水 利 学 报

2021年6月 SHUILI XUEBAO 第52卷 第6期

文章编号:0559-9350(2021)06-0689-11

基于IGMM-Copula的入库径流过程预报误差随机模拟模型

张验科，张佳新，邰雨航，纪昌明，马秋梅

（华北电力大学 水利与水电工程学院，北京 102206）

摘要： 为揭示入库径流过程预报误差的统计特征及其变化规律，进而为水电站水库优化调度提供更为准确的输入，基于AIC与BIC准则选取最优高斯混合数，同时引入K-means++算法确定高斯混合模型(Gaussian Mixture Model，GMM)的初始参数值，对GMM-Copula模型中的GMM部分进行了改进，建立了基于IGMM-Copula的入库径流过程预报误差随机模拟模型，该模型不仅在单一预见时刻径流预报误差的量化估计上更具优势，而且能通过建立误差的多维联合分布函数实现对误差序列的随机模拟。以锦屏一级水电站水库为例，应用IGMM-Copula模型对预见时刻为6 h、12 h、18 h、24 h的径流预报误差进行随机模拟。结果表明，IGMM-Copula所得拟合曲线的图形效果及适用性检验结果均优于GMM-Copula模型，且其模拟预报误差的统计参数更贴近于实测预报误差，验证了其合理性与可行性，为入库径流过程预报误差的估计与模拟提供了一种更为精确有效的方法。

关键词： 入库径流预报误差；GMM-Copula；IGMM-Copula；随机模拟；锦屏一级水电站

中图分类号： P338 **文献标识码：** A **doi：** 10.13243/j.cnki.slxb.20200681

1 研究背景

入库径流预报误差是水文预报不确定性的一种表现，目前的研究一般将量化其随机性作为研究 重点，多采用参数估计[1-2]、非参数估计[3]、最大熵[4]等方法对其分布形式近似描述，取得了大量研究 成果[5-9]。考虑到入库径流预报误差往往包含多种统计特征，若采用单一分布，例如正态分布、*t* Location-Scale分布等对其进行拟合，某些情况下虽然可呈现出较好的拟合效果，但由于各分布函数有时在对称性和尾部特征上的表现存在差异，并不能保证总可以找到一种分布形式刚好能够达到预期的要求；同时，已有的研究成果多集中于单一变量或单一预见时刻的径流预报误差的量化分析[10-11]， 而对于多个变量或多个预见时刻的径流过程预报误差的分析相对较少。对于预报调度工作而言，前期的入库径流预报作业多在某一固定时刻以某一固定间隔对预见期内多个后续时刻的入库流量进行预报[12]，而根据对应时刻的历时实测入库流量序列，即可分析得到相应的径流过程预报误差序列。这一预报误差序列反映了整个调度期的预报不确定性，而仅对单一预见时刻的预报误差进行研究，则忽略了调度期内不同预见时刻预报误差间的关联性。因此，需要在寻求单一预见时刻预报误差分布的基础上，综合考虑分析不同预见时刻预报误差之间的相关性，才能更为全面地揭示入库径流过程预报误差的变化规律。

为解决上述入库径流预报误差分布研究中存在的问题，纪昌明等[13]依据高斯混合模型(Gaussian Mixture Model，GMM)在密度函数估计中的高度自适应性，结合高维meta-student t Copula函数在耦合

收稿日期：2020-06-20；网络首发时间：2021-04-02

网络首发地址：https://kns.cnki.net/kcms/detail/11.1882.TV.20210401.2123.003.html

基金项目：国家自然科学基金项目(51709105)；中央高校基本科研业务费专项资金(2020MS026；2019MS031)；中国博士后科学基金(2020M680487)

作者简介：张验科(1983-)，博士，副教授，主要从事水资源与能源科学、风险管理与决策理论等研究。E-mail：ykzhang2008@163.com

通讯作者：马秋梅(1988-)，博士，主要从事流域水文模拟遥感水文应用等研究。E-mail：qiumeima@ncepu.edu.cn

多个类型边缘分布上的优势，建立了多个预见时刻入库径流预报误差的GMM-Copula随机模型，可实现对入库径流预报误差序列的随机模拟。在该项研究中，GMM-Copula模型是以单一预见时刻径流预报误差分布的拟合结果为基础建立的多维径流预报误差联合分布，在应用过程中仍存在两个关键问题：(1)高斯分布混合个数的确定，这是一个权衡的过程，个数过少会影响拟合的精度，相反则会导致模型过于复杂不利于分析；(2)参数初始值的确定，由于采用的是EM(Expectation-Maximization algorithm，EM)法[14]求解模型拟合需要的参数，它是一类通过迭代进行极大似然估计的优化算法，需要对高斯混合模型的权重等参数进行初始化，不同的初始参数值会对迭代的结果造成较大的影响，从而决定最终的拟合精度。由于原GMM-Copula模型对于拟合过程的两个关键问题都未做细致的研究，各个预见时刻的径流预报误差的高斯分布混合个数以及初始参数值的确定存在较大的主观性，难以保证拟合结果的精度，从而得到的模拟误差序列也存在进一步优化的空间。

本文基于AIC(Akaike Information Criterion，AIC)与BIC(Bayesian Information Criterion，BIC)准则[15]选取最优高斯分布混合个数，同时引入数据挖掘中的K-means++算法确定高斯混合模型的初始参数值，对GMM-Copula模型中的GMM部分进行改进，以此建立基于IGMM-Copula(Improved Gaussian Mixture Model Copula)的入库径流过程预报误差随机模拟模型。以雅砻江流域锦屏一级水库短期入库径流过程预报误差分析为例，基于IGMM-Copula模型，首先对预见时刻为6 h、12 h、18 h和24 h的入库径流预报误差分布进行拟合分析；其次，将拟合结果作为边缘分布从而建立四维入库径流预报误差的联合分布，据此对误差序列进行随机模拟与统计分析，并与GMM-Copula模型得到的结果进行了对比，验证模型的可行性与合理性。

2 GMM-Copula中高斯混合模型描述及其存在的问题

现假设预报误差为 $e_{x(i)t(j)}$，表示当预报起始时刻为 $x(i)$ 时，对未来 $t(j)$ 时刻进行径流预报所产生的预报误差，其中 i 表示历史预报的次数，$i\in(1,2,\cdots,n)$；j 表示预报的时刻数，$j\in(1,2,\cdots,m)$。可定义为：

$$e_{x(i)t(j)}=\frac{S_{x(i)t(j)}-H_{x(i)t(j)}}{H_{x(i)t(j)}}\times 100\% \tag{1}$$

式中：$S_{x(i)t(j)}$、$H_{x(i)t(j)}$ 分别为预报流量值和实测流量值。

当进行 n 次预报后，可得到由 n 个预报误差序列组成的误差数据矩阵 $\boldsymbol{E}$：

$$\boldsymbol{E}=\begin{bmatrix} e_{x(1)t(1)} & e_{x(1)t(2)} & \cdots & e_{x(1)t(m)} \\ e_{x(2)t(1)} & e_{x(2)t(2)} & \cdots & e_{x(2)t(m)} \\ \cdots & \cdots & \cdots & \cdots \\ e_{x(n)t(1)} & e_{x(n)t(2)} & \cdots & e_{x(n)t(m)} \end{bmatrix} \tag{2}$$

针对时刻 $t(j)$ 的预报误差 $e_{x(i)t(j)}$，可利用高斯混合模型求其分布，设模型中待估计的参数为 θ，则高斯混合模型的概率密度函数可表示为：

$$p\left(e_{x(i)t(j)};\ \theta\right)=\sum_{k=1}^{K}\alpha_k N\left(e_{x(i)t(j)}\middle|u_k,\ \sigma_k^2\right) \tag{3}$$

式中：$p\left(e_{x(i)t(j)};\ \theta\right)$ 为高斯混合分布的概率密度函数；α_k、u_k、σ_k^2 分别为第 k 个高斯分布的参数，分别代表权重、均值和方差，且 $\sum_{k=1}^{K}\alpha_k=1$，K 表示高斯混合模型中的高斯混合个数。

从式(3)中可知，若要求解单一预见时刻误差分布拟合的表达式，则需要对参数 α_k、u_k、σ_k^2 进行估计，而直接对原式进行最大似然估计则会产生对数的和，求解十分困难，因此采用EM法[14]对参数进行推求。具体的求解步骤如下。

(1)进行EM算法中的E步(求期望值)，提出一个隐变量 γ_{ik} ， $\gamma_{ik}\in\{0,1\}$ ，表示第 i 个样本是否来自于第 k 个高斯分布， $k=1,\cdots,K$ 。将 γ_{ik} 引入式(3)，得到单个样本完全数据的对数似然函数，如式(4)。对参数 θ 进行初始化，其中 $a_k=1/K$ ， u_k 设为随机数， σ_k^2 取1，根据利用贝叶斯定理求解 γ_{ik} 的后验概率 $P\left(\gamma_{ik}=1\middle|e_{x(i)t(j)},\ \theta\right)$ ，见式(5)。

$$P\left(e_{x(i)t(j)},\ \gamma_{ik}|\theta\right)=\prod_{k=1}^{K}\alpha_k^{\gamma_{ik}}N\left(e_{x(i)t(j)}\middle|u_k,\ \sigma_k^2\right)^{\gamma_{ik}} \tag{4}$$

$$P\left(\gamma_{ik}=1\middle|e_{x(i)t(j)},\ \theta\right)=\frac{\alpha_k N\left(e_{x(i)t(j)}|u_k,\ \sigma_k^2\right)}{\sum_{k=1}^{K}\alpha_k N\left(e_{x(i)t(j)}|u_k,\ \sigma_k^2\right)} \tag{5}$$

(2)进行EM算法中的M步(求最大值)，由式(4)得到全部样本完全数据的对数似然函数，见式(6)。依据式(6)分别对 α_k 、 u_k 、 σ_k^2 求偏导，令似然函数值为0，求出更新后的参数分别如式(7)所示。

$$P\left(e_{x(i)t(j)},\ \gamma_{ik}|\theta\right)=\prod_{i=1}^{n}\prod_{k=1}^{K}\alpha_k^{\gamma_{ik}}N\left(e_{x(i)t(j)}\middle|u_k,\ \sigma_k^2\right)^{\gamma_{ik}} \tag{6}$$

$$\begin{cases}\alpha_k^{\text{new}}=\dfrac{n_k}{n}\\ u_k^{\text{new}}=\dfrac{\sum_{k=1}^{K}\gamma_{ik}e_{x(i)t(j)}}{n}\\ \sigma_k^{2\text{new}}=\dfrac{\sum_{k=1}^{K}\gamma_{ik}\left(e_{x(i)t(j)}-u_k^{\text{new}}\right)^2}{n}\end{cases} \tag{7}$$

(3)不断迭代步骤(1)和(2)，重复更新 α_k 、 u_k 、 σ_k^2 ，直到前后两次迭代结果的变化幅度小于一个设定值 ε ，则终止迭代，即 $|\theta-\theta^{\text{new}}|\leqslant\varepsilon$ ， θ^{new} 代表更新后的参数值， ε 通常取 10^{-4}。

GMM-Copula模型在对单一预见时刻误差分布进行估计的过程中，由于涉及到高斯混合个数 K 的选取，而通过人为确定的常规方法主观性太强，尤其在于相邻时段的预报误差分布可能具有不同的特征，若高斯混合个数选取不当，势必会对最终拟合的精度造成影响；同时，模型采用随机抽取确定初始参数值的方式同样不严谨，初始值的不同会影响最终迭代得到的参数，从而影响最终分布拟合的结果。本文基于AIC与BIC准则选取最优的高斯混合个数，同时采用K-means++算法确定高斯混合模型参数的初始取值，建立基于IGMM-Copula的入库径流过程预报误差随机模拟模型。模型在优化单一预见时刻径流预报误差分布拟合方法的基础上，考虑了各个预见时刻之间的相关关系，并结合随机抽样理论，可实现对多个预见时刻入库径流过程误差的随机模拟。

3 基于IGMM-Copula的入库径流过程预报误差随机模拟模型

模型主要从两个方面叙述，提出GMM-Copula模型改进的两个方面；2.3节描述了如何利用IGMM-Copula模型建立多个预见时刻的径流预报误差联合分布以及误差序列随机模拟的步骤。

3.1 基于AIC与BIC准则的最优高斯分布混合数选取 AIC与BIC都作为衡量统计模型拟合优良性的标准，与AIC不同的是，BIC引入了与模型参数个数相关的惩罚项，考虑了样本数量，更适合样本数量较多时的情况。对应的计算公式如下：

$$AIC=2K-2\ln(L) \tag{8}$$

$$BIC=K\ln(n)-2\ln(L) \tag{9}$$

式中：K 为模型的参数个数，在本文中表示为高斯混合模型中的高斯混合个数；n 为预报误差样本个数；L 为模型参数求解中所对应的似然函数。

当模型的参数变动时，AIC及BIC的值越小则表示模型的拟合效果越好。由于高斯混合模型中$K \geqslant 2$，因此选择以$K=2$作为起点，通过逐一列举的方式计算模型拟合预报误差分布的AIC及BIC的值，并选取对应AIC及BIC值最小的K值作为当前预见时刻误差分布最优的高斯分布混合个数。

3.2 基于K-means++的参数初始值确定 K-means++算法是一种确定聚类迭代初始起点的算法[16]，一般用于初始化K-means++聚类算法的聚类中心。高斯混合模型本质上作为一种聚类模型，其原理是通过计算每个数据所属的高斯类从而实现聚类的效果，因此同样可以利用K-means++算法进行预分类，并根据预分类的结果计算高斯混合模型的初始参数值，其具体步骤如下：

(1)从单一预见时刻误差数据集中随机选定与最优高斯分布混合数K相同数目的初始聚类中心；

(2)计算每个误差数据$e_{x(i)t(j)}$与之最近一个聚类中心的马氏距离$D\left(e_{x(i)t(j)}\right)$；

(3)计算每个误差数据被选为下一个聚类中心的概率$\dfrac{D\left(e_{x(i)t(j)}\right)}{\sum_{i \in n} D\left(e_{x(i)t(j)}\right)^2}$，按照轮盘法选出下一个聚类中心；

(4)重复(2)(3)，直至每一个聚类中心不再变化为止；

(5)依据得到的聚类中心将误差数据聚类，并计算每个类的均值u'_k与方差$\left(\sigma_k^2\right)'$作为迭代的初始值，其中权值的初始值$\alpha'_k=\dfrac{1}{K}$。

当初始参数值确定后，利用EM算法迭代计算则可得到最终的α_k、u_k、σ_k^2，带入式(3)即可得到单一预见时刻的误差分布表达式。

3.3 多个预见时刻径流预报误差联合分布建立及应用 meta-elliptic Copula函数族中的高维meta-student t Copula[17]、高维meta-Gaussian Copula[18]是水文中常用的两类分布，由于在预报模型在预报过程中受到的干扰因素难以量化，可能会在某一时间点产生与均值偏离较大的预报误差，而高维meta-Gaussian Copula不具有尾部相关性，因此本文在推求各个预见时刻误差分布的前提下，选用高维meta-student t Copula建立多个预见时刻径流预报误差间的联合分布，可表示为：

$$C\left(\varphi_1,\ \varphi_2,\ \cdots,\ \varphi_m\right)=t_{\boldsymbol{\Sigma}\nu}\left[t_\nu^{-1}\left(\varphi_1\right),\ t_\nu^{-1}\left(\varphi_2\right),\ \cdots,\ t_\nu^{-1}\left(\varphi_m\right)\right] \tag{10}$$

将其展开可得：

$$C\left(\boldsymbol{\varphi};\ \boldsymbol{\Sigma};\ \nu\right)=\int_{-\infty}^{t_\nu^{-1}(\varphi_1)}\int_{-\infty}^{t_\nu^{-1}(\varphi_m)}\frac{\Gamma\left(\dfrac{\nu+m}{\nu}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)\sqrt{(\pi\nu)^m\left|\boldsymbol{\Sigma}\right|}}\left(1+\frac{1}{\nu}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}X\right)^{\frac{\nu+m}{2}}\mathrm{d}x \tag{11}$$

式中：$C\left(\varphi_1,\ \varphi_2,\ \cdots,\ \varphi_m\right)$为$m$维随机变量联合分布，$\varphi_1,\ \varphi_2,\ \cdots,\ \varphi_m$分别表示$m$个预见时刻的入库径流预报误差；$t_{\boldsymbol{\Sigma}\nu}$为$\nu$个自由度的$t$分布函数，其协方差矩阵为$\boldsymbol{\Sigma}$；$t_\nu^{-1}(\cdot)$为自由度为$\nu$的$t$分布的逆函数；$\Gamma(\cdot)$为伽马分布；$\boldsymbol{X}$为不同预见时刻的预报误差变量矩阵，$\boldsymbol{X}=\left(\varphi_{x(i)t(1)},\ \varphi_{x(i)t(2)},\ \cdots,\ \varphi_{x(i)t(m)}\right)$；$\boldsymbol{\varphi}$为被积函数变量矩阵，$\boldsymbol{\varphi}=\left[\varphi_1,\ \varphi_2,\ \cdots,\ \varphi_m\right]^{\mathrm{T}}$。

当对整个入库径流预报过程进行不确定性分析时，需要依据入库径流过程预报误差联合分布对误差序列进行随机抽样，主要分为如下两步。

(1)依据联合分布$C\left(\varphi_1,\ \varphi_2,\ \cdots,\ \varphi_m\right)$生成随机序列矩阵$\boldsymbol{\omega}=\left[\omega_1,\ \omega_2,\ \cdots,\ \omega_m\right]$，$\omega_j$表示对应预见时刻误差分布的累积分布概率。

(2)推求对应预见时刻误差分布的逆累积分布函数$F_j^{-1}\left(\varphi_i\right)$，将随机序列矩阵$\boldsymbol{\omega}$带入该函数，可得到模拟的入库径流过程误差序列，表示为$e'=\left(F_1^{-1}\left(\omega_1\right),\ F_2^{-1}\left(\omega_2\right),\ \cdots,\ F_m^{-1}\left(\omega_m\right)\right)$。

基于IGMM-Copula的入库径流过程预报误差随机模拟模型的研究流程如图1所示。

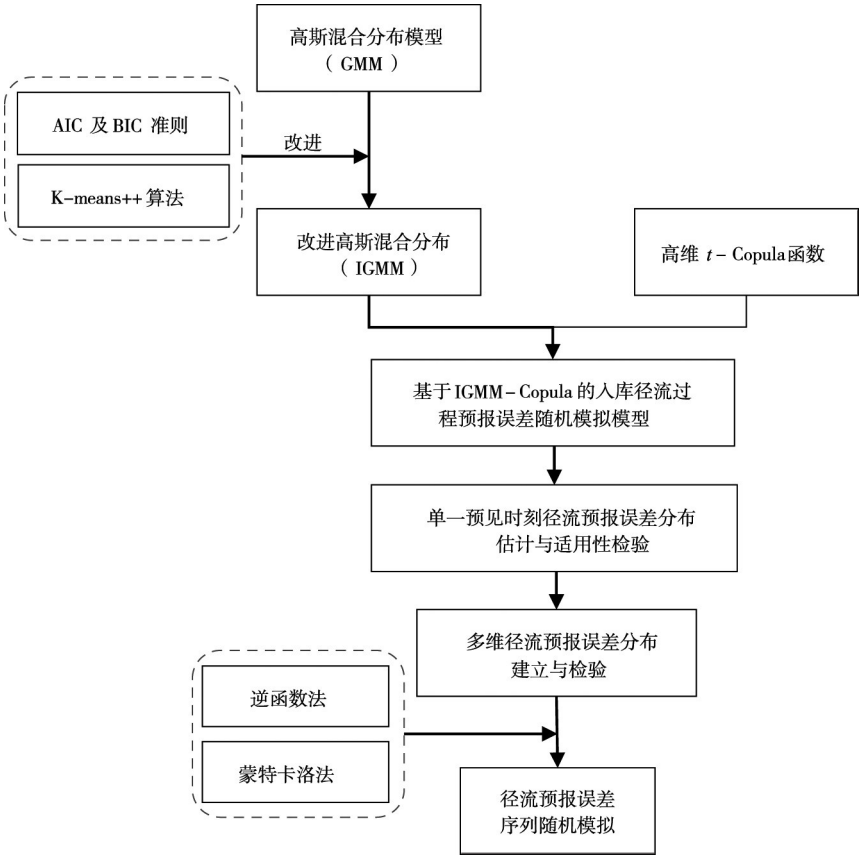

图1　基于IGMM-Copula的入库径流过程预报误差随机模拟模型的流程

4　基于IGMM-Copula的入库径流过程预报误差随机模拟模型

锦屏一级水电站水库在雅砻江下游“锦官电源组(锦屏一级、锦屏二级、官地)”梯级水电站水库群中起控制性作用，其入库径流预报采用的是三水源新安江模型[19]，具体是以每天早上的8点作为预报起点，以连续预报的形式对当天14点、20点、次日2点及次日的8点进行径流预报(依次对应预见时刻为6 h、12 h、18 h和24 h)，一次预报可得到未来一天的预报流量序列，以此制订水电站水库的调度计划[20-21]，预报的实施过程见图2。自投产以来，存在因来水预测误差造成水库运行水位超出控制范围而需要调整发电计划的情况[22]，本文选用锦屏一级水电站水库2014年1月1日至2017年9月23日每日的实测及预报径流资料，根据式(1)计算得到未来24 h内的预报误差序列(每个序列包含4个预见时刻的预报误差)，共得到1362条预报误差序列样本，而后利用本文所提方法建立基于IGMM-Copula的入库径流误差随机模拟模型，一方面通过求解各预见时刻的径流预报误差分布，分析了确定性预报模型产生预报误差的统计特征和分布规律，同时也考虑了不同预见时刻下预报误差的相关性，实现了对预报误差的多元联合模拟，为根据预报资料制订水库调度计划，提供了更多参考信息。

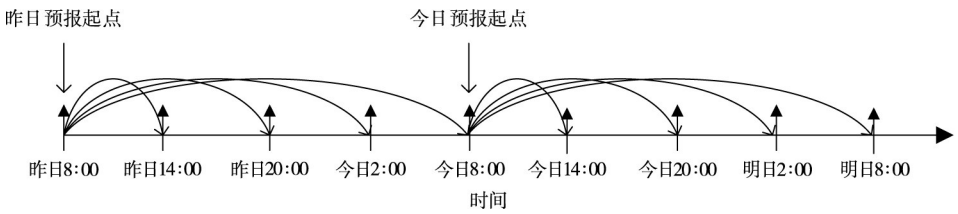

图2　锦屏一级水电站水库径流预报实施过程

4.1 各预见时刻径流预报误差分布求解 基于IGMM-Copula模型中的分布拟合部分(此处称为IGMM)对各预见时刻径流预报误差经验分布进行拟合，首先需要确定各时刻预报误差的最优高斯混合数，在原GMM模型中，由于没有相应的指标作为参考，作者偏向于参数较少，更容易求解的模型结构，因此最终选取的高斯混合数为K=2。现以K=2为起点，根据逐一列举法计算不同K值下的AIC及BIC的值。其计算结果如表1。

表1 不同高斯混合数下各预见时刻径流预报误差IGMM的AIC及BIC的值

K	*AIC*				*BIC*			
	6h	12h	18h	24h	6h	12h	18h	24h
2	9607.04	9826.18	9636.05	9889.08	9633.13	9871.78	9672.53	9915.16
3	9573.79	9812.80	9627.70	9872.16	9615.52	9854.54	9669.43	9913.90
4	9578.30	9817.34	9632.83	9874.52	9647.88	9881.38	9700.24	9939.17
5	9588.31	9822.19	9642.83	9883.94	9659.87	9890.50	9718.69	9958.27

从表1可以看出，以6 h、12 h、18 h、24 h径流预报误差为基础建立的IGMM在AIC及BIC的值取得最小时，对应的最优高斯混合数为K=3，据此可建立对应权重的IGMM对各预见时刻径流预报误差进行拟合，表2为基于IGMM的各预见时刻径流误差分布参数估计值。为体现IGMM在误差分布估计方面的优势，将原GMM与单高斯分布(SGM)作为比较模型。其中，GMM高斯混合个数的选取及参数初始值的确定均按照原文献里介绍的方法进行计算。

表2 基于IGMM的各预见时刻径流误差分布参数估计值

预见时刻/h	K	权重(α)	均值(μ)	方差(σ^2)
6	3	α_1=0.4635，α_2=0.1769，α_3=0.3596	μ_1=0.9872，μ_2=-1.6250，μ_3=0.1835	σ_1^2=11.2367，σ_2^2=268.2060，σ_3^2=84.7914
12	3	α_1=0.3573，α_2=0.0021，α_3=0.6406	μ_1=-3.5116，μ_2=59.4452，μ_3=-0.6708	σ_1^2=206.3921，σ_2^2=23.6007，σ_3^2=23.4097
18	3	α_1=0.3350，α_2=0.0033，α_3=0.06617	μ_1=-0.6037，μ_2=-39.7471，μ_3=0.5424	σ_1^2=167.5890，σ_2^2=17.4578，σ_3^2=27.2638
24	3	α_1=0.6347，α_2=0.2540，α_3=0.1113	μ_1=0.8350，μ_2=-2.1944，μ_3=-0.8019	σ_1^2=35.0500，σ_2^2=122.9527，σ_3^2=335.5607

图3是3种模型对不同预见时刻径流预报误差的拟合效果图，其中实测预报误差样本个数$n \geqslant 1000$，据此做出的频率直方图总体呈现出对称的形状，误差数据大部分集中在0值附近，SGM拟合得到的概率密度函数较为矮胖，在峰部的集中度与频率直方图相差较远，而IGMM与GMM由于自适应性较强，两者的拟合结果在数据集中的位置更契合频率直方图，在图中体现为概率密度函数的形状更为尖瘦，因此能更大程度的反映实测预报误差样本所展现的规律，这也印证了高斯混合模型在概率密度估计中的优越性。同时，由于不同预见时刻的径流预报误差具有不同的特征，而人为确定GMM中高斯混合数与参数初始值难以达到最佳的拟合效果，尤其对于预报误差数据峰部的拟合，基于IGMM得到的拟合曲线的拟合效果明显优于GMM，其原因主要在于IGMM对高斯混合数进行了优选，并对参数初始值进行了预计算，因此拟合曲线更贴近于频率直方图。

4.2 模型适用性检验 适用性检验包括拟合优度检验与后验检验，拟合优度检验的目的是用历史数据来检验模型是否能够反映径流误差的随机特性[23]，具体包括χ^2检验、t-检验、K-S检验等，其中K-S检验是一种检验单个总体是否服从某一理论分布的非参数检验方法，相比于χ^2检验能够更充分地利用样本信息，适用于连续和定量数据[24]。K-S检验统计量D_n定义见下式：

$$D_n = \max_{i \in (1,\ 2,\cdots,\ n)} \left(F_n\left(e_{x(i)t(j)}\right) - F_t\left(e_{x(i)t(j)}\right) \right) \tag{12}$$

式中：D_n越小说明拟合优度越高；$F_n(\cdot)$为经验分布函数；$F_t(\cdot)$为假设的理论分布函数。

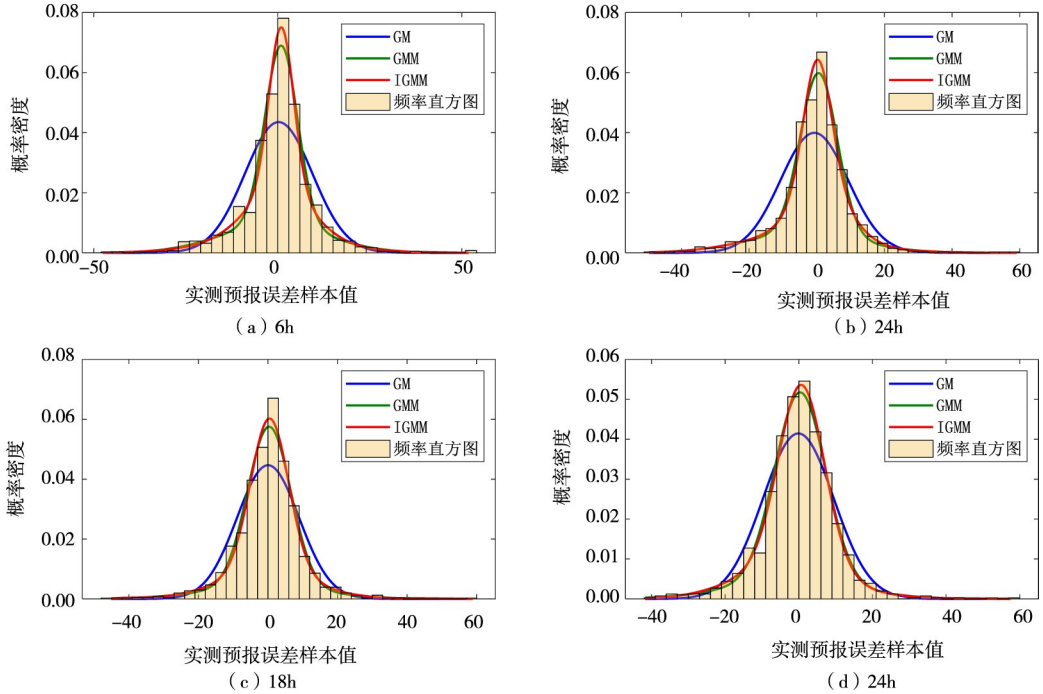

图3　三种模型对不同预见时刻径流预报误差的拟合效果

分别对改进高斯混合模型、核密度估计模型以及单高斯模型进行K-S检验，显著性水平设置为0.01，其计算结果见表3。

表3　模型K-S检验指标值

模型	显著性水平	临界值	D_n			
			6h	12h	18h	24h
IGMM			0.0158	0.0123	0.0186	0.0178
GMM	0.01	0.0440	0.0216	0.0156	0.0311	0.0213
SGM			0.0947	0.1031	0.0663	0.0660

从表3可看出，IGMM及GMM都通过K-S检验，即两者与经验分布在显著性水平为0.01的前提下没有显著性差异，验证了IGMM与GMM模型在径流预报误差密度估计中的可行性，同时IGMM在各预见时刻下误差拟合曲线的 D_n 值都为最小，说明其对于单一预见时刻径流预报误差经验分布的拟合优度最高。

后验检验的目的是定量评估概率模型与数据观测分布（即统计直方图）之间的差异[25]，采用式（13）的均方根误差 ε_{RMSE} 与式（14）的平均误差百分数 ε_{MAPE} 作为指标对求得的各预见时刻径流误差分布进行检验，得到的指标越小，说明拟合模型与数据观测分布之间的差异越小，其计算结果见表4。

$$\varepsilon_{RMSE} = \sqrt{\sum_{i=1}^{N}\left(F_n\left(e_{x(i)t(j)}\right) - F_t\left(e_{x(i)t(j)}\right)\right)^2 \Big/ N} \tag{13}$$

表4　模型后验检验指标值

预见时刻/h	ε_{RMSE}			ε_{MAPE}		
	IGMM	GMM	SGM	IGMM	GMM	SGM
6	0.0042	0.0083	0.0589	1.7573	2.4622	17.8822
12	0.0041	0.0071	0.0586	1.3031	1.9966	18.8490
18	0.0060	0.0096	0.0432	1.6941	2.4455	13.7431
24	0.0057	0.0075	0.0366	1.7389	2.1126	12.2327

$$\varepsilon_{MAPE}=\frac{1}{n}\sum_{i=1}^{N}\left|\frac{F_t\left(e_{x(i)t(j)}\right)-F_n\left(e_{x(i)t(j)}\right)}{F_n\left(e_{x(i)t(j)}\right)}\times 100\%\right| \tag{14}$$

由表4可见，对于所有预见时刻而言，基于IGMM得到的误差分布的 ε_{RMSE} 都保持在0.06及以下，ε_{MAPE} 都保持在1.8以下，相较于SGM与GMM，在两指标上的表现最好，说明基于IGMM得到的各预见时刻误差分布与数据观测分布之间的差异相差最小。

为分析不同预见时刻间径流预报误差分布的差异，图4将IGMM拟合得到的各径流预报误差分布曲线汇总，从图中可观察到，随着预见时刻的增加，误差分布形状逐渐从“尖瘦型”变化为“矮胖型”，表面预报不确定在随时间的增加而增大，符合确定性预报模型的预报规律。

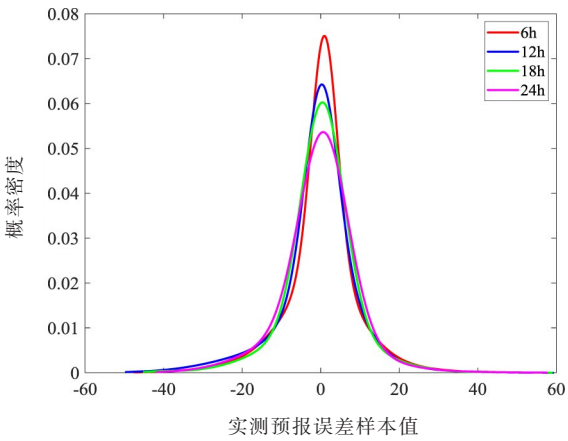

图4　各预见时刻径流预报误差分布曲线

4.3 径流预报误差序列随机模拟　首先分析预见期内各时刻径流预报误差两两间的相关性，$e_{t(6)}$、$e_{t(12)}$、$e_{t(18)}$、$e_{t(24)}$ 分别表示6 h、12 h、18 h、24 h的径流预报误差，采用Kendall相关系数作为相关性的度量并进行双侧显著性检验，设置零假设为：变量 $e_{t(6)}$、$e_{t(12)}$、$e_{t(18)}$、$e_{t(24)}$ 没有相关性，显著性水平为0.05，其计算及检验结果如表5所示。

表5　各预见时刻径流预报误差间的Kendall相关系数

		$e_{t(6)}$	$e_{t(12)}$	$e_{t(18)}$	$e_{t(24)}$
$e_{t(6)}$	相关系数	1	0.309	0.28	0.218
	双侧显著性检验 P 值		0.000	0.000	0.000
$e_{t(12)}$	相关系数	0.309	1	0.329	0.305
	双侧显著性检验 P 值	0.000		0.000	0.000
$e_{t(18)}$	相关系数	0.280	0.329	1	0.428
	双侧显著性检验 P 值	0.000	0.000		0.000
$e_{t(24)}$	相关系数	0.218	0.305	0.428	1
	双侧显著性检验 P 值	0.000	0.000	0.000	

从表5可知，双侧假设检验的P值均小于显著性水平，因此拒绝零假设，认为各预见时刻的预报误差均存在正相关性，考虑利用高维 t–Copula建立径流过程预报误差的联合分布并据此进行误差序列抽样，实现对预报误差序列的随机模拟。

为检验高维 t–Copula对多维误差的适用情况，根据图形分析法原理，以IGMM求得的6 h、12 h、18 h、24 h径流预报误差分布为例，分别按式(11)与式(15)计算误差的理论联合分布概率和经验联合分布概率，并点绘两者的关系图，见图5。

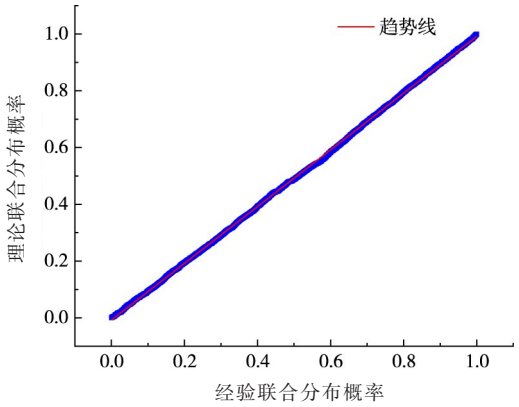

图5　理论联合分布概率与经验联合分布概率关系

$$\begin{aligned}&F\left(e_{x(1)t(j)},\ e_{x(2)t(j)},\ \cdots,\ e_{x(n)t(j)}\right)\\&=P\left(X_1\leqslant e_{x(1)t(j)},\ X_2\leqslant e_{x(2)t(j)},\ \cdots,\ X_n\leqslant e_{x(n)t(j)}\right)=\frac{N_q-0.44}{N+0.22}\end{aligned} \tag{15}$$

式中：N_q 为同时满足 $X_1 \leqslant e_{x(1)t(j)}$，$X_2 \leqslant e_{x(2)t(j)}$，…，$X_n \leqslant e_{x(n)t(j)}$ 的样本个数；N 为样本总量。

其中，由于实测预报误差样本个数较大，因此绘出的散点图比较密集（图5中蓝色数据点），形状趋向于一条线，由趋势线的斜率可看出，图中的点都大致落在45°对角线两侧，理论联合分布概率和经验联合分布概率的 *RMSE* 为0.00907，表明利用高维 *t*-Copula 建立多个预见时刻径流预报误差的联合分布是合理可行的。

现根据IGMM求出的各预见时刻径流预报误差分布为边缘分布，以高维 *t*-Copula 作为连接函数，建立基于IGMM-Copula的入库径流过程预报误差随机模拟模型，并与GMM-Copula模型进行对比，按照2.3节的步骤分别利用两模型对预报误差序列进行500 000次随机模拟，计算IGMM-Copula与GMM-Copula模拟预报误差的统计参数，计算结果见表6和表7。

表6　GMM-Copula模拟预报误差与实测预报误差的统计参数值

预见时刻	均值 u		方差 σ^2		变差系数 c_v	
	模拟预报误差	实测预报误差	模拟预报误差	实测预报误差	模拟预报误差	实测预报误差
6h	0.2453	0.2361	83.1753	84.0809	37.1839	38.8375
12h	−0.6753	−0.7046	97.6418	99.7676	−14.6324	−14.1763
18h	0.0288	0.0252	76.9179	79.8370	304.5237	354.5920
24h	−0.1098	−0.1167	90.3502	92.6293	−87.0764	−82.4656

表7　IGMM-Copula模拟预报误差与实测预报误差的统计参数值

预见时刻	均值 u		方差 σ^2		变差系数 c_v	
	模拟预报误差	实测预报误差	模拟预报误差	实测预报误差	模拟预报误差	实测预报误差
6h	0.2375	0.2361	83.5746	84.0809	38.4885	38.8375
12h	−0.7014	−0.7046	99.1580	99.7676	−14.1965	−14.1763
18h	0.0269	0.0252	78.8725	79.8370	330.0785	354.5920
24h	−0.1136	−0.1167	91.6564	92.6293	−84.2483	−82.4656

由表中可知，相较于GMM-Copula模型，IGMM-Copula模型各预见时刻模拟预报误差的统计参数更贴近于实测预报误差，而在18 h时，由于误差均值较小，模型模拟预报误差与实测预报误差的变差系数产生了较大的差异，而IGMM-Copula模型明显更优于GMM-Copula模型，这也反应了不同的预报误差边缘分布会对抽样结果造成一定程度的影响。从统计参数的整体变化趋势分析，IGMM-Copula模型的模拟预报误差能与实测预报误差保持一致，说明模型能够很好的反映径流预报误差序列的统计特性，可作为对预报结果进行修正的参考依据。

5　结论

本文在GMM-Copula模型的基础上，对高斯混合数的选取以及初始参数值的确定进行了改进，建立了基于IGMM-Copula的入库径流过程预报误差随机模拟模型，模型可运用于单个预见时刻及多个预见时刻的径流预报误差分析与模拟。以锦屏一级水库为例的结果表明：对于单一预见时刻的径流预报误差的描述，IGMM-Copula模型在拟合优度检验及后验检验中皆优于单高斯与GMM-Copula模型；通过IGMM-Copula模型模拟得到的误差序列的均值、方差和变差系数相较于GMM-Copula模型更贴近于实测误差序列，统计参数的变化规律基本一致，验证了模型的可行性，不仅为研究径流预报误差提供一种新的方法，也为水库短期优化调度计划的制订提供了更为丰富的参考信息。

参 考 文 献：

[1] 王晶晶，梁忠民，王辉．基于熵法的入流预报误差规律研究[J]．水电能源科学，2010，28(7)：12-14，126．

[2] 翟劭燚，张晓雪，刘九夫，等．大伙房水库洪水预报误差分布规律分析[J]．水文，2011，31(6)：14-17．

[3] 刘心愿，郭生练，李响，等．考虑水文预报误差的三峡水库防洪调度图[J]．水科学进展，2011，22(6)：771-779．

[4] 何洋，纪昌明，田开华，等．基于最大熵原理的入库径流预报误差分布规律研究[J]．中国农村水利水电，2016(11)：115-120．

[5] 张验科，刘源，纪昌明，等．考虑水流时滞的梯级水电站水库群短期发电优化调度[J]．电力系统自动化，2020，44(6)：45-53．

[6] 马超，崔喜艳．水库月平均流量滚动预报及其不确定性研究[J]．水力发电学报，2018，37(2)：59-67．

[7] 蒋志强，陈育赟，徐杨，等．基于预报难度的径流预报水平综合评定[J]．水力发电学报，2020，39(2)：81-93．

[8] ZHAO T，CAI X，YANG D．Effect of streamflow forecast uncertainty on real-time reservoir operation[J]．Advances in Water Resources，2011，34(4)：495-504．

[9] ZHAO T，ZHAO J，YANG D，et al．Generalized martingale model of the uncertainty evolution of streamflow forecasts [J]．Advances in Water Resources，2013，57(9)：41-51．

[10] LI X，GUO S，LIU P，et al．Dynamic control of flood limited water level for reservoir operation by considering inflow uncertainty[J]．Journal of Hydrology (Amsterdam)，2010，391(1/2)：124-132

[11] ZHOU Y，GUO S．Risk analysis for flood control operation of seasonal flood-limited water level incorporating inflow forecasting error[J]．International Association of Scientific Hydrology Bulletin，2014，59(5)：1006-1019．

[12] 纪昌明，李荣波，田开华，等．基于来水不确定性的梯级水电站负荷调整耦合模型——以锦官电源组梯级水电站为例[J]．水利学报，2017，48(1)：1-12．

[13] 纪昌明，梁小青，张验科，等．入库径流预报误差随机模型及其应用[J]．水力发电学报，2019，38(10)：75-85．

[14] LAW MARTIN H C，FIGUEIREDOMÁRIO A T，JAIN ANIL K．Simultaneous feature selection and clustering using mixture models[J]．Ieee Transactions on Pattern Analysis and Machine Intelligence，2004，26(9)：1154-1166．

[15] 舒晓娟，陈洋波，任启伟．模型选择准则在洪水频率分析中的应用[J]．水利学报，2010，41(1)：80-85．

[16] 张良均，杨坦，肖刚．MATLAB数据分析与挖掘实战[M]．北京：机械工业出版社，2015．

[17] 马超，崔喜艳．水库月平均流量滚动预报及其不确定性研究[J]．水力发电学报，2018，37(2)：59-67．

[18] 杜奕良，涂新军，杜晓霞，等．基于高维Gaussian Copula函数的区域农田灌溉需水分析[J]．水利学报，2018，49(3)：323-331．

[19] 李匡，何朝晖，梁犁丽．雅砻江流域水情预报方案[J]．人民珠江，2017，38(8)：33-38．

[20] 何洋，纪昌明，田开华，等．基于最大熵原理的径流预报误差分布规律研究[J]．中国农村水利水电，2016(11)：115-120．

[21] 李荣波．梯级电站优化调度与负荷调整模型及决策方法研究[D]．北京：华北电力大学，2017．

[22] 蒋志强，武文杰，覃晖，等．考虑预报误差的水电站短期调度模糊风险研究[J]．水力发电学报，2019，38(2)：38-48．

[23] 颜伟，任洲洋，赵霞，等．光伏电源输出功率的非参数核密度估计模型[J]．电力系统自动化，2013，37(10)：35-40．

[24] 寇燕明，迟晓晴，王艳慧．一种基于假设检验的数据筛选算法[J]．统计与决策，2020，36(8)：21-26．

[25] QIN Z，LI W，XIONG X．Estimating wind speed probability distribution using kernel density method[J]．Electric Power Systems Research，2011，81(12)：2139-2146．

Stochastic simulation model of forecast errors in the process of reservoir runoff based on IGMM-Copula

ZHANG Yanke, ZHANG Jiaxin, TAI Yuhang, JI Changming, MA Qiumei

(*School of Water Resources and Hydropower Engineering*, *North China Electric Power University*, *Beijing* 102206, *China*)

Abstract: In order to reveal the statistical characteristics and changing laws of forecast errors in the process of inflow runoff, so as to provide more accurate input conditions for the optimal operation of hydropower stations and reservoirs. The optimal Gaussian mixture number is selected based on the AIC and BIC criteria, and the K-means++ algorithm in data mining is introduced to determine the initial parameter values of the Gaussian mixture model. The GMM part of the GMM-Copula model is improved, based on which the stochastic simulation model of the forecast error in the reservoir runoff process is established. The model not only has more advantageous in the quantitative estimation of the runoff forecast error at a single foreseeable moment, but also can realize the stochastic simulation of the error sequence by establishing a multi-dimensional distribution function of the error. The results applied to the reservoir of Jinping I Hydropower Station show that the model in this paper fits the runoff forecast errors at 6h, 12h, 18h, 24h foreseeing moments, and the graphical effect and applicability test results of the fitted curve are better than those of GMM -Copula model, and the statistical parameters of the simulation error of the IGMM-Copula model are closer to the actual error, which verifies the rationality and feasibility of the model, and provides a more accurate and effective method for the estimation and simulation of the forecast error of the reservoir runoff process method.

Keywords: reservoir runoff forecast error; GMM-Copula; IGMM-Copula; stochastic simulation; Jinping I Hydropower Station

（责任编辑：杨　虹）

论文摘要编写要点

论文摘要十分重要，它是沟通读者和作者之间的桥梁。在今天信息时代，读者不可能阅读刊物的每一篇论文去查找所需的信息，只有通过摘要了解论文的主要内容，从而判断有无必要阅读全文。国内外的检索系统为了信息交流，更建立了各种二次文献数据库，帮助读者通过查找论文摘要，以便提取原文。而二次文献数据库的基础是作者的论文摘要。但许多作者却往往忽视了论文摘要的这一重要意义，没有下工夫把论文摘要写好，尤其是英文摘要。当今科技领域，英文已经成为国际交流语言，世界各国学者想追踪了解某一学科的发展情况，多会用英文工具书、数据库进行检索。而国外的检索系统也主要通过英文摘要判断论文是否被收录进数据库。为了帮助作者写好摘要，我编辑部在给作者的论文修改通知中都附有《摘要编写要点》，供作者参考。现再刊登于下，以便作者查阅。

1. 摘要是论文内容不加注释和评论的简短陈述，应包括正文的要点，具有独立性和自含性，让读者不阅读全文就能了解论文的基本内容，以判断有无必要阅读全文，也可供二次文献采用。

2. 摘要应说明研究工作的目的、方法与手段、结果和结论，要尽量简短，尽可能省略课题的背景信息。

3. 摘要中的内容应在正文中出现，不能对正文进行补充和修改。

4. 摘要中不用图、表、非公用共知的符号和术语，不能引用文献；缩写名称在第一次出现时要有全称(包括中文和英文)。

5. 摘要中不要多列数据，出现的应该是最重要的、最关键数据。

6. 中文摘要一般300字左右，中英文摘要应基本对照，不能因为某些内容不好翻译就略去。

7. 关键词是为了文献标引工作，是从论文中选取出来用以表示全文主题内容信息的单词或术语，一般为3～8个词，尽量用规范词。

《水利学报》编辑部