

基于聚类法筛选历史相似气象数据的玉米产量 DSSAT-CERES-Maize 预测

陈 上^{1,2}, 窦子荷^{1,2}, 蒋腾聪^{1,2}, 李华龙^{1,2}, 马海姣^{1,2},
冯 浩^{2,3}, 于 强³, 何建强^{1,2,3*}

(1. 西北农林科技大学旱区农业水土工程教育部重点实验室, 杨凌 712100;

2. 西北农林科技大学中国旱区节水农业研究院, 杨凌 712100; 3. 中国科学院水利部水土保持研究所, 杨凌 712100)

摘 要: 根据陕西杨凌、合阳、长武 3 个站点各 2 a 玉米试验, 在对玉米生长模拟模型 CERES-Maize 进行调试、验证的基础上, 探索在生育期内进行动态产量预测的方法并验证。研究将目标生育期内未知气象数据分别用试验地的多年历史同期数据代替, 结合生育期实时数据对应生成多个完整的气象数据序列运行模型预测产量。随着生育期的推进, 逐日在气象数据序列中融入目标年实测的气象数据, 从播种至收获动态模拟玉米产量。此外该研究使用改进前后的 *K*-NN 算法从历史气象年份中筛选目标年的气象相似年份进而预测产量。通过对 3 种方法预测精度及预测效率对比, 确定改进的 *K*-NN 算法最优。研究表明, 玉米生育前期产量预测可靠性和准确率均较差, 抽雄后预测精度迅速提高; 利用改进的 *K*-NN 算法在 3 个站点全生育期预测产量的平均绝对相对误差的均值分别为 9.9%、19.8%、17.9%, 抽雄后预测产量的平均绝对相对误差在 0.2%~12.6% 之间, 相比于使用全部历史年份数据进行全生育期产量预测, 模拟所需时间从 61 min 缩短至 25 min。对该方法中降雨因子的筛选进一步改进可提高预报精度, 未来有望达到业务应用水平。

关键词: 聚类; 气象预报; 模型; 玉米; 产量预测; CERES-Maize; *K*-NN

doi: 10.11975/j.issn.1002-6819.2017.19.019

中图分类号: S165+.27

文献标志码: A

文章编号: 1002-6819(2017)-19-0147-09

陈 上, 窦子荷, 蒋腾聪, 李华龙, 马海姣, 冯 浩, 于 强, 何建强. 基于聚类法筛选历史相似气象数据的玉米产量 DSSAT-CERES-Maize 预测[J]. 农业工程学报, 2017, 33(19): 147-155. doi: 10.11975/j.issn.1002-6819.2017.19.019
<http://www.tcsae.org>

Chen Shang, Dou Zihe, Jiang Tengcong, Li Hualong, Ma Haijiao, Feng Hao, Yu Qiang, He Jianqiang. Maize yield forecast with DSSAT-CERES-Maize model driven by historical meteorological data of analogue years by clustering algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(19): 147-155. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.2017.19.019
<http://www.tcsae.org>

0 引 言

玉米是中国最主要的粮食和经济作物之一, 2015 年以来连续 4 a 为中国产量第 1 位的粮食作物^[1], 对其长势和最终产量进行动态预测非常重要。

作物生长是光合同化物累积的过程, 籽粒产量依赖于其生育期内的气象要素以及人为管理措施^[2], 准确量化产量形成过程较为困难。自 20 世纪 80 年代以来, 气象、农业等多个部门开始应用计算机模拟和遥感技术等手段进行农业产量预测研究, 主要方法包括气象学方法^[3]、农业生物学方法^[4]、经济计量因子预测法^[5]、基于遥感技术的综合性产量预测方法^[6]以及作物生长模型模拟^[7]等。其

中, 前 3 种方法可以较为快捷地进行产量预测, 但是预测精度取决于已建经验模型的因子, 机理性较差且需要进行大量的调研; 遥感手段结合其他方法可以在较大范围进行快速预测, 但是其监测的是电磁波信号而非直接的作物生长状态, 因此预测精度受天气状况和反演方式等因素的干扰较大^[8]。作物生长模型(简称作物模型)具有较强的机理性以及模拟研究低成本等特点, 近年来田间试验结合作物模型模拟的研究方法受到了广泛的关注^[9-10]。

DSSAT (decision support system for agrotechnology transfer) 是目前使用最广泛的作物模型之一, 可逐日模拟多种不同作物的生长发育过程, 计算各影响因子对产量的影响^[11-12]。DSSAT 模拟系统中专门用于玉米生长模拟的 CERES-Maize 模型^[13]被广泛用于气象状况对玉米产量影响的研究^[14-18]。作物模型的本质是建立作物生长发育(包括最终产量的形成)与天气、土壤等环境条件以及管理措施之间的量化函数关系^[7], 应用其进行产量预测可以弥补统计类模型生物意义不明确的缺陷^[19]。现有的主流作物模型在模拟作物生长时通常是估测品种、管理、环境 3 因素之间的相互作用, 而种植品种、管理措施相对确定, 生育期内未知气象数据的生成是利用作物模型进行产量

收稿日期: 2017-02-21 修订日期: 2017-08-10

基金项目: 国家高技术研究发展计划(863 计划)(2013AA102904); 陕西省科技统筹创新工程计划项目(2016KTZDNY03-06); 黄土高原土壤侵蚀与旱地农业国家重点实验室开放基金(A314021402-1611); 西北农林科技大学人才专项资金(千人计划项目); 高等学校学科创新引智计划(111 计划)(B12007)

作者简介: 陈 上, 江苏淮安人, 博士生, 研究方向为农业生态系统模拟。Email: ching@nwsuaf.edu.cn

*通信作者: 何建强, 甘肃天水人, 教授, 博士生导师, 研究方向为农业生态系统模拟。Email: jianqiang_he@nwsuaf.edu.cn

预测的关键^[20]。利用统计学模型进行产量预测时,使用历史多年气象数据的均值作为模型的输入参数,但由于作物生长对环境响应的非线性,使用该方法进行产量预测存在一定问题^[21-22]。近年来利用大气—海洋—陆面模式的气象预测研究广泛开展,相应开发基于过程的农学、生态学和农水文学综合模拟模型为产量预测提供了新的解决方案^[23-24],但是这类模型生成的气象数据与目标年份通常存在时空不匹配的问题^[25]。此外,还有研究利用气象发生器生成一般作物模型需要的日步长气象数据序列^[26-28]。现有的气象发生器主要分为参数型和非参数型 2 大类,参数型气象发生器通常以降雨因子为驱动变量,先模拟降雨发生时间以及降雨量,再模拟其余气象因子^[29-30],但其主要缺点是无法准确模拟干旱以及长时间的降雨^[31],尽管很多人对其进行改进但仍不理想^[32-33],非参数型气象发生器并不需要对气象因子的分布进行假设且各气象因子之间相互独立^[34],故成为气象要素预测研究中一类非常重要的方法。

气象数据作为典型的时空数据,在时间上具有时间序列特征,在空间上又存在着地理空间相似性^[35],而区域气象数据时空相似性是农业气候区划的主要依据^[36],因此在同站点多年历史情境中可以找到与目标年份气象要素相似的年型,基于相似年型的气象数据进行产量预测已经开展大量工作^[37-40]。此类非参数方法需要计算目标年已有气象数据与该站点多年历史同日期气象数据之间的欧氏距离 (Euclidian distance) 并进行几何聚类分析,从历史年份中挑选出与目标年已有天气模式类似的相近年型 (analogue years),再使用其数据进行产量预测。本研究使用 K 最近邻算法 (K nearest neighbor, K -NN) 筛选相似年型,目前该方法已经在多地应用于天气预测的研究^[41-43]。不设定 K 值时,即用研究站点所有的历史气象数据作为未来气象的可能情境来补充目标年生育期内的未知气象数据,但这种方法计算量较大、预测结果的不确定性较高;此外,有研究发现 K -NN 算法可以从历史年份中较好地筛选出与目标年最高温度、最低温度以及阳光辐射较为相似的年份,但是对降雨的筛选效果较差^[44],而降雨对作物生长发育和产量形成至关重要,所以有必要对现有的 K -NN 算法进行改进以提高产量预测的精度。

本研究将 DSSAT-CERES-Maize 模型引入到玉米产量预测研究中,主要目的包括:1) 建立基于历史气象数据和 DSSAT-CERES-Maize 模型的玉米产量预测方法;2) 改进现有的 K -NN 方法以提高上述产量预测方法的效率,并降低预测的不确定性。

1 模型及数据筛选

1.1 DSSAT-CERES-Maize 模型

DSSAT 模型是目前使用最为广泛的作物模型系统之一^[11, 45],能够模拟作物营养生长和生殖生长发育过程、作物光合作用、呼吸作用、干物分配和植株生长以及衰老等基本生理生态过程。DSSAT 包括一维土壤水平衡模块和一维土壤氮素平衡模块,其内嵌的 CERES 系列模型最早由美国农业部农业研究服务署 (USDA-ARS) 主

导于 20 世纪 70 年代开始研发,已在全球得到广泛的应用^[46-47]。CERES-Maize 是 CERES (crop environment resource syntheses system) 系列模型中专门用于模拟玉米生长发育和产量形成的模型,该模型以作物系统模块的形式在 DSSAT-CSM (cropping system model) 公共平台上运行。为提高模拟精度,该模型特别考虑了玉米受基因型以及天气因素影响物候发育阶段、叶茎根的生长、生物量的累积和分配、土壤水分平衡,以及作物水分利用等重要过程^[48]。

DSSAT 模型主要输入参数分为环境和人为管理 2 大类,环境参数包括天气数据、土壤理化特征、水肥初始状态等;人为管理措施主要包括播种、灌溉、施肥、耕作方式等。由于该模型以日为步长模拟作物生长过程,因此其输出结果包括干物质、叶面积指数、籽粒质量、土壤水分等参数的逐日变化动态以及最终产量、物候期、最大叶面积指数、总叶片数等。

1.2 气象文件的生成

1.2.1 利用历史同期气象数据直接替代未知数据

对于生育期内未知的气象数据,分别利用历史年份的相应数据进行替代,结合已实测数据生成对应年份数量的气象文件(即播种至预报日使用当年实测气象数据,预报日至收获日用历史各年气象数据替代^[37]),随着生育期的推进将逐日实测的气象数据动态融入,不断生成产量模拟所需的气象数据序列,即可对该季玉米产量进行动态预测。

1.2.2 利用 K -NN 算法筛选气象相似年份

K -NN 植根于模式识别,基于该方法可以在目标空间(如具有观测值的历史年份)中寻找一个与定义了矢量特征的目标对象类似的模式^[49],通过设置不同的判断标准,可以筛选出具有不同相似程度的结果。Bannayan 等^[44]已经较详细地描述了应用 K -NN 算法进行气象预测的原理以及计算流程。对于选定的站点,计算目标年份已有的每天气象数据与历史同日气象数据的欧氏距离 d_j ,由于 CERES-Maize 模型运行需要逐日的 4 个不同的气象因子(最高气温 T_{max} 、最低气温 T_{min} 、降雨 Rain、太阳辐射 SRAD)的数据且量纲不同,因此在计算欧氏距离时首先要对各气象因子数据进行标准化:

$$V_s = (V - V_{min}) / (V_{max} - V_{min}) \quad (1)$$

式中 V 为气象变量; V_s 为标准化后的气象变量; V_{max} 、 V_{min} 分别为该气象变量在整个数据空间中的最大、最小值。在本研究中 V_{max} 、 V_{min} 分别为已有历史气象数据中该变量所有值的最大和最小值。

欧氏距离计算如下:

$$d_j = \sqrt{\sum_{n=1}^4 (V_{sijn} - V_{smjn})^2} \quad (2)$$

式中 d_j 为玉米生育期第 j 天目标年份 i 和历史年份 m 的 4 个不同气象变量之间的总体欧氏距离; V_{sijn} 为第 j 天目标年份 i 的第 n 个标准化气象变量; V_{smjn} 为第 j 天历史年份 m 的第 n 个标准化气象变量。

本研究设置的邻数 K 为 10,即从 1 月 1 日起每天挑选欧氏距离最小的 10 个历史年份作为该日的气象近似年

份, 然后从预测日期之前各日挑选的近邻中筛选 10 个出现频率最高的年份, 作为截止该日目标年的近似年份。

1.2.3 利用改进 K-NN 算法筛选气象相似年份

上述一般的 K-NN 算法通过计算逐日 4 个气象变量的综合欧氏距离来判断不同年份天气模式的接近程度, 从而筛选出气候相似年份。但对作物生长而言, 作物的生长发育通常受阶段性的积温、累积降雨等因素的影响较大 (冷害、冰雹等极端瞬时天气事件除外), 例如降雨发生时间早、晚 1 d 对作物生长的影响几乎可以忽略不计, 但是使用 K-NN 算法筛选的相似年份却会有很大的区别。因此, 本研究对一般的 K-NN 算法进行改进, 使用某一研究日期前若干天 (本研究取 7 d) 气象数据的均值作为该日的实际气象数据 (式 (3)), 进而计算新的欧氏距离 d'_j , 再据此从历史年份中筛选目标年的相似年份。

$$V'_{sijn} = \sum_{j=j-7}^j (V_{sijn}) / 7 \quad (j > 7) \quad (3)$$

$$d'_j = \sqrt{\sum_{n=1}^4 (V'_{sijn} - V'_{smjn})^2} \quad (4)$$

式中 V'_{sijn} 为目标年 i 内第 j 天的 4 个气象因子中的第 n 个因子在该日前 7 d 的均值。 V'_{smjn} 为历史 m 年内第 j 天的 4 个气象因子中的第 n 个因子在该日前 7 d 的均值。

2 模型应用

2.1 试验区概况

本研究选取陕西省杨凌西北农林科技大学旱区农业水土工程教育部重点实验室节水灌溉试验站 (34°17'N, 108°04'E, 506 m)、合阳县甘井镇西北农林科技大学旱作农业试验站 (35°19' N, 105°05' E, 900 m) 以及长武县中国科学院长武黄土高原农业生态试验站 (35°14'N, 107°41'E, 1 220 m) 3 个试验站。以这 3 个站点各自 2 a 的试验数据及历史已有气象实测值 (表 1 和表 2) 来建立并验证产量预测方法。

3 个试验站玉米种植均无灌溉措施, 全部肥料作为基肥在播种前施入, 其余耕作措施和当地习惯一致。生育期以全小区 50% 植株达到发育要求为标准, 记录玉米出苗、抽穗、成熟的日期。生物量和产量通过将玉米各器官分离称鲜质量, 而后分别装于纸袋, 在 105 °C 下杀青 30 min 后, 在 80 °C 下烘干至质量恒定, 分别测定各器官干物质质量。收获期在各小区随机选取 10 株测定穗质量、行数、穗粒数以及百粒质量从而确定产量构成要素。本研究所需的气象数据由国家气象科学数据共享服务平台获取, 土壤参数通过实地取样测定, 管理措施根据当地实际情况进行。

表 1 不同站点历史气象数据

Table 1 Historical weather data of different experimental sites

试验站点 Site	历史年份 Historical years	平均蒸发量 Mean evaporation/mm	平均温度 Mean temperature/°C	阳光辐射 SRAD/(MJ·m ⁻²)		降雨 Precipitation/mm			最高气温 Max temperature/°C		最低气温 Min temperature/°C	
				最大 Max	最小 Min	最大 Max	最小 Min	平均 Mean	最大 Max	最小 Min	最大 Max	最小 Min
杨凌 Yangling	1956—2015	884	12.3	30.5	3.5	140.8	0	637	42	-7.7	28.6	-19.4
长武 Changwu	1957—2014	1 335	9.5	30.5	4.3	142.2	0	575	37.6	-11.6	24	-26.2
合阳 Heyang	1962—2014	1 833	105	30.5	4.2	114.7	0	550	40.1	-8.3	27.4	-21.2

表 2 不同试验站点信息

Table 2 Information of different experimental sites

站点 Site	试验 年份 Year	播种 日期 Sowing date	收获 日期 Harvest date	品种 Cultivar	土壤类型 Soil type	容重 Bulk density (g·cm ⁻³)	田间持水率 Field capacity (cm ³ ·cm ⁻³)	凋萎系数 Permanent wilting point (cm ³ ·cm ⁻³)	有机碳 Organic carbon (g·kg ⁻¹)	pH 值 pH value	播种密度 Sowing density (株·hm ⁻²)	施肥 Fertilizers/(kg·hm ⁻²)		
												N	P	K
杨凌 Yangling	2014	06—12	09—27	郑单-958	壤土	1.31	26	12	8.2	8.4	55 000	210	96	77
	2015	06—12	09—24											
长武 Changwu	2010	04—24	09—13	先玉-335	黑垆土	1.29	27	13	8.5	8.1	60 000	138	38	
	2011	04—24	09—13											
合阳 Heyang	2009	04—18	09—10	豫玉-22	壤土	1.33	22	11	11.4	8.2	49 500	200	85	70
	2011	04—18	09—10											

注: 田间持水量、凋萎系数和容重为 0~100 cm 各层土壤平均值, pH 值、有机碳质量分数为 0~20 cm 土层平均值。

Note: The values of field capacity, permanent wilting point and bulk density are mean across soil depths of 0-100 cm; and pH value and organic carbon are means across 0-20 cm soil depth.

2.2 产量预测结果评价指标

由于生育期内气象数据的未知性, 通常无法得到一个确切的预测结果, 因此本研究将逐日预测的多个产量的分布范围作为该日产量预测的结果, 采用 3 个指标对预测结果进行评价: 1) 绝对相对误差 (absolute relative

error, ARE), 即相对误差的绝对值; 2) 变异系数 (coefficient of variation, CV) 评价每日预测产量的离散程度即可靠度; 3) 平均绝对相对误差 (mean absolute relative error, MARE) 来评价产量预测值和实际值之间的差异, 式 (5)。

$$\text{MARE}_i = \sum_{j=1}^{n'} |Y_{ij} - Y| / (Y \cdot n) \times 100\% \quad (5)$$

式中 i 含义同上式; n' 为已有历史气象数据的年份数; j 为第 j 个历史年份; Y 为目标年玉米实际产量; Y_{ij} 为生育期内第 i 天使用第 j 年的历史气象数据生成气象的气象文件运行模型所预测的玉米产量。

3 结果与分析

3.1 模型适用性评价

在 CERES-Maize 模型中分别输入 3 个站点的土壤、管理以及品种数据, 结合试验年份的气象数据分别模拟

个品种玉米各 2 季的生长, 结合田间实测物候期以及生物量等数据对模型的适用性进行评价 (表 3)。总体而言, 3 个品种所估计的参数在各自站点模拟精度均较好, 各参数模拟值与实测值的 ARE 在 0~13.3% 之间。相比于抽雄期, 成熟期的模拟结果略差, 主要因为玉米成熟期的监测较为困难, 且在各站点试验过程中每年播种、收获基本为相同日期, 可能与玉米的实际生育状态存在差异。模拟误差最大的是生物量, 因为在参数估计时并未使用生物量这一参数, 但是 ARE 仍不超过 15%。因此, CERES-Maize 模型适用于杨凌、长武以及合阳这 3 个站点。

表 3 不同站点基于 CERES-Maize 模型的模拟值 (Sim.) 与实测值 (Obs.) 对比

Table 3 Comparison between observed and simulated values based on CERES-Maize model at different sites

站点 Sites	年份 Year	抽雄期 Tasseling/d			成熟期 Maturity/d			生物量 Biomass/(kg·hm ⁻²)			籽粒产量 Grain yield/(kg·hm ⁻²)		
		Sim.	Obs.	ARE/%	Sim.	Obs.	ARE/%	Sim.	Obs.	ARE/%	Sim.	Obs.	ARE/%
杨凌 Yangling	2014	62	62	0.0	106	109	2.8	10 882	12 545	13.3	6 150	6 309	2.5
	2015	61	63	3.2	102	105	2.9	13 019	14 213	8.4	6 999	6 750	3.7
合阳 Heyang	2009	90	88	2.3	138	145	4.8	13 733	15 836	13.3	7 725	8 212	5.9
	2011	89	87	2.3	135	144	6.3	12 526	13 860	9.6	6 608	6 650	0.6
长武 Changwu	2010	87	86	1.2	145	147	1.4	16 377	18 094	9.5	10 563	10 618	0.5
	2011	87	87	0.0	148	147	0.7	15 513	17 365	10.7	9 176	9 292	1.2

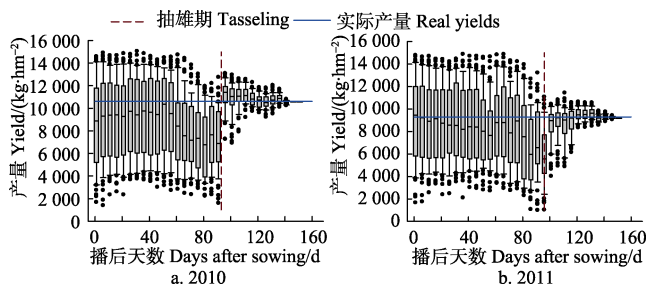
注: ARE 为绝对相对误差。Note: ARE is absolute relative error.

3.2 直接利用历史气象数据产量预测结果分析

利用陕西省 3 个不同站点 (杨凌、合阳、长武) 各 2 个不同年份的试验数据来检验本研究所建立的夏玉米产量预测方法, 以长武春玉米产量预测为例, 如图 1 所示。玉米生育期内预报日之后的未知气象数据分别用各站点历史同期气象数据替代生成完整气象数据序列, 结合 CERES-Maize 模型逐日预测产量 (图 1, 数据点时间间隔为 5 d)。

而是在约 90 d 的时候才开始迅速收敛, 如播后 100、130 d 预测产量的范围为 4 933~10 826、8 484~10 565 kg/hm²。在长武地区播后 90 d 对应春玉米的抽雄期, 此时玉米完成营养生长, 由气象的未知性而导致的产量预测的未知性降低, 可预测性大大提高。另一方面, 模拟的产量直接受驱动模型的气象数据决定, 部分年份的气象情形过于理想或者过于恶劣即属于低频年型 (图 1 中的异常点) 均对预测产量的范围产生影响, 因此可以对逐日预测产量取不同百分位数进行范围筛选。

为了进一步分析生育期内产量预测精度的变化状况, 计算 3 个站点生育期内每日多个模拟产量的 CV 以及 MARE, 为了表述简洁设定数据点间隔为 10 d, 如图 2 所示。

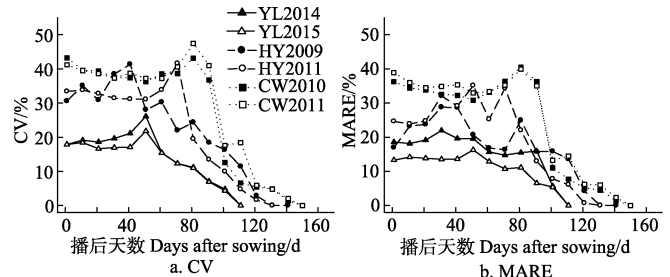


注: 箱图表示从播种之日起每隔 5 d 的预测产量。箱图上下边缘分别表示模拟产量的 75% 和 25% 百分位数, 上下误差棒分别表示 90% 和 10% 百分位数。Note: Box plots show forecast yields on every 5 days from sowing day during growth seasons. Edges of boxes are 75% and 25% percentiles and solid vertical bars are 90% and 10% percentiles, respectively.

图 1 长武春玉米产量动态预测

Fig.1 Dynamic of yield forecast of spring maize at Changwu

整体来看 (图 1), 2 a 的实际产量均在逐日预测结果的范围内, 逐日预测的产量在生育前期的分布范围较广, 随着生育期的推进, 合成的气象序列中实测气象数据所占的比例不断增加, 逐日预测产量的分布范围并没有明显随之减小, 如 2011 年播后 30、60、90 d 预测产量的范围分别为 3 531~14 461、3 413~14 828、961~13 210 kg/hm²,



注: YL, 杨凌; HY, 合阳; CW, 长武
Note: YL, Yangling; HY, Heyang; CW, Changwu

图 2 不同地点玉米产量预测的变异系数和平均相对绝对误差
Fig.2 Coefficient of variation (CV) and mean absolute relative error (MARE) of forecast yields at different sites

图 2 表明, 合阳站点的春玉米生育期与长武相近, 抽雄期约为播后 80 d; 杨凌的夏玉米抽雄期约为播后 60 d。整体而言 3 个站点每日模拟产量的 CV 和 MARE 在生育

前期没有明显规律、在抽雄附近时段均有一个迅速下降的过程, 抽雄后逐日预测产量的 CV 和 MARE 均小于 20%。杨凌 2 a 产量模拟的 2 个指标(CV 最大值为 26.2%, MARE 最大值为 22.0%) 明显小于另外 2 个站点, 这主要是因为合阳和长武地区处于渭北旱塬, 降雨量年际变化大且蒸发强烈, 其余气象因子的变化空间大于杨凌; 相比之下杨凌处于关中平原腹地, 农业生产系统的稳定性高于另外 2 个站点。对于模拟所需时间, 以长武 2011 年春玉米全生育期产量模拟为例, 该年全生育期为 150 d, 使用历史气象年份为 1957—2014 共 58 a, 合计生成 8 700 个气象文件并逐次调用 DSSAT 模型、筛选模拟结果, 就笔者使用的计算机处理器 (Intel (R) Core (TM) i7-7700 CPU @ 3.60 GHz, 下同) 而言, 完成 1 个生育期的产量模拟需要 61 min, 因此有必要对运算过程进行优化并寻找提高预测精度的产量预测方法。

3.3 基于改进前后的 K-NN 算法预测的产量结果分析

对 3 个站点共 6 个生长季分别应用全部历史年份、改进前后的 K-NN 算法从历史多年气象数据中挑选的相似年型数据用于产量模拟, 分别计算逐日产量预测值的 CV 和 MARE (表 4)。

表 4 基于不同气象数据替代方法在 3 个站点预测产量精度对比
Table 4 Comparison of yield forecast precision among different methods of weather generation at 3 sites

站点 Sites	预测方法 Forecast method	CV/%			MARE/%		
		最大 Max	最小 Min	均值 Average	最大 Max	最小 Min	均值 Average
杨凌 Yangling	所有历史年份	26.1	4.3	15.1	22.0	5.4	14.6
	一般 K-NN	27.4	3.7	14.7	20.3	5.7	12.9
	改进 K-NN	26.6	4.5	11.7	16.8	4.7	9.9
合阳 Heyang	所有历史年份	41.6	1.6	24.0	34.4	0.7	19.6
	一般 K-NN	33.8	3.2	22.0	35.1	0.7	20.3
	改进 K-NN	51.7	1.8	23.8	39.5	0.2	19.8
长武 Changwu	所有历史年份	47.5	1.4	29.1	40.5	1.4	25.8
	一般 K-NN	47.3	1.1	26.9	37.2	1.2	19.6
	改进 K-NN	36.3	1.2	23.3	34.4	1.3	17.9

注: 最小值为距收获最近一次的模拟结果。

Note: The minimum value is related to the last simulation before harvest.

从 6 个生育期的整体模拟结果来看 (表 4), 改进后的 K-NN 算法模拟精度最高: 3 个站点全生育期不同时段预测产量的 CV 的均值为 11.7%、23.8%、23.3%, MARE 的均值分别为 9.9%、19.8%、17.9%, 即通过在历史年份中筛选与目标年份一段时间内气象要素平均值得到的相似年型的数据进行产量预测, 在玉米生长的大部分年份可以取得较好的产量预测结果, 但其中合阳站模拟效果较差, 比较该站点 2 季玉米产量预测发现, CV 和 MARE 最大值出现在 2011 年播后 70 d, 合阳该年播后前 70 天累计降雨 63.2 mm, 远低于同时段该地区多年平均降雨量 114.4 mm, 即出现严重的伏旱, 为相似年型的筛选造成影响。此外杨凌站点模拟结果的 CV 和 MARE 最大值远小于其余站点, 但最小值大于其余站点。这主要因为杨凌地区多年气候要素较为稳定, 且种植的夏玉米生育期

只有春玉米的 2/3 左右, 而模拟结果的最小值取得是收获前 10 d 的模拟结果, 此时夏玉米生长仍处于旺盛时期、产量可变范围较大, 而对于春玉米而言此时生长已非常缓慢, 所以杨凌模拟的最优精度低于另外 2 个站点。

为了进一步分析 3 种产量预测方法在不同时段预测效果的优劣, 分析长武 2011 年全生育期 16 个时段共进行的 48 次产量预测 (图 3), 其中应用改进的 K-NN 预测精度最优 (CV 和 MARE 均最小) 占 13 个时段, 而应用一般 K-NN 算法预测精度优于使用全部历史年份占 8 次, 此外 3 种方法预测结果的 CV 的最大值、最小值、均值分别为: 47.5%、47.3%、41.4%; 2.0%、2.2%、1.5%; 29.9%、29.3%、22.9%; MARE 的最大值、最小值、均值分别为: 40%、35.7%、34.4%; 2.4%、1.9%、1.5%; 26.6%、21.6%、18.1%, 即改进后的 K-NN 算法能够对使用全部历史气象数据进行产量预测的方法进行优化, 但是一般 K-NN 算法却不能, 这主要是因为一般的 K-NN 算法只是一种简单的机器学习算法, 使用该方法得到的是与目标年每日 4 个气象因子数值上较为相似的年份, 但是气象要素对作物生长发育过程的影响通常是阶段性的而非瞬时性的 (冰雹、冷害等极端气象事件除外)。例如, 降雨发生的时间提前或者推迟 1~2 d 对作物生长的影响几乎可以忽略不计, 真正产生影响的是某一生长阶段的总降雨量; 影响作物发育速度的也不是具体某一天的最高气温和最低气温, 而是某一阶段的平均气温与积温。使用一般的 K-NN 算法挑选的相似年份并不能真正体现气象要素对作物生长的影响, 因此使用其筛选年份的气象数据预测产量的精度总体较差; 而本研究使用改进的 K-NN 算法依据连续 7 d 的气象因子的平均值筛选相似年型, 故整体上提高了产量预测精度, 但该方法同时弱化了极端气象事件在实际中的低频型, 所以当生育期内某一时段内出现此类气象状况时, 在对应时段筛选的相似年型将与实际年份存在较大差异, 继而导致在该阶段预测产量变差。其次, 使用 3 种方法在玉米抽雄前预测产量的误差均较大, 但抽雄后 CV 和 MARE 的最大值分别为 18.5% 和 14.5%, 远低于生育前期。此外, 改进前后的 K-NN 算法预测全生育期逐日的产量所用的时间分别为 14 min、25 min, 远低于使用全部历史气象数据进行预测所需时间。因此改进的 K-NN 算法相比于使用全部历史年份的产量预测方法而言, 提高预测精度的同时能够明显缩短产量预测时间。

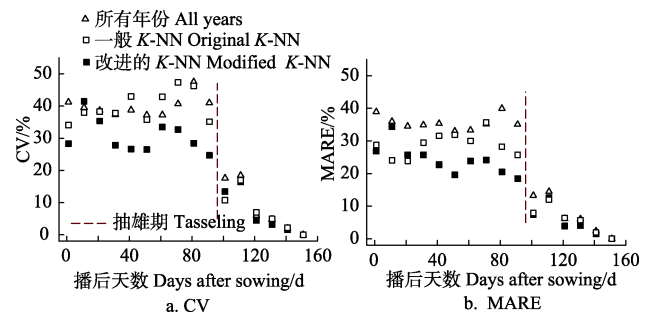


图 3 基于不同气象数据替代方法的长武 2011 年春玉米预测产量的变异系数和平均绝对相对误差的动态

Fig.3 Dynamic of coefficients of variation (CV) and mean absolute relative error (MARE) of forecast yields with different methods of weather generation in Changwu in 2011

4 讨论

4.1 预测产量的分布规律

作物生长是光合产物累积的过程, 生育期内每天的天气状况都会对最终的作物特性产生影响^[50]。从播种到收获, 随着合成的气象数据序列中实测数据所占比例逐渐增加, 由气象因素变异性所产生的产量预测不确定性将逐渐减少, 即产量预测精度逐渐提高^[51]。但是通过本研究3个站点(杨凌、合阳、长武)各自2a共计6个不同生长季的玉米产量预测结果可以发现整个生育期内预测产量的分布并非一开始就随着已知气象数据的增加而立即开始收敛, 而是前期分布较为分散, 直到抽雄之后逐渐开始收敛。

玉米产量形成涉及到“源”、“流”和“库”3个方面, 其中“源”即光合器官提供的光合产物是籽粒产量形成的根本来源^[52]。玉米籽粒有机物的“源”主要有2个部分, 约2/3来由光合器官在灌浆期合成有机物的直接输入, 1/3主要通过茎、叶等器官在营养生长阶段存储有机物活化后输入籽粒^[53]。抽雄后, 茎秆等存储的有机物含量随之确定下来, 同时抽雄至乳熟末期玉米群体叶面积也处于高而稳定的时期, 此时气象因素对作物产量的限制性影响已经越来越小, 此时逐日模拟产量的分布随着实测气象数据增加而逐渐收敛趋近于实际模拟产量。

4.2 应用K-NN算法的玉米产量预测结果

传统产量预测方法主要是基于多站点实地取样分析建立统计模型, 此类方法一方面需要大量数据用于建模, 需要大量时间以及人力; 另一方面机理性较差, 对生育期内实测气象数据的响应较弱。用以日为步长模拟作物生长的模型可以根据生育期内实测气象数据进行逐日产量预测, 但是作物模型通常在站点尺度上进行作物生长发育过程的模拟, 而产量预测往往需要在较大的空间尺度进行, 因此有必要将作物模型与卫星、无人机等遥感手段获取的数据进行同化, 以对预测的产量升尺度。

气象数据具有典型的时空双属性, 特定区域的气候状况存在明显的特征^[35], 基于这一特征进而可以对农业气候区进行划分^[36]。通过对目标年份的气象数据与同站点历史数据进行聚类分析是寻找相似年型的常用方法, 本研究采用的聚类方法是K-NN算法。一般的K-NN算法通过计算目标年与历史各年同期气象数据之间的欧氏距离, 筛选相似年型, Bannayan等^[44]比较分析了基于K-NN算法筛选出相似年份的气象数据, 结果发现最高气温、最低气温、阳光辐射这3个因子的筛选结果较好、降雨因子的筛选结果较差, 这主要是因为降雨事件的发生具有随机性, 并非每天都有数值, 因此有必要对总体欧式距离的计算方法进一步进行改进, 如对4个气象要素采用最小值法、乘法、不同权重法等。此外, K-NN算法进行聚类分析的分类结果受设定的近邻数量K的影响, 下一步有必要设置不同的K值进行比较。

4.3 玉米产量预测方法的进一步改进

本研究尝试了3种获取生育期内未知气象数据的非参数型方法, 相较于气候模拟和天气发生器生成的结果,

此类方法不需要提前对未知气象情境的特征进行假设以及区域数据降尺度, 而是直接利用目标站点的历史实测气象数据替代未知气象数据, 继而减少了人为误差。由于气象数据的未知性, 预测的产量通常在一定范围内或者伴随一定的概率而非某一个定值, 但由于已知历史气象数据总量有限, 很多地方从1953年以后才有较为系统的气象观测数据, 因此每日产量预测仅能得到有限数量的结果, 并不能完整代表未来所有可能出现的气象情境。有研究使用气象发生器结合Sirius模型模拟小麦产量, 结果发现当有超过300个气象情境用于填充未知气象数据时, 预测产量的分布将不会发生改变^[51], 因此下一步考虑与气象发生器结合进行玉米产量预测。

此外, 本研究采用生育期内已知气象数据与历史气象数据结合的方法生成气象数据序列, 考虑到气象部门已经能够在一定时间内对天气状况及气象灾害进行高精度预测^[54], 因此有必要采用“生育期内实测+预报+历史气象数据”的三段式方法生成气象数据序列, 以提高CERES-Maize模型预测产量的准确性和时效性。

5 结论

本研究利用历史气象数据、K-NN算法和CERES-Maize模型相结合的方法, 构建了玉米生育期内产量动态预测方法, 然后利用陕西杨凌、合阳、长武3个站点不同玉米品种各2a雨养试验数据对上述产量预测方法的可靠性和精度进行了验证。根据研究结果得出以下结论:

1) 3种方法在玉米抽雄前产量预测精度均较差, 长武站春玉米抽雄后模拟产量的CV和MARE不超过18.5%和14.5%; 使用改进的K-NN算法在3个站点全生育期预测产量的MARE的均值分别为9.9%、19.8%、17.9%, 该方法预测精度最佳且运算所需时间较短, 便于在较大尺度开展的产量预测研究。

2) 在降雨正常年份, 使用改进的K-NN算法可以有较高预测精度且耗时较短; 降雨异常年份宜使用所有年份气象数据进行产量预测。

在下一步研究中有必要进一步改进K-NN算法中总体欧式距离的计算方法, 增强降雨因素的权重, 此外引入气象发生器、实时气象预报数据等方式, 进一步提高本研究提出玉米产量预测方法的准确性和时效性。

[参 考 文 献]

- [1] 中华人民共和国国家统计局. 中国统计年鉴[M]. 北京: 中国统计出版社, 2016.
- [2] 代立芹, 吴炳方, 李强子, 等. 作物单产预测方法研究进展[J]. 农业网络信息, 2006(4): 24-27.
Dai Liqin, Wu Bingfang, Li Qiangzi, et al. Overview of technique used by crop yield for ecasting[J]. Agricultural Network Information, 2006(4): 24-27. (in Chinese with English abstract)
- [3] 甘一忠, 刘流. 特征展开模糊推理模式在农作物产量预报中的应用[J]. 模糊系统与数学, 2002, 16(1): 104-109.
Gan Yizhong, Liu Liu. Application of character spread faintness consequence model to forecast the output of crop[J].

- Fuzzy Systems and Mathematics, 2002, 16(1): 104–109. (in Chinese with English abstract)
- [4] 许清海, 孟令尧, 阳小兰, 等. 应用花粉分析预报板栗产量的研究[J]. 植物生态学报, 1999, 23(4): 370–378.
Xu Qinghai, Meng Lingyao, Yang Xiaolan, et al. Prediction yield of castanea mollissima with pollen concentration[J]. Acta Phytocologica Sinica, 1999, 23(4): 370–378. (in Chinese with English abstract)
- [5] 陈锡康, 杨翠红. 农业复杂巨系统的特点与全国粮食产量预测研究[J]. 系统工程理论与实践, 2002, 22(6): 108–112.
Chen Xikang, Yang Cuihong. Characteristic of agricultural complex giant system and national grain output prediction[J]. Systems Engineering Theory and Practice, 2002, 22(6): 108–112. (in Chinese with English abstract)
- [6] 吴炳方. 全国农情监测与估产的运行化遥感方法[J]. 地理学报, 2000, 55(1): 25–35.
Wu Bingfang. Operational remote sensing methods for agricultural statistics[J]. Acta Geographica Sinica, 2000, 55(1): 25–35. (in Chinese with English abstract)
- [7] 刘布春, 刘文萍, 梅旭荣, 等. 我国农业气象业务引入作物生长模型的前景[J]. 气象, 2006, 32(12): 10–15.
Liu Buchun, Liu Wenping, Mei Xurong, et al. Prospects for crop growth models introduced into agrometeorology services in china[J]. Meteorological Monthly, 2006, 32(12): 10–15. (in Chinese with English abstract)
- [8] 吴蕾, 柏军华, 肖青, 等. 作物生长模型与定量遥感参数结合研究进展与展望[J]. 农业工程学报, 2017, 33(9): 155–166.
Wu Lei, Bai Junhua, Xiao Qing, et al. Research process and prospect on combining crop growth models with parameters derived from quantitative remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(9): 155–166. (in Chinese with English abstract)
- [9] 曹宏鑫, 赵锁劳, 葛道阔, 等. 作物模型发展探讨[J]. 中国农业科学, 2011, 44(17): 3520–3528.
Cao Hongxin, Zhao Suolao, Ge Daokuo, et al. Discussion on development of crop models[J]. Scientia Agricultura Sinica, 2011, 44(17): 3520–3528. (in Chinese with English abstract)
- [10] 王亚莉, 贺立源. 作物生长模拟模型研究和应用综述[J]. 华中农业大学学报, 2005, 24(5): 529–535.
Wang Yali, He Liyuan. A review on the research and application of crop simulation model[J]. Journal of Huazhong Agricultural University, 2005, 24(5): 529–535. (in Chinese with English abstract)
- [11] Jones J W, Hoogenboom G, Porter C H, et al. The DSSAT cropping system model[J]. European Journal Agronomy, 2003, 18(3): 235–265.
- [12] Hoogenboom G, Jones J, Wilkens P, et al. Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.5[M]. Honolulu: University of Hawaii, 2010.
- [13] Jones C A, Kiniry J R, Dyke P T. CERES-Maize: A Simulation Model of Maize Growth and Development[M]. Texas: Texas A and M University, 1986.
- [14] He J, Dukes M D, Hochmuth G J, et al. Evaluation of sweet corn yield and nitrogen leaching with CERES-Maize considering input parameter uncertainties[J]. Transactions of the ASABE, 2011, 54(4): 1257–1268.
- [15] Jing Q, Shang J, Huffman T, et al. Using the CSM–CERES–Maize model to assess the gap between actual and potential yields of grain maize[J]. Journal of Agricultural Science, 2016, 155(2): 1–22.
- [16] Kassie B T, Ittersum M K V, Hengsdijk H, et al. Climate-induced yield variability and yield gaps of maize (*Zea mays* L.) in the Central Rift Valley of Ethiopia[J]. Field Crops Research, 2014, 160(4): 41–53.
- [17] Jin Zhiqing, Zhu Dawei. Impacts of changes in climate and its variability on food production in northeast china[J]. Acta Agronomica Sinica, 2008, 34(9): 1588–1597.
- [18] Guo Ruiping, Lin Zhonghui, Mo Xingguo, et al. Responses of crop yield and water use efficiency to climate change in the North China Plain[J]. Agricultural Water Management, 2010, 97(8): 1185–1194.
- [19] 黄晚华, 薛昌颖, 李忠辉, 等. 基于作物生长模拟模型的产量预报方法研究进展[J]. 中国农业气象, 2009, 30(S1): 140–143.
Huang Wanhua, Xue Changyin, Li Zhonghui, et al. Research progresses in yield forecasting method based on crop growth simulation model in china[J]. Chinese Journal of Agrometeorology, 2009, 30(S1): 140–143. (in Chinese with English abstract)
- [20] Zinyengere N, Mhizha T, Mashonjowa E, et al. Using seasonal climate forecasts to improve maize production decision support in Zimbabwe[J]. Agricultural and Forest Meteorology, 2012, 151(12): 1792–1799.
- [21] Porter J R, Semenov M A. Crop responses to climatic variation[J]. Philosophical Transactions of the Royal Society of London, 2005, 360(1463): 2021.
- [22] Semenov M A, Barrow E M. Use of a stochastic weather generator in the development of climate change scenarios[J]. Climatic Change, 1997, 35(4): 397–414.
- [23] Hoogenboom G. Contribution of agrometeorology to the simulation of crop production and its applications[J]. Agricultural and Forest Meteorology, 2000, 103(1): 137–157.
- [24] Hyndman R J, Koehler A B. Another look at measures of forecast accuracy[J]. International Journal of Forecasting, 2005, 22(4): 679–688.
- [25] Goel S, Dash S K. Response of model simulated weather parameters to round-off-errors on different systems[J]. Environmental Modelling and Software, 2007, 22(8): 1164–1174.
- [26] Geng S, Penning de Vries F W T, Supit I. A simple method for generating daily rainfall data[J]. Agricultural and Forest Meteorology, 1986, 36(4): 363–376.
- [27] Kilsby C G, Jones P D, Burton A, et al. A daily weather generator for use in climate change studies[J]. Environmental Modelling and Software, 2007, 22(12): 1705–1719.

- [28] Jones P G, Thornton P K. MarkSim: Software to generate daily weather data for Latin America and Africa[J]. *Agronomy Journal*, 2000, 92(3): 445—453.
- [29] Richardson C W. Stochastic simulation of daily precipitation, temperature, and solar radiation[J]. *Water Resources Research*, 1981, 17(1): 182—190.
- [30] Nicks A D, Harp J F. Stochastic generation of temperature and solar radiation data[J]. *Journal of Hydrology*, 1980, 48(1/2): 1—17.
- [31] Hartkamp A D, White J W, Hoogenboom G. Comparison of three weather generators for crop modeling: A case study for subtropical environments[J]. *Agricultural Systems*, 2003, 76(2): 539—560.
- [32] Semenov M A, Brooks R J, Barrow E M, et al. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates[J]. *Climate Research*, 1998, 10(2): 95—107.
- [33] Sharif M, Burn D H. Simulating climate change scenarios using an improved K-Nearest Neighbor model[J]. *Journal of Hydrology*, 2006, 325(1): 179—196.
- [34] Brandsma T, Können G P. Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level[J]. *International Journal of Climatology*, 2010, 26(26): 75—89.
- [35] 彭思岭. 气象要素时空插值方法研究[D]. 长沙: 中南大学, 2010.
Peng Siling. Developments of Spatio-temporal Interpolation Methods for Meteorological Elements[D]. Changsha: Zhongnan University, 2010. (in Chinese with English abstract)
- [36] 王连喜, 陈怀亮, 李琪, 等. 农业气候区划方法研究进展[J]. *中国农业气象*, 2010, 31(2): 277—281.
Wang Lianxi, Chen Huailiang, Li Qi, et al. Research progress on agricultural climatic division method in China[J]. *Chinese Journal of Agrometeorology*, 2010, 31(2): 227—281. (in Chinese with English abstract)
- [37] 王雪姣, 潘学标, 王森, 等. 基于 COSIM 模型的新疆棉花产量动态预报方法[J]. *农业工程学报*, 2017, 33(8): 160—165.
Wang Xuejiao, Pan Xuebiao, Wang Sen, et al. Dynamic prediction method for cotton yield based on COSIM model in Xinjiang[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2017, 33(8): 160—165. (in Chinese with English abstract)
- [38] 易雪, 王建林, 宋迎波, 等. 早稻产量动态集成预报方法研究[J]. *中国水稻科学*, 2011, 25(3): 307—313.
Yi Xue, Wang Jianlin, Song Yingpo, et al. Study on dynamic integrated prediction of early rice yield[J]. *Chinese Journal of Rice Science*, 2011, 25(3): 307—313. (in Chinese with English abstract)
- [39] 王建林, 宋迎波. 棉花产量动态预测方法研究[J]. *中国棉花*, 2002, 29(9): 5—8.
Wang Jianlin, Song Yingbo. Study on dynamic prediction method for cotton yield[J]. *China Cotton*, 2002, 29(9): 5—7. (in Chinese with English abstract)
- [40] 陈艳春, 李鸿怡, 赵红. 山东棉花引种的农业气候相似诊断分析[J]. *气象*, 1997, 23(3): 42—45.
Chen Yanchun, Li Hongyi, Zhao Hong. Agroclimatic similarity diagnostic analysis for introduction of cotton varieties in Shandong province[J]. *Meteorological Monthly*, 1997, 23(3): 42—45
- [41] Young K C. A multivariate chain model for simulating climatic parameters from daily data[J]. *Journal of Applied Meteorology*, 2010, 33(6): 661—671.
- [42] Rajagopalan B, Lall U. A k-nearest-neighbor simulator for daily precipitation and other weather variables[J]. *Water Resources Research*, 1999, 35(10): 3089—3101.
- [43] Buishand T A, Brandsma T. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling[J]. *Water Resources Research*, 2001, 37(11): 2761—2776.
- [44] Bannayan M, Hoogenboom G. Weather analogue: A tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach[J]. *Environmental Modelling and Software*, 2008, 23(6): 703—713.
- [45] Hoogenboom G, Jones J, Wilkens P, et al. Decision Support System for Agrotechnology Transfer (DSSAT) version 4.6.1.0[M] Washington: DSSAT Foundation, 2015.
- [46] Jones C A, Ritchie J T, Kiniry J R, et al. The CERES wheat and maize models[C]//Proceedings of the International Symposium on Minimum Data Sets for Agrotechnology Transfer, ICRISAT Center, India, 1983: 95—100.
- [47] Ritchie J R, Otter S. Description and performance of CERES-Wheat: A user-oriented wheat yield model[R]. USA: ARS-United States Department of Agriculture, 1985: 159—175.
- [48] 刘海波, 袁靖. CERES 玉米模拟模式评价及其应用[J]. *黑龙江气象*, 1997(1): 22—25.
- [49] Cover T, Hart P. Nearest neighbor pattern classification[J]. *IEEE Transactions on Information Theory*, 1967, 13(1): 21—27.
- [50] 王石立, 马玉平. 作物生长模拟模型在我国农业气象业务中的应用研究进展及思考[J]. *气象*, 2008, 34(6): 3—10.
Wang Shili, Ma Yuping. The progress in application of crop growth simulation models to agro-meteorological services in china[J]. *Meteorological Monthly*, 2008, 34(6): 3—10. (in Chinese with English abstract)
- [51] Lawless C, Semenov M A. Assessing lead-time for predicting wheat growth using a crop simulation model[J]. *Agricultural and Forest Meteorology*, 2005, 135(1): 302—313.
- [52] 于振文. 作物栽培学各论[M]. 北京: 中国农业出版社, 2003.
- [53] 刘铁梅, 谢国生. 农业系统分析与模拟[M]. 北京: 科学出版社, 2010.
- [54] 郭建平. 农业气象灾害监测预测技术研究进展[J]. *应用气象学报*, 2016, 27(5): 620—630.
Guo Jianping. Research progress on agricultural meteorological disaster monitoring and forecast[J]. *Journal of Applied Meteorological Science*, 2016, 27(5): 620—630. (in Chinese with English abstract)

Maize yield forecast with DSSAT-CERES-Maize model driven by historical meteorological data of analogue years by clustering algorithm

Chen Shang^{1,2}, Dou Zihé^{1,2}, Jiang Tengcong^{1,2}, Li Hualong^{1,2}, Ma Haijiao^{1,2},
Feng Hao^{2,3}, Yu Qiang³, He Jianqiang^{1,2,3*}

(1. Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, China; 2. Institute of Water-Saving Agriculture in Arid Areas of China, Northwest AandF University, Yangling 712100, China; 3. Institute of Water and Soil Conservation, Chinese Academy of Science and Ministry of Water Resource, Yangling 712100, China)

Abstract: Crop growth simulation models can simulate the processes of crop growth, development, yield formation, and its response to environment, which provides an effective method for crop yield forecast. However, how to select suitable weather data for the forecast is still an open question. In this study, we established a method for maize yield forecast based on maize growth simulation model of CERES-Maize and historical weather data from the year of 1956 to 2015. Two year's experimental data from 3 sites of Yangling (2014 and 2015), Heyang (2009 and 2011) and Changwu (2010 and 2011) in Shaanxi Province were used to test the reliable and accuracy of the method established. The weather data needed for model simulation were divided into 2 different groups including the known weather data and unknown weather data during the whole growth season of spring maize. The known weather data were obtained from local weather stations, while unknown data were supplemented with historical weather data of multiple years in the local experimental sites. Multiple complete climatic data series were then created and used to run the CERES-Maize model to forecast maize yield for a given year. As the advancing of maize growth season, the daily weather data were gradually merged into the observed weather data in a target year. Consequently, the daily maize yield was forecasted from sowing day to harvest. In addition, in order to reduce the times of model runs and reduce the uncertainties in yield forecasts, this study compared the daily meteorological data of historical and target years with normal K nearest neighbor (K -NN) and a modified K -NN algorithm to select several historical analogue years whose weather data were similar to the target year. The results showed that: 1) the model was suitable for the yield simulation since the absolute relative error was smaller than 15%; 2) the data distribution of predicted yields began to converge and the uncertainty decreased rapidly after the tasseling stage. For example, the predicted yield after 30, 60 and 90 days (the tasseling stage) of sowing was 3 531-14 461, 3 413-14 828 and 961-13 210 kg/hm², respectively. But, the yield was 49 33-10 826, 8 484-10 565 kg/hm², respectively after 100 and 130 days of sowing. The coefficient of variation had a sudden fall around the tasseling stage; 3) Yield forecast accuracy was generally lower than expectation for the method based on all historical data and climatic analogue years selected with historical data. The model run cost 61 min for a yield prediction during a complete growth stage of spring maize, indicting a necessary change in the prediction method optimization; 4) Among the 3 methods, the modified K -NN method showed a higher prediction accuracy and shorter run time than the other methods. The coefficient of variation was 11.7%-23.8% for the modified K -NN method, 15.1%-29.1% for the historical data, and 14.7%-26.9% for the K -NN method, respectively. To complete the yield prediction of a growth stage of spring maize, the modified K -NN method only took 14 min, which was shorter than the normal K -NN method. Thus, the modified K -NN method in this study had a big potential for the yield prediction by the CERES-Maize model. The study provides an effective method for selecting precipitation factor used for the yield prediction by crop models.

Keywords: clustering algorithms; weather forecasting; models; maize; yield forecasting; CERES-Maize; K -NN