因子分析—方法简介

因子分析法是指从研究指标相关矩阵内部的依赖关系出发，把一些信息重叠、具有错综复杂关系的变量归结为少数几个不相关的综合因子的一种多元统计分析方法。 基本思想是：根据相关性大小把变量分组，使得同组内的变量之间相关性较高，但不同组的变量不相关或相关性较低，每组变量代表一个基本结构一即公共因子

传统的基于 ANOVA 的敏感性分析方法概括为：

$$SST = \sum_{i=1}^{k} SS_i + \sum_{i=1}^{k}\sum_{j>1}^{k} SS_{ij} + \ldots + SS_{1,\,2,\,\ldots,\,k}$$

其中 SS_i 代表单因子效应，SS_{ij} 至 $SS_{1,\,2,\,\ldots,\,k}$ 代表 k 个因子之间的交互作用，SST 代表所有因子和交互的方差和。

单二次抽样 ANOVA 方法指只对模型中一个参数进行二次抽样，其表达公式如下所示：

$$SST^{j} = \sum_{t_1=1}^{T_1}\sum_{t_2=1}^{T_2}\cdots\sum_{h=1}^{2}\cdots\sum_{t_k=1}^{T_k}\left(Y^{t_1,\,t_2\cdots g(h,\,j)\cdots t_k} - Y^{o,\,o\cdots g(o,\,j)\cdots o}\right)^2$$

当 i = n 时，

$$SS_i^{j} = T_1T_2\cdots T_{n-1}T_{n+1}\cdots T_k\sum_{h=1}^{2}\left(Y^{t_1,\,t_2\cdots g(h,\,j)\cdots t_k} - Y^{o,o\cdots g(o,\,j)\cdots o}\right)^2$$

当 $i \neq n$，

$$SS_i^{j} = 2\times T_1T_2\cdots T_{i-1}T_{i+1}\cdots T_{n-1}T_{n+1}\cdots T_k\sum_{t_i=1}^{T_i}\left(Y^{o,\,o\cdots t_i\cdots g(o,\,j)\cdots o} - Y^{o,\,o\cdots g(o,\,j)\cdots o}\right)^2$$

多二次抽样方法指多个模型参数同时被二次抽样，可以表达为：

$$SST^{j} = \sum_{t_1=1}^{T_1}\sum_{t_2=1}^{T_2}\cdots\sum_{h_p=1}^{2}\cdots\sum_{h_q=1}^{2}\cdots\sum_{t_k=1}^{T_k}\left(Y^{t_1,\,t_2\cdots g(h_p,\,j_p)\cdots g(h_q,\,j_q)\cdots t_k} - Y^{o,o\cdots g(o,\,j_p)\cdots g(o,\,j_q)\cdots o}\right)^2$$

当 i = p, …, q 时，

$$SS_i^j = T_1 \times T_2 \times \cdots \times T_k \sum_{h_p=1}^{2} \cdots \sum_{h_q=1}^{2} (Y^{t_1, t_2 \cdots g(h_p, j_p) \cdots g(h_q, j_q) \cdots t_k} - Y^{o,o \cdots g(o, j_p) \cdots g(o, j_p) \cdots o})^2$$

当$i \neq p, ..., q$，

$$SS_i^j = 2 \times \cdots \times 2 \times T_1 \times T_2 \cdots T_{i-1} \times T_{i+1} \cdots T_k \sum_{t_i=1}^{T_i} (Y^{o, o \cdots t_i \cdots g(o, j_p) \cdots g(o, j_q) \cdots o} - Y^{o,o \cdots g(o, j_p) \cdots g(o, j_q) \cdots o})^2$$

全二次抽样方法指模型中所有参数都被二次抽样，数学表达式如下：

$$SST^j = \sum_{t_1=1}^{2} \sum_{t_2=1}^{2} \cdots \sum_{t_k=1}^{2} (Y^{g(h_1, j_1), g(h_2, j_2) \cdots g(h_k, j_k)} - Y^{g(o, j_1) g(o, j_2) \cdots g(o, j_k)})^2$$

$$SS_i^j = \sum_{h_1=1}^{2} \sum_{h_2=1}^{2} \cdots \sum_{h_k=1}^{2} (Y^{g(h_1, j_1), g(h_2, j_2) \cdots g(h_k, j_k)} - Y^{g(o, j_1), g(o, j_2), \cdots g(o, j_k)})^2$$

案例 1：基于因子分析的气候变化和人类活动对珠三角径流变化的影响。

为了定量分析气候变化和人类活动对珠三角地区径流变化的影响，本研究借助多水平-析因分析技术，选择不同气候因子种类，并考虑不同提前时间的前期气候，开展不同组合情景下的降水-径流模拟分析。首先，为了识别影响珠三角地区径流的最主要气象因子，本研究分析了 9 种气象因子和径流之间 95%置信度下的相关性，如

表 1 所示。由

表 1 可以看出，气压和风速和径流之间存在负的相关性，其余 7 种气象因子和径流之间显示正的相关性。与此同时，气象因子和径流的关系存在地域差异。例如，西江梧州站的径流和降雨之间的相关性是 0.52，而北江横石站的径流和降雨之间的相关性是 0.79。水汽压对径流的影响在西江比较大（相关性 0.77），对北江径流的影响相对比较小（相关性 0.52）。相关性的分析是为了筛选出最主要的气候预测因子，为建立统计模型的提供理论依据。

表 1 气候预测因子和径流之间的相关性

站点	气候预测因子								
	降雨	气压	风速	平均温度	水汽压	相对湿度	日最低温	日高温度	相对湿度
西江梧州	0.52	-0.75	-0.27	0.71	0.77	0.41	0.72	0.69	0.08
北江石角	0.42	-0.56	-0.19	0.50	0.55	0.47	0.51	0.46	0.01
北江横石	0.79	-0.53	-0.18	0.42	0.52	0.59	0.47	0.36	-0.05
东江博罗	0.71	-0.61	0.04	0.54	0.61	0.55	0.58	0.48	-0.08

（注：灰色背景框代表相关性没有通过 95%置信度的显著性检验。）

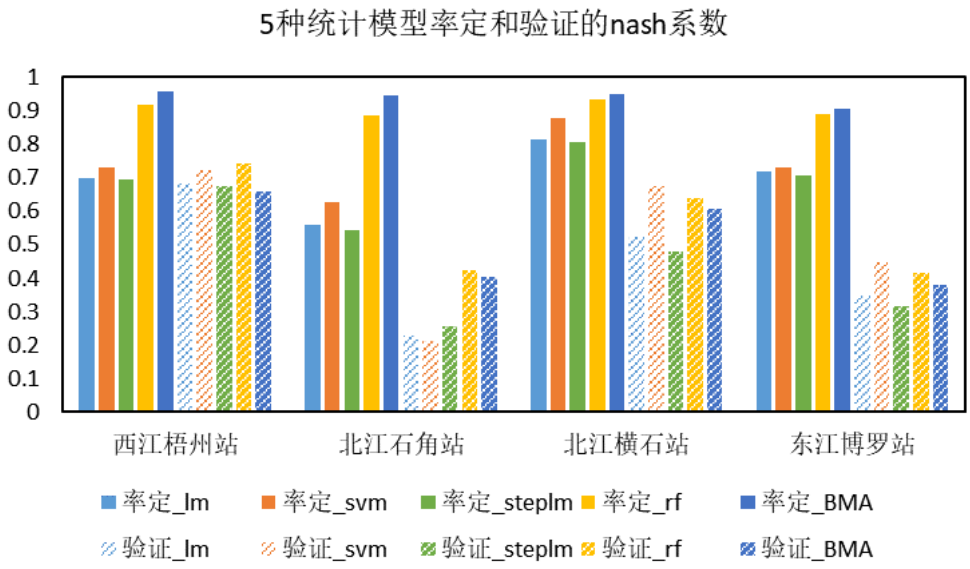

图 1 五种统计模型在 4 个站点的率定和验证效果

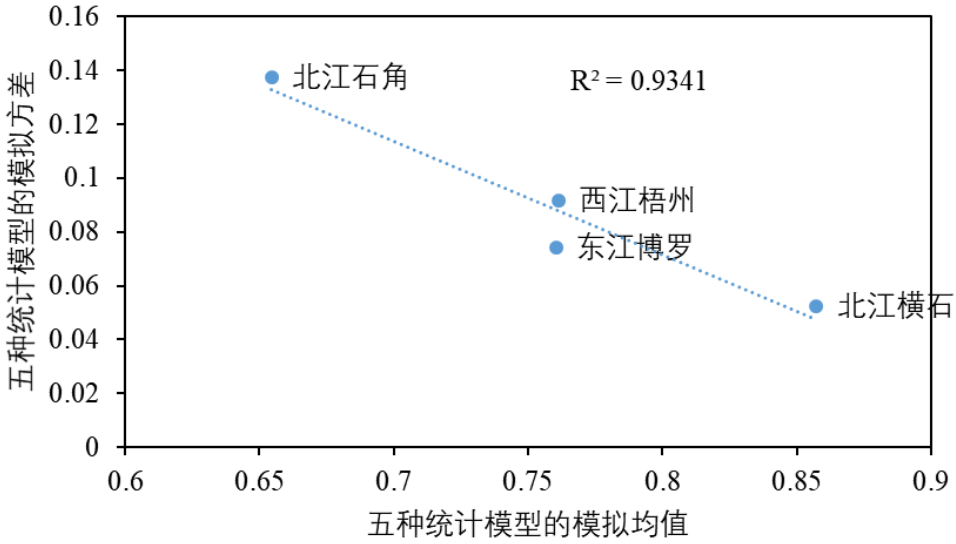

图 2 五种统计模型在 4 个站点模拟效果的均值和方差。

其次，本研究采用五种统计模型建立珠三角地区径流预测模型。这五种统计模型分别为线性回归模型（lm），支持向量机模型（svm），逐步回归模型（steplm），随机森林模型（rf）以及基于贝叶斯模型平均的集合模型（BMA）。本研究用 25 年历史数据（1955 - 1979）做模型率定，5 年历史数据（1980 -1984）做验证。 图 1 展示了五种统计模型在 4 个站点的率定和验证效

果（以 nash 系数为例）。从图 1 可以看出，5 种统计模型在不同站点模拟效果不同，其中随机森林模型和贝叶斯集合模型在率定期表现最好。然而在验证期，这两种模型并不一定是最优的模型。例如在北江横石站和东江博罗站，支持向量机模型在率定期有最优的模拟效果（nash 最高）。但整体而言，率定期的模拟效果优于验证期的模拟效果，这说明气候因子和径流的关系在验证期和率定期发生了变化。相比于西江，北江和东江的降水径流关系年际变化较大、且更易受模型不确定性的影响。另外，分析五种模拟模型在 4 个站点模拟效果的均值和方差，如图 2 所示。可以看出，统计模型之间的差异和模拟效果之间存在显著的线性关系（R2=0.93）。这说明，模型准确度由站点气候因子-径流关系决定，受模型结构差异的影响较弱。不同站点，气候因子对径流的解释度不同。

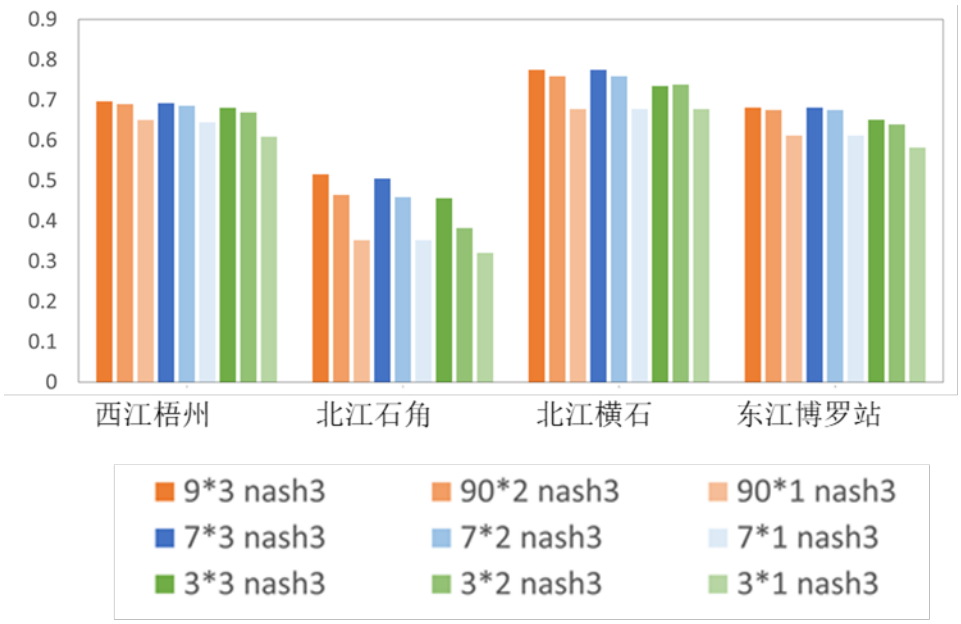

图 3 不同预测因子组合下集合预报模型在 4 个站点模拟效果

最后，为了定量分析当月气候因子，前期气候因子，以及下垫面和人类活动的影响，本研究选择不同气候因子种类，并考虑不同提前时间的前期气候，针对 9 种组合情景进行模拟分析，结果如图 3 所示。

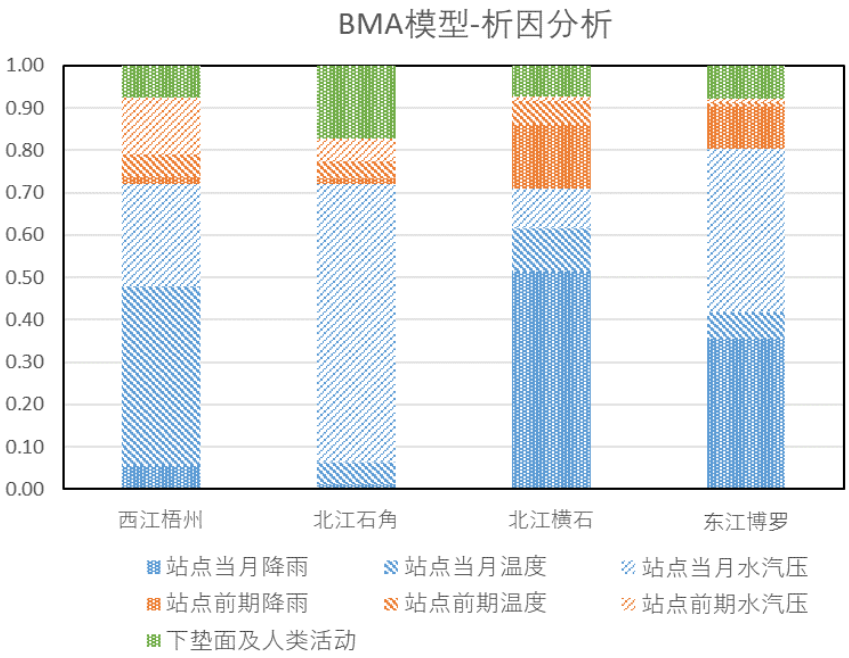

图 4 基于 BMA 多模型集合的多水平-析因分析。

此外，进一步利用多水平-析因分析技术，对图 3 模拟效果进行解析，结果如图 4 所示。由图可以看出，降雨对径流变化的贡献随地域变化。如北江横石站，本地当月降雨对径流变化的贡献是 51%，而在西江梧州站，本地当月降雨对径流的贡献仅为 5%。对北江石角站而言，当地的水汽压对径流变化的影响最大。整体而言，在珠三角地区，站点当月气候因子对径流变化起主要贡献（71%-80%），其次是前期气候条件（11%-22%），最后是下垫面变化和人类活动的影响（7%-17%）。

案例 2：基于因子分析的三参数模型敏感性分析

以一个简单的三参数模型为例，该模型表达式为：

$$F_3(x_1, x_2, x_3) = x_1 \times x_3 + x_1 \times \sin\left(\frac{\pi}{2} \times x_2\right) + x_2 \times e^{|x_3|} + x_1 \times x_2 \times x_3$$

其中，x_1，x_2，x_3是独立变量，且在[0, 1]之间均匀分布。

本研究采用单二次抽样、多二次抽样和全二次抽样方法对三个参数及其交互作用的敏感性进行定量分析，敏感性结果如图 5 所示。结果表明，应用三种方法得到的参数敏感性具有显著差异。以 Sobol’s 敏感性结果为基准，可以看出，参数的单因子敏感性会因为二次抽样而显著降低。这说明单二次抽样方法和多二次抽样方法会显著低估参数的敏感性。此外，对于全二次抽样方法而言，三个参数的个体敏感性和相互作用敏感性随参数水平的变化而变化。随着参数水平从 222 增加到 555，x1 和 x3 的个体敏感性分别从 11.7%和 19.4%逐渐增加到 19.1%和 24.1%。同时，交互参数灵敏度从 18.1%逐渐下降到 5.5%。x2 的个体灵敏度保持相对稳定，从 50.9%到 52.2%不等。单参数和参数间交互作用的敏感性会受到二次抽样参数水平的影响。增加抽样参数的水平数会略微增加该参数的敏感性。

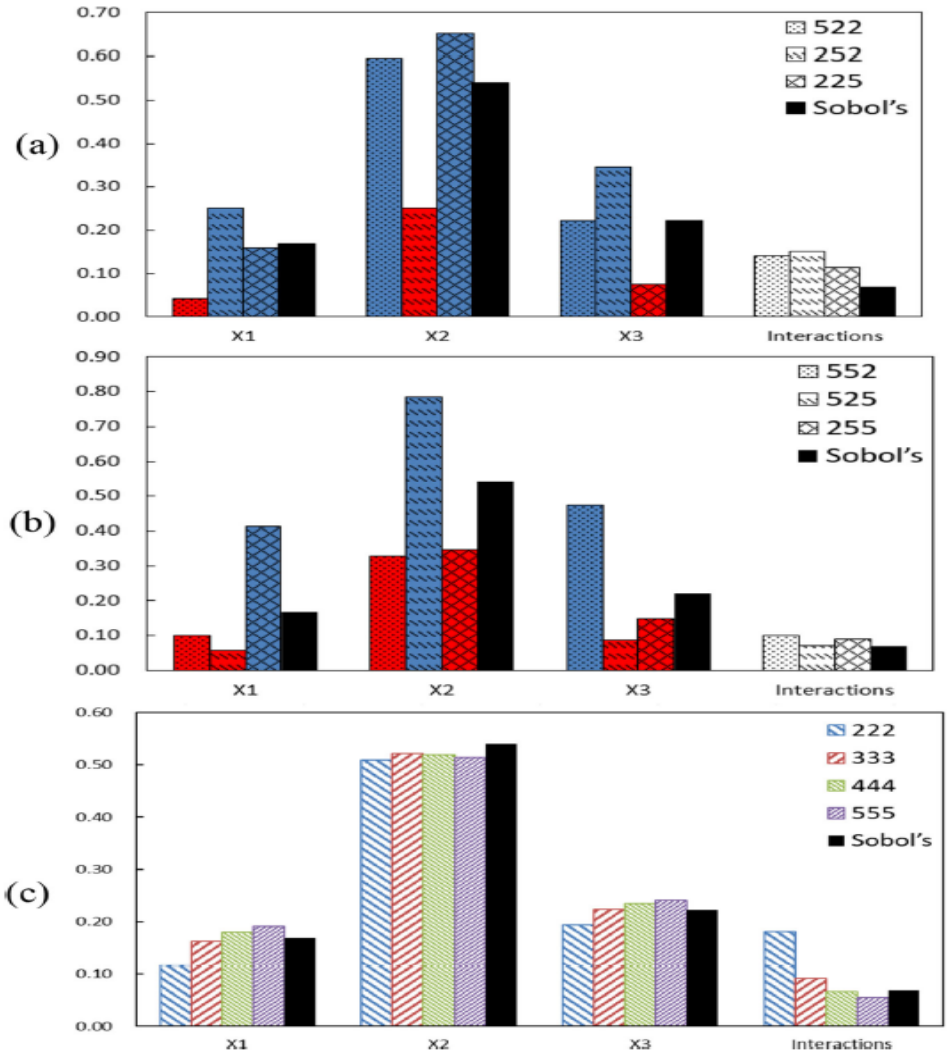

图 5. 三个参数及其交互的敏感性：（a）单二次抽样、（b）多二次抽样、（c）全二次抽样。

案例 3：基于再抽样因子分析的水文模型参数敏感性分析

为进一步研究上述三种方法在水文模拟中的可用性，这三种方法被应用于概念式水文模型 GR4J 中（详见图 6），以模拟增江流域（详见图 7）的径流量。GR4J 是降雨径流模型之一，在流域水文建模方面具有很强的基础性和有效性。该模型采用产流水库、汇流水库这两个线性水库进行产汇流计算，对湿润地区进行洪水预报和水资源规划，具有简便、准确等特点。该模型只包含 4 个参数，分别是：x_1 产流水库容量(mm)、 x_2 地下水交换系数(mm)、x_3 汇流水库容量(mm)、x_4 单位线汇流时间(day)。本文选用 GR4J 模型进行水文模拟和预报研究。其次，将该模型应用于增江，其为东江支流，是珠三角地区重要的流域之一。本研究所使用的数据（日蒸发量、日降水量和日径流）均来自于麒麟咀水文站，数据采集的时间范围为 2009-2015 年。麒麟咀水文站以上流域总面积为 2866 km^2，占增江流域（3160 km^2）的 91%。流域年平均气温和降水量分别为 21.6 °C 和 2188 mm。

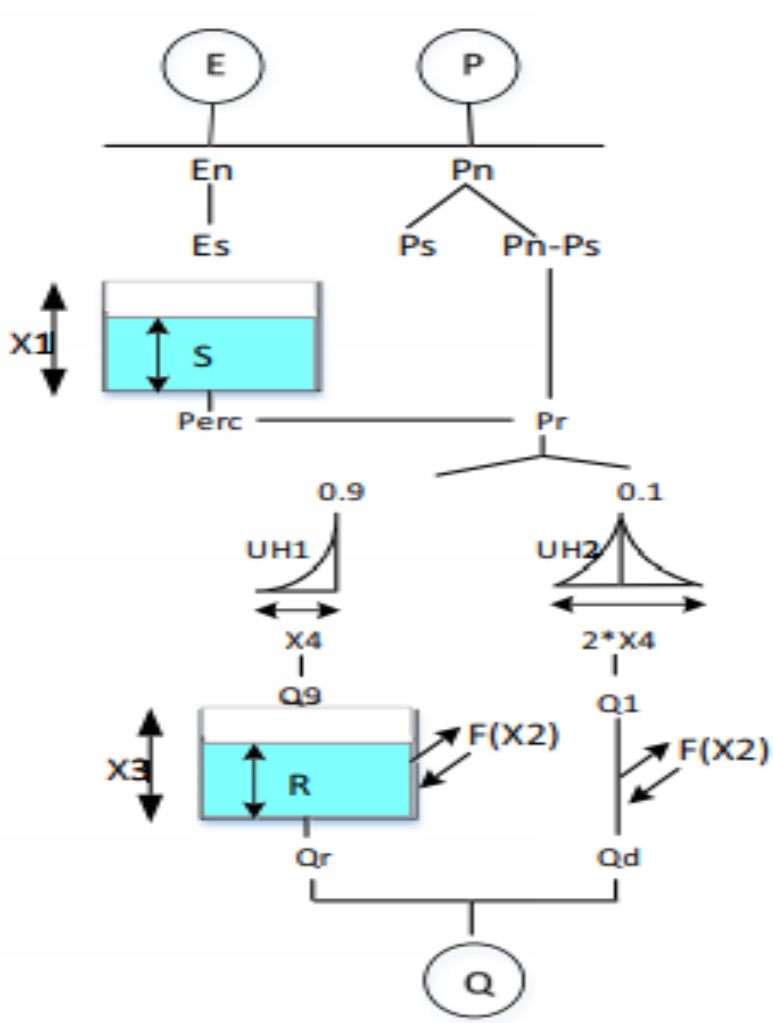

图 6 GR4J 模型结构图。

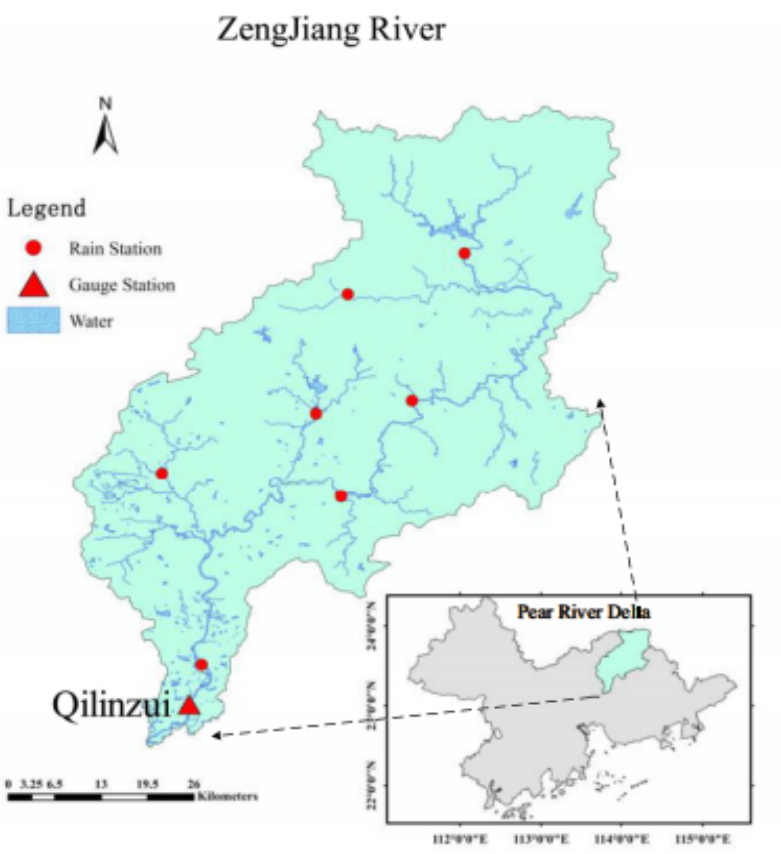
ZengJiang River
N
Legend
Rain Station
Gauge Station
Water
Qilinzui
Pear River Delta
0 3.25 6.5 13 19.5 26
Kilometers
112°0'0"E
113°0'0"E
114°0'0"E
115°0'0"E

图 7 增江流域示意图。

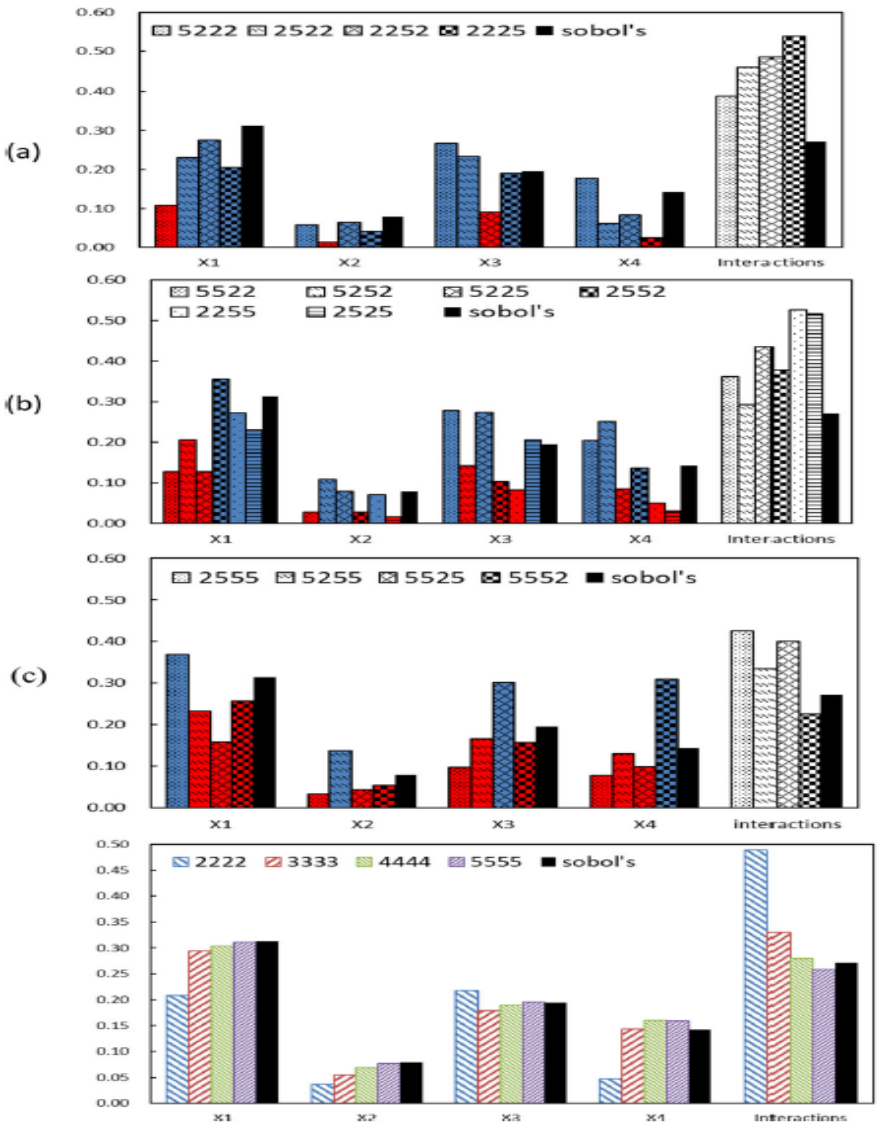

图 8 GR4J 模型参数及其交互在不同二次抽样方法下的敏感性。

模型参数及其交互效应的敏感性如图 8 所示。结果表明，增江流域的径流量对 x_1 的变化最为敏感。与案例研究 1 结果类似，二次抽样会在一定程度上降低参数的敏感性，且各参数及其交互作用的敏感性排序也受到抽样方法的影响。此外，在全采样方差分析方法中，所有参数都在其变化范围内以不同级别进行子采样。在这项研究中，将测试四种情景，每个参数分别具有 2、3、4 或 5 个级别。如图 8 所示，4 个参数的个体敏感性和相互作用敏感性随参数水平的变化而变化。对于全二次抽样方法而言，随着参数水平的增加（从 2 水平至 5 水平），相应参数的敏感性也逐渐增加。参数 x_1，x_2 和 x_4 的敏感性分别从

20.1、3.7 和 4.7%增长到 31.0、7.6 和 15.8%。与此同时，参数 x_3 的敏感性从 21.7%减少到 17.8%，而交互效应的敏感性从 48.9%减少到 25.9%。结果表明，参数水平将影响全采样方差分析方法中的个体和交互敏感性。具体来说，最敏感的参数和交互作用的灵敏度一般会降低，而其他参数的灵敏度随着参数级别的增加而增加。

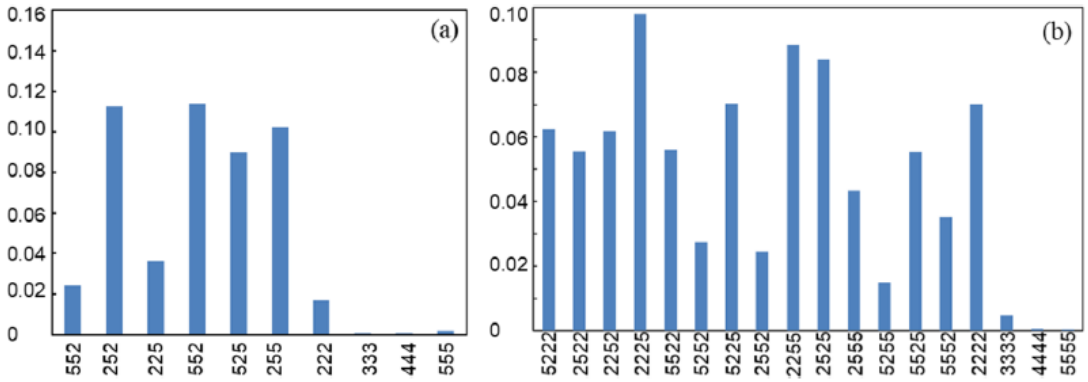

图 9 基于 Sobol’s 法与二次抽样法的水文模型参数敏感性的对比分析。

本研究还对比了上述三种二次抽样 ANOVA 算法与 Sobol’s 算法在计算模型参数敏感性方面的差异（如图 9 所示）。结果表明，与单二次抽样 ANOVA 算法和多二次抽样 ANOVA 算法相比较，全二次抽样 ANOVA 算法可以产出更为可靠的结果。此外，为得到较为可靠的敏感性分析结果，建议每个参数至少设置 3 至 4 个水平。许多研究表明 Sobol’s 算法计算负担较重。相比而言，二次抽样 ANOVA 算法是一种更高效的算法。例如，对于案例 1 来说，Sobol’s 算法需要运行 10000 次以得到稳定的结果，而二次抽样 ANOVA 算法只需要 256 次。尽管相较于 Sobol’s 算法，二次抽样 ANOVA 算法并不能提供更好的结果，但是它极大地降低了运算要求，且可用于非数值化分析。