水　利　学　报

2022 年 12 月　　SHUILI　XUEBAO　　第 53 卷　第 12 期

文章编号：0559-9350(2022)12-1456-11

基于两阶段粒子群优化算法的新型逐步分解集成径流预测模型

郭田丽[1,2]，宋松柏[1,2]，张　特[1,2]，王慧敏[1,2]

（1. 西北农林科技大学 水利与建筑工程学院，陕西 杨凌　712100；
2. 西北农林科技大学 旱区农业水土工程教育部重点实验室，陕西 杨凌　712100）

摘要：传统分解集成径流预测模型首先将整个径流序列分解成若干个子序列，再将这些子序列划分为训练期和验证期进行建模，错误地将验证期内预报因子数据视作已知数据处理，难以应用于实际径流预报工作中。并且，这类模型的预测结果仅为一个确定数值，难以描述由于径流序列随机性和波动性而导致的预测不确定性。为解决以上问题，本文结合变分模态分解方法、支持向量机模型和核密度估计方法，提出了一种可同时进行点预测和区间预测的新型逐步分解集成(VMD-SVM-KDE)模型，并提出了一种两阶段粒子群优化(TSCPSO)算法来优化模型参数。选用黄河流域月径流数据评估模型性能，研究结果表明：(1)VMD-SVM-KDE 模型将单一 SVM-KDE 模型的确定系数(R^2)和纳什效率系数(*NSE*)值由 0.145~0.630 提升至 0.872~0.921，区间平均偏差(*INAD*)值由 0.046~95.844 降低至 0.005~0.034，说明 VMD-SVM-KDE 模型显著改进了单一 SVM-KDE 模型的点预测和区间预测性能；(2)相较于一阶段 PSO 算法，TSCPSO 优化算法将单一模型的 R^2 和 *NSE* 值由 0.145~0.480 提升至 0.309~0.630，*INAD* 值由 48.813~95.844 降低至 0.046~0.195，将分解集成模型的 R^2 和 *NSE* 值由 0.872~0.912 提升至0.876~0.921，*INAD* 值由 0.007~0.034 降低至 0.005~0.014，说明 TSCPSO 优化算法可以克服 SVM 的过拟合问题，并能提高单一模型和分解集成模型的预测精度；(3)VMD-SVM-KDE-TSCPSO 有效解决了传统分解集成预测模型存在的错误使用验证期内预报因子数据的问题，并在各站的 R^2 和 *NSE* 值均约为 0.9，*INAD* 值的范围为 0.005~0.014，具有更高的点预测和区间预测精度。文中模型可为优化径流预测模型和非平稳非线性水文序列预报提供新思路。

关键词：径流预测；区间预测；分解集成模型；两阶段粒子群优化算法；变分模态分解；支持向量机

中图分类号：P338　　**文献标识码：**A　　**doi：**10.13243/j.cnki.slxb.20220349

1　研究背景

径流预测可为各类涉水工程的规划设计、运行管理以及水资源管理提供依据。由于全球气候变化和高强度人类活动的双重作用与影响，径流序列表现出复杂的非线性和非平稳性，加大了径流预报工作难度[1]。

现有径流预测模型可分为过程驱动和数据驱动两大类。相比于过程驱动模型，数据驱动模型可操作性强，无需考虑径流发生的物理机制，仅需对时间序列进行数学分析，建立输入变量和输出变量之间的函数关系。近年来，机器学习因其强大的学习能力，在水文领域得到了长足的发展，并广泛应用于径流预测工作中[2-3]。其中，支持向量机(Support vector machine，SVM)模型具备结构简单、容错性强、能克服维数灾以及过拟合问题等优点，在大量径流预测研究中得到成功应用[4-6]。但由于径流序列呈现出高度复杂的非平稳性、非线性和不确定性，单一预测模型的预测能力受限。耦合奇异谱分析、集合经验模态分解、小波分解和变分模态分解(Variational modal decomposition，VMD)等数据分解

收稿日期：2022-05-05；网络首发日期：2022-11-03

网络首发地址：https://kns.cnki.net/kcms/detail/11.1882.TV.20221102.1042.001.html

基金项目：国家自然科学基金项目(52079110)

作者简介：郭田丽(1994-)，博士生，主要从事流域水文模拟预报研究。E-mail：guotianli@nwsuaf.edu.cn

通讯作者：宋松柏(1965-)，教授，主要从事水文水资源研究。E-mail：ssb6533@nwafu.edu.cn

方法与机器学习模型来建立分解集成模型可有效改进单一机器学习模型的预测精度[7-9]。其中，VMD方法能够控制中心频率混叠现象和噪声水平，更易于提高分解集成模型的预测性能[10-12]。

建立分解集成模型来提高径流预测精度已成为当前研究热点之一[13-15]。但是，这类模型首先将整个径流序列分解成若干个子序列，再将这些子序列划分为训练期和验证期，这种先分解后划分法会导致研究者在训练模型之前，提前使用了验证期数据，难以满足实际预报需求[16]。对于单变量分解集成模型，模型是将当期预测径流和前期径流作为预测模型的下一步预报因子，显然在验证模型的预测性能时，当预测径流需要依靠预测模型预测。若将上述先分解后划分法用于构建单变量分解集成模型，会导致验证期内的预报因子数据被视为已知数据处理。Du 等[17]和 Fang 等[18]对比分析了多种分解集成水文预测模型的结果，发现将验证期内预报因子数据当作已知数据去构建分解集成模型会得到“虚假”的高精度预测结果。分解集成模型的正确构建过程需先将实测序列划分为训练期和验证期，再将训练期数据进行分解建模，但其预测精度相对降低[19]。为此，Fang 等[18]提出一种逐步分解采样技术来构建正确的分解集成模型，以期提高正确分解集成模型的预测精度。该逐步分解采样技术可作为分解集成模型的替代采样技术，但使用该技术构建的分解集成模型的预测精度依旧较低。因此，如何正确构建能应用于现实径流预报工作中的较高精度分解集成径流预测模型尚需深入研究。

上述研究均旨在提高和改进模型的点预测性能，但点预测结果仅为一个确定数值，难以描述由于径流序列随机性和波动性而导致的预测不确定性。把握未来径流数据的波动范围和预测的不确定性，有助于研究者做出更加科学合理的分析、评估和决策。区间预测可通过参数或非参数密度估计方法估计点预测误差的分布特性，进而构建一定置信水平下的预测区间，可在一定程度上体现预测的不确定性。与参数估计模型相比，非参数核密度估计(Kernel density estimation，KDE)方法不需提前假设误差样本服从的分布函数，避免因误差分布选取错误导致出现较大预测误差问题[20]。KDE 方法已广泛应用于区间预测研究中，例如，Li 等[21]提出了基于分数阶灰色模型与 KDE 方法的混合模型，并将该模型用于农业用水量预测，Guo 等[22]构建了基于自激励自回归模型与 KDE 方法的地下水埋深区间预测模型，王晓东等[23]将分位数回归、长短记忆神经网络与 KDE 方法进行结合，构建了风电功率概率预测模型。这些研究结果均表明，KDE 方法可通过结合适宜的点预测模型产生较高精度的区间预测结果。

本文提出一种可同时进行点预测和区间预测的 VMD-SVM-KDE 逐步分解集成模型，并提出两阶段粒子群优化(TSCPSO)算法来优化模型参数。该模型可解决现有分解集成预测模型错误使用验证期内预报因子数据的问题，并能提高分解集成模型的预测精度。VMD-SVM-KDE-TSCPSO 模型可有效解决传统分解集成模型被错误开发和预测精度较低的问题，拓展该类模型在实际预报业务中的可应用性。

2 方法

2.1 变分模态分解方法 VMD 分解方法是一种新型自适应、完全非递归的数据分解方法[24]。该方法通过搜索约束变分问题的最优解，将原始信号 x 分解为一系列固有模态函数(Intrinsic mode functions，IMFs)。

$$\begin{cases} \underset{\{\mu_k,\ \omega_k\}}{\text{minimize}} \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{\mathrm{j}}{\pi t} \right) \otimes \mu_k(t) \right] \mathrm{e}^{-\mathrm{j}\omega_k t} \right\|_2^2 \\ \text{subject to} \sum_{k=1}^{K} \mu_k = x \end{cases} \tag{1}$$

式中：K 为 IMFs 的个数；$\{\mu_k(t)\}=\{\mu_1,\ \mu_2,\ \cdots,\ \mu_K\}$ 为第 k 个模态分量；$\mu_k(t)$ 为第 k 个模态分量在第 t 时刻的值；$\{\omega_k\}=\{\omega_1,\ \omega_2,\ \cdots,\ \omega_K\}$ 为第 k 个模态分量对应的中心频率；t 为时间；∂_t 为函数对时间 t 的一阶偏导数；$\delta(t)$ 为单位脉冲函数；j 为虚数单位；$\otimes$ 为卷积运算。

利用增广拉格朗日函数得到式(1)变分问题的最优解，如式(2)所示。

$$L(\{\mu_k\},\ \{\omega_k\},\ \lambda) = \alpha \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{\mathrm{j}}{\pi t} \right) \otimes \mu_k(t) \right] \mathrm{e}^{-\mathrm{j}\omega_k t} \right\|_2^2 + \left\| x(t) - \sum_{k=1}^{K} \mu_k(t) \right\|_2^2 + \left\langle \lambda(t),\ x(t) - \sum_{k=1}^{K} \mu_k(t) \right\rangle \tag{2}$$

式中：α 为二次惩罚因子；λ 为拉格朗日乘子。$\lambda(t)$ 为 λ 在第 t 时刻的值；$x(t)$ 为 x 在第 t 时刻的值。利用交替方向乘子迭代算法求解式(2)的鞍点。

2.2 支持向量机 SVM 模型建立在 VC 维(Vapnik-Chervonenkis Dimension)理论的基础上，并基于结构风险最小化原理求解分类和回归问题。假设一组给定的训练样本集：$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$，其中 $\boldsymbol{x}_i$ 为第 i 个输入向量，$\boldsymbol{y}_i$ 为输出值，n 为训练样本集的样本容量。SVM 回归函数为：

$$y=\boldsymbol{w}\varphi(x)+\boldsymbol{b} \tag{3}$$

式中：$\boldsymbol{w}$ 为权重向量；$\varphi(x)$ 为映射函数；$\boldsymbol{b}$ 为偏置向量。

根据结构风险最小化原理，式(3)的求解可转化为求解以下最小化问题：

$$\min J(\boldsymbol{w}, e)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\boldsymbol{e}_i^2 \tag{4}$$

式中：C 为惩罚因子；$\boldsymbol{e}=[e_1, e_2, \cdots, e_n]^{\mathrm{T}}$ 为误差变量。

引入拉格朗日函数和对偶理论，并用核函数代替式(3)的映射函数，可得到：

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K_1(\boldsymbol{x}_i, \boldsymbol{x})+b \tag{5}$$

式中：α_i 和 α_i^* 为拉格朗日乘子；$K_1(\cdot)$ 为核函数。

核函数的选取对 SVM 的计算精度至关重要。常用的核函数有线性核函数、多项式核函数、径向基(Radial Basis Function，RBF)核函数和 Sigmoid 函数。RBF 核函数对于大小样本都具有较高的计算精度，学习能力强，参数也相对较少。因此，本文选用 RBF 核函数，如式(6)所示。

$$K_1(\boldsymbol{x}_i, \boldsymbol{x})=\exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}\|^2) \tag{6}$$

式中 γ 为 RBF 核函数参数。

因此，对于 SVM 模型，主要的待求参数有惩罚因子 C 和核函数参数 γ。

2.3 核密度估计 KDE 是一种非参数估计方法，可直接计算样本的概率密度，无需事先假设其分布。对于一组样本数据 $\{x_j\}_{j=1}^{N}$，KDE 的表达式为：

$$\hat{f}(x)=\frac{1}{Nh}\sum_{j=1}^{N}K_2\left(\frac{x-x_j}{h}\right) \tag{7}$$

式中：N 为样本长度；h 为带宽；$K_2(\cdot)$ 为核函数。

核函数和带宽的选择对 KDE 计算影响较大。本文选用高斯核函数，公式为：

$$K_2\left(\frac{x-x_j}{h}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-x_j)^2}{2h^2}\right) \tag{8}$$

对应的带宽公式为：

$$h=1.06SN^{-\frac{1}{2}} \tag{9}$$

式中 S 为样本的标准差。

3 模型构建

本文提出的 VMD-SVM-KDE-TSCPSO 模型的总体框架如图 1 所示。基于 TSCPSO 优化算法的 VMD-SVM-KDE 逐步分解集成(VMD-SVM-KDE-TSCPSO)模型的构建步骤包括以下两步。第一步，根据 VMD-SVM 模型获取点预测结果。具体步骤为：将径流序列划分为率定期、测试期和验证期，使用 VMD 方法逐步分解率定期和测试期的径流序列获得子序列，根据这些子序列确定 SVM 模型的输入输出序列，并用 TSCPSO 优化算法综合求解 VMD-SVM 模型的未知参数，最后，根据 VMD-SVM 模型计算点预测值。第二步，根据点预测误差，采用 KDE 方法计算区间预测结果。VMD-SVM-KDE-TSCPSO 模型的具体构建步骤如下。

3.1 数据划分 首先，将径流序列 $x(t)(t=1, 2, \cdots, N)$ 划分为训练期 $x_T(t)(t=1, 2, \cdots, T)$ 和验

证期 $x_v(t)$ ($t=T+1$, $T+2$, …, N)，其中 N 为整个径流序列长度 $x(t)$ 的长度，T 为训练期序列 $x_T(t)$ 的长度。为防止模型过拟合，进一步将训练期 $x_T(t)$ 划分为率定期 $x_c(t)$ ($t=1$, 2, …, c) 和测试期 $x_t(t)$ ($t=c+1$, $c+2$, …, T)，其中 c 为率定期序列 $x_c(t)$ 的长度。率定期、测试期和验证期径流序列长度依次占整个数据的比例为 60%、20% 和 20%。

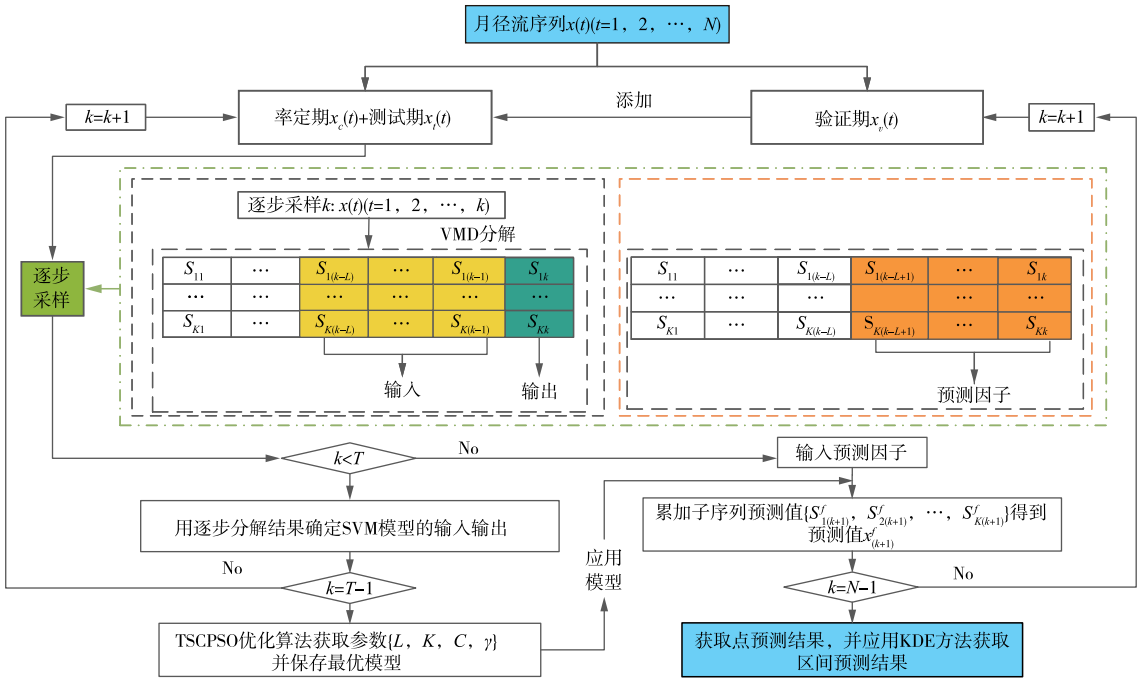

图 1　VMD-SVM-KDE-TSCPSO 模型框架

3.2　模型训练　模型的输入输出对于模型的训练来说至关重要。对于采样 k，令 k 的初始值为 $L+1$，应用 VMD 分解方法将径流序列子集 $x_T(t)$ ($t=1$, 2, …, k) 分解为 K 个子序列 $\{S_{1(t)}\}$，$\{S_{2(t)}\}$，…，$\{S_{K(t)}\}$ ($t=1$, 2, …, k)，其中 $\{S_i(t)\}$ ($i=1$, 2, …, k) 代表第 i 个子序列 $\{S_{i1}, S_{i2}, \cdots, S_{ik}\}$。子序列第 k 列之前的前 L 列序列 $\{S_{i(k-L)}\}$，$\{S_{i(k-L+1)}\}$，…，$\{S_{i(k-1)}\}$ ($i=1$, 2, …, K) 作为单个 SVM 模型的输入变量，子序列的第 k 列序列 $\{S_{ik}\}$ ($i=1$, 2, …, k) 作为单个 SVM 模型的输出变量，其中 $\{S_i(t)\}$ ($t=k-L$, $k-L+1$, …, k) 代表第 t 列序列 $\{S_1(t), S_2(t), \cdots, S_k(t)\}$。以上的输入输出用来训练 SVM 模型的惩罚因子 C 和核函数 γ。将元素 $\{S_{i(k-L+1)}\}$，$\{S_{i(k-L+2)}\}$，…，$\{S_{i(k)}\}$ ($i=1$, 2, …, K) 作为训练好的 SVM 模型的预测因子，用来求出子序列第 $k+1$ 列的预测值 $\{S^f_{1(k+1)}, S^f_{2(k+1)}, \cdots, S^f_{K(k+1)}\}$，累加该值得到第 $k+1$ 的预测值 $x^f(k+1)$。令 $k=k+1$，重复以上步骤，直至 $k=T-1$，依次得到率定期预测结果 $\{x^f_c(t)\}$ ($t=L+2$, $L+3$, …, c) 以及测试期预测结果 $\{x^f_t(t)\}$ ($t=c+1$, $c+2$, …, T)。为消除量纲单位不同造成的影响，同时提高模型的训练和收敛速度，对输入、输出和预测样本数据采取的归一化处理，如式(10)所示。

$$x_{\text{norm}}(t)=\frac{x(t)-x_{\min}}{x_{\max}-x_{\min}} \tag{10}$$

式中：x_{norm} 为样本的归一化结果；$x_{\min}$ 和 $x_{\max}$ 分别为样本的最小值和最大值。

因此，在 VMD-SVM-KDE-TSCPSO 模型中，待求的未知参数分别为：模型输入序列的长度 L、VMD 算法的分解个数 K、SVM 模型的惩罚因子 C 和核函数参数 γ。其中，模型输入序列的长度 L 和 VMD 的分解个数 K 影响模型的输入和输出，这对模型的确定至关重要。一般情况下，VMD 的分解个数由试错法确定，序列的输入长度 L 通过自相关分析和偏相关分析等方法确定。SVM 模型的惩罚因子 C 和核函数参数 γ 都是由智能优化算法优化确定。为了简化现有模型对所有未知参数推求的复杂性，本文提出 TSCPSO 优化算法可以在一个目标下同时对混合模型的 4 个未知参数(L, K, C, γ)进行寻

优。首先，根据以上计算得到的率定期预测结果$\{x_c^f(t)\}$($t=L+2$，$L+3$，…，c)以及测试期预测结果$x_t^f(t)$($t=c+1$，$c+2$，…，T)，分别计算率定期的均方误差($RMSE^c$)和测试期的均方根误差($RMSE^t$)，然后，应用 PSO 优化算法搜寻当 $RMSE^c$ 和 $RMSE^t$ 中的最大值取得最小值时的待求参数。TSCPSO 算法适应度函数的计算公式为：

$$\underset{L,K,C,\gamma}{\text{minimize}}\{\text{maximize}(RMSE^c,\ RMSE^t)\} \tag{11}$$

求出未知参数(L，K，C，γ)后，保存训练好的分解集成模型等待预测。

3.3 模型预测 令 $k=T$，重复 3.2 小节中的分解步骤，直至 $k=N-1$。不同于训练期的是，只需将预测因子输入训练好的混合模型中，然后获得验证期的所有预测值，也就是点预测值。

区间预测的构建步骤为：首先计算训练期的误差，然后根据 KDE 方法计算误差的概率密度函数，并积分得到相应的累积密度函数，据此获取一定置信水平 a 的分位数值$F_{1-(1-a)/2}(t)$和$F_{(1-a)/2}(t)$，最后根据式(12)可得预测区间$[x_{\text{low}}(t),\ x_{\text{up}}(t)]$。

$$\begin{cases} x_{\text{up}}(t)=x^f(t)+F_{1-(1-a)/2}(t) \\ x_{\text{low}}(t)=x^f(t)+F_{(1-a)/2}(t) \end{cases} \tag{12}$$

式中：$x_{\text{low}}(t)$和 $x_{\text{up}}(t)$分别为 t 时刻预测区间的下限和上限；$F_{1-(1-a)/2}(t)$和 $F_{(1-a)/2}(t)$为累积密度函数的分位数；给定的置信水平 a 也称为名义置信水平。

3.4 模型评价 为了验证 VMD-SVM-KDE 模型的预测优势，将 VMD-SVM-KDE 与 SVM-KDE 模型进行比较。同样，为了验证 TSCPSO 优化算法的优势，将 TSCPSO 算法与传统的一阶段 PSO 优化算法进行比较，传统 PSO 算法没有将训练期进一步划分为率定期和测试期。具体地，首先将 PSO 优化的 SVM(SVM-PSO)模型、TSCPSO 优化的 SVM(SVM-TSCPSO)模型、PSO 优化的 VMD-SVM(VMD-SVM-PSO)模型以及 TSCPSO 优化的 VMD-SVM(VMD-SVM-TSCPSO)模型进行对比，分析评价模型的点预测性能。然后将 PSO 优化的 SVM-KDE(SVM-KDE-PSO)模型、TSCPSO 优化的 SVM-KDE(SVM-KDE-TSCPSO)模型、PSO 优化的 VMD-SVM-KDE(VMD-SVM-KDE-PSO)模型以及 TSCPSO 优化的 VMD-SVM-KDE(VMD-SVM-KDE-TSCPSO)模型进行对比，分析评价模型的区间预测性能。

模型的点预测性能均由平均相对误差(MAE)、最大均方根误差($RMSE$)、确定系数(R^2)和纳什效率系数(NSE)4 个误差指标进行评估。MAE 和 $RMSE$ 越小且 R^2 和 NSE 越大，表示模型的点预测性能越好。区间预测性能由预测区间覆盖率($PICP$)、预测区间平均宽度($PINAW$)和区间平均偏差($INAD$)3 个误差指标进行评估[25]。$PICP$ 小于预测区间名义置信水平时，说明区间预测结果不可信。在 $PICP$ 达到预测区间名义置信水平的前提下，$PICP$ 越大且 $PINAW$ 越小代表区间预测性能越好。当不同模型的 $PICP$ 和 $PINAW$ 值相差不大时，可用 $INAD$ 进行比较择优。$INAD$ 越小代表模型性能越好。

4 实例应用

4.1 数据分析 为验证文中模型的适用性，选取黄河流域民和、兰州、白马寺和黑石关 4 个水文站的月径流数据进行预报模型研究。各水文站月径流数据资料均来源于水利部黄河水利委员会，其统计特征如表 1 所示。由表 1 可知，兰州站的月径流量均值依次小于民和站、黑石关站和白马寺站，离差系数依次大于民和站、黑石关站和白马寺站。各水文站月径流数据的偏度系数均为正数，民和、兰州和白马寺 3 站月径流数据的峰度系数均小于 3，表明曲线峰值比正态分布低，而黑石关站的峰度系数大于 3，表明曲线峰值比正态分布高。按照整个径流序列长度的 60%、20%和 20%及时间先后顺序依次划分率定期、测试期和验证期。因此，民和、兰州和黑石关 3 站的率定期长度为 30 a，白马寺站率定期长度为 31 a，测试期和验证期长度均为 10 a。

4.2 模型的点预测结果 将 SVM-PSO、SVM-TSCPSO、VMD-SVM-PSO 和 VMD-SVM-TSCPSO 这 4 种模型分别记为模型 1、模型 2、模型 3 和模型 4，表 2 为以上模型的参数识别结果。由表 2 可知，各站 SVM 模型参数 C 和 γ 以及模型的输入参数 L 随着模型种类变化。对于模型的输入参数 L，模型 1 在

表 1　月径流数据的统计特征

水文站	起止年份	时长/a	均值×10^8/m^3	标准差×10^8/m^3	离差系数	偏度系数	峰度系数
民和	1961—2010	50	28.867	33.228	1.151	1.455	2.179
兰州	1961—2010	50	24.861	29.862	1.201	1.591	2.457
白马寺	1955—2005	51	48.185	52.984	1.100	1.585	2.293
黑石关	1961—2010	50	47.783	54.372	1.138	2.001	5.808

各站中的识别结果均为 10，模型 2 在民和、兰州和黑石关站的识别结果分别为 80、69 和 94，模型 3 在兰州站的识别结果为 18，模型 4 在黑石关站的识别结果为 50。除此之外，其余结果都大于 100。对于 VMD 分解个数的识别结果，除了模型 4 在兰州站为 5 之外，其他均为 4。

表 2　模型参数结果

水文站	模型 1			模型 2			模型 3				模型 4			
	C	*γ*	*L*	*C*	*γ*	*L*	*C*	*γ*	*L*	*K*	*C*	*γ*	*L*	*K*
民和	11.459	15.533	10	10.757	0.160	80	16	16	121	4	16	16	358	4
兰州	7.548	14.735	10	5.574	0.279	69	1.693	0.063	18	4	16	16	354	5
白马寺	3.211	11.991	10	0.325	0.063	221	16	0.945	275	4	16	1.128	264	4
黑石关	14.584	15.950	10	6.473	0.477	94	16	0.660	358	4	16	16	50	4

表 3 列出了各模型在训练期和验证期点预测性能的评价结果。在整个训练期，民和站各模型拟合精度由高至低依次为：模型 1、模型 3、模型 4、模型 2；兰州站各模型拟合精度由高至低依次为：模型 1、模型 2、模型 4、模型 3；白马寺站各模型拟合精度由高至低依次为：模型 1、模型 4、模型 3、模型 2；黑石关站各模型拟合精度由高至低依次为：模型 1、模型 3、模型 2、模型 4。由此可见，研究区各站均是模型 1 的拟合性能最优，且模型 1 的 R^2 和 *NSE* 值均接近于 1，*MAE* 和 *RMSE* 值也非常小，而其余三种模型的拟合性能在研究区水文站点表现各不相同。进一步比较两种单一模型的拟合性能发现，对于各站径流序列均是模型 1 的拟合性能优于模型 2。对比混合模型的拟合性能发现，对于民和和黑石关站，模型 3 优于模型 4，而在兰州和白马寺两站，模型 4 的拟合性能则优于模型 3。

表 3　各模型在训练期和验证期的点预测评价结果

水文站	模型	训练期				验证期			
		MAE	*RMSE*	R^2	*NSE*	*MAE*	*RMSE*	R^2	*NSE*
民和	模型 1	0.077	0.079	0.999	0.999	18.653	25.238	0.267	0.265
	模型 2	4.291	11.023	0.887	0.887	13.970	21.469	0.520	0.468
	模型 3	6.915	9.768	0.909	0.907	7.770	10.544	0.874	0.872
	模型 4	8.452	11.292	0.902	0.898	7.599	10.377	0.879	0.876
兰州	模型 1	0.075	0.077	0.999	0.999	17.838	22.508	0.200	0.150
	模型 2	3.795	9.689	0.904	0.902	14.657	20.295	0.419	0.309
	模型 3	7.763	10.464	0.892	0.886	6.938	8.694	0.873	0.873
	模型 4	7.410	10.166	0.901	0.896	5.545	7.845	0.905	0.897
白马寺	模型 1	0.140	0.143	0.999	0.999	38.964	49.437	0.148	0.145
	模型 2	14.640	27.397	0.740	0.706	26.260	38.317	0.518	0.486
	模型 3	11.027	14.542	0.907	0.906	14.098	18.720	0.879	0.877
	模型 4	10.929	14.535	0.909	0.908	14.055	18.656	0.880	0.878
黑石关	模型 1	0.213	0.218	0.999	0.999	30.921	39.337	0.480	0.444
	模型 2	7.725	19.517	0.883	0.882	23.102	32.553	0.630	0.619
	模型 3	13.313	17.460	0.887	0.885	12.167	15.837	0.912	0.910
	模型 4	13.223	20.010	0.873	0.868	11.492	15.469	0.921	0.914

表 4 为 TSCPSO 优化模型在率定期和测试期的模型精度评价结果。在率定期，模型 2 在民和、兰州和黑石关三站均优于模型 4，而白马寺站则相反。模型 2 在民和、兰州和黑石关站的 R^2 和 *NSE* 值均接近 1，且 *MAE* 和 *RMSE* 值较低。在测试期，模型 4 在各站的性能均优于模型 2。而在训练期，模型 4 在民和和白马寺两站均优于模型 2，而在兰州和黑石关站则相反，表明 TSCPSO 优化模型(模型 2 和模型 4)在率定期、测试期和训练期对于各站的性能表现存在差异。

表 4　TSCPSO 优化模型(模型 2 和模型 4)在率定期和测试期的评定结果

水文站	模型	率定期				测试期			
		MAE	*RMSE*	R^2	*NSE*	*MAE*	*RMSE*	R^2	*NSE*
民和	模型 2	0.081	0.083	0.999	0.999	14.116	20.125	0.687	0.679
	模型 4	8.954	10.128	0.584	0.331	8.440	11.319	0.904	0.899
兰州	模型 2	0.079	0.081	0.999	0.999	12.805	17.931	0.682	0.670
	模型 4	5.092	6.718	0.970	0.967	7.545	10.332	0.895	0.890
白马寺	模型 2	8.050	23.640	0.896	0.819	22.933	31.495	0.464	0.464
	模型 4	10.520	14.760	0.922	0.921	11.300	14.327	0.891	0.889
黑石关	模型 2	0.212	0.218	0.999	0.999	24.377	35.002	0.545	0.543
	模型 4	13.097	20.625	0.873	0.865	13.547	18.318	0.893	0.875

为更好分析各模型在验证期的预测性能，图 2 展示了实测值和预测值的散点图，图中黑色线条和红色线条分别代表 1∶1 线和线性拟合线。对比模型 1 和模型 2 在验证期各站预测结果可知，模型 2 的散点分布较 1∶1 线更紧凑，线性拟合线偏离 1∶1 线也更近，*MAE* 和 *RMSE* 值更小，R^2 和 *NSE* 值更大，且模型 2 将模型 1 的 *MAE* 值由 17.838～38.964 降低至 13.970～26.260，*RMSE* 值由 22.058～49.437 降低至 20.295～38.317，R^2 和 *NSE* 值由 0.145～0.480 提升至 0.309～0.630，表明 TSCPSO 优化的 SVM 模型性能优于普通 PSO 优化的 SVM 模型。对比模型 3 和模型 4 的预测结果可知，模型 4 将模型 3 的 *MAE* 值由 6.938～14.098 降低至 5.545～14.055，*RMSE* 值由 8.694～18.720 降低至 7.845～18.656，R^2 和 *NSE* 值由 0.872～0.912 提升至 0.876～0.921，表明 TSCPSO 优化的 VMD-SVM 模型性能优于普通 PSO 优化的 VMD-SVM 模型。以上结果表明相较于普通 PSO 算法，TSCPSO 算法可以提高单一模型和分解集成模型的点预测精度。与两种单一模型(模型 1 和模型 2)相比，混合模型(模型 3 和模型 4)在各站均表现出更高的 *NSE* 和 R^2 值及更低的 *MAE* 和 *RMSE* 值，且混合模型将各站单一模型的 *MAE* 值由 13.970～38.964 降低至 5.545～14.098，*RMSE* 值由 20.295～49.437 降低至 7.845～18.720，R^2 和 *NSE* 值由 0.145～0.630 提升至 0.872～0.921。

从图 2 可以进一步看出，两种单一模型的散点分布分散，散点拟合线偏离 1∶1 线较远，表明模型的整体预测能力较差，尤其是对极值的预测能力。而两种混合模型的散点分布集中，拟合线距离 1∶1 线较近，整体预测与极值的预测结果均表现较好，表明混合模型较单一模型的预测性能有较大提升。整体来看，各模型预测性能由高至低依次为：模型 4、模型 3、模型 2、模型 1；其中，模型 4 在各站的预测精度最高，R^2 和 *NSE* 值均为 0.9。

分析模型在训练期和验证期性能可知：模型 1 对训练期径流数据拟合度较高，误差较小，但该模型在验证期的预测精度较低，表明模型存在过拟合问题。模型 2 在训练期的拟合性能较差，但验证期的预测精度却优于模型 1，而模型 3 和模型 4 的预测精度也表现出类似的规律，表明相比 PSO 算法，TSCPSO 算法可以改善模型过拟合问题，进而提高模型的预测性能。

4.3　模型的区间预测结果　表 5 列出了各模型在 90%置信水平下的区间预测评价结果。图 3 进一步给出了模型的点预测和区间预测结果，图中蓝色散点代表实测径流值，黑色线条代表预测值，灰色区间为预测区间。

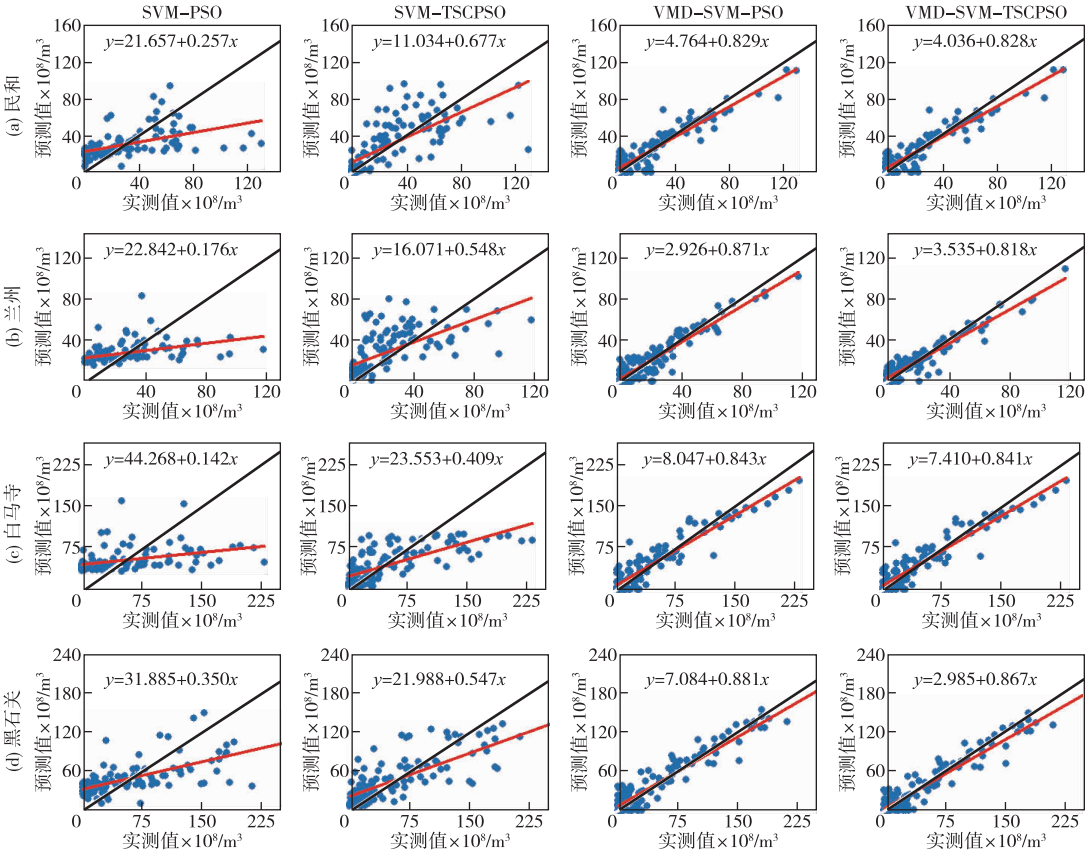

图 2　各站径流序列的预测结果散点图

表 5　各模型的区间预测性能评价结果

水文站	SVM-KDE-PSO			SVM-KDE-TSCPSO			VMD-SVM-KDE-PSO			VMD-SVM-KDE-TSCPSO		
	PICP	*PINAW*	*INAD*	*PICP*	*PINAW*	*INAD*	*PICP*	*PINAW*	*INAD*	*PICP*	*PINAW*	*INAD*
民和	0.008	0.002	71.980	0.708	0.294	0.119	0.900	0.269	0.017	0.942	0.311	0.009
兰州	0.017	0.002	80.757	0.675	0.229	0.195	0.950	0.293	0.007	0.958	0.302	0.005
白马寺	0	0.002	95.844	0.792	0.426	0.046	0.817	0.222	0.034	0.908	0.297	0.014
黑石关	0.008	0.003	48.813	0.733	0.227	0.127	0.908	0.249	0.009	0.967	0.270	0.006

对于单一模型，SVM-KDE-TSCPSO 模型的 *PICP* 值远大于 SVM-KDE-PSO 模型，而 *PICP* 均小于名义置信水平，表明单一模型的区间预测结果不可信。对于混合模型，仅 VMD-SVM-KDE-PSO 模型在白马寺站的 *PICP* 值小于名义置信水平，其区间预测结果不可信，而其余 *PICP* 值均达到名义置信水平，且混合模型的 *PICP* 值均大于单一模型，说明混合模型提高了单一模型的区间预测覆盖率，且混合模型将单一模型的 *INAD* 值由 0.046～95.844 降低至 0.005～0.034。对比混合模型的 *PICP* 和 *PINAW* 值可知，VMD-SVM-KDE-TSCPSO 模型的 *PICP* 和 *PINAW* 值均大于 VMD-SVM-KDE-PSO 模型，表明 VMD-SVM-KDE-TSCPSO 模型的覆盖率略高于 VMD-SVM-KDE-PSO 模型，但区间宽度却略宽于 VMD-SVM-KDE-PSO 模型。由此可见，两种模型的区间预测性能相近。进一步比较混合模型的 *INAD* 值可知，VMD-SVM-KDE-TSCPSO 模型在各站的 *INAD* 值均为最小，范围为 0.005～0.014，说明 VMD-SVM-KDE-TSCPSO 模型的区间预测结果最优。

以上分析结果也表明：相较于传统的一阶段 PSO 算法，TSCPSO 优化算法有效提高了单一模型和分解集成模型的区间预测精度，并将普通 PSO 算法优化模型的 *INAD* 值由 0.007～95.844 降至 0.005～0.195，其中将单一模型的 *INAD* 值由 48.813～95.844 降低至 0.046～0.195，将分解集成模型的 *INAD* 值由 0.007～0.034 降低至 0.005～0.014。

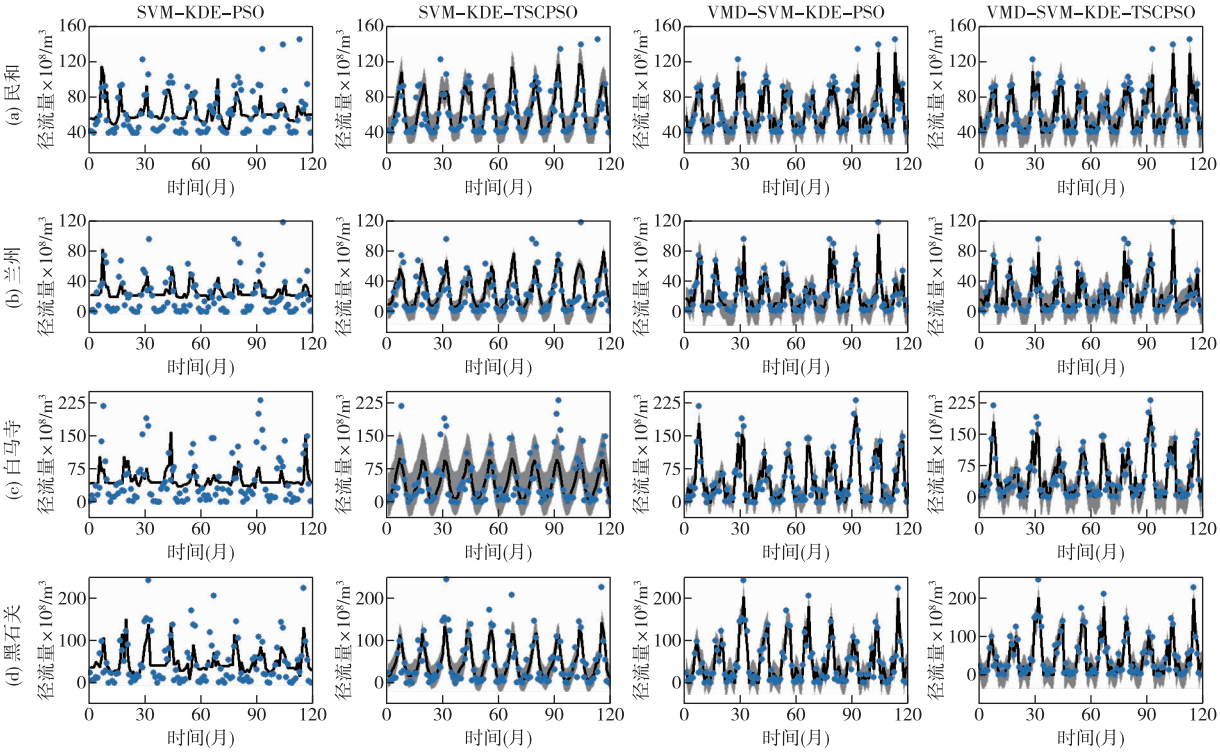

图 3　各站径流序列的预测结果图

4.4　讨论　Du 等[17]和 Fang 等[18]的研究表明：将验证期内预报因子数据当作已知数据，构建分解集成模型会导致“虚假”的高精度预测结果，使该类模型难以满足实际预报工作。而本文提出的 VMD-SVM-KDE-TSCPSO 模型先将径流序列分解为率定期、测试期和验证期，然后逐步分解径流序列得到子序列，最后根据率定期和测试期的子序列数据，应用 TSCPSO 优化算法训练模型的 4 个未知参数。由此可见，该模型在训练模型时并未使用验证期数据，这说明该模型可解决传统分解集成预测模型存在的错误使用未来预测因子数据的问题。并且，本文构建的 VMD-SVM-KDE-TSCPSO 模型在正确开发的基础上，在各站的 R^2 和 *NSE* 值均达到约 0.9，且预测区间覆盖率高，宽度窄，说明该模型具有较高的点预测和区间预测精度。因此，可将 VMD-SVM-KDE-TSCPSO 模型应用到实际预报工作中。

5　结论

为提高径流预测精度并量化预测结果的不确定性，本文提出一种新型 VMD-SVM-KDE 逐步分解集成模型和新型 TSCPSO 优化算法进行径流的点预测和区间预测。选用黄河流域 4 个水文站的月径流数据对模型性能进行评估。通过对比 SVM-PSO、VMD-SVM-PSO、SVM-TSCPSO 和 VMD-SVM-TSCPSO 模型的点预测结果，以及 SVM-KDE-PSO、VMD-SVM-KDE-PSO、SVM-KDE-TSCPSO 和 VMD-SVM-KDE-TSCPSO 模型的区间预测结果，研究结果表明：(1) VMD-SVM-KDE 模型将单一 SVM-KDE 模型的 R^2 和 *NSE* 值由 0.145～0.630 提升至 0.872～0.921，*INAD* 值由 0.046～95.844 降低至 0.005～0.034，说明 VMD-SVM-KDE 模型可以显著改进单一 SVM-KDE 模型的点预测和区间预测性能；(2) 相较于传统的一阶段 PSO 算法，TSCPSO 优化算法将单一模型的 R^2 和 *NSE* 值由 0.145～0.480 提升至 0.309～0.630，*INAD* 值由 48.813～95.844 降低至 0.046～0.195，将分解集成模型的 R^2 和 *NSE* 值由 0.872～0.912 提升至 0.876～0.921，*INAD* 值由 0.007～0.034 降低至 0.005～0.014，说明 TSCPSO 优化算法可以克服 SVM 的过拟合问题，并能提高单一模型和分解集成模型的预测精度；(3) VMD-SVM-KDE-TSCPSO 有效解决了传统分解集成预测模型存在的错误使用验证期内预报因子数据的问题，并在各站的 R^2 和 *NSE* 值均约为 0.9，*INAD* 值的范围为 0.005～0.014，具有更高的点预测和区间预测精度，可为

分解集成水文预测模型的正确建立提供参考，也可为非平稳非线性水文序列的实际预报工作提供依据。

参 考 文 献：

[1] YASEEN Z M, JAAFAR O, DEO R C, et al. Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq[J]. Journal of Hydrology, 2016, 542: 603-614.

[2] BARZEGAR R, FIJANI E, MOGHADDAM A A, et al. Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models[J]. Science of the Total Environment, 2017, 599: 20-31.

[3] 王树威，李建林，崔延华，等. 混沌理论与 BPNN 耦合的径流中长期预测模型[J]. 水资源与水工程学报，2021，32(3)：73-79.

[4] 康艳，杨沁瑜，张芳琴，等. 基于 Box-Cox 变换与 Lasso 回归的支持向量回归模型在月径流预报中的应用[J]. 应用基础与工程科学学报，2022，30(1)：27-39.

[5] ZUO G G, LUO J G, WANG N, et al. Two-stage variational mode decomposition and support vector regression for streamflow forecasting[J]. Hydrology and Earth System Sciences, 2020, 24(11): 5491-5518.

[6] ADNAN R M, KISI O, MOSTSFA R R, et al. The potential of a novel support vector machine trained with modified mayfly optimization algorithm for streamflow prediction[J]. Hydrological Sciences Journal-Journal des Sciences Hydrologiques, 2022, 67(2): 161-174.

[7] FENG Z K, NIU W J, TANG Z Y, et al. Monthly runoff time series prediction by variational mode decomposition and support vector machine based on quantum-behaved particle swarm optimization[J]. Journal of Hydrology, 2020, 583: 124627.

[8] MENG E H, HUANG S Z, HUANG Q, et al. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model[J]. Journal of Hydrology, 2019, 568: 462-478.

[9] 梁浩，黄生志，孟二浩，等. 基于多种混合模型的径流预测研究[J]. 水利学报，2020，51(1)：112-125.

[10] ZUO G G, LUO J G, WANG N, et al. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting[J]. Journal of Hydrology, 2020, 585: 124776.

[11] NIU W J, FENG Z K, CHEN Y B, et al. Annual streamflow time series prediction using extreme learning machine based on gravitational search algorithm and variational mode decomposition[J]. Journal of Hydrologic Engineering, 2020, 25(5): 04020008.

[12] HE X X, LUO J G, ZUO G G, et al. Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks[J]. Water Resources Management, 2019, 33(4): 1571-1590.

[13] 王斌，张洪波，辛琛，等. 基于贝叶斯模型加权平均法的径流序列高频分量预测研究[J]. 水力发电学报，2016，35(5)：75-83.

[14] JIANG Z Q, LI R B, JI C M, et al. Wavelet analysis-based projection pursuit autoregression model and its application in the runoff forecasting of Li Xiangjiang basin[J]. Hydrological Sciences Journal-Journal des Sciences Hydrologiques, 2018, 63(12): 1817-1830.

[15] XU Z X, ZHOU J Z, MO L, et al. A novel runoff forecasting model based on the decomposition-integration-prediction framework[J]. Water, 2021, 13(23): 3390.

[16] ZHANG X L, PENG Y, ZHANG C, et al. Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences[J]. Journal of Hydrology, 2015, 530: 137-152.

[17] DU K C, ZHAO Y, LEI J Q. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series[J]. Journal of Hydrology, 2017, 552: 44-51.

[18] FANG W, HUANG S Z, REN K, et al. Examining the applicability of different sampling techniques in the development of decomposition-based streamflow forecasting models[J]. Journal of Hydrology, 2019, 568: 534-550.

[19] KARTHIKEYAN L, KUMAR D N. Predictability of nonstationary time series using wavelet and EMD based ARMA models[J]. Journal of Hydrology, 2013, 502: 103-119.

[20] 杨锡运，张艳峰，叶天泽，等. 基于朴素贝叶斯的风电功率组合概率区间预测[J]. 高电压技术，2020，

46(3): 1099-1108.
[21] LI J, SONG S S, KANG Y, et al. Prediction of urban domestic water consumption considering uncertainty[J]. Journal of Water Resources Planning and Management, 2021, 147(3): 05020028.
[22] GUO T L, SONG S S, MA W J. Point and interval forecasting of groundwater depth using nonlinear models[J]. Water Resources Research, 2021, 57(12): e2021WR030209.
[23] 王晓东, 鞠邦国, 刘颖明, 等. 基于 QR-NFGLSTM 与核密度估计的风电功率概率预测[J]. 太阳能学报, 2022, 43(2): 479-485.
[24] DRAGOMIRETSKIY K, ZOSSO D. Variational mode decomposition[J]. IEEE Transactions on Signal Processing, 2014, 62(3): 531-544.
[25] 王若恒. 基于 LSTM 的风电功率区间预测研究[D]. 武汉: 华中科技大学, 2018.

A new stepwise decomposition ensemble model based on two-stage particle swarm optimization algorithm for the runoff prediction

GUO Tianli[1,2], SONG Songbai[1,2], ZHANG Te[1,2], WANG Huimin[1,2]

(1. *College of Water Resources and Architectural Engineering*, *Northwest A&F University*, *Yangling* 712100, *China*;
2. *Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas*, *Ministry of Education*, *Northwest A&F University*, *Yangling* 712100, *China*)

Abstract: The traditional decomposition ensemble runoff prediction model firstly decomposes the entire runoff series into several subseries, and then divides the subseries into training and validation periods for modeling, which wrongly treats the predictor data of validation period as known data and is difficult to be applied to actual runoff forecasting. Moreover, the prediction results of such models are only definite values, which is difficult to describe the prediction uncertainty caused by the randomness and volatility of runoff series. To solve the above problems, this study proposes a stepwise decomposition ensemble (VMD-SVM-KDE) model combining variable mode decomposition method, support vector machine model and kernel density estimation method, which performs both point prediction and interval prediction, and proposes a two-stage particle swarm optimization (TSCPSO) algorithm. The monthly runoff series of the Yellow River Basin is used to evaluate the model performance, and the study results show that: (1) the VMD-SVM-KDE model improves the coefficient of determination (R^2) and Nash efficiency coefficient (*NSE*) values of the single SVM-KDE model from the range of 0.145 to 0.630 to the range of 0.872 to 0.921, and reduces the interval average deviation (*INAD*) values from the range of 0.046 to 95.844 to the range of 0.005 to 0.034, indicating that the VMD-SVM-KDE model significantly improves the point prediction and interval prediction performance of a single SVM-KDE model; (2) compared with the traditional one-stage PSO algorithm, the TSCPSO algorithm improves the R^2 and *NSE* values of the single model from the range of 0.145 to 0.480 to the range of 0.309 to 0.630, and reduces the *INAD* value from the range of 48.813 to 95.844 to the range of 0.046 to 0.195, and also improves the R^2 and *NSE* values of the decomposition ensemble model from the range of 0.872 to 0.912 to the range of 0.876 to 0.921, and reduces the *INAD* values from the range of 0.007 to 0.034 to the range of 0.005 to 0.014, indicating that the TSCPSO optimization algorithm overcomes the overfitting problem of support vector machine models and effectively improves the prediction accuracy of the single and decomposition ensemble models; (3) the VMD-SVM-KDE-TSCPSO model addressed the mistakes of traditional decomposition ensemble models that forecast factor data of validation period, and has higher accuracy of point prediction and interval prediction with R^2 and *NSE* values of about 0.9 and the *INAD* values ranging from 0.005 to 0.014. The VMD-SVM-KDE-TSCPSO model can provide a basis for practical forecasting of non-stationary and non-linear hydrological series.

Keywords: runoff prediction; interval prediction; decomposition ensemble model; two-stage particle swarm optimization algorithm; variable mode decomposition; support vector machine

(责任编辑: 于福亮)