J. Lake Sci.(湖泊科学), 2019, **31**(5): 1248-1258
DOI 10.18307/2019.0512

太湖富营养化条件下影响蓝藻水华的主导气象因子*

罗晓春[1],杭　鑫[1],曹　云[2],杭蓉蓉[3],李亚春[1**]
(1:江苏省气象服务中心, 南京 210008)
(2:南京市水利规划设计院股份有限公司,南京 210001)
(3:南京航天宏图信息技术有限公司,南京 210006)

摘　要: 利用2004－2018年卫星遥感解译的太湖蓝藻水华信息构建蓝藻综合指数, 采用随机森林机器学习算法分析同期气象因子与蓝藻水华综合指数的关系, 定量评估影响蓝藻水华的主要气象因子特征变量的重要性度量和贡献率. 结果表明, 在光、温、水、风等主要气象要素中, 气温对蓝藻水华综合指数起着主导的作用, 其次是风速和降水, 日照时间的影响或可忽略. 其中气温条件中重要性度量最大的是年平均气温, 其次是冬、春季节的平均气温; 风速因子中影响较大的是7月份的平均风速; 水分条件中主导因子是9月累计降水量. 优选的随机森林模型模拟值与实际蓝藻水华综合指数的变化趋势基本一致, 拟合优度为0.91, 通过0.01显著性检验, 随机森林模型模拟效果较好. 用随机森林模型模拟值对太湖蓝藻水华分等级评估, 模型模拟精度达到了86.7%, 其中5个重度等级年份模型模拟结果完全一致, 中度等级的6个年份模型模拟值有5年与之相符, 中度以上等级的模拟精度达90.9%, 模型能够反映气象因子对蓝藻水华综合指数的综合影响, 对中、重度蓝藻水华的模拟效果更好. 随机森林模型有助于理解富营养化状态下影响蓝藻水华的主导气象因子, 利用气象因子的可预测性可以促进蓝藻水华预测预警能力的提升.
关键词: 蓝藻水华;主导气象因子;随机森林;太湖

Dominant meteorological factors affecting cyanobacterial blooms under eutrophication in Lake Taihu

LUO Xiaochun[1], HANG Xin[1], CAO Yun[2], HANG Rongrong[3] & LI Yachun[1**]
(1: *Jiangsu Meteorological Service Center*, *Nanjing* 210008, *P.R.China*)
(2: *Nanjing Water Planning and Designing Institute Co.*, *Ltd*, *Nanjing* 210001, *P.R.China*)
(3: *Nanjing Piesat Information Technology Co.*, *Ltd.*, *Nanjing* 210006, *P.R.China*)

Abstract: Based on the cyanobacteria comprehensive index (I_c) constructed by the satellite imageries of Lake Taihu in 2004－2018, the random forest machine learning algorithm was used to analyze the relationship between meteorological factors and I_c, and quantitatively evaluate the importance measures and contribution rate of the main meteorological features. The results show that among the main meteorological elements such as light, temperature, water and wind, temperature plays a leading role in cyanobacterial comprehensive index, followed by wind speed and precipitation, and the influence of sunshine hours may be neglected. Among them, the most important measure of importance in temperature conditions is the annual average temperature, followed by the average temperature in winter and spring. The most important in the wind factors is the average wind speed in July. The dominant factor in the water condition is the cumulative precipitation in September. The optimal random forest model simulation value is basically consistent with the actual cyanobacteria comprehensive index, and the determination coefficient is 0.91. The random forest model simulation effect is better by the 0.01 significance test. Using the random forest model simulation value to evaluate the cyanobacteria blooms in Lake Taihu, the model simulation accuracy reached 86.7%. The simulation results of the five severe grades of the year model are completely consistent. The simulation values of the six grade models of the medium grade are consistent with the five

* 江苏省科技支撑计划项目(BE2011840)、江苏省气象局重点项目(KZ201403)和江苏省气象局青年基金项目(KQ201819)联合资助. 2019－01－23收稿;2019－03－05收修改稿. 罗晓春(1970 ~),男,高级工程师;E-mail: 295646190@qq.com.
** 通信作者;E-mail: jsqxlyc@163.com.

years, and the simulation accuracy of the medium and above grades is 90.9%. The model can reflect the comprehensive effects of meteorological factors on the cyanobacteria comprehensive index, and the simulation effect on medium and severe cyanobacteria blooms is better. The random forest model is helpful to understand the dominant meteorological factors affecting cyanobacterial blooms under eutrophication conditions. The predictability of meteorological factors can promote the improvement of cyanobacterial bloom prediction and early warning ability.

Keywords: Cyanobacteria bloom; dominant meteorological factors; random forest; Lake Taihu

环境污染及由此引起的富营养化是中国淡水湖泊当前面临的主要问题和挑战[1-2].富营养化的直接后果之一是蓝藻的大量繁殖和水华频繁出现. 太湖的水质富营养化始于 1980 年,2000 年后呈加重趋势,蓝藻水华随之趋重[3-5],至 2007 年引发饮用水危机[6]. 治污力度的持续加大[7],使得营养盐浓度总体上也持续下降,但富营养化状况并未根本改变,蓝藻水华面积下降趋势不明显[8-9]. 虽然有关蓝藻水华发生机理的研究较多,但有些影响因素与蓝藻水华之间的相互作用机制尚未十分明确[10-11],从而影响了蓝藻水华的预测、尤其是中长期预测的准确率. 在营养盐浓度充足情况下,其他环境因子的影响已越来越受到关注,其中气象条件可能是主要的限制因子[5,12-13],其影响甚至可能超过营养盐[9,14]. 气象因子对蓝藻水华的影响十分重要且较为复杂. 其中,气温在蓝藻的休眠、复苏和增长等不同阶段所起的作用都较大[15-17],但蓝藻复苏后气温就不再是主要影响因子,且高温有一定的抑制作用[18-19]; 微风条件有利于蓝藻水华形成[20],风向则主要影响蓝藻水华的移动方向和空间分布格局[21]; 光照对于藻类水华也必不可少,但日照时间仅是湖泊表层蓝藻水华形成的一个非必要条件[22-23]; 持续性的或大量降水会对藻密度产生稀释作用,不利于形成蓝藻水华[24-26],但另一方面降水可能将较多的营养盐带到湖水中[27]. 对于太湖这样一个大型、复杂的生态系统,蓝藻水华也是外界环境因子的综合而复杂的影响结果,观测数据是多维的且可能有缺失,变量之间可能存在非线性的影响关系且影响过程复杂,传统的统计方法在定量反映不同气象因子及在不同阶段对蓝藻水华发生发展的影响程度存在困难,不能很好地描述蓝藻生长水华形成的复杂过程和影响格局,因此,寻找新的度量特征因子重要性的方法显得非常必要.

随着人工智能技术的快速发展,随机森林等机器学习算法在特征重要性评估和预测等应用中开始显示出优势. 随机森林(random forest,RF)是 Breiman 于 2001 年提出的一种组合分类智能算法,是一种现代分类与回归技术,它在不增加原样本集样本的情况下利用 Bootstrap 重抽样方法和节点随机分裂技术进行分类树构建,然后通过投票方式得到最终分类结果[28]. 随机森林具有很多优点,如极高的准确率、极强的数据挖掘能力、分析复杂相互作用分类特征的能力以及可以给出变量重要性估计等,甚至被誉为当前最好的机器学习算法之一[29]. RF 模型已应用于灾害风险评估[30-31]、医学影像检测[32-33]、作物分类[34-35]和生物物种分布影响因素评估[36-37]等领域,在分析变量指标的重要度和确定主导因素等方面均取得了较好的效果[38-40]. 作为高维数据的有效的特征选择工具,随机森林模型具有识别、量化特征变量重要性的功能,理论上为评估太湖蓝藻水华气象影响因子的重要性、帮助理解富营养化状态下影响蓝藻水华的主导气象因子提供了一种新的思路和方法,但相关的研究却鲜见报道.

为此,本文利用卫星遥感监测的太湖蓝藻水华面积和次数等信息构建蓝藻水华综合指数,选取同期光、温、水和风等主要气象要素的观测数据为评价因子,尝试用 RF 机器学习算法为基本工具,基于 RF 算法的变量重要性度量进行特征重要性排序,分析和评价影响蓝藻水华综合指数的主要气象因子的重要性和贡献率,确定太湖蓝藻水华的主导气象因子,并用模型模拟蓝藻综合指数进行验证,以期为提升蓝藻水华的预测、预警能力提供科技支撑.

1 资料与方法

1.1 气象资料

根据太湖蓝藻水华发生发展特点并参考相关文献研究结果,气象资料选用太湖区域的苏州、无锡、宜兴、吴江和东山 5 个基本站 2004－2018 年的气温、降水、风速、日照共 4 类气象因子,分别计算 1－12 月各月平均气温、年平均气温、1－12 月各月平均风速、年平均风速、1－12 月各月平均日照时数、年平均日照时数、1－12 月各月累计降水量、年累计降水量和年高温日数,并统计影响蓝藻生长和水华形成主要阶段的上述各

气象因子的不同组合,共选取 82 组变量.

1.2 卫星数据及蓝藻水华综合指数

卫星资料选用 2004－2018 年 EOS/MODIS 卫星的 Terra/Aqua 传感器和 FY-3 卫星的 MERSI 传感器观测的影像数据,空间分辨率均为 250 m. 蓝藻水华的反演采用常用的归一化植被指数法(*NDVI*)[17],共得到 1370 幅蓝藻水华面积≥1 km^2 的卫星遥感影像,并提取面积和次数等定量信息.

为更客观地反映蓝藻水华的影响程度,综合考虑蓝藻水华的面积和次数,构建太湖蓝藻水华综合指数,具体方法如下:

$$I_c = w_A \cdot NA_{total} + w_F \cdot NF_{total} \quad (1)$$

式中,I_c 即为蓝藻水华综合指数,NA_{total} 和 NF_{total} 分别为各年蓝藻水华累计面积和累计次数经归一化处理后的数值,w_A 和 w_F 分别为权重系数,权重系数采用客观的信息量权数法[41]确定. 共计算得到 2004－2018 年 15 个蓝藻综合指数样本:0.057、0.247、0.486、0.800、0.258、0.188、0.370、0.243、0.276、0.318、0.252、0.338、0.374、0.724、0.503. 在此基础上利用随机森林机器学习算法分析光、温、水和风等气象要素与蓝藻水华综合指数的关系,定量评价影响蓝藻水华的气象因子特征变量的重要性度量和贡献率.

1.3 随机森林基本原理

随机森林是一个集成学习模型,以决策树为基分类器,由多个 Bagging 集成学习技术训练得到的决策树构成,通过单个决策树的输出结果投票决定最终的分类结果. RF 的基本思路和生成步骤:(1)对于一个原始训练样本集 D,利用 Bootstrap 采样法从中选取 N_{tree} 个与 D 中样本数量相同的子训练样本集分别为 D1、D2、…、Dn,分别建立 N_{tree} 个分类树模型,将未被抽取到的袋外数据作为测试样本;(2)在每一分类树的每个节点上随机抽取 M_{try} 个特征变量(M_{try} 须小于原始数据变量个数 P),依据优选法则在 M_{try} 个特征变量中选择高分类能力的特征进行节点分裂;(3)每棵树都不做任何裁剪,最大限度地生长;(4)形成随机森林,再用随机森林对新数据进行分类,分类结果按树分类器的投票多少而定. 参数 N_{tree} 为森林中树的数目,参数 M_{try} 决定在随机森林中决策树的每次分支时所选择的变量个数,N_{tree} 和 M_{try} 是两个非常重要的自定义参数,也是决定随机森林预测能力的两个重要参数,必须进行优化. 通常 N_{tree} 最好设定为 500 或者 1000, M_{try} 值要在模型构建过程中通过逐次计算来挑选最优值,在回归模型中一般为变量个数的三分之一.

1.4 变量重要性评估

特征选择的目的是从成百上千个特征变量中选取对最终结果影响较大的数目较少的特征变量,通过特征变量的筛选,可以删除一些和任务无关或者冗余的特征变量,简化的特征数据集也常会得到更精确的模型,增强对特征和特征值的理解. 利用随机森林算法本身所具有的变量重要性度量可以对特征重要性进行排序,从而选出重要性靠前的特征. 在 RF 模型中变量重要性度量的主要评价指标为精度平均减少值(IncMSE)和节点不纯度减少值(IncNodePurity). IncMSE 是指变量随机取值后 RF 模型估算的误差相对于原来误差的升高幅度,IncMSE 值越大,说明该变量越重要. IncNodePurity 是指变量对各个决策树节点的影响程度,值越大,说明该变量越重要,反之则相对不重要. 本文采用 IncNodePurity 作为变量重要性的评价指标.

2 结果与分析

2.1 特征变量的筛选

即使同类因子不同时间段的变量对蓝藻水华的作用也不相同,为剔除这些重要性不够的因子变量,分别对气温、降水、风速和日照等 4 类气象因子的 82 组特征变量进行筛选,根据变量的重要性排序选取影响蓝藻水华的相对重要的特征变量.

1)首先利用不同类型的气象因子特征变量与蓝藻水华综合指数 I_c 分别构建 RF 模型. 每类气象因子重复建模 50 次,计算每个特征变量的节点不纯度减少值的平均值,按照重要性及频次进行分类排序统计,分别得到气温、风速、降水和日照 4 类因子不同时间组合特征变量的重要度评估曲线(图 1),将曲线出现拐点前的因子特征变量认为是相对重要的变量,初步筛选出 4 类因子的 18 组特征变量如下:

气温因子共有 7 组特征变量:T_{01-02}、T_y、T_{01-04}、T_{12-02}、T_{03-05}、T_{12-03}、T_{12-04},分别表示 1－2 月的平均气温、年平均气温、1－4 月的平均气温、12 月至次年 2 月的平均气温、3－5 月的平均气温、12 月至次年 3 月的平均气

温和 12 月至次年 4 月的平均气温,单位:℃.

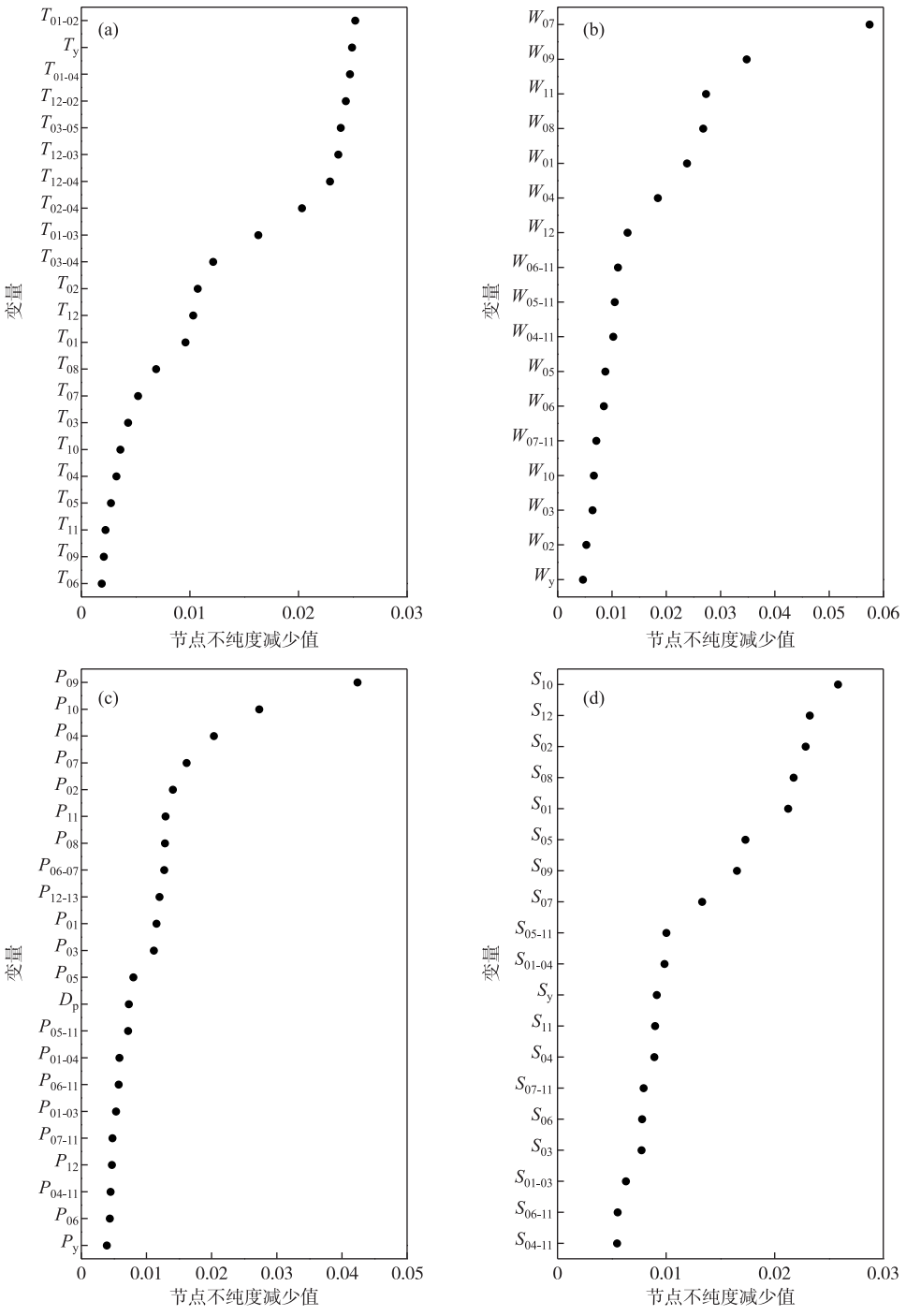

图 1 变量重要性排序(a:气温变量;b:风速变量;c:降水变量;d:日照时间)

Fig.1 Variable importance ordering

(a: Temperature variable; b: Wind speed variable; c: Precipitation variable; d: Sunshine time)

风速因子共有 2 组特征变量：W_{07}、W_{09}，分别表示 7 和 9 月的平均风速，单位：m/s.

降水因子共有 3 组特征变量：P_{09}、P_{10}和 P_{04}，分别表示 9、10 和 4 月的降水量，单位：mm.

日照因子共有 5 组特征变量：S_{10}、S_{12}、S_{02}、S_{08}、S_{01}，分别表示 10、12、2、8 和 1 月的日照时数，单位：h.

高温天数 1 组特征变量：D_{Tmax}，表示一年中日最高气温≥35℃的总天数，单位：d.

2）对筛选后的 4 类气象因子共 18 组特征变量，再次与蓝藻水华综合指数 I_c 样本进行 RF 建模. RF 回归模型中树节点预选的变量个数 M_{try} 一般为变量数的三分之一，因此选取 5、6、7、8 四个数值，对应不同的随机森林中树的个数 N_{tree} 分别进行建模和模型优化. 每个参数重复建模 20 次，筛选出验证精度较高的以下 7 个模型（表 1）.

表 1 模型精度

Tab.1 Model accuracy

模型序号	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6	模型 7
M_{try}	5	6	6	7	7	8	8
N_{tree}	800	650	650	750	500	400	400
模型精度（相关系数）	0.976	0.925	0.977	0.980	0.961	0.979	0.904

针对这筛选出的 7 种模型，根据节点不纯度减少值选择排序占比较重的因子特征变量，重新进行 RF 模拟筛选特征变量，结果如下：

模型（1）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、P_{09}、W_{07}；

模型（2）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、P_{09}、P_{10}、W_{07}、S_{12}；

模型（3）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、D_{Tmax}、P_{09}、P_{10}、W_{07}、W_{09}；

模型（4）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、D_{Tmax}、P_{09}、P_{10}、W_{07}；

模型（5）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、P_{09}、W_{07}、W_{09}、S_{12}；

模型（6）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、D_{Tmax}、P_{09}、W_{07}；

模型（7）选择的特征变量有：T_{12-02}、T_{12-03}、T_{12-04}、T_{01-04}、T_{y}、T_{01-02}、T_{03-05}、P_{09}、P_{10}、W_{07}、S_{01}.

3）根据重要性度量和频次排序筛选特征变量. 针对上述 7 种模型选择的特征变量，分别采用不同的树节点预选变量个数 M_{try} 和随机森林中树的个数 N_{tree} 重新建模，从中各选择约 20 个精度相对较高（相关系数 R>0.85）的模型，共生成 243 组模型. 将 243 组模型中的所有变量的节点不纯度减少值进行排序，分别统计不同名次中各特征变量出现的次数，重要性度量排第一的特征变量为在重要性排序中第一位次出现的频次最高的，重要性度量排第 2 的特征变量为在重要性排序中第 1、2 位累计出现频次最高的（除第 1 位），按照上述规则依次统计第 3、4 位特征变量等，计算公式为：

$$p_{ij} = \sum_{k=1}^{j} a_{ik} \tag{2}$$

式中，p_{ij} 为第 i 个特征变量 a_i 在重要性度量排序第 j 位的累计频次. 将各模型中的特征变量重要性度量排序由高到低分成 12 个位次，第 1 位次重要性程度最高，第 12 位次重要性程度最低，根据式（1）统计各因子特征变量第 1~12 位次的累计频次 p_{ij}（表 2），将所有因子特征变量中在各位次中累计频次最高的特征变量作为本位次最重要的特征变量（排除此前已选变量），如在所有的特征变量中，重要性排序第 1 位次中累计频次最高的变量为 T_{y}，累计频次为 148，T_{y} 即被选为重要性排序第 1 位次中排在最前面的特征变量，也即最重要的特征变量；在重要性排序第 2 位次中，累计频次排在前两位的分别是 T_{y} 的 194 和 T_{03-05} 的 80，由于 T_{y} 已经被选为最重要（排第 1 位）的变量，因此 T_{03-05} 被选作重要性度量排在第 2 位的特征变量，以此类推，得到各因子特征变量的重要性度量排序分别为：T_{y}、T_{03-05}、T_{12-04}、T_{12-02}、W_{07}、T_{01-02}、T_{01-04}、T_{12-03}、P_{09}、P_{10}、D_{Tmax}、S_{12}. 然后在 243 组模型中进行匹配，将模型中特征变量重要排序与上述变量重要性排序基本一致的模型选为最优模型，考察这一模型中特征变量的重要性度量.

2.2 特征变量重要性评价

将上一步优选出来的模型认为是最符合蓝藻水华综合指数与气象因子关系的随机森林模型，该模型的

表 2 特征变量的重要性排序及累计频次

Tab.2 Order importance and cumulative frequency of variables

变量	1	2	3	4	5	6	7	8	9	10	11	12
T_{01-02}	14	38	59	101	130	168	200	227	239	242	243	243
T_{03-05}	28	80	118	155	191	211	226	235	242	243	243	243
T_{y}	148	194	210	219	227	235	240	243	243	243	243	243
T_{12-04}	13	35	104	130	158	178	204	231	243	243	243	243
T_{12-03}	1	6	17	36	61	97	135	184	231	240	241	243
T_{12-02}	11	32	62	112	148	185	210	233	240	243	243	243
T_{01-04}	4	11	24	37	57	98	141	184	229	240	243	243
W_{07}	11	38	73	101	136	162	195	222	240	242	243	243
W_{09}	0	0	0	0	0	2	2	2	6	35	50	60
S_{01}	0	0	0	1	3	3	4	4	4	7	41	41
S_{12}	0	0	0	0	0	0	0	0	0	21	81	81
P_{09}	13	51	61	77	96	105	127	155	233	241	242	243
P_{10}	0	1	1	3	6	8	8	13	23	108	140	142
D_{Tmax}	0	0	0	0	2	6	9	11	14	62	96	101

因子特征变量重要程度用相对重要性度量图表示,同时模型还计算了各因子特征变量对蓝藻水华综合指数的贡献率大小(图 2),从随机森林模型给出的变量重要性估计表明,共有 4 组气温变量的重要度排在前 4 位,分别是 T_{y}、T_{03-05}、T_{12-04}、T_{12-02},表明影响蓝藻水华综合指数的主导气象因子是气温,其次是风速,W_{07} 重要性度量排名第 5,降水因子特征变量 P_{09} 的重要性度量排名第 7,年高温日数 D_{Tmax} 和 P_{10} 的重要性度量排在最后(图 3). 在气温因子中,重要性最大的特征变量是年平均气温 T_{y},其次分别是 3－5 月平均气温 T_{03-05}、12 月至次年 4 月平均气温 T_{12-04}、12 月至次年 2 月平均气温 T_{12-02}等,表明冬、春季节的平均气温高低对蓝藻水华的影响较大. 从图 2 还可以看出,前 9 项特征变量重要度对蓝藻水华综合指数的贡献率均超过 5%,累计

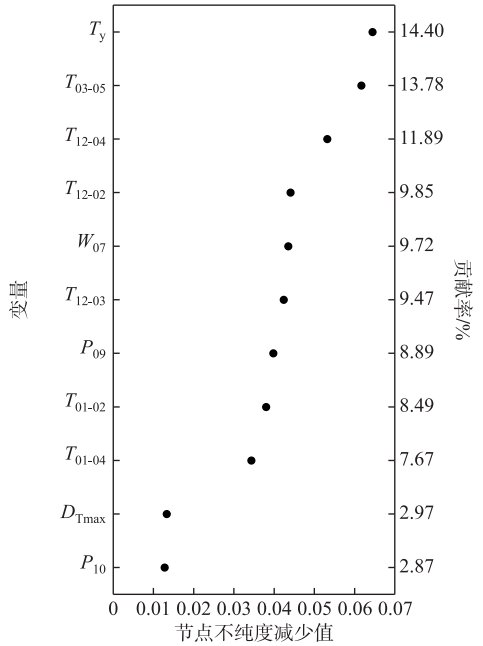

图 2 变量重要性度量和贡献率

Fig.2 Variable importance measure contribution rate

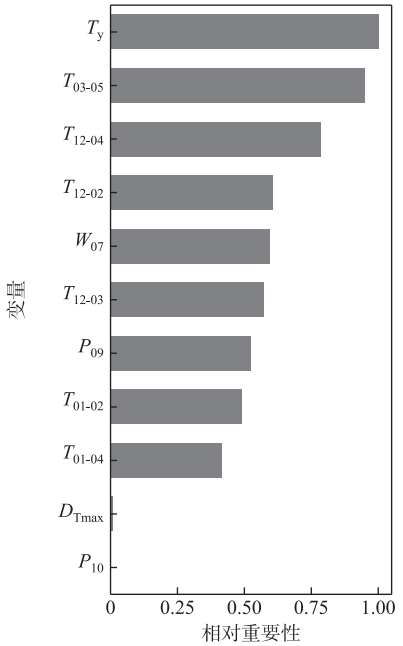

图 3 变量相对重要性

Fig.3 Relative importance of variables

达到总数的 94.2%，表明这些变量已对蓝藻水华的形成起到决定性的作用，而年高温日数和 10 月份降水量的作用几乎可以忽略不计.

2.3 RF 模型模拟验证

将优选 RF 模型模拟值跟蓝藻水华综合指数进行 Pearson 相关分析，拟合优度为 0.91，通过 0.01 显著性检验，表明模拟值跟蓝藻水华综合指数高度相关. 从模型模拟值与实际蓝藻水华综合指数的拟合曲线可以看出，模型模拟值和实际蓝藻水华综合指数的变化趋势基本吻合，仅 2015 年有偏差，RF 模型模拟效果较好（图 4）.

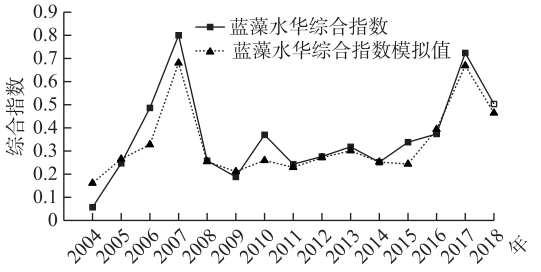

图 4 2004－2018 年蓝藻水华综合指数与模型模拟值

Fig.4 Cyanobacterial index and model simulation values from 2004 to 2018

根据百分位数法[42]确定蓝藻综合指数和模型模拟的蓝藻水华综合指数等级阈值，将 2004－2018 年的蓝藻水华综合指数和模拟值分别以百分位法计算 30%、70% 对应的百分位数，分别分成 3 个等级，由低到高依次为轻度、中度和重度，结果见表 3. 蓝藻水华综合指数达重度等级的分别为 2006、2007、2016、2017 和 2018 年共 5 个年份，模型模拟值均为重度等级，结果完全一致；蓝藻水华综合指数为中度的分别为 2008、2010 和 2012－2015 年共 6 个年份，模型模拟值有 5 年相符，仅 2015 年偏小一个等级，原因可能是除气象因子外，其他因子如水文、水质参数也都会对蓝藻水华产生作用. 模型模拟总精度达 86.7%，其中中度以上等级的模拟精度达 90.9%，表明此模型能够反映气象因子对蓝藻水华的综合影响，对中、重度蓝藻水华的模拟效果更好.

表 3 蓝藻水华综合指数等级划分及模型验证

Tab.3 Grade division of I_c and test of model

年份	蓝藻水华综合指数	蓝藻等级	模拟指数	模拟等级	模型检验
2004	0.057	轻度	0.161	轻度	一致
2005	0.247	轻度	0.266	中度	偏大一个等级
2006	0.486	重度	0.328	重度	一致
2007	0.800	重度	0.681	重度	一致
2008	0.258	中度	0.254	中度	一致
2009	0.188	轻度	0.211	轻度	一致
2010	0.370	中度	0.260	中度	一致
2011	0.243	轻度	0.230	轻度	一致
2012	0.276	中度	0.272	中度	一致
2013	0.318	中度	0.301	中度	一致
2014	0.252	中度	0.252	中度	一致
2015	0.338	中度	0.244	轻度	偏小一个等级
2016	0.374	重度	0.394	重度	一致
2017	0.724	重度	0.669	重度	一致
2018	0.503	重度	0.464	重度	一致

3 讨论

客观定量评价富营养化条件下影响蓝藻水华的气象因子的重要性,明确影响蓝藻水华的主导气象因子,利用气象因子的可预测性,可以提升蓝藻水华预测预警的能力. 随机森林模型的变量重要性评价结果表明,在光、温、水、风等主要气象要素中,气温对蓝藻水华起着主导的作用,其次是风速和降水,日照时间的影响最小. 虽然已有的研究证实了气温[17-18]、降水[25]、风速风向[21]和日照时间[23]在蓝藻生长和水华形成过程中的作用,但能够定量、客观地明确这些主要气象因素的重要性度量及贡献率的研究报道尚不多见. 在气温因子中,在蓝藻生长和水华形成的不同阶段,气温的影响并不相同,在所有 243 组 RF 模型中年平均气温 T_y 重要性度量排在第 1 位的达 148 次,占比为 60.9%,排在前两位累计 194 次,占比近 80%,远超其他变量,因此,可以认为影响年际间蓝藻水华程度差异的最主要的气象因子是年平均气温 T_y. Rigosi 等[11]研究证实了气候变暖是国内外蓝藻水华增多增强的重要原因,而年平均气温正是反映年际间气候冷暖的重要指标. 比较 2004 年以来卫星监测到的蓝藻水华面积和年平均气温可以发现,太湖蓝藻水华大面积暴发的 2007 年年平均气温达 17.7℃,为 1961 年以来的最高值,同样,2017 年年平均气温也高达 17.5℃,当年蓝藻水华之严重程度也仅次于 2007 年,实际情况证实了年平均气温 T_y 的重要性. 紧随 T_y 的分别是 3－5 月平均气温 $T_{03\sim05}$、12 月至次年 4 月平均气温 $T_{12\sim04}$ 和 12 月至次年 2 月平均气温 $T_{12\sim02}$,表明与其他时间段相比,冬春季节的气温对蓝藻水华的影响显得更为重要. 秦伯强等[43]分析认为在蓝藻细胞生长阶段,气温等环境因素的影响较为显著,吴晓东等[44]发现气温是影响太湖蓝藻复苏进程的关键因素之一,谢小萍等[17]认为气温对蓝藻的复苏和休眠都有影响. 但复苏后气温条件已经基本满足蓝藻生长需要,不再是主要限制因子[18],因此模型中没有出现 5 月以后的平均气温. 风则主要通过风浪的作用影响蓝藻的聚集和移动,决定蓝藻水华的规模、范围及位置[5,43]. 对于单次蓝藻水华的形成来说,风的作用较大,一般认为适宜范围为日平均风速 3~4 m/s[19-20],而前 6 h 平均风速更小,为 0~2.0 m/s[21],但较长时间尺度如月、季等,由于求算平均值的过程平滑了波动,平均风速都处于适宜范围. 仅有 7 月的平均风速 W_{07} 的重要性度量排名靠前,可能是因为 7 月是梅雨结束的月份,梅雨期后的微风晴热天气适宜蓝藻水华形成,平均风速大小对蓝藻水华的形成相对较为敏感. 降水对蓝藻水华的影响机制目前尚不十分清楚,刘心愿等[25]分析认为降雨对蓝藻水华的抑制作用是阶段性的,但同时会将营养盐带入湖中产生中长期的影响. RF 模型给出的 9 月降水量 P_{09} 的重要性排名靠前,可能是因为 9 月是太湖蓝藻水华容易暴发的月份[45],巢湖也有类似现象[46],因此 9 月降水量的多少会影响到当年蓝藻水华程度. 优选的 RF 模型中没有出现日照时数变量,这与孔繁翔等[12]和张海春等[47]发现在低光照及黑暗环境下蓝藻仍能生存、张晓忆等[48]得到的蓝藻水华与日照时数无显著相关关系等结论是一致的,说明在太湖这样的大型浅水湖泊,光照调节机制难以解释蓝藻水华的形成与消散.

利用随机森林机器学习算法的重要性度量可以作为估计影响蓝藻水华的因子重要性和选择重要特征变量的工具,尽管可能受到太湖蓝藻水华样本数的限制,但本文的研究依然显示了随机森林算法的有效性,随机森林算法在特征因子重要性评估和重要特征选择方面应会具有广阔的前景.

4 结论

1)随机森林算法重要性度量表明,影响蓝藻水华综合指数的主导气象因素是气温,其次是风速和降水,日照时间的影响或可忽略. 其中重要性度量最大的是年平均气温 T_y,其次是冬、春季平均气温; 7 月平均风速 W_{07} 和 9 月降水量 P_{09} 排名也靠前. 前 9 项变量对蓝藻水华综合指数的贡献累计达 94.2%,表明这些变量对蓝藻水华综合指数起决定性作用.

2)将随机森林模型模拟值跟蓝藻水华综合指数进行相关分析,拟合优度为 0.91,通过 0.01 显著性检验,表明模型模拟值与实际蓝藻水华综合指数的变化趋势基本一致,仅有一年有所偏差,RF 模型模拟效果较好.

3)用随机森林模型模拟值与太湖蓝藻水华综合指数进行分等级评价,模拟精度达到 86.7%,其中 5 个重度等级年份模拟结果完全一致,6 个中度等级年份中仅有 1 年偏小一个等级,中度以上等级的模拟精度达 90.9%,表明 RF 模型能够反映气象因子对蓝藻水华的综合影响,对中、重度蓝藻水华的模拟效果更好.

5 参考文献

[1] Liu J, Yang W. Water sustainability for China and beyond. *Science*, 2012, **337**(6095): 649-650.

[2] Cheng XY, Li SJ. An analysis on the evolvement processes of lake eutrophication and their characteristics of the typical lakes in the middle and lower reaches of Yangtze River. *Chinese Science Bulletin*, 2006, **51**(13):1603-1613.

[3] Qin B, Xu P, Wu Q *et al*. Environmental issues of lake Taihu, *China*. *Hydrobiologia*, 2007, **58**(1): 3-14.

[4] Zhu GW. Eutrophic status and causing factors for a large, shallow and subtropical Lake Taihu, China. *J Lake Sci*, 2008, **20**(1): 21-26. DOI: 10.18307/2008.0103. [朱广伟. 太湖富营养化现状及原因分析. 湖泊科学, 2008, **20**(1): 21-26.]

[5] Zhang M, Yang Z, Shi XL. Expansion and drivers of cyanobacterial blooms in Lake Taihu. *J Lake Sci*, 2019, **31**(2): 336-344. DOI: 10.18307/2019.0203. [张民, 阳振, 史小丽. 太湖蓝藻水华的扩张与驱动因素. 湖泊科学, 2019, **31**(2): 336-344.]

[6] Qin B, Zhu G, Gao G *et al*. A drinking water crisis in Lake Taihu, China: Linkage to climatic variability and lake management. *Environmental Management*, 2010, **45**(1): 105-112. DOI: 10.1007/s00267-009-9393-6.

[7] Meng W. To achieve safe status of lake water environment and ecology still a long way to go. *Science & Technology Review*, 2017, **35**(9): 1. [孟伟. 湖泊"水环境与生态安全"依然任重而道远. 科技导报, 2017, **35**(9): 1.]

[8] Dai XL, Qian PQ, Ye L *et al*. Changes in nitrogen and phosphorus concentrations in Lake Taihu, 1985-2015. *J Lake Sci*, 2016, **28**(5): 935-943. DOI: 10.18307/2016.0502. [戴秀丽, 钱佩琪, 叶凉等. 太湖水体氮、磷浓度演变趋势(1985—2015 年). 湖泊科学, 2016, **28**(5): 935-943.]

[9] Zhu GW, Qin BQ, Zhang YL *et al*. Variation and driving factors of nutrients and chlorophyll-a concentrations in northern region of Lake Taihu, China, 2005-2017. *J Lake Sci*, 2018, **30**(2): 279-295. DOI: 10.18307/2018.0201. [朱广伟, 秦伯强, 张运林等. 2005－2017 年北部太湖水体叶绿素 a 和营养盐变化及影响因素. 湖泊科学, 2018, **30**(2): 279-295.]

[10] Hans WP, Timothy GO. Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microbial Ecology*, 2013, **65**(4): 995-1010.

[11] Rigosi A, Carey CC, Ibelings BW *et al*. The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnology and Oceanography*, 2014, **59**(1): 99-114.

[12] Kong FX, Gao G. Hypothesis on cyanobacteria bloom-forming mechanism in large shallow eutrophic lakes. *Acta Ecologica Sinica*, 2005, **25**(3): 589-595. [孔繁翔, 高光. 大型浅水富营养化湖泊中蓝藻水华形成机理的思考. 生态学报, 2005, **25**(3): 589-595.]

[13] Zhao QH, Sun GD, Wang JJ *et al*. Coupling effect of water temperature and light energy on the algal growth in Lake Taihu. *J Lake Sci*, 2018, **30**(2): 385-393. DOI: 10.18307/2018.0210. [赵巧华, 孙国栋, 王健健等. 水温、光能对春季太湖藻类生长的耦合影响. 湖泊科学, 2018, **30**(2): 385-393.]

[14] Jessica R, Claire M, Stephen CM *et al*. Effects of multiple stressors on cyanobacteria abundance vary with lake type. *Glob Change Biol*, 2018, **24**: 5044-5055. DOI: 10.1111/gcb.14396.

[15] Thomas MK, Litchman E. Effects of temperature and nitrogen availability on the growth of invasive and native cyanobacteria. *Hydrobiologia*, 2016, **763**(1): 357-369. DOI: 10.1007/s10750-015-2390-2.

[16] Zhang M, Yu Y, Yang Z *et al*. Photochemical responses of phytoplankton to rapid increasing-temperature process. *Phycological Research*, 2012, **60**(3): 199-207.

[17] Xie XP, Li YC, Hang X *et al*. The effect of air temperature on the process of cyanobacteria recruitment and dormancy in Lake Taihu. *J Lake Sci*, 2016, **28**(4): 818-824. DOI: 10.18307/2016.0415. [谢小萍, 李亚春, 杭鑫等. 气温对太湖蓝藻复苏和休眠进程的影响. 湖泊科学, 2016, **28**(4): 818-824.]

[18] Li YC, Xie XP, Zhu XL *et al*. Applying remote sensing techniques in analysis of temperature features causing cyanobacteria bloom in Lake Taihu. *J Lake Sci*, 2016, **28**(6): 1256-1264. DOI: 10.18307/2016.0611. [李亚春, 谢小萍, 朱小莉等. 结合卫星遥感技术的太湖蓝藻水华形成温度特征分析. 湖泊科学, 2016, **28**(6): 1256-1264.]

[19] Baig SA, Huang L, Sheng T *et al*. Impact of climate factors on cyanobacterial dynamics and their interactions with water quality in South Taihu Lake, China. *Chemistry & Ecology*, 2017, **33**(1): 76-87. DOI: 10.1080/02757540.2016.1261122.

[20] Wu T, Qin B, Brookes JD *et al*. The influence of changes in wind patterns on the areal extension of surface cyanobacterial blooms in a large shallow lake in China. *Science of the Total Environment*, 2015, **518/519**(15): 24-30. DOI: 10.1016/j.scitotenv.2015.02.090.

[21] Li YC, Xie XP, Hang X *et al*. Analysis of wind field features causing cyanobacteria bloom in Taihu Lake combined with remote sensing methods. *China Environmental Science*, 2016, **36**(2): 525-533. .[李亚春, 谢小萍, 杭鑫等. 结合卫星遥感技术的太湖蓝藻水华形成风场特征. 中国环境科学, 2016, **36**(2): 525-533.]

[22] Zhou Q, Zhang Y, Lin D *et al*. The relationships of meteorological factors and nutrient levels with phytoplankton biomass in a shallow eutrophic lake dominated by cyanobacteria, Lake Dianchi from 1991 to 2013. *Environmental Science and Pollution Research*, 2016, (23): 15616-15626. DOI: 10.1007/s11356-016-6748-4.

[23] Wu J, Chen XC, Kong HN *et al*. The effect of light intensity on the cell density and chain length of *Anabaena flos-aquae*. *China Environmental Science*, 2012, **32**(5): 875-879.[巫娟, 陈雪初, 孔海南等. 光照度对水华鱼腥藻细胞比重与藻丝长度的影响研究. 中国环境科学, 2012, **32**(5): 875-879.]

[24] Lu WK, Yu LX, Ou XK *et al*. Relationship between occurrence frequency of cyanobacteria bloom and meteorological factors in Lake Dianchi. *J Lake Sci*, 2017, **29**(3): 534-545. DOI: 10.18307/2017.0302.[鲁韦坤, 余凌翔, 欧晓昆等. 滇池蓝藻水华发生频率与气象因子的关系. 湖泊科学, 2017, **29**(3): 534-545.]

[25] Liu XY, Song LX, Ji DB *et al*. Effect of the rainfall on extinction of cyanobacteria bloom and its mechanism analysis. *Environmental Science*, 2018, **39**(2): 774-782.[刘心愿, 宋林旭, 纪道斌等. 降雨对蓝藻水华消退影响及其机制分析. 环境科学, 2018, **39**(2): 774-782.]

[26] Zhan X, Zou LY. Study advances on technology of nutrient control in lake water body. *Environmental Science and Technology*, 2009, **22**(4): 60-64.[詹旭, 邹路易. 湖泊水体中营养盐控制技术研究进展. 环境科技, 2009, **22**(4): 60-64.]

[27] Simić SB, Đorđević N, Milošević D *et al*. The relationship between the dominance of Cyanobacteria species and environmental variables in different seasons and after extreme precipitation. *Fundamental & Applied Limnology*, 2017, **190**(1): 1-11. DOI: 10.1127/fal/2017/0975.

[28] Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5-32.

[29] Iverson LR, Prasad AM, Matthews SN *et al*. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management*, 2008, **254**(3): 390-406. DOI: 10.1016/j.foreco.2007.07.023.

[30] Lai CG, Chen XH, Zhao SW *et al*. A flood risk assessment model based on random forest and its application. *Journal of Hydraulic Engineering*, 2015, **46**(1): 58-66.[赖成光, 陈晓宏, 赵仕威等. 基于随机森林的洪灾风险评价模型及其应用. 水利学报, 2015, **46**(1): 58-66.]

[31] Wu XQ, Lai CG, Chen XH *et al*. A landslide hazard assessment based on random forest weight: a case study in the Dongjiang River Basin. *Journal of Natural Disasters*, 2017, **26**(5): 119-129.[吴孝情, 赖成光, 陈晓宏等. 基于随机森林权重的滑坡危险性评价: 以东江流域为例. 自然灾害学报, 2017, **26**(5): 119-129.]

[32] Zhao RC, Chen ZL, Duan XC *et al*. Automated glaucoma detection based on multi-channel features from color fundus images, *Journal of Computer-Aided Design & Computer Graphics*, 2017, **29**(6): 998-1006.[赵荣昌, 陈再良, 段宣初等. 聚合多通道特征的青光眼自动检测. 计算机辅助设计与图形学学报, 2017, **29**(6): 998-1006.]

[33] Cai JX, Feng GC, Tang X *et al*. Human action recognition based on local image contour and random forest. *Acta Optica Sinica*, 2014, **34**(10): 204-213. DOI: 10.3788/AOS201434.1015006.[蔡加欣, 冯国灿, 汤鑫等. 基于局部轮廓和随机森林的人体行为识别. 光学学报, 2014, **34**(10): 204-213.]

[34] Wang LM, Liu J, Yang LB *et al*. Application of random forest method in maize-soybean accurate identification. *Acta Agronomica Sinica*, 2018, **44**(4): 569-580. DOI: 10.3724/SP.J.1006.2018.00569.[王利民, 刘佳, 杨玲波等. 随机森林方法在玉米-大豆精细识别中的应用. 作物学报, 2018, **44**(4): 569-580.]

[35] Shi LJ, Lu J. Automatic measurement method for maize ear development degree based on random forest. *Transactions of the Chinese Society for Agricultural Machinery*, 2017, **48**(1): 169-174.[石礼娟, 卢军. 基于随机森林的玉米发育程度自动测量方法. 农业机械学报, 2017, **48**(1): 169-174.]

[36] Zhang L, Liu SR, Sun PS *et al*. Partitioning and mapping the sources of variations in the ensemble forecasting of species distribution under climate change: a case study of *Pinus tabulaeformis*. *Acta Ecologica Sinica*, 2011, **31**(19): 5749-5761.

DOI：10.3724/SP.J.1011.2011.00110.［张雷，刘世荣，孙鹏森等. 气候变化对物种分布影响模拟中的不确定性组分分割与制图——以油松为例. 生态学报，2011，**31**(19)：5749-5761.］

［37］ Zhang L，Liu SR，Sun PS *et al*. Comparative evaluation of multiple models of the effects of climate change on the potential distribution of *Pinus massoniana*. *Chinese Journal of Plant Ecology*，2011，**35**(11)：1091-1105. DOI：10.3724/SP.J.1258.2011.01091.［张雷，刘世荣，孙鹏森等. 气候变化对马尾松潜在分布影响预估的多模型比较. 植物生态学报，2011，**35**(11)：1091-1105.］

［38］ Zhang L，Wang LL，Zhang XD *et al*. The basic principle of random forest and its applications in ecology：a case study of *Pinus yunnanensis*. *Acta Ecologica Sinica*，2014，**34**(3)：650-659.［张雷，王琳琳，张旭东等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例. 生态学报，2014，**34**(3)：650-659.］

［39］ Smith PF，Ganesh S，Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*，2013，**220**(1)：85-91. DOI：10.1016/j.jneumeth.2013.08.024.

［40］ Ai FF，Bin J，Zhang ZM *et al*. Application of random forests to select premium quality vegetable oils by their fatty acid composition. *Food Chemistry*，2014，**143**：472-478. DOI：10.1016/j.foodchem.2013.08.013.

［41］ Song YH，Yu K，Lv W *et al*. Research on case similarity matching of city bus fire incidents. *China Safety Science Journal*，2017，**27**(4)：163-168.［宋英华，余侃，吕伟等. 城市公交车火灾事件案例相似度匹配研究. 中国安全科学学报，2017，**27**(4)：163-168.］

［42］ Xu M，Wu HY，Zhang P *et al*. Long-term prediction method of rice annual agricultural climate status in Jiangsu province based on climatic suitability. *Meteor Mon*，**44**(9)：1200-1207.［徐敏，吴洪颜，张佩等. 基于气候适宜度的江苏水稻气候年景预测方法. 气象，2018，**44**(9)：1200-1207.］

［43］ Qin BQ，Yang GJ，Ma JR. Dynamics of variability and mechanism of harmful cyanobacteria bloom in Lake Taihu，China. *Chinese Science Bulletin*，2016，**61**(7)：759-770.［秦伯强，杨桂军，马健荣等. 太湖蓝藻水华"暴发"的动态特征及其机制. 科学通报，2016，**61**(7)：759-770.］

［44］ Wu XD，Kong FX. The determination of in situ growth rates of the bloomed *Microcystis* in Meiliang Bay，Lake Taihu. *China Environmental Science*，2008，**28**(6)：552-555 .［吴晓东，孔繁翔. 水华期间太湖梅梁湾微囊藻原位生长速率的测定. 中国环境科学，2008，**28**(6)：552-555.］

［45］ Yang XH，Chen J，Zhou L *et al*. South Tai Lake's main lake inlet water bloom time distribution rule and related blue-green alga spa and factor response analysis. *Environmental Monitoring in China*，2011，**27**(2)：92-96.［杨晓红，陈江，周李等. 南太湖入湖口蓝藻水华时空分布规律及相关响应因子分析. 中国环境监测，2011，**27**(2)：92-96.］

［46］ Fan YX，Jin SJ，Zhou P *et al*. Analysis on the distributions and meteorological conditions of cyanobacteria bloom in Chaohu Lake. *Journal of Anhui Agri Sci*，2015，**43**(4)：191-193.［范裕祥，金社军，周培等. 巢湖蓝藻水华分布特征和气象条件分析. 安徽农业科学，2015，**43**(4)：191-193.］

［47］ Zhang HC，Chen XC，Li CJ. The effect of light intensity on cyanophytes vertical distribution. *Environmental Pollution and Control*，2010，**32**(5)：64-67.［张海春，陈雪初，李春杰. 光照度对蓝藻垂直迁移特性影响研究. 环境污染与防治，2010，**32**(5)：64-67.］

［48］ Zhang XY，Jing YS，Chen F *et al*. Effect and forecast of weather conditions on cyanobacterial bloom outbreaks based on RDALR model in Taihu Lake，China. *Chinese Journal of Environmental Engineering*，2016，**10**(10)：5722-5729. DOI：10.12030/j.cjee.201504022.［张晓忆，景元书，陈飞等. 基于 RDALR 模型分析气象条件对太湖蓝藻水华发生的影响及预报. 环境工程学报，2016，**10**(10)：5722-5729.］