普通线性回归、Lasso 与 Ridge 及其融合模型 Elastic Net 在径流预测中的应用与评估

小　组　成　员：　包一宁 赵知微 林月磊

林　谷 唐玉彬

摘要

本研究致力于评估普通线性回归及其正则化改进模型 Lasso、Ridge 以及融合模型 Elastic Net 在径流预测中的应用与性能。

本研究为应对提供的水文数据存在的严重右偏分布和异常值问题，采用了严格的数据预处理策略，包括对径流量和降雨量进行$ln(x+1)$对数变换以校正偏度，并对所有特征进行 Z-Score 标准化以消除量纲差异。

本研究基于已处理的数据集，通过 Python 中 SGDRegression 类构建了四种模型并进行评估。评估结果显示，在经过有效的数据预处理后，普通线性回归模型表现出最优异的预测性能。普通线性回归模型在测试集上取得了最高的决定系数达 0.715，同时具有最低的均方根误差 2051 和平均绝对误差 1330。

研究发现，强大的特征工程已充分缓解了数据中的共线性和过拟合风险，使得正则化技术引入的L_1或L_2惩罚项未能显著提升模型性能，介于其中的 Elastic Net 模型也一样表现不佳。这一发现强调了在水文径流预测中，数据预处理对线性模型的稳定性和泛化能力的重要性。尽管普通线性回归模型性能最佳，但其在预测径流极高值（例如洪水事件）时仍存在一定的低估现象。本研究为数据驱动的水文预测提供了一种高效且可解释的方法，并指出了未来可通过引入时滞和非线性特征进一步优化的方向。

关键词：**SGDRegression**、L_1和L_2正则化、**Elastic Net**

第一章 问题背景及意义

1.1 研究背景

水资源是维系生态平衡与支撑社会经济发展的关键自然要素，其可持续利用已成为全球关注的焦点[1]。径流作为水文循环的核心组成部分，直接关系到区域水资源的分布、可利用量及灾害风险程度。随着全球气候变化与人类活动干扰的加剧，径流过程呈现出更强的非线性和时空异质性，使得传统水文模型在预测精度与适应性方面面临严峻挑战[2]。在此背景下，数据驱动模型因其不依赖复杂的物理机制、仅通过历史数据构建预测关系的优势，逐渐成为径流预报的重要研究方向之一[3]。

线性回归作为一类基础且广泛使用的数据驱动方法，因其模型结构简单、解释性强，被广泛应用于径流预测中。然而，普通线性回归模型在处理高维特征或存在多重共线性的数据时，容易出现过拟合问题，导致模型泛化能力下降[4]。为提升预测稳定性与准确性，正则化技术被引入线性回归框架中，如 Lasso 回归通过L_1正则化实现特征选择，Ridge 回归通过L_2正则化抑制参数膨胀，而 Elastic Net 则结合二者优势，在保持模型简洁性的同时增强鲁棒性[5]。此外，通过集成学习与多模型耦合策略，进一步整合不同回归模型的优势，已成为提高径流预测性能的有效途径。

当前，尽管已有研究尝试将正则化线性回归应用于水文预报，但多数工作仍集中于单一模型的对比，缺乏系统性的多模型耦合与集成策略探讨，尤其在面对不同流域特性与水文情势时，模型的适应性与解释性仍有待深化[6]。因此，开展基于线性回归的集成与正则化改进研究，对提升径流预测的准确性与可靠性具有重要理论与实用价值。

1.2 问题意义

本研究围绕线性回归模型在径流预测中的优化与集成展开，重点探讨普通线性回归、Lasso、Ridge 及其耦合模型的表现差异与适用条件，具有如下意义：

在理论层面，本研究通过系统比较不同正则化策略与模型融合方式，有助于揭示线性回归模型在水文序列预测中的泛化机制与抗干扰能力，推动数据驱动水文模型向更稳健、更可解释的方向发展。同时，模型耦合策略的研究为多算法协同建模提供新思路，丰富了水文预报的方法体系[7]。

在实践层面，径流预测的准确性直接关系到水资源调度、洪水防控、干旱应对及水利工程运行等关键决策。本研究提出的正则化线性回归及其集成模型，能够有效提升预测精度与时效性，为区域水资源管理、灾害预警系统优化提供技术支撑，尤其在数据稀缺或异质性强的流域中展现出良好的应用潜力。此外，通过模型可解释性增强，有助于决策者理解预测结果背后的驱动因素，提升水资源管理的科学性与透明度。

综上所述，本研究不仅对线性回归模型在水文领域的应用进行了深化与拓展，也为应对复杂环境下的水资源挑战提供了方法论支持，具有较强的现实意义。

第二章 数据预处理

2.1 数据校验

本研究基于文件“qingshandataforregression.xlsx”中的降雨、蒸发和径流数据进行线性回归模型拟合。为确保数据质量并提升模型拟合的准确性和效率，在建模之前，必须对数据进行必要的预处理。这将有助于优化拟合过程，并保证最终模型的可靠性和良好性能。

2.1.1 缺失值校验

缺失值处理是数据预处理的首要步骤。本研究利用 Python 程序对文件中三个变量（降雨量、蒸发量、径流量）进行了校验。得到表 1：

表 1：降雨量、蒸发量和径流量缺失值检验结果表

数据类型	缺失值数目	数据类型
降雨量	0	dtype: int64
蒸发量	0	dtype: int64
径流量	0	dtype: int64

校验结果显示，三个变量的缺失值数目均为 0。这表明数据集的完整性高，无需进行缺失值填充或删除处理，可以直接进入下一步的异常值校验。

2.1.2 异常值校验

依据统计学原理，在进行异常值校验之前，先要对数据集进行偏度和正态性检测，即分布特征分析，通过 Python 得到图 1：

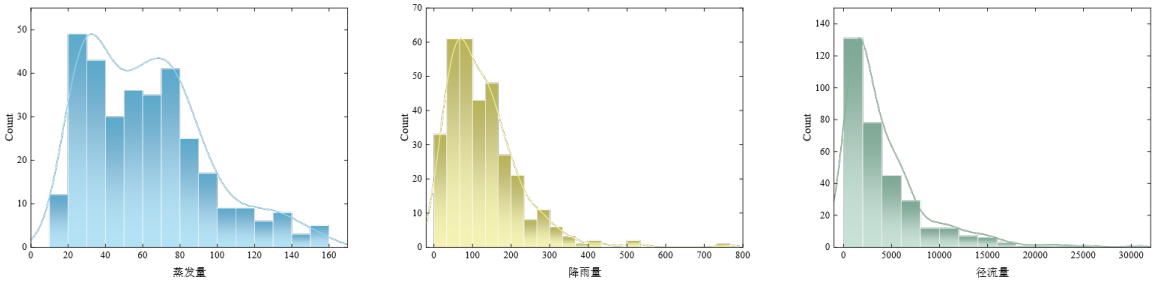

图 1：数据分布特征图

结果表明，三项数据均存在一定程度的右偏，其中降雨量与径流量数据右偏较为严重。

基于此判断，本研究后续采用四分位数间距（Interquartile Range, IQR）方法对径流预测模型的输入变量进行异常值检测。IQR 方法基于数据的分位数特征，具有对极端值不敏感的优点，适用于非正态分布数据的异常值识别。检测标准为：

$$Q_1 - 1.5 \times IQR < \text{正常值} < Q_3 - 1.5 \times IQR$$

其中，Q_1和Q_3分别为第一四分位数和第三四分位数，$IQR = Q_3 - Q_1$为四分位数间距。

本研究利用 IQR 方法绘制箱线图对数据进行可视化以定位异常值，结果如图 2 所示：

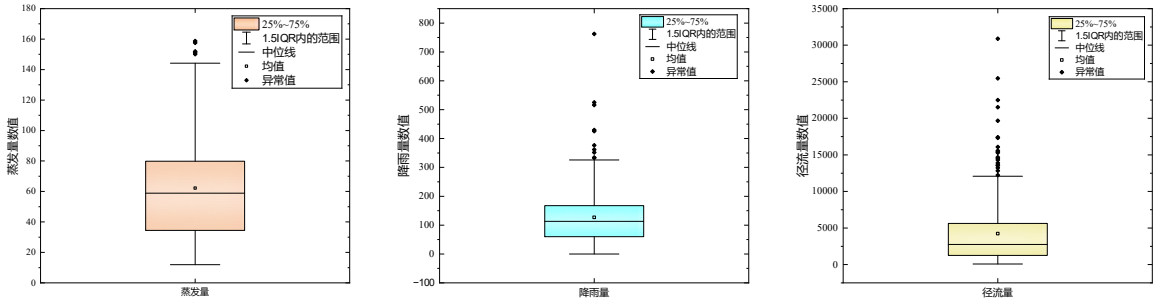

图 2：蒸发量、降雨量和径流量箱线图

显然在不同数据列中均存在一定量的异常值，这也符合偏度预期，进一步分析得到表 2：

表 2：异常值分析表

变量	异常值数量	异常值比例(%)	正常值范围	最大异常值	最小异常值
蒸发量(mm)	5	1.52	[-33.02, 147.57]	158.70	150.10
降雨量(mm)	10	3.05	[-99.48, 326.93]	762.50	333.10
径流量(m^3/s)	20	6.10	[-5284.50, 12173.50]	30890.00	12230.80

其中，**蒸发量变量**表现出较好的数据质量，仅有 5 个异常值（占比 1.52%），均为上界异常值，数值范围为 150.10-158.70mm。这些异常值主要出现在极端高温或干旱条件下，符合气象学规律，属于自然现象的极端表现。**降雨量变量**检测出 10 个异常值（占比 3.05%），最大异常值达 762.50 mm，为均值的 6.02 倍。这些异常值反映了研究区域极端降水事件的存在，如暴雨、台风等极端天气现象。降雨量的变异系数为 73.56%，表明其

时空变异性较大。**径流量变量**异常值数量最多，共 20 个（占比 6.10%），最大异常值为 30890.00m³/s，约为均值的 7.31 倍。径流量的变异系数高达 103.71%，表明其具有显著的不均匀分布特征。这些异常值主要对应于洪水事件，是降雨量异常值的直接响应。

2.2 特征工程

数据校验结果显示，径流量和降雨量均存在明显的右偏分布，特别是径流量的偏度值接近 2.315。同时，为保留水文系统极端事件的完整信息，本研究采取了保留异常值的策略。鉴于这些特性，必须对数据进行变换，以缓解异常值对模型拟合的不利影响，并使数据分布更接近线性回归模型所要求的正态分布假设。此外，特征变量之间的量纲和取值范围差异必须消除，以确保后续 Lasso 和 Ridge 等正则化模型的良好拟合。本研究采用双重策略进行数据变换：

（1）首先对于具有严重右偏特性的径流量和降雨量，采用对数变换进行处理。对数变换能有效将右偏分布拉向对称，显著降低偏度，使数据分布更接近正态。同时，它能够压缩极端高值的尺度，从而减轻保留的异常值对模型系数的过度影响，提升模型的稳健性。具体而言，即采用$x' = ln(x + 1)$的形式，其中“+1”是为了处理数据中可能存在的零值，以避免$ln(0)$无意义。

表 3：对数变换后的数据偏度表

变量	偏度
蒸发量	0.7649
$ln\left(\text{降雨量}\right)$	-1.0561
$ln\left(\text{径流量}\right)$	-0.0984

根据表 3 可以发现经过对数变换后径流量偏度得到了有效减少，接近理想的对称状态，极大地优化了模型的拟合条件。虽然降雨量偏度从 1.9855(高度右偏)到 1.0561(中度左偏)显著改善了数据分布，但距离精确的正态分布仍存在一定偏差。不过，考虑到气象数据保留极端值信息的需求，-1.0561 的偏度仍在可接受范围内。

（2）由于不同变量之间存在量纲区别，并且蒸发量仍保持中度右偏（0.7649），为了使其与经过变换的降雨量和径流量特征保持尺度一致性，并满足正则化模型的要求，需要对其进行 Z-Scroe 标准化。

Z-Score 标准化，也称为标准化或 Z 标准化，它能将原始数据集中的数据点转换成标准分数，从而使转换后的数据集具有均值为 0 和标准差为 1 的特性。Z-Score 的计算公式非常直观：

$$Z = \frac{x - \mu}{\sigma}$$

其中：Z 是原始数据点 x 对应的 Z-Score，x 是数据集中的原始数据点（或称为观测值），μ是数据集的平均值，σ是数据集的标准差。

通过 Python 可以直接计算出标准化后的数据的均值和标准差均符合预期。至此，数据集已进行一定的特征工程，可进行数据划分为训练集和测试集以供后续模型拟合。

第三章 多模型集成径流预测方法构建

在第二章中，我们对原始径流、降雨和蒸发数据进行了彻底的校验、特征工程和标准化处理。特别是通过对径流量和降雨量进行对数变换，有效校正了数据的严重右偏分布，并结合 Z-Score 标准化消除了变量间的量纲差异，确保了输入特征集和目标变量具备满足线性模型假设的良好统计特性。

本章将基于已处理的数据集，正式构建用于径流预测的核心模型。本研究选取了普通线性回归、Lasso 回归、Ridge 回归作为基础模型。其中，正则化线性模型（Lasso 和 Ridge）的引入旨在通过范数约束来优化参数估计，以缓解普通线性回归在面对特征共线性和过拟合时的局限性。

此外，为充分发挥各模型在特征选择和参数收缩方面的互补优势，并进一步提升预测的稳健性和准确性，本研究将探索一种多模型融合或集成策略。本章将首先详细阐述模型的选取与集成思路，随后深入剖析普通线性回归、Lasso 和 Ridge 这三种核心模型的数学原理与正则化机制。

3.1 基础模型选取

本研究首先选取了三种基础的线性回归模型，它们都基于普通线性回归假设，但在参数约束上各有侧重。

线性回归假设因变量和自变量之间存在以下线性关系：

$$Y = \beta_0 + \beta_2 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

其中，Y是目标变量，X_i是特征变量，β_0是截距项，β_i是回归系数。

3.1.1 普通线性回归模型

普通线性回归的核心是利用最小二乘法来估计模型中的系数$(\beta_0, \beta_1, \dots)$。目标是找到一条最能拟合现有数据的“最佳拟合直线”（或高维空间中的平面）。最小二乘法能够直接通过求解解析解的方式最小化残差平方和（Residual Sum of Squares, RSS），即损失函数，表示为：

$$\min_{\beta}\sum_{i=1}^{m}(y_i-\widehat{y_i})^2=\min_{\beta}\sum_{i=1}^{m}\left(y_i-\left(\beta_0+\sum_{j=1}^{n}\beta_j X_{ij}\right)\right)^2$$

其中，m是是样本观测值的数量，y_i是第i个观测的真实径流量（因变量），$\hat{y}_i$是模型对该观测的预测值。

普通线性回归模型的优点在于计算简单和结果具有良好的可解释性。然而当数据集中的特征变量数量较多或特征之间存在显著的多重共线性时，本方法估计的系数方差可能会变得很大，导致模型过拟合，泛化能力下降。因此本研究后续引入了 Lasso 和 Ridge 等正则化线性模型来提升预测稳定性和准确性。

3.1.2 Lasso 线性回归模型

Lasso（Least Absolute Shrinkage and Selection Operator）回归是一种通过在普通最小二乘法的目标函数中加入L_1范数惩罚项来实现参数估计和特征选择的正则化方法 。它的引入旨在解决普通线性回归在处理高维特征或存在多重共线性数据时，容易出现过拟合、泛化能力下降的局限性。其数学原理同样是最小化损失函数，在 RSS 基础上中加入了系数向量β的L_1范数，其形式表示为：

$$\min_{\beta}\left\{\sum_{i=1}^{m}\left(y_i-\left(\beta_0+\sum_{j=1}^{n}\beta_j X_{ij}\right)\right)^2+\lambda\sum_{j=1}^{n}|\beta_j|\right\}$$

其中$\lambda\sum_{j=1}^{n}|\beta_j|$是$L_1$正则化项，$\lambda(\lambda\geq 0)$是正则化参数（或惩罚系数），控制着惩罚项对模型的约束强度。λ 越大，系数被压缩得越多。Lasso 回归的主要优势在于其独特的L_1惩罚机制，这使其倾向于将不重要特征的系数直接压缩至零，从而实现了自动化的特征选择，能够剔除次要特征，增强模型的可解释性和简洁性；抑制膨胀系数，而减轻了多重共线性的影响；限制模型的复杂度，更好地平衡模型的方差和偏差，降低了模型在训练集上过拟合的风险，提升了模型在未见过数据上的预测鲁棒性和准确性。

3.1.3 Ridge 线性回归模型

Ridge（岭）回归是一种通过在普通最小二乘法的目标函数中加入L_2范数惩罚项来限制模型系数大小的正则化技术。与 Lasso 一样，它的主要目的是解决普通线性回归在存在多重共线性或过拟合时不稳定和泛化能力差的问题。

Ridge 回归的损失函数是最小化 RSS 与系数向量β的L_2范数的平方之和，其形式为：

$$\min_{\beta}\left\{\sum_{i=1}^{m}\left(y_i-\left(\beta_0+\sum_{j=1}^{n}\beta_j X_{ij}\right)\right)^2+\lambda\sum_{j=1}^{n}\beta_j^2\right\}$$

Ridge 回归与 Lasso 回归在参数收缩，提升模型鲁棒性上基本一致，但处理相关特征与 Lasso 不同，当面对一组高度相关的特征时，Ridge 回归倾向于将这组特征的系数平均分配并进行同等程度的收缩，而不是仅选择其中一个，这在所有相关特征都对预测结果有贡献时更加合理。

3.2 Elastic Net 弹性网格模型

Elastic Net（弹性网格）模型是一种巧妙地融合了 Lasso 回归和 Ridge 回归的惩罚项的优势。这种融合策略旨在克服 Lasso 在处理高度相关特征时可能只选择其中一个的局限性，同时保留 Lasso 的特征选择能力和 Ridge 的系数收缩稳定性。

Elastic Net 的核心思想是在 RSS 中，同时引入L_1和L_2范数惩罚项。其损失函数形式表示为：

$$\min_{\beta}\left[\sum_{i=1}^{m}\left(y_i-\beta_0-\sum_{j=1}^{n}\beta_j X_{ij}\right)^2+\lambda\left((1-\alpha)\sum_{j=1}^{n}\beta_j^2+\alpha\sum_{j=1}^{n}|\beta_j|\right)\right]$$

α是混合参数，通过α参数的调整，Elastic Net 可以灵活地在 Lasso 和 Ridge 之间进行切换：当$\alpha=1$时，模型退化为 Lasso 回归；当$\alpha=0$时，模型退化为 Ridge 回归；当$0<\alpha<1$时，模型为 Elastic Net，同时具有L_1和L_2范数的优势。在本研究中，我们采用$\alpha=0.5$来训练模型。

在径流预测的应用中，Elastic Net 模型的引入使其能够灵活的处理特征的重要性并保持其他回归的优势。因此，在本研究中引入 Elastic Net 模型，是为了探索其在结合正则化和特征集成方面，能否提供比单一 Lasso 或 Ridge 模型更优异的径流预测性能。

第四章 模型评估及结果分析

本章旨在对第三章构建的四种径流预测模型——普通线性回归、Lasso、Ridge 及其集成模型 Elastic Net 进行定量评估与比较。所有模型均基于对数变换后的径流量进行训练，但在评估阶段，所有指标均在原始径流空间计算，以便更贴合实际水文情势的预测需求。

本研究通过 Python 中 Scikit-learn 库中的 SGDRegressor 类，可以同时实现四种方式的训练及其评估。

4.1 模型评估

为全面衡量模型的预测能力和泛化性能，本研究采用了水文领域常用的纳什效率系数（NSE），其取值范围$(-\infty, 1]$，越接近 1 模型性能越理想，是衡量预测值与实际值序列的吻合程度重要的评估标准之一。除此之外还使用了机器学习中经典的指标：决定系数（R^2），用于评价拟合效果；均方根误差（RMSE），对大误差（异常高值）敏感，衡量预测值的集中趋势和模型稳定性；平均绝对误差（MAE），误差的平均值，直观反映预测值与实际值的偏离程度。

4.1.1 四模型评估对比

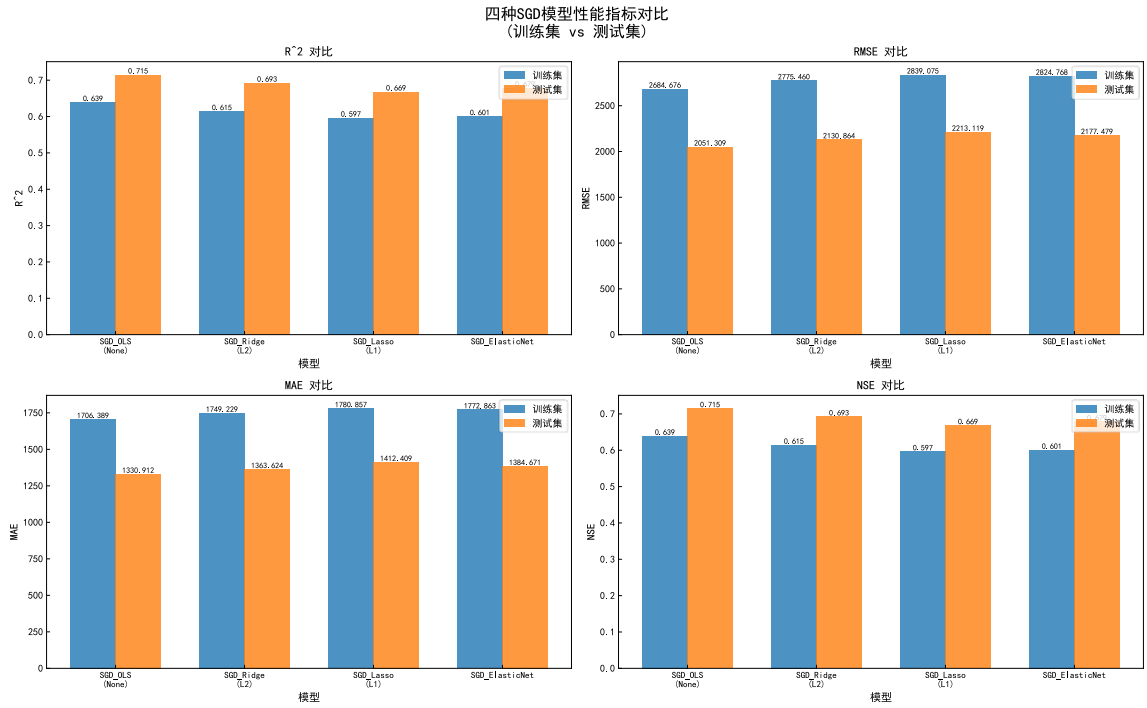

图 3：四种 SGD 模型性能指标对比

在明确评估指标后，本研究在符合一般机器学习流程的基础上，同时评估了训练集和测试集的性能，如图 3 所示。

4.2 结果分析

4.2.1 性能对比与最佳模型确定

根据图 3 评估结果，普通线性回归模型在所有指标上均略优于其他三个正则化模型，其测试集R^2达到了 0.715，并同时具有最低的 RMSE 和 MAE。这一结果表明，在径流预测任务中，经过对数变换和 Z-Score 标准化预处理的数据集，其特征间的共线性和过拟合风险已经得到有效缓解，使得普通最小二乘法在泛化性能上优于引入L_1或L_2惩罚项的模型。**正则化技术**在本研究中并未对模型性能带来显著提升，这侧面印证了**数据预处理**对线性模型的重要性。

此外，无论哪种模型，测试集的R^2均高于训练集，并且所有模型在所有误差指标（RMSE、MAE）上，训练集的误差都明显低于测试集的误差，这可能是随机且单一的数据集划分导致的，或者是训练过程中参杂了过多噪声，在径流预测中可能是一些异常天气情况。

4.2.2 超参数优化过程分析

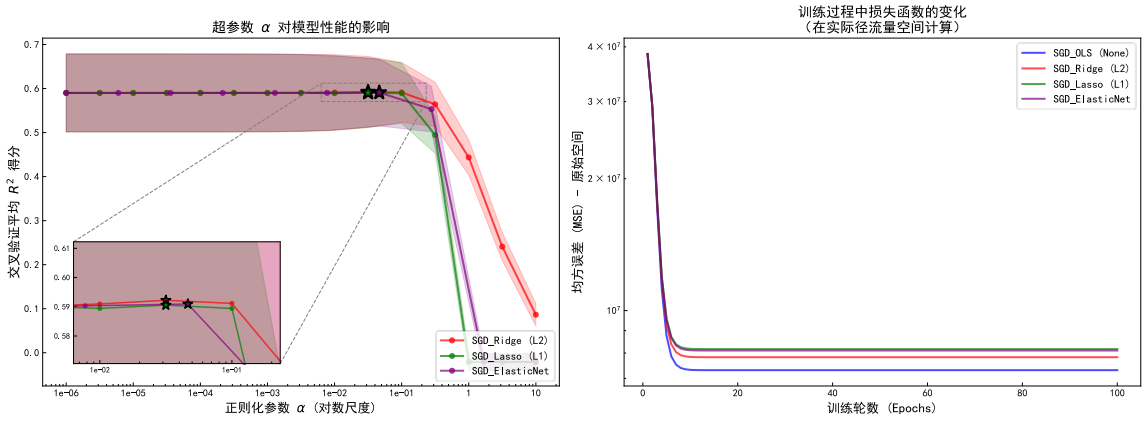

图 4：超参数寻找过程图

图 4 展示了正则化模型超参数对模型在交叉验证上的R^2得分的影响。其中三种模型的R^2曲线基本一致，最优点也十分接近。最佳α均集中在 1e-02 到 1e-01 的小范围中，这印证了前述观点：数据预处理后的数据集结构良好，对正则化的需求极低，微小的值即可满足模型优化要求，使其性能无限接近于普通线性回归模型。

损失函数变化过程图展示了在经历数轮便找到了最小 MSE，与超参数寻找表现的性状一致。

4.2.3 梯度下降收敛性分析

我们利用最优超参数配置的 SGRegression 跟踪了四种模型在训练集上的损失函数变化，如图 5 右图所示：

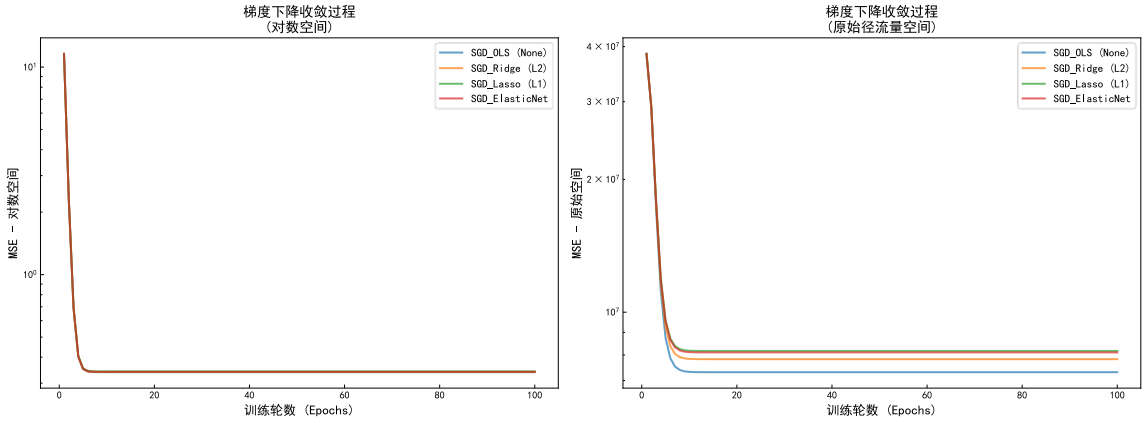

图 5：SGD 梯度下降对比

结果显示，四种模型均展现出快速的初期收敛，在前 10 个 Epochs 内，损失函数值急剧下降，表明模型参数迅速调整到接近最优解的位置。尽管所有模型最终都趋于收敛，但在对数空间和原始空间中观察，普通线性回归模型的损失函数收敛速度稍快，并在稳定后维持在最低水平，这与前文表现出其最优的性能指标相一致。

Ridge、Lasso 和 Elastic Net 在收敛过程中，由于正则化项的引入，对权重的约束使得其损失值略高于普通线性回归模型，这也解释了其最终预测性能略逊于普通线性回归模型的结果。三种正则化模型在收敛过程中彼此差异不显著，表明在本研究的特定数据集和特征工程处理后，L_1与L_2范数对于收敛速度的影响差异不大。

4.2.4 最佳模型预测表现与残差分析

根据 4.2.1 的评估结果，**普通线性回归**模型被确定为最佳模型，其在测试集上取得了最高的R^2（0.715）和最低的 RMSE 与 MAE，为进一步探索最佳模型的性质，本研究绘制了普通线性回归预测散点图和残差分析图，即图 6 和图 7，均包含训练集和测试集两部分。图 6 显示，在测试集的实际值与预测值散点图中，数据点主要集中在对角线（Y=X）附近，表明模型的预测结果与实际径流量具有高度的一致性。然而，在径流量极高值（对应洪水事件）区域，散点图显示部分预测值低于实际值，这反映了模型对极端高值预测的能力有限，可能存在一定的低估现象。

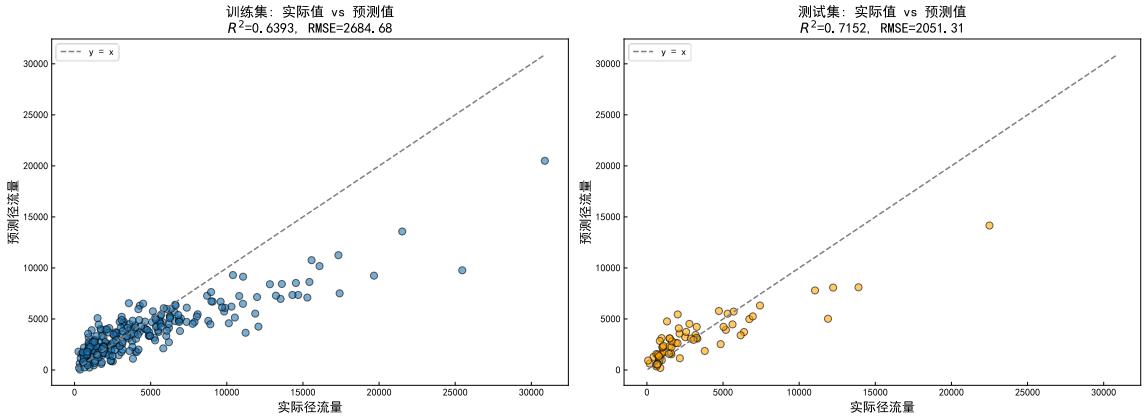

图 7：普通线性回归预测散点图

图 7 中残差分布表现出径流原始数据分布一样的右偏特性，符合散点预测图的特征。由于极端天气在默认模型假设的保留，导致了该结果的出现。图中残差随预测值没有展现出明显的异方差性（残差的离散程度没有随预测值的增大或减小而规律性变化），但可以观察到：在低径流量区间，残差波动较小。在高径流量区间，残差波动显著增大，再次佐证了模型在预测极端洪水事件时误差较大。

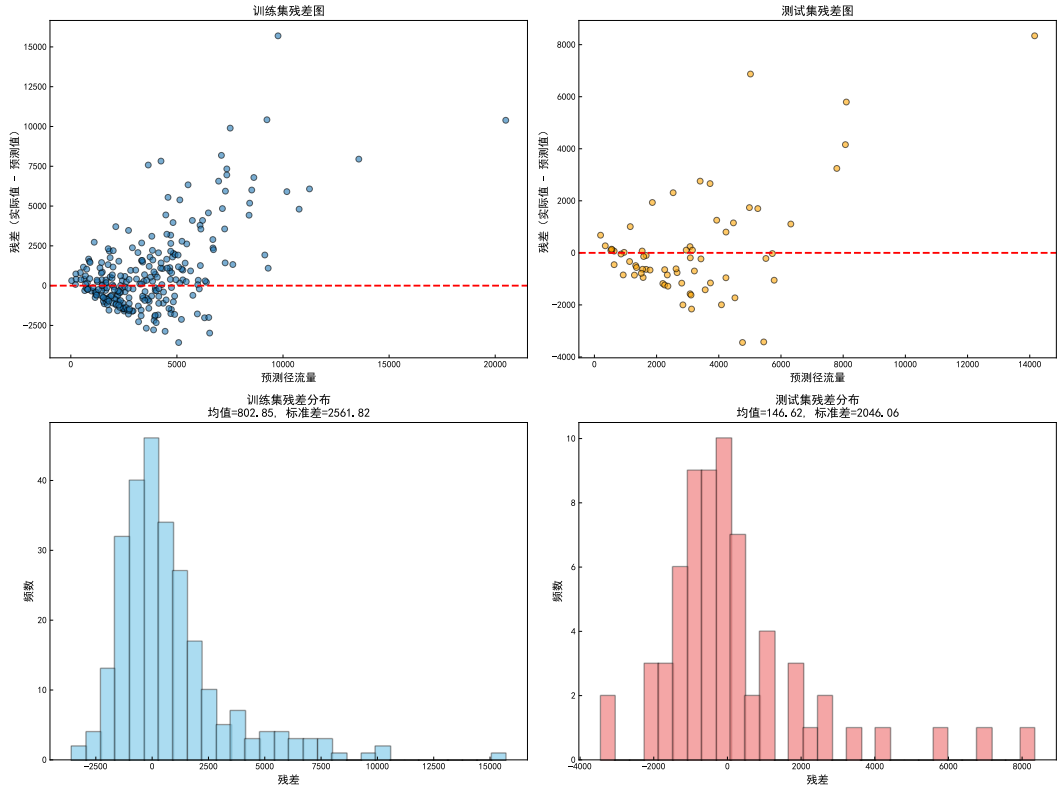

图 6：普通线性和回归残差分析图

第五章 模型讨论

5.1 结果成因分析

经过上述结果分析，本研究发现了普通线性回归表现最优，而正则化模型普遍性能较弱的结果，本章节我们将进一步探索其结果成因。

经过彻底的对数变换和 Z-Score 标准化预处理后，普通线性回归模型在测试集上取得了最佳性能，其R^2达到了 0.715，而正则化模型普遍偏弱。这一结果暗示了以下几点关键成因：

（1）**数据分布有效校正：**径流量和降雨量严重的右偏分布经过$ln(x+1)$变换后，数据分布形态被有效校正，极大地缓解了模型对极端高值的敏感性，使数据更接近线性模型所需的正态分布假设。这种预处理策略从根本上降低了数据中的多重共线性风险和异常值带来的干扰，使得普通线性回归的系数估计方差保持在可接受的低水平，从而避免了过拟合。

（2）**正则化需求降低：**Lasso、Ridge 和 Elastic Net 等正则化模型通过引入惩罚项来约束系数，以应对特征共线性和过拟合。由于强大的预处理工作已经“清洗”了数据结构，原始数据集对正则化约束的需求极低。这可以从超参数优化结果中得到印证：最佳正则化强度集中在 1e−02 到 1e−01 的微小范围内，导致正则化模型性能无限接近但低于普通线性回归模型，并未带来显著提升。

（3）**模型复杂度较低：**径流预测仅使用了降雨量和蒸发量两个特征变量。在特征数量极少且特征工程到位的情况下，引入正则化惩罚项反而可能对模型造成轻微的欠拟合，略微牺牲了普通线性回归在训练集上的拟合能力，从而导致正则化模型的性能略逊于普通线性回归，这也是为什么训练集结果全部低于测试集的重要成因之一。

5.2 模型局限性

尽管普通线性回归在经过数据预处理后表现优异，但其基于线性假设的本质仍存在以下局限性：

（1）**无法有效应对极端气候的径流事件**，比如洪水。第五章中预测散点图和残差图都表现出径流量极高区域预测值普遍偏低的情况，这表明线性模型难以完全捕捉径流过程中降雨-径流转换的高度非线性和时滞效应，尤其是在极端天气条件下。

（2）**缺乏时间依赖性考量**。本研究采用的线性回归模型是静态模型，它假设输入变量与目标变量之间存在瞬时、固定的线性关系。然而，真实的径流系统是一个复杂的动态

系统，当前径流不仅取决于当前的降雨和蒸发，还取决于前期的土壤湿度、地下水储量等时间积累效应。缺乏对时间序列依赖性的考量是预测精度的主要限制之一。

（3）**特征数量受限：**模型仅使用了降雨量和蒸发量。虽然是径流形成的主要特征变量，但径流过程还受到气温、融雪、下垫面性质和地形等多种因素的影响。模型的预测能力受到输入特征的显著限制。

5.3 模型优化方向

基于上述局限性，本研究提出以下优化方向：1、引入非线性特征，比如降雨量×蒸发量来部分模拟降雨-径流关系的非线性，以提升模型在极端值处的拟合能力。2、引入前一时刻或前几时刻的降雨量和径流量作为新的特征变量，将静态模型扩展为动态回归模型(DLM)，以捕捉水文系统的记忆和时滞效应。3、改进异常值处理策略，修改模型假设，对数据集进行进一步拆分。对径流量的极端高值，可以考虑采用分位数回归或混合模型。例如，使用一个模型预测常规径流，另一个模型专注于预测极端高值，以避免对数变换平滑掉重要的洪水信息。

结论

本研究系统地比较了普通线性回归、Lasso 回归、Ridge 回归及其集成模型 ElasticNet 在径流预测中的性能表现，并深入探讨了数据预处理对模型预测结果的影响，结果表明：

（1）数据预处理是性能优化的关键。针对径流量和降雨量存在的严重右偏分布，采用$ln(x+1)$对数变换和 Z-Score 标准化的特征工程策略被证明是极其有效的。它成功地将径流量的偏度从高度右偏显著减少至接近对称状态，极大地优化了线性模型的拟合条件。

（2）普通线性回归模型表现最优。在经过优化的特征工程后，普通线性回归模型在测试集上的R^2达到 0.715，并在 RMSE 和 MAE 指标上均优于其他正则化模型。这一结果表明，对于本研究的特定数据集，正则化技术（Lasso,Ridge,Elastic Net）带来的参数约束并未提供额外的泛化性能优势。更进一步探索反映出该数据集和基础模型正则化需求极低。

尽管普通线性回归模型性能最佳，但其本质上的线性假设限制了对水文系统中非线性和时滞效应的捕捉能力。模型在预测极端高径流值时存在普遍的低估的倾向。

综上所述，本研究不仅验证了线性回归模型在径流预测中的潜力，更突显了在应用数据驱动模型时，特征工程和数据质量管理的重要性不亚于单纯的模型选择。未来的工作应着重于通过引入非线性或时滞特征，或探索非线性集成模型，以提高模型对水文过程复杂性的模拟能力，尤其是在应对极端水文事件的预测精度上。

参考文献

[1] 刘昌明, 陈志恺. 中国水资源现状评价和供需发展趋势分析[M]. 北京: 中国水利水电出版社, 2001.

[2] 夏军, 谈戈. 全球变化与水文科学新的进展与挑战[J]. 水科学进展, 2002, 13(5): 667-674.

[3] 程慧先. 基于傅里叶变换、近似熵和线性回归的数据驱动径流预测模型及机理揭示[D]. 北京: 中国水利水电科学研究院, 2020.

[4] 黄瑾. 岷江上游生态水遥感定量反演及径流预测模型研究[D]. 成都: 成都理工大学, 2017.

[5] TIBshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[6] 王宁,张永玲,肖让.水文水资源模型研究进展与展望[J].河南科技,2025,52(06):93-98.DOI:10.19968/j.cnki.hnkj.1003-5168.2025.06.017.

[7] 赵立杰,仕玉治,李福林,等.基于云耦合协调模型的水资源关联系统协调性动态评价[J].南水北调与水利科技(中英文),2024,22(04):747-758.DOI:10.13476/j.cnki.nsbdqk.2024.0076.