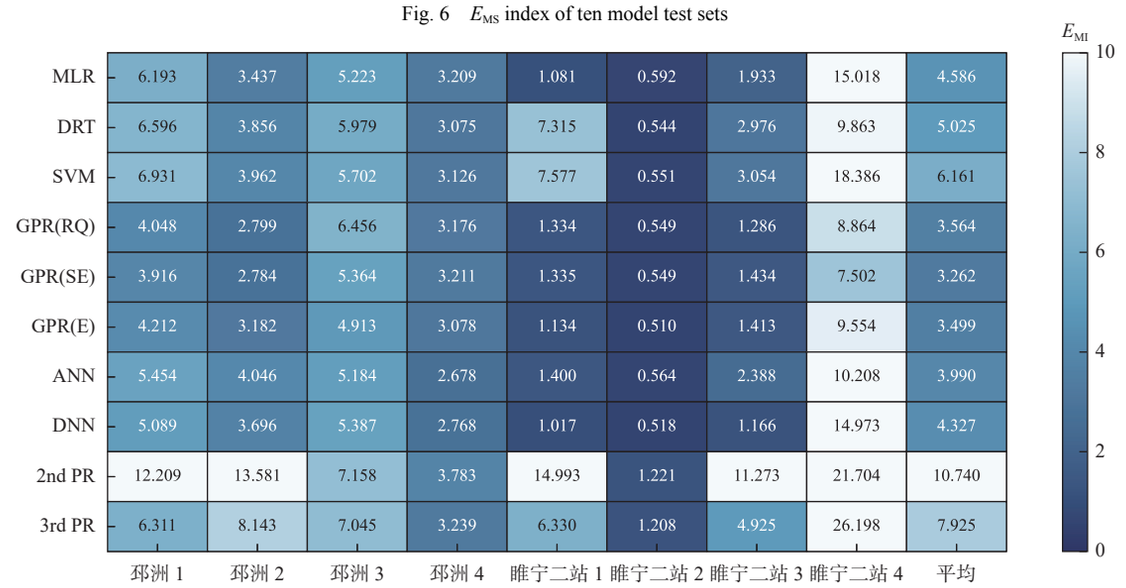
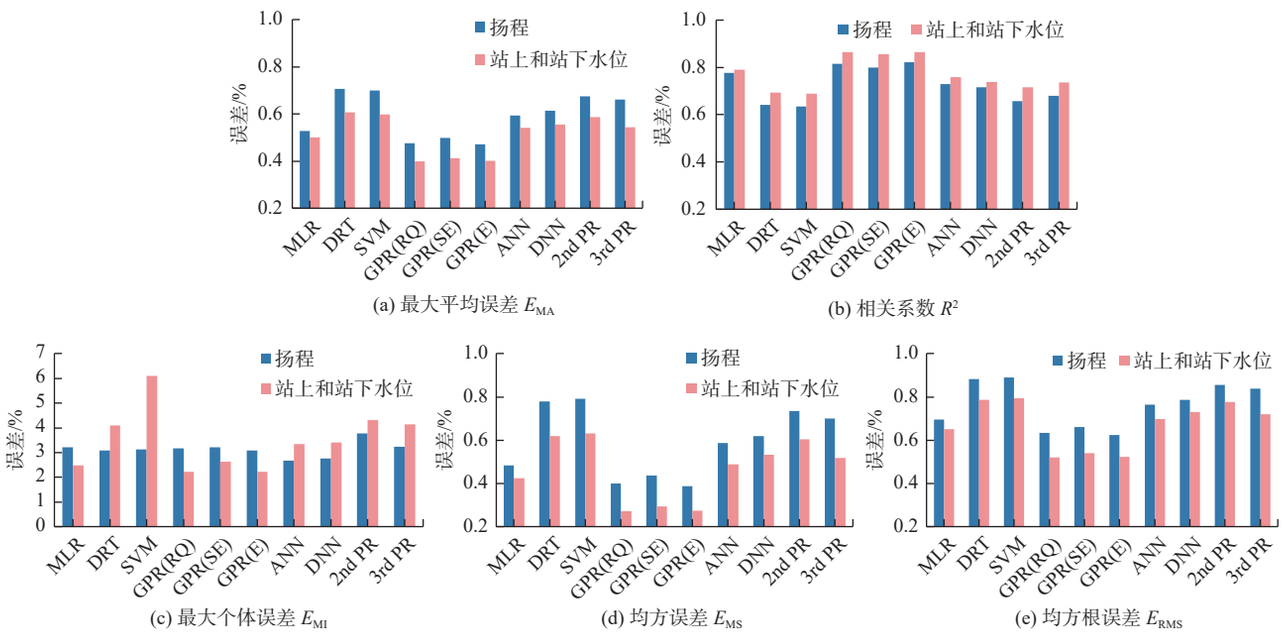


图 7　10 种模型测试集的 E_{MI} 指标

Fig. 7　The E_{MI} index of ten model test sets

进一步分析原因可能是：扬程与站上水位和站下水位为非线性关系，而传统的扬程直接为站上水位减去站下水位的水位差和水头损失计算得到，没有考虑到管道摩擦损失、水头损失等因素，缺少对中间过程可能存在的误差分析，误差较大；且站上水位和站下水位上的分布可能不均匀，导致某些区域的数据权重较大，而其他区域的权重较小。扬程为站上水位减去站下水位的表示方式可能与实际的物理意义有所差异，因此使用机器学习和深度学习模型时，直接使用站上水位和站下水位作为特征输入，可以让算法模型更直观地分析其中潜在影响因素对结果的影响。



(a) 最大平均误差 E_{MA}　(b) 相关系数 R^2

(c) 最大个体误差 E_{MI}　(d) 均方误差 E_{MS}　(e) 均方根误差 E_{RMS}

图 8　两种训练方法的测试集 5 种指标对比

Fig. 8　Comparison of five indicators of the test set of the two training methods

4　结 论

通过对 10 种回归模型模拟泵站效率的研究，找到回归效果最好的模型，且分析对比传统多项式模型的效果提升。根据评估指标结果显示，传统效率模拟方法 PR 效果较差，而在其他回归模型中 GPR 模型表现最佳，在对邳州站和睢宁二站共 8 台泵站机组效率进行训练后的测试集误差指标 E_{RMS}、E_{MA}、R^2、E_{MS} 和 E_{MI} 的平均值均优于其他模型。

以邳州泵站 4 台机组为例，使用站上、站下水位代替传统的扬程作为输入特征数据进行训练，通过观察 E_{MA}、E_{MS}、E_{RMS} 和 R^2指标发现，10 种模型的误差均有所减少，而对于最大个体误差 E_{MI} 指标，GPR 模型的表现效果提升明显。

以邳州站 4 号机为例：使用传统的流量-扬程-转角进行泵站机组效率模拟时，GPR 模型可将模拟效率的平均绝对误差和最大绝对误差控制在 0.50% 和 3.20% 以内；进一步使用站上、站下水位代替扬程模拟后，模拟效率的平均绝对误差和最大绝对误差控制在 0.41% 和 2.30%；可将模拟效率的 E_{MA} 和 E_{MI} 分别从现状的 16.49% 和 20.40% 减少至 0.41% 和 2.30%，为泵站站内经济运行提供效率精准模拟的技术支撑。

参考文献：

[1] 周建中, 何中政, 贾本军, 等. 水电站长中短期嵌套预报调度耦合实时来水系统动力学建模方法研究及应用[J]. 水利学报, 2020, 51(6): 642-652. DOI: 10.13243/j.cnki.slxb.20190664.

[2] 王本德, 周惠成, 卢迪. 我国水库(群)调度理论方法研究应用现状与展望[J]. 水利学报, 2016, 47(3): 337-345. DOI: 10.13243/j.cnki.slxb.20150940.

[3] 王富喜. 流量调节后的泵站节能效果分析[J]. 水利科技与经济, 2019, 29(3): 60-63.

[4] KUMAR S, BHATNAGAR V. A review of regression models in machine learning[J/OL]. Journal of Intelligent Systems and Computing, 2022, 3(1): 40-47.

[5] 闻昕, 黄抒艺, 谭乔凤, 等. 江苏省南水北调多工程多目标联合优化调度方法[J]. 水资源保护, 2023, 39(5): 118-124,134. DOI: 10.3880/j.issn.1004-6933.2023.05.014.

[6] KOOR M, VASSILJEV A, KOPPEL T. Optimization of pump efficiencies with different pumps characteristics working in parallel mode[J/OL]. Advances in Engineering Software, 2016, 101: 69-76. DOI: 10.1016/j.advengsoft.2015.10.010.

[7] 方国华, 曹蓉, 刘芹, 等. 改进遗传算法及其在泵站优化运行中的应用[J]. 南水北调与水利科技, 2016, 14(2): 142-147. DOI: 10.13476/j.cnki.nsbdqk.2016.02.025.

[8] 曹子恒, 李永坤, 胡义明, 等. 基于机器学习模型的数值降雨预报校正[J]. 南水北调与水利科技 (中英文), 2023, 21(5): 843-861,950. DOI: 10.13476/j.cnki.nsbdqk.2023.0083.

[9] 疏杏胜, 王子茹, 李福威, 等. 基于机器学习模型的短期降雨多模式集成预报[J]. 南水北调与水利科技(中英文), 2020, 18(1): 42-50. DOI: 10.13476/j.cnki.nsbdqk.2020.0006.

[10] 杨朔. 基于机器学习的调洪演算方法研究[J]. 山东水利, 2021(6): 84-86. DOI: 10.16114/j.cnki.sdsl.2021.06.039.

[11] 陈建国, 武俞, 军马. 基于多项式回归算法的电量水量转换模型构建: 以宁夏西干渠管理处马场滩灌溉站为例[J]. 应用数学进展, 2022, 11(6): 3230-3238. DOI: 10.12677/AAM.2022.116342.

[12] SAHOO S, JHA M. Groundwater-level predicition using multiple linear regression and artificial neural network techniques: A comparative assessment[J]. Hydrogeology Journal, 2013(21): 1865-1881. DOI: 10.1007/s10040-013-1029-5.

[13] SUN N, ZHANG S, PENG T, et al. Multi-Variables-Driven model based on random forest and Gaussian process regression for monthly streamflow forecasting[J/OL]. Water, 2022, 14(11): 1828. DOI: 10.3390/w14111828.

[14] 曹桃云, 张日权. 非对称误差分布的贝叶斯累加回归树模型研究及应用[J]. 系统科学与数学, 2022, 42(11): 3119-3133. DOI: 10.12341/jssms21570.

[15] 张研, 廖逸夫, 王鹏鹏, 等. 基于相关向量机的调水工程调蓄水位预测模型[J]. 南水北调与水利科技(中英文), 2021, 19(4): 814-821. DOI: 10.13476/j.cnki.nsbdqk.2021.0085.

[16] XIONG B, LI R P, REN D, et al. Prediction of flooding in the downstream of the Three Gorges Reservoir based on a back propagation neural network optimized using the AdaBoost algorithm[J]. Natural Hazards, 2021, 107(2): 1559-1575. DOI: 10.1007/S11069-021-04646-4.

[17] BAEK S S, PYO J C, CHUN J A. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach[J]. Water, 2020, 12(12): 3399. DOI: 10.3390/W12123399.

[18] ALSUMAIEI A A. A nonlinear autoregressive modeling approach for forecasting groundwater level fluctuation in urban aquifers[J]. Water, 2020, 12(3): 12030820. DOI: 10.3390/w12030820.

[19] ZHANG J F, ZHU Y, ZHANG X P, et al. Developing a Long Short-Term Memory(LSTM) based model for predicting water table depth in agricultural areas[J]. Journal of Hydrology, 2018, 561: 918-929. DOI: 10.1016/j.jhydrol.2018.04.065.

[20] HUANG R, ZHANG Z, ZHANG W, et al. Energy performance prediction of the centrifugal pumps by using a hybrid neural network[J/OL]. Energy, 2020, 213: 119005[2023-12-27]. DOI: 10.1016/j.energy.2020.119005.

[21] 金容鑫, 娄岱松, 黄华德, 等. 水电机组状态监测数据清洗方法[J]. 中国农村水利水电, 2022, 477(07): 187-192.

[22] 位文涛, 靳燕国, 张召, 等. 南水北调中线工程流量监测站点倒挂数据清洗模型及应用[J]. 南水北调与水利科技 (中英文), 2022, 20(6): 1158-1167. DOI: 10.13476/j.cnki.nsbdqk.2022.0114.

[23] 王超, 张睿, 张诚, 等. 时间尺度对梯级水电站发电调度建模准确度影响分析[J]. 工程科学与技术, 2017, 49(6): 19-29. DOI: 10.15961/j.jsuese.201601294.

[24] 王方成, 刘玉敏, 崔庆安. 基于高斯过程回归的混合型参数建模及优化[J]. 统计与决策, 2023, 39(1): 34-39. DOI: 10.13546/j.cnki.tjyjc.2023.01.006.

[25] ALSUGAIR A M, AL-GAHTANI K S, ALSANABANI N M, et al. Artificial neural network model to predict final construction contract duration[J/OL]. Applied Sciences, 2023, 13(14): 8078. DOI: 10.3390/app13148078.

[26] GALELLI S, CASTELLETTI A. Tree-based iterative input variable selection for hydrological modeling[J/OL]. Water Resources Research, 2013, 49(7): 4295-4310. DOI: 10.1002/wrcr.20339.

[27] 李治军, 姚蓉. 基于主成分分析和多元线性回归的黑龙江省用水效率研究[J]. 水利科技与经济, 2023, 29(2): 60-64. DOI: 10.3969/j.issn.1006-7175.2023.02.013.

[28] 何中政,方丽,刘万,等. 基于指数核函数高斯过程回归的短期径流预测研究[J/OL]. 中国农村水利水电:1-14[2023-08-10]. DOI: 10.12396/znsd.230394.

Simulation model for operational efficiency of typical pumping stations in the Jiangsu section of the Eastern Route of the South-to-North Water Transfers Project

YANG Jingren[1,2], WANG Chao[2], LEI Xiaohui[1,2], HE Zhongzheng[3]

(1. *School of Water Conservancy and Hydropower*, *Hebei Engineering University*, *Handan* 056002, *China*; 2. *China Institute of Water Resources and Hydropower Research*, *Beijing* 100038, *China*; 3. *School of Architectural Engineering*, *Nanchang University*, *Nanchang* 330031, *Jiangxi*, *China*)

Abstract: With China's extensive water transfers projects underway, the focus has shifted towards optimizing their operation, highlighting the significance of pumping station efficiency studies. The efficiency characteristic curve, a fundamental feature, plays a crucial role in optimizing station operation by utilizing measured head-flow data. However, long-term operation introduces efficiency errors, stemming from design inaccuracies, mechanical losses, fluid friction, operational errors, and improper maintenance. This is evident in cases like pumping Unit 4 at the Pizhou station, where substantial disparities between actual and theoretical efficiency exist, necessitating precise efficiency simulation models to align optimization schemes with the actual optimal state. Recent endeavors have integrated machine learning algorithms like polynomial regression, Gaussian process regression, and neural networks into hydraulic forecasting and simulation, offering promising avenues for pumping station efficiency simulation. Therefore, employing artificial intelligence techniques to investigate pumping station efficiency simulation was proposed, focusing on a representative station of the Eastern Route of the South-to-North Water Transfers Project.

The efficiency simulation of pumping units in water management systems is a critical task, demanding meticulous preprocessing of operational data and the selection of appropriate modeling techniques. Initially, data preprocessing involves aligning time-stamped measurements, clustering data into windows, and handling anomalies to ensure data quality. Various influencing factors, such as flow rates, water levels, and blade rotating angles, are scrutinized to optimize efficiency modeling. Traditional methods, like Polynomial Regression and Multivariate Linear Regression, are contrasted with advanced techniques including decision regression trees, support vector regression, Gaussian process regression, and neural networks. Each method offers unique advantages, such as the interpretability of decision trees and the flexibility of neural networks. Training these models involves careful parameter selection and validation using established metrics like root mean squared error and determination coefficient. Python and MATLAB are prominent tools used for implementation, offering libraries and functions tailored for regression tasks.

The average indicators of the eight pumping units indicate that GPR (Gaussian process regression) models with three different kernel functions (RQ, SE, E) exhibit the best overall performance in simulating the efficiency of the four units at the Pizhou station and the four units at the Suining station. The indicator shows the three GPR models are around 0.34 to 0.36, while ANN, DNN, and MLR are slightly above 0.5, with other models showing poorer performance. In terms of the R^2 indicator, except for DRT and SVM models, which are approximately between 0.7 and 0.8, all other models score above 0.9. Regarding the E_{MI} indicator, traditional polynomials (2nd PR and 3rd PR) perform the worst, while other models are within approximately 5, with the three GPR models ranging from 3.2 to 3.5, showing better performance. Considering the five metrics E_{RMS}, E_{MA}, R^2, E_{MS} and E_{MI}, the GPR models demonstrate the best overall performance among various traditional and machine learning methods in the comprehensive testing. Comparing efficiency simulation methods, incorporating station upstream and downstream water levels alongside traditional head as features yielded superior results, notably enhancing model performance in various metrics. This approach, particularly evident in GPR models, addresses non-linear relationships and potential error sources in head calculations. Utilizing station water levels directly improves model accuracy, offering a more intuitive analysis of influencing factors.

In conclusion, after analyzing ten regression models for pump efficiency simulation, the GPR model emerged as

the most effective, outperforming traditional polynomial methods. Evaluation metrics showed significantly superior performance of GPR over other models, evidenced by reduced errors across various indicators when applied to training datasets of eight pump units at two stations. Substituting station water levels for traditional head as input features yielded notable improvements in model accuracy, particularly evident in GPR models. For instance, using GPR for efficiency simulation at one pump unit resulted in average and maximum absolute errors within 0.50% and 3.20% respectively, while employing water levels instead of head further reduced these errors to 0.41% and 2.30%. This enhancement signifies a substantial improvement over current methods, offering precise efficiency simulation crucial for optimizing pump station operations.

Key words: machine learning**;** deep learning**;** Gaussian process regression**;** pumping station efficiency simulation**;** Eastern Route of South-to-North Water Transfers Project

(上接第 347 页)

The results indicate that, although the overall intensity of extreme precipitation events exhibited a decrease during the recent period, there was a discernible upward trend in the frequency of these events, specifically between the years 2015 and 2020. Subsequently, distinct computations of the HEV were carried out. The findings revealed higher hazard values concentrated in the eastern regions of the Guanzhong zones and southern Shaanxi, while relatively lower hazard values in the southern part of northern Shaanxi. Notably, Xi'an City was the highest exposure index in the province, while Yulin in northern Shaanxi had the lowest exposure index. The central parts of both the Guanzhong zones and Hanzhong showed higher vulnerability values, in contrast to the lower vulnerability values in the northern part of Hanzhong and the southern part of Baoji. Considering the comprehensive attributes encompassing hazard, exposure, and vulnerability, the risk analysis pinpointed several notable zones. Xianyang City's southern part, the northern part of Xi'an City, and the central part of Hanzhong were identified as high-risk zones, significantly susceptible to the impacts of extreme precipitation. On the other hand, the southern part of Xi'an City and the northern part of Hanzhong, characterized by higher exposure but lower hazard and vulnerability, fell within the medium-risk category. Areas with relatively lower socio-economic development, such as the northern regions of Shaanxi and the southern part of Baoji, were designated as low-risk zones within the context of this comprehensive risk assessment framework. Guided by the research findings, relevant authorities should emphasize specific strategies when addressing the impacts of extreme precipitation on agricultural production and livelihoods.

In regions predominantly influenced by extreme precipitation intensity and frequency, like the southern part of Xianyang City, it is essential to enhance the precision and accuracy of early warnings and forecasts and refining responsive mechanisms. In high-risk areas primarily shaped by topography, such as the central part of Hanzhong City, appropriately increasing vegetation cover based on local water carrying capacity and improving agricultural infrastructure are recommended. In areas marked by high-density rural economies, such as the northern part of Xi'an City, efforts should be directed towards enhancing the disaster resilience of agricultural activities through crop optimization and management. This multidimensional risk analysis provides a theoretical foundation, enabling agricultural and water resource management authorities to effectively address the increasingly frequent and intense extreme precipitation disasters in a changing environment.

Key words: climate change**;** Shaanxi Province**;** extreme precipitation**;** agricultural production**;** risk assessment.