网络首发时间：2020-08-14 14:39:24
网络首发地址：https://kns.cnki.net/kcms/detail/42.1673.TV.20200813.1540.002.html
2020 年 8 月 水 利 水 运 工 程 学 报 No. 4
HYDRO-SCIENCE AND ENGINEERING Aug. 2020
DOI:10.12170/20190626001
李伶杰, 王银堂, 胡庆芳, 等. 基于随机森林与支持向量机的水库长期径流预报 [J]. 水利水运工程学报, 2020(4): 33-40. (LI Lingjie, WANG Yintang, HU Qingfang, et al. Long-term inflow forecast of reservoir based on Random Forest and support vector machine[J]. Hydro-Science and Engineering, 2020(4): 33-40. (in Chinese))

基于随机森林与支持向量机的水库长期径流预报

李伶杰[1], 王银堂[1], 胡庆芳[1], 刘定忠[2], 张安富[2], 巴亚荃[3]
(1. 南京水利科学研究院 水文水资源与水利工程科学国家重点实验室, 江苏 南京 210029; 2. 云南龙江水利枢纽开发有限公司, 云南 德宏 678400; 3. 深圳市深水水务咨询有限公司, 广东 深圳 518003)

摘要： 水库长期径流预报对于研判水文情势变化和指导水库调度管理具有重要意义。针对云南龙江水库年、汛期和枯水期平均入库径流, 利用随机森林从环流指数、海温、气压和前期月径流中选取关键预报因子, 基于粒子群与交叉验证相结合的算法优选参数, 建立随机森林与支持向量机模型, 开展龙江水库入库径流预报研究。结果表明: 太平洋中北部与西部气候因子对径流预报的影响较大, 前期月径流对年、汛期径流的重要性偏低, 但对枯水期的影响程度与部分气候因子相当。随机森林与支持向量机模型总体精度较高, 模拟与预报的合格率均达到 85% 以上, 平均绝对百分比误差均低于 15%, 支持向量机的泛化能力强于随机森林, 但二者在局部极值流量处的预报精度尚有待提升。

关 键 词： 龙江水库; 长期径流预报; 随机森林; 支持向量机
中图分类号： P338 **文献标志码：** A **文章编号：** 1009-640X(2020)04-0033-08

云南龙江水库是龙江-瑞丽江流域的重要防洪控制性工程, 兼有防洪和发电等综合效益。在保障水库自身及下游防洪安全的前提下, 开展水库中长期优化调度是实现综合效益最大化的重要途径。而中长期优化调度效果又十分依赖于入库径流预报, 因此开展中长期入库径流预报具有重要的实际应用价值。

对于中长期径流预报, 根据预报机理的不同, 可划分为水文循环过程驱动和相关影响因子(前期径流与气候因子)驱动两类[1]。前者通过将降雨预报信息输入到具有明确产汇流机制的水文模型实现预报, 后者通过构建径流与影响因子的统计学习模型, 以影响因子的历史实测值作为驱动, 从而实现中长期径流预报。在机器学习算法迅速发展与海量大尺度气候信息(大气环流指数、海温、气压等)迅速累积的背景下, 基于相关影响因子驱动的中长期径流预报方法逐渐成为研究重点, 这类方法涉及的预报因子筛选、预报模型优化等已经取得了长足的发展。在预报因子筛选方面, 最优子集回归[2]、逐步回归[3]、LASSO(Least Absolute Shrinkage and Selection Operator)回归[4] 等算法的引入为识别影响径流预报的关键因子提供了丰富途径。在预报模型优选方面, 基于人工神经网络[5-6]、支持向量机(Support Vector Machine, SVM)[7-9]、随机森林(Random Forest, RF)[10-11] 等机器学习算法的预报研究大量开展。如刘勇等[2] 将最优子集回归和神经网络耦合, 建立了预报精度与稳定性均令人满意的丹江口秋汛期入库径流量预报模型。赵铜铁钢等[10] 应用随机森林模型开展了长江上游枯水期径流预报及不确定性分析, 取得了较好的应用效果。崔东文[9] 研究发现利用智能优化算法估计参数条件下, 支持向量机对中长期月径流的预测精度较高。此外, 何国栋[12]、赵鹏雁[13] 等开展了多种算法预测性能的比较。然而随着多源气候气象信息的引入, 预报因子空间向超高维度发展, 传统的回归类因子筛选方法已不能适应这种发展趋势, 神经网络、支持向量机等预报模型参数确定及避免过拟合问题仍然有待于进一步解决。另外不同模型的精度随数据特性变化而不同, 对于具体研究对象需强化

收稿日期： 2019-06-26
基金项目： 国家重点研发计划资助项目(2016YFC0400902; 2016YFC04009010); 国家自然科学基金资助项目(51809252); 中央级公益性科研院所基本科研业务费专项资金资助项目(Y519007)
作者简介： 李伶杰 (1992—), 男, 山西吕梁人, 工程师, 主要从事水文水资源研究。E-mail: ljli@nhri.cn

预报模型的评估和筛选工作。

鉴于此，本文以云南龙江水库年、汛期和枯水期平均入库径流为长期径流预报的研究对象，利用随机森林模型能有效评估影响因子重要性的特点，从大气环流指数、海温、气压和前期月径流等高维度影响因子空间中筛选预报因子。在此基础上，以随机森林和支持向量机 2 种机器学习算法为预报工具，采用粒子群优化算法与交叉验证相结合的方法估计模型参数，对比评估各模型预报效果，为龙江水库入库径流预报及水库优化调度提供技术支撑，同时对机器学习算法在中长期径流预报中的应用提供有益借鉴。

1 研究区域概况

云南龙江水库位于龙江-瑞丽江流域(瑞丽江一级水电站以上集水区)的龙江干流下游河段(图 1)，是流域内规划的第 13 座梯级水电站，于 2009 年正式投入使用，兼有防洪、发电、灌溉、旅游等综合效益。坝址以上龙江河段长约 300 km，河床平均坡降约 5‰，控制流域面积 5 758 km^2，占龙江-瑞丽江流域面积的 49%[14]。

龙江水库无入库径流水文站，本文利用 1960—2010 年腾龙桥站逐月流量，按照径流自上游向下形成演化流程推算。首先采用水文比拟法得到弄另水库的长系列入库径流和弄另-龙江区间天然径流，其次利用弄另水库多年平均调节系数(出库与入库流量的比值)计算出库流量，最后以弄另水库出库流量与弄另-龙江区间天然径流之和作为龙江水库入库径流，从而得到了 1960—2010 年龙江水库入库径流序列。将其与采用水量平衡法反推的 2011—2018 年入库径流连接得到 1960—2018 年水库入库径流序列。鉴于两段序列计算方式不同可能导致的非一致问题，采用 Mann-Kendall 检验法诊断(图 2(a))，发现入库径流序列在 2010 年后发生跳跃变异，年径流均值降幅达到了 19.8%，而其附近其他水文站序列均满足一致性假设，且同期未出现明显径流减少的现象，因此认为龙江水库入库径流序列的非一致性问题主要是由水量平衡法中水库渗漏及水面蒸发存在较大误差引起。将 2011—2018 年序列年径流均值修正到与 1960—2010 年序列相同，按照修正前后年径流均值的比例对径流年内分配同倍比缩放，并按照 6 月—翌年 5 月的顺序计算，得到满足一致性假设且考虑弄另水电站调节影响的龙江水库 1960—2017 年水文年入库径流序列。本文以年、汛期(6—11 月)与枯水期(12 月—翌年 5 月)平均入库径流作为长期径流预报的研究对象(图 2(b))。经统计，多年平均年、汛期和枯水期径流量分别为 198.9，316.8 和 81.4 m^3/s。

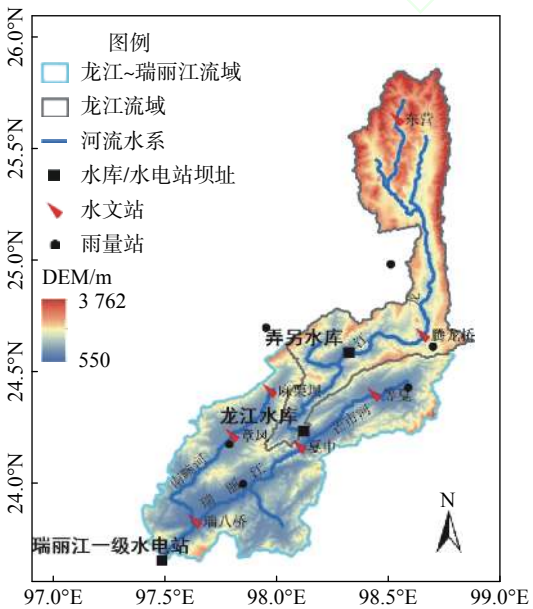

图 1 龙江水库位置示意

Fig. 1 Schematic map of the Longjiang Reservoir location

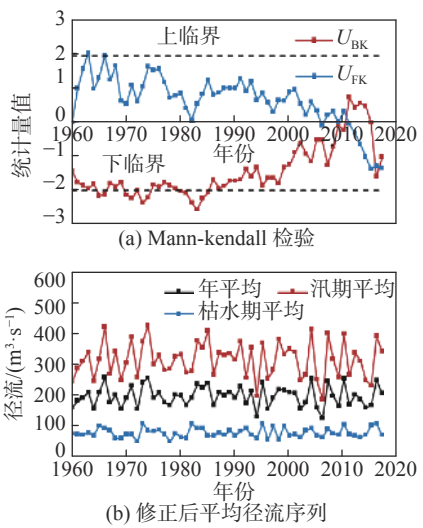

图 2 修正前龙江水库年径流 Mann-Kendall 检验与修正后平均径流序列

Fig. 2 Mann-Kendall test of annual runoff of the Longjiang Reservoir before the correction and the revised mean runoff sequence of the annual, flood and dry seasons

2　研究方法

2.1　径流预报模型

本文重点研究机器学习领域经典的随机森林和支持向量机对于龙江水库入库径流的预报效果。以下对 2 种模型的计算原理与特点进行简要介绍。

2.1.1　随机森林　文献 [15-16] 研究表明大气环流指数、海温、气压和前期径流等对中长期径流变化具有较好的指示作用, 丰富的预报因子有助于提高预报精度, 但也给识别这些因子影响径流的关键时空区域增加了难度。传统的最优子集回归方法等需要针对不同因子组合建立预报模型, 而可能的因子组合数随着预报因子的丰富呈现指数级增加, 常常出现计算灾难。

随机森林具有优秀的高维和非线性数据集处理能力, 为这一问题的解决提供了一种可行途径。对于连续型变量的预报问题, 随机森林是由一组相互独立的回归决策树(决策树规模 N_{tree} 为模型主要参数)构成的集合预报模型, 其中每一棵决策回归树对应由原始样本有放回自助抽取的一个样本集, 回归决策树的构建过程即为样本集根据预报因子完成二分裂的过程。首先, 从预报因子集合中按照一定规模随机筛选子集(子集规模 M_{tyr} 为另一重要参数)。其次, 进入分裂程序, 每次分裂历遍子集中各因子的所有数值, 依次尝试分裂, 以分裂节点两端样本平方误差之和最小为准则, 确定最优切分因子和对应数值, 并完成分裂。重复上述步骤至分裂次数达到上限或决策树末端节点最大样本数小于某一阈值, 即完成回归决策树的构建[17]。应用随机森林模型进行预测时, 将预报因子值输入到各决策回归树得到对应的预测值, 对所有回归树预测值取算数平均即为预测结果。

上述建模过程中样本及其预报因子的两层随机抽样设计, 保证了决策树之间的独立性与随机性; 同时采用有放回抽样方法后, 部分未被抽取的余留样本(也称为袋外数据, 数据量约为原始样本的 1/3)可用于决策树预测效果的验证, 利用余留样本对预报因子进行重要性评估的具体步骤如下: (1)对于随机森林中某一棵回归决策树, 使用相应余留样本计算预测均方误差, 记为 E_{MS1}; (2)随机扰动所有余留样本预报因子 X 的数值(随机改变样本 X 的数值, 或者更换 X 数据的顺序), 再次计算决策树的预测均方误差, 记为 E_{MS2}; (3)对于 N_{tree} 棵决策树均重复步骤(1)和(2), 以 $\Delta E_{\text{MSo}}=\Sigma(E_{\text{MS2}}-E_{\text{MS1}})/N_{\text{tree}}$ 作为 X 重要性的度量。计算所有预报因子的 ΔE_{MSo}, 若 ΔE_{MSo} 值较大, 则表明对应的预报因子随机扰动后, 余留样本预报误差大幅增加, 即该因子对于预报结果影响较大, 重要程度较高; 反之, 该因子的重要程度较低[10]。与传统方法相比, 预报因子随机分布于随机森林模型中不同决策树与分裂点, 不需要针对不同因子组合分别建模, 仅 1 次建模即可对不同因子的重要性进行评估, 从而降低了计算资源的开销。

2.1.2　支持向量机　给定的中长期径流预报样本集 $\{(x_i,y_i),i=1,2,\cdots,n\}$, x_i 表示预报因子向量(L 维), y_i 表示实测径流。对于这种高维非线性预测问题, SVM 通过变换函数 $\Phi(x)$ 将原始空间映射到高维特征空间, 在高维特征空间建立线性回归函数见式(1), 进而引入松弛变量 ξ_i、ξ_i^* 和惩罚因子 C, 依据结构风险最小化原则构建凸二次规划模型见式(2), 求解得到线性回归函数的 ω 与 b, 即可应用其实现径流预报[7-9]。式(2)的目标函数前半部分代表回归函数的泛化能力, 后半部分表征拟合误差, 二者之和最小时, 模型既具备较强的泛化能力、又具有较高的拟合精度[8]。另外求解过程中涉及样本 x_i 和 x_j 特征空间的内积 $\Phi(\boldsymbol{x_i})^{\mathrm{T}}\Phi(\boldsymbol{x_j})$, 为降低计算复杂度, 以原始空间中核函数 $\kappa(\boldsymbol{x_i},\boldsymbol{x_j})$ 的计算结果代替。鉴于径向基函数(式(3))处理高维复杂样本的性能优于其他核函数, 且所需参数较少, 本文选择径向基函数为核函数。

$$f(\boldsymbol{x_i})=\omega\Phi(\boldsymbol{x_i})+b\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{1}$$

$$\text{s.t.}\begin{cases} f(\boldsymbol{x_i})-y_i\leqslant\varepsilon+\xi_i \\ y_i-f(\boldsymbol{x_i})\leqslant\varepsilon+\xi_i^* \\ \xi_i\geqslant 0,\xi_i^*\geqslant 0,i=1,2,\cdots,n \end{cases} \tag{2}$$

$$\kappa(\boldsymbol{x_i},\boldsymbol{x_j})=\exp(-g\left\|\boldsymbol{x_i}-\boldsymbol{x_j}\right\|^2) \tag{3}$$

式中：ω为超平面的法向量；b 为超平面的偏移量；ε 为不敏感损失系数；g 为核函数参数。

2.2 模型参数优选

径流预报模型的实际预测效果与模型结构和参数密切相关。随机森林模型的待选参数是决策树个数 N_{tree} 与预报因子子集规模 M_{tyr}，N_{tree} 越大会导致模型过拟合，M_{tyr} 过大则使不同决策树差异过小。支持向量机模型的待选参数是惩罚因子 C、核函数参数 g 和不敏感损失系数 ε，C 过大、g 过小均会使模型过拟合，ε 对模型的影响较小。参数优化过程中单纯追求建模样本高精度拟合，常常导致模型在预报检验期应用效果较差，出现过拟合现象。

因此，本文采用交叉验证与粒子群相结合的算法优选参数[8]。步骤如下：将所有样本按一定比例划分为建模期和预报检验期样本；将建模期样本随机均分为 S 组，依次以 1,2,…,S 组作为验证样本，剩余 S-1 组作为训练样本，从而建立 S 个待选模型。采用粒子群优化算法优选参数，以 S 个待选模型的平均误差及相应方差之和最小作为目标函数，优选平均预报性能与稳定性综合最高的参数组合，以此作为预报模型参数。这种方法实现了建模期样本规模的扩展，并且在建模期增加了互斥的验证数据，模型对于验证数据的平均预报精度能够更好地反映其预测性能。

3 结果分析

3.1 预报因子重要性评估

对环流指数、海温、气压位势(气候因子)和前期月径流(水文因子)等潜在预报因子，分析其对径流预报的重要性。收集国家气候中心发布的 130 项逐月环流指数、NOAA 公开的全球月平均海温格点数据(2°×2°)和 NCEP/NCAR Reanalysis 1 数据集中 500 hPa 和 100 hPa 月平均气压位势格点数据(2.5°×2.5°)，开展龙江水库年、汛期和枯水期平均径流与前 12 个月环流指数、各格点海温、气压位势的相关性普查。以空间连续的显著相关格点区(置信水平为 0.05)为关键影响区，并以最高相关系数所在格点的气候因子作为预报因子。在此基础上，参考文献 [18-19] 的研究结果，剔除无物理背景的相关因子，得到各月径流的基础预报因子集合。表 1 给出了年径流基础预报因子集合，其中环流因子包括前 1 月亚洲经向环流指数等 7 项，海温与气压位势显著相关的空间区域主要为太平洋中北和西北部，时间上相对分散。

表 1 云南龙江水库年平均径流基础预报因子集

Tab. 1 Basic predictors for annual mean inflow of the Longjiang Reservoir in Yunnan Province

类别	预报因子
环流指数	前1月亚洲经向环流指数、前1月欧亚经向环流指数、前2月北极涛动指数、前7月东大西洋遥相关型指数、前6月欧亚纬向环流指数、前7月极地-欧亚遥相关型指数、前12月北大西洋-欧洲环流W型指数
海温	前6月第3 402格点海温(太平洋中西部)、前6月第5 040格点海温(西伯利亚北部)、前8月第1 246格点海温(西伯利亚东部)、前11月第6 297格点海温(太平洋中北部)、前11月第1 229格点海温(日本北部)
气压位势	前1月第2 956格点500 hPa位势(地中海)、前2月第3 876格点500 hPa位势(日本东南)、前5月第3 385格点500 hPa位势(地中海)、前7月第2 903格点500 hPa位势(太平洋东部)、前7月第3 027格点500 hPa位势(太平洋中北部)、前7月第2 956格点500 hPa位势(地中海)、前7月第1 862格点500 hPa位势(鄂霍次克海北部)、前11月第1 269格点500 hPa位势(乌拉尔山北部)、前2月第2 307格点100 hPa位势(白令海南部)、前3月第1 393格点100 hPa位势(乌拉尔山)、前7月第2 614格点100 hPa位势(太平洋东部)、前7月第3 029格点100 hPa位势(太平洋中北部)

注：表中格点为与年径流具有显著物理成因相关的海温和气压位势关键区域中相关系数最高的格点，格点序号从全球经纬网格(海温与气压的分辨率不同)的左上角开始，按照Z字型顺序递增。

将基础预报因子集与前 12 个月逐月径流合并，得到预报因子全集，年、汛期、枯水期平均径流预报因子总数分别为 36, 34 和 31 项，需合理评估各因子重要性，进一步缩减因子规模。本文基于 1961—2018 年全部样本建立随机森林模型(预报因子涉及前期月径流，因而建模滞后 1 年)，该模型仅用于预报因子重要性评估。经测试，当 N_{tree} 较大时，N_{tree} 与 M_{tyr} 的变化对预报因子重要性评估的影响基本可以忽略，为此取

N_{tree} 为 2 000、M_{tyr} 为预报因子总数的 1/3。基于建立的随机森林模型, 对各因子的重要性进行评估。图 3 给出了所有预报因子的 ΔE_{MSo}, 最后 12 个因子表示前 12 个月入库径流, 其余为气候因子。由图 3 可知, 不同因子对于径流预报的重要性存在明显差异, 太平洋中北部气压位势(21 号因子)对于年入库径流预报的影响程度最大, 前 3 月径流重要性明显高于其他月份。汛期入库径流预报因子重要性评估结果同样显示太平洋中北部气压位势(19 号因子)的重要性程度最高, 前 3 月径流的影响较大。对于枯水期径流, 台湾东部海温(8 号因子)对于预报结果的影响最大, 前 5 月至前 8 月径流(主要为前一水文年枯水期月径流)对径流预报较为重要。总体上, 前期月径流对年和汛期平均径流的影响弱于气候因子, 但前期枯水期月径流对枯水期平均径流预报的影响程度与部分气候因子相当。

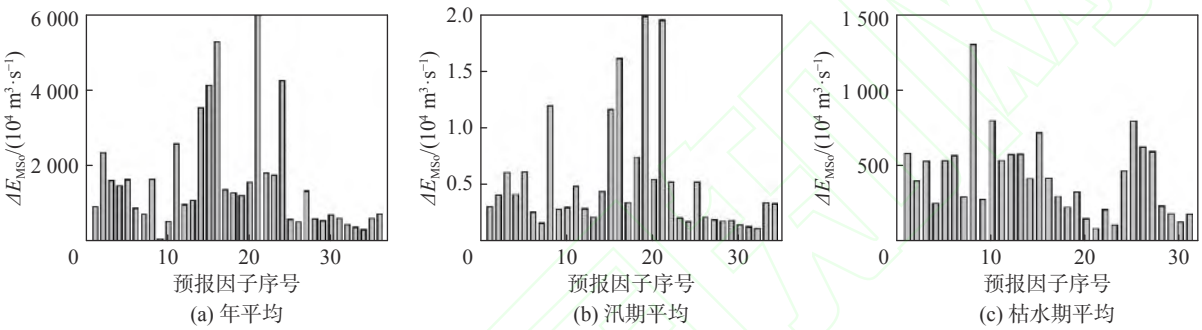

图 3　云南龙江水库入库径流预报因子重要性评估结果

Fig. 3　Evaluation on the importance of factors for forecasting the inflow of the Longjiang Reservoir in Yunnan Province

3.2　模型构建与预报结果分析

以 1961—2002 年为建模期, 以 2003—2017 为预报检验期, 预见期为 1 月, 采用交叉验证与粒子群相结合的方法构建 RF 和 SVM 模型。粒子群优化算法的迭代次数为 500 次、种群数量为 50、学习因子为 1.5。根据样本序列长度, 将建模期样本随机均分为 4 组进行交叉验证, 以待选模型对验证样本的平均误差与相应方差之和最小作为目标函数, 优化确定预报精度与稳定性综合效果最佳的模型参数组合。建模期与预报检验期的模型精度采用合格率 R_Q 和平均绝对百分比误差 E_{MAP}[20] 评估, 见式(4)~(5)。

$$R_Q = m/n_0 \times 100\% \tag{4}$$

$$E_{MAP} = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\left|Q_{p,i} - Q_{obs,i}\right|}{Q_{obs,i}} \times 100\% \tag{5}$$

式中: m 为建模期或预报检验期预报值合格(根据 SL 250—2000《水文情报预报规范》, 取预报值与实际值相对误差在±20% 之间为合格)的月份数, n_0 为建模期或预报检验期月份总数; $Q_{p,i}$, $Q_{obs,i}$分别为建模期或预报检验期第 i 个月份径流预报值、实际值。

考虑到预报因子重要性, 评估仅给出了各因子重要性排序结果, 究竟选择哪些因子尚不可知, 文献[10-11] 等按照一定数量直接选取, 缺乏科学依据。本文按照预报因子重要性降低的顺序, 逐步扩充因子数量, 即针对不同规模的输入变量建立预报模型, 分析新因子引入对于模型性能的影响, 从而确定最佳因子组合。图 4 给出了预报因子数量对建模期入库径流模拟误差的影响。分析发现, 无论何种模型, 随着预报因子规模的扩大, 合格率的变化不大, 但对定量误差的影响较为显著, SVM 模型的 E_{MAP} 总体呈现减小趋势, 而 RF 模型由于决策树分裂过程中随机选用部分预报因子, 导致 E_{MAP} 呈现震荡变化, 趋势性特征不明显。以较少预报因子获取较强模型性能为原则, 确定年平均径流 RF 模型的最优预报因子组合为按重要性降序排列的前 20 个因子, SVM 模型为前 17 个预报因子, 汛期平均径流 RF 与 SVM 的最佳因子集规模分别为 12 和 17, 对于枯水期平均径流分别为 14 和 17。

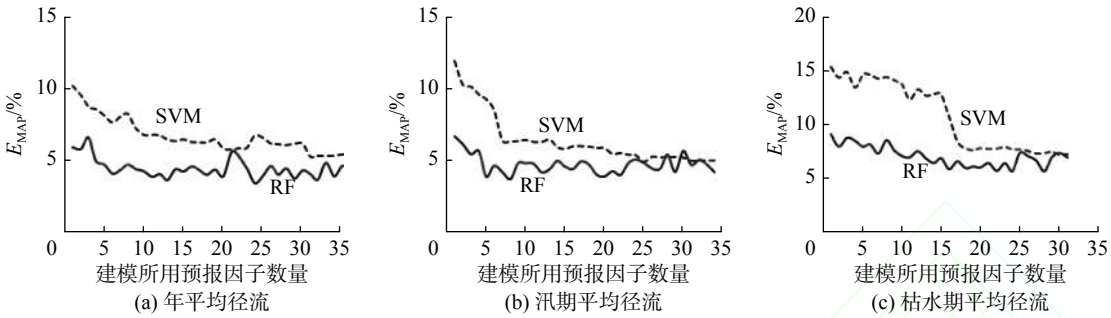

图 4 预报因子数量对建模期入库径流模拟误差的影响

Fig. 4 Influence of number of forecasting factors on the simulation error of the reservoir inflow in the modeling period

以最优因子组合及相应最佳参数建立预报模型并进行试报，模拟与预报精度如表 2 所示。由表 2 可知，无论何种时间尺度，RF 与 SVM 模型合格率均超过了 85%，E_{MAP} 均在 15% 以内，模拟与预报的精度总体较高。由建模期到预报检验期，RF 模型的 E_{MAP} 有不同程度的增大，其中枯水期平均径流在预报阶段较好地保持了模型性能；SVM 模型对年平均径流的预报误差较模拟误差的增幅明显小于 RF 模型，甚至对于枯水期平均径流的预报精度有所提升。因此，SVM 模型的泛化能力强于 RF。对比定量误差，发现建模期 RF 优于 SVM，而预报检验期恰好相反。图 5 给出了两种模型的模拟和预报径流过程，可见二者对实际径流时程变化的跟踪效果较好，但存在局部高流量低估和低流量高估的问题；建模期 RF 的模拟效果较 SVM 更贴近实际值，而 SVM 在预报检验期的优势更加明显，这与 E_{MAP} 的比较结果吻合。上游梯级水库调节影响在一定程度上扰动了径流与气候和水文因子的关系，影响了极值流量预报的不确定性；而 RF 与 SVM 在两阶段性能的相异性与样本统计特性的变化密切相关，因此有必要扩充预报模型库，考虑以多模型集合预报降低预报的不确定性。

表 2 龙江水库年、汛期、枯水期平均径流的模拟与预报精度

Tab. 2 Accuracy of simulation and forecast for mean inflow in annual, flood and dry seasons of the Longjiang Reservoir in Yunnan Province

径流类型	建模期				预报检验期			
	RF模型		SVM模型		RF模型		SVM模型	
	R_Q/%	E_{MAP}/%	R_Q/%	E_{MAP}/%	R_Q/%	E_{MAP}/%	R_Q/%	E_{MAP}/%
年平均径流	100.0	3.83	97.6	6.77	93.3	11.45	93.3	9.39
汛期平均径流	97.6	4.15	95.2	6.09	86.7	11.34	86.7	8.91
枯水期平均径流	97.6	6.76	97.6	10.99	100.0	7.24	100.0	7.43

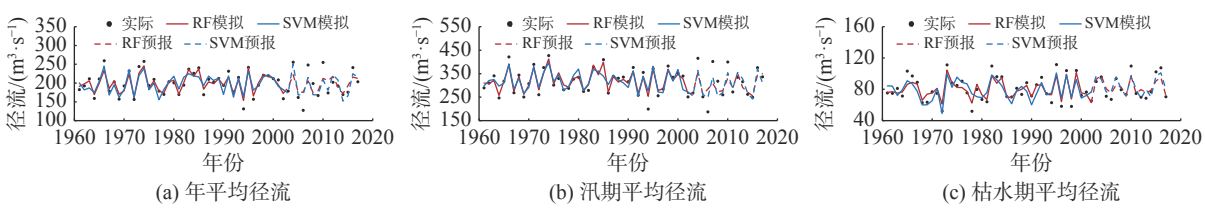

图 5 云南龙江水库年、汛期和枯水期平均径流模拟与预报结果

Fig. 5 Simulation and forecast results of mean inflow in annual, flood and dry seasons of the Longjiang Reservoir in Yunnan Province

4 结 语

本文利用随机森林与支持向量机模型开展了云南龙江水库长期入库径流的预报研究。随机森林在训练样本与预报因子方面的双层随机抽样设计为预报因子重要性评估提供了可行途径。评估结果表明：太平

洋中北部与西部的气候因子对入库径流预报的影响较大, 前期月径流对年、汛期径流的重要性偏低, 但对枯水期的影响程度与部分气候因子相当。在对预报因子重要性排序的基础上, 分析发现预报因子规模对模型性能有明显影响, 应合理分析确定。以最佳因子组合与融合交叉验证的粒子群算法优选的模型参数, 建立的随机森林与支持向量机径流预报模型总体精度较高, 模拟与预报的合格率均高于 85%, 平均绝对百分比误差均不超过 15%, 但对于局部极值流量的预报效果相对较差。建模期与预报检验期定量误差的对比结果表明支持向量机模型预报龙江水库入库径流的泛化能力优于随机森林。

本研究建立的预报模型在局部高、低流量处的预报精度尚有较大提升空间, 引入上游电站出库流量、表征流域下垫面变化的因子等是模型预报性能改善的潜在增长点; 关于多模型集合预报在有效应对径流统计特性非平稳变化、降低预报不确定性方面的效益, 有待深入研究。另外, 在月尺度上影响龙江水库入库径流预报的关键因子如何变化, 随机森林等机器学习模型的预报效果如何, 也值得进一步探讨。

参　考　文　献:

[1] SANG Y F. A review on the applications of wavelet transform in hydrology time series analysis[J]. Atmospheric Research, 2013, 122: 8-15.

[2] 刘勇, 陈元芳, 王银堂, 等. 基于OSR-BP神经网络的丹江口秋汛期径流长期预报研究[J]. 水文, 2010, 30(6): 32-36. (LIU Yong, CHEN Yuanfang, WANG Yintang, et al. Long-term forecasting for autumn flood season in Danjiangkou Reservoir Basin based on OSR-BP neural network[J]. Journal of China Hydrology, 2010, 30(6): 32-36. (in Chinese))

[3] 谢帅, 黄跃飞, 李铁键, 等. LASSO回归和支持向量回归耦合的中长期径流预报[J]. 应用基础与工程科学学报, 2018, 26(4): 709-722. (XIE Shuai, HUANG Yuefei, LI Tiejian, et al. Mid-long term runoff prediction based on a Lasso and SVR hybrid method[J]. Journal of Basic Science and Engineering, 2018, 26(4): 709-722. (in Chinese))

[4] 张素琼, 张艳军, 刘佳明, 等. 基于逐步回归-LMBP算法的大通站旬径流与月径流预报[J]. 水电能源科学, 2014, 32(6): 13-15, 4. (ZHANG Suqiong, ZHANG Yanjun, LIU Jiaming, et al. Ten-days and monthly runoff forecasting in Datong station based on stepwise regression and LMBP algorithm[J]. Water Resources and Power, 2014, 32(6): 13-15, 4. (in Chinese))

[5] 汪哲荪, 袁潇晨, 金菊良, 等. 基于集对分析的年径流自组织预测模型[J]. 水利水运工程学报, 2010(4): 33-37. (WANG Zhesun, YUAN Xiaochen, JIN Juliang, et al. GMDH network forecast model for annual runoff based on set pair analysis[J]. Hydro-Science and Engineering, 2010(4): 33-37. (in Chinese))

[6] 崔东文. 改进Elman神经网络在径流预测中的应用[J]. 水利水运工程学报, 2013(2): 71-77. (CUI Dongwen. An improved Elman neural network and its application to runoff forecast[J]. Hydro-Science and Engineering, 2013(2): 71-77. (in Chinese))

[7] 林剑艺, 程春田. 支持向量机在中长期径流预报中的应用[J]. 水利学报, 2006, 37(6): 681-686. (LIN Jianyi, CHENG Chuntian. Application of support vector machine method to long-term runoff forecast[J]. Journal of Hydraulic Engineering, 2006, 37(6): 681-686. (in Chinese))

[8] 周婷, 金菊良, 李荣波, 等. 基于小波支持向量机的径流预测性能优化分析[J]. 水力发电学报, 2017, 36(10): 45-55. (ZHOU Ting, JIN Juliang, LI Rongbo, et al. Performance optimization analysis for inflow prediction using wavelet Support Vector Machine[J]. Journal of Hydroelectric Engineering, 2017, 36(10): 45-55. (in Chinese))

[9] 崔东文. 几种智能算法与支持向量机融合模型在中长期月径流预测中的应用[J]. 华北水利水电大学学报(自然科学版), 2016, 37(5): 51-57. (CUI Dongwen. Application of several intelligent algorithms and Support Vector Machine fusion model in medium and long term runoff forecasting[J]. Journal of North China University of Water Resources and Electric Power (Natural Science Edition), 2016, 37(5): 51-57. (in Chinese))

[10] 赵铜铁钢, 杨大文, 蔡喜明, 等. 基于随机森林模型的长江上游枯水期径流预报研究[J]. 水力发电学报, 2012, 31(3): 18-24, 38. (ZHAO Tongtiegang, YANG Dawen, CAI Ximing, et al. Predict seasonal low flows in the upper Yangtze River using random forests model[J]. Journal of Hydroelectric Engineering, 2012, 31(3): 18-24, 38. (in Chinese))

[11] 赵文秀, 张晓丽, 李国会. 基于随机森林和RBF神经网络的长期径流预报[J]. 人民黄河, 2015, 37(2): 10-12. (ZHAO Wenxiu, ZHANG Xiaoli, LI Guohui. Research on the long-term runoff forecast based on random forest model and RBF network[J]. Yellow River, 2015, 37(2): 10-12. (in Chinese))

[12] 何国栋, 崔东文. 基于阴阳对算法优化的随机森林与支持向量机组合模型及径流预测实例[J]. 人民珠江, 2019, 40(3): 33-38. (HE Guodong, CUI Dongwen. Runoff prediction examples based on random forest of Yin-yang optimization algorithm and

Support Vector Machine model[J]. Pearl River, 2019, 40(3): 33-38. (in Chinese))
[13] 赵鹏雁, 张利平, 王旭, 等. 澜沧江流域中长期径流预报方法研究[J]. 武汉大学学报(工学版), 2018, 51(7): 565-569, 595. (ZHAO Pengyan, ZHANG Liping, WANG Xu, et al. Study of medium and long term runoff forecasting method for Lancang River Basin[J]. Engineering Journal of Wuhan University, 2018, 51(7): 565-569, 595. (in Chinese))
[14] 龚学贤. 云南省龙江水库水情自动测报系统的建设与应用[J]. 技术与市场, 2016, 23(2): 137-139. (GONG Xuexian. Construction and application of automatic runoff regime forecasting system for Longjiang Reservoir in Yunnan Province[J]. Technology and Market, 2016, 23(2): 137-139. (in Chinese))
[15] 冯小冲. 水库中长期水文预报模型研究[D]. 南京: 南京水利科学研究院, 2010. (FENG Xiaochong. Study on mid-long term hydrological forecasting model of reservoir[D]. Nanjing: Nanjing Hydraulic Research Institute, 2010. (in Chinese))
[16] 刘勇, 王银堂, 陈元芳, 等. 基于物理成因的中长期水文预报方法与应用研究[M]. 南京: 河海大学出版社, 2011. (LIU Yong, WANG Yintang, CHEN Yuanfang, et al. Methods and application for mid-long term hydrological forecast based on physical cause[M]. Nanjing: Hohai University Press, 2011. (in Chinese))
[17] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 68-69. (LI Hang. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012: 68-69. (in Chinese))
[18] 赵永晶, 钱永甫. 全球海温异常对中国降水异常的影响[J]. 热带气象学报, 2009, 25(5): 561-570. (ZHAO Yongjing, QIAN Yongfu. Analysis of the impacts of global SST on precipitation anomaly in China[J]. Journal of Tropical Meteorology, 2009, 25(5): 561-570. (in Chinese))
[19] 阮成卿, 李建平, 冯娟. 中国西南地区后冬降水的统计降尺度模型[J]. 中国科学: 地球科学, 2015, 58(10): 1827-1839. (RUAN Chengqing, LI Jianping, FENG Juan. Statistical downscaling model for late-winter rainfall over Southwest China[J]. Science China: Earth Sciences, 2015, 58(10): 1827-1839. (in Chinese))
[20] 张岩, 杨明祥, 雷晓辉, 等. 基于PCA-PSO-SVR的丹江口水库年径流预报研究[J]. 南水北调与水利科技, 2018, 16(5): 35-40. (ZHANG Yan, YANG Mingxiang, LEI Xiaohui, et al. Research on annual runoff forecast of Danjiangkou Reservoir based on PCA-PSO-SVR[J]. South-to-North Water Transfers and Water Science & Technology, 2018, 16(5): 35-40. (in Chinese))

Long-term inflow forecast of reservoir based on Random Forest and support vector machine

LI Lingjie[1], WANG Yintang[1], HU Qingfang[1], LIU Dingzhong[2], ZHANG Anfu[2], BA Yaquan[3]

(1. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering*, *Nanjing Hydraulic Research Institute*, *Nanjing* 210029, *China*; 2. *Yunnan Longjiang Water Conservancy Project Development Co., Ltd.*, *Dehong* 678400, *China*; 3. *Shenzhen Shenshui Water Resources Consulting Co., Ltd.*, *Shenzhen* 518003, *China*)

Abstract: Long-term runoff forecasting for the reservoir is of great significance for studying the hydrological regime and guiding the regulations. In this paper, the mean inflow of annual, flood and dry seasons of the Longjiang Reservoir are selected as forecast elements. Random Forest (RF) is utilized to filter key predictors from circulation indices, sea temperature, air pressure and previous monthly runoff. Afterwards, models based on RF and support vector machine (SVM), which are calibrated using particle swarm optimization algorithm combined with cross-validation, are established to predict the inflow of the Longjiang Reservoir. Results show that climate factors in the north-central and western Pacific have generally implemented a greater influence on prediction, while the effect of the pre-monthly runoff is relatively low, however it can be comparable to some climate factors when used to predict runoff in the dry season. The average accuracy of RF and SVM is generally satisfactory, with the qualification rate of simulation and forecast exceeding 85% and the average absolute percentage error less than 15%. SVM shows stronger generalization ability compared to RF in this study case, while the ability of both models in predicting partial extreme inflow remains to be improved.

Key words: Longjiang Reservoir; long-term inflow forecast; Random Forest; support vector machine