基于气象因子的东北地区大豆产量年际变化预测：机器学习方法的应用

胡岳[1,2,3] 李芳[3,4] 苏秦[1,2] 林中达[3]

1 云南大学大气科学系，昆明 650500

2 云南省大湄公河次区域气象灾害与气候资源重点实验室，昆明 650500

3 地球系统数值模拟与应用全国重点实验室，中国科学院大气物理研究所，北京 100029

4 国际气候与环境科学中心，中国科学院大气物理研究所，北京 100029

摘要 大豆是全球最重要的油料作物和植物蛋白来源以及全球四大粮食作物之一。中国是排名世界第四的大豆生产国和最大消费国，而东北地区是中国大豆主产区，其大豆产量约占全国总产量的一半。东北大豆产量的年际变化主要受气象因素驱动，准确预测气象因素主导的产量年际变化对保障粮食安全和市场稳定具有重要意义。已有的统计预测研究多集中于东北地区局部区域、部分县、或单个省，且预测时段不超过 5 年，或仅评估了拟合效果。针对上述问题，本研究采用岭回归、Lasso 回归、支持向量机回归、K 近邻回归、决策树回归和随机森林方法构建了 1981~2018 年东北地区省级尺度气象因子预测大豆产量年际变化模型，并评估和比较其交叉验证技巧。主要结论如下：（1）在六种机器学习方法中，岭回归在三省整体表现最优，其交叉验证相关系数在黑龙江、吉林和辽宁分别达到 0.48（P＜0.01）、0.58（P＜0.001）和 0.72（P＜0.001）；（2）相较于逐步线性回归，岭回归在相关系数（R）和均方根误差（RMSE）上均表现更好，仅在吉林省和辽宁省的幅度预测准确性上略低。（3）因子选择与样本叠加处理在多数情况下能提升机器学习模型的交叉验证预测技巧。（4）气象因子对产量形成的关键作用窗口集中在 7 至 8 月的开花结荚期，此期间温度、降水和日照时长的正向协同效应显著，充足的水热条件有利于花荚形成、籽粒发育及光合作用增强，从而提升最终产量。研究结果为东北大豆产量预测和农业风险管理提供了科学依据。

关键词 大豆产量预测 机器学习 气象因子 中国东北 年际变化

doi:10.3878/j.issn.1006-9585.2025.25029

Meteorological factors-driven prediction of interannual variability of soybean yield in Northeast China based on machine learning methods

HU Yue[1,2,3] LI Fang[3,4] SU Qing[1,2] LIN Zhongda[3]

1 Department of Atmospheric Sciences, Yunnan University, Kunming 650500

2 Yunnan Key Laboratory of Meteorological Disasters and Climate Resources in the Great Mekong Subregion, Yunnan University, Kunming 650500

收稿日期 2025-02-21；**收修定稿日期** 2025-04-17

第一作者 胡岳，女，1998 年出生，硕士研究生，主要从事作物产量预测研究。E-mail: huyue_1998@163.com

通讯作者 李芳，E-mail: lifang@mail.iap.ac.cn

资助项目 广东省基础与应用基础研究重大项目 2021B0301030007、国家重点研发计划项目 2017YFA0604302、国家自然科学基金项目 42175074、云南省自然科学基金 202302AN360006、国家重大科技基础设施项目“地球系统数值模拟器”（EarthLab）

Funded by Guangdong Major Project of Basic and Applied Basic Research (2021B0301030007), National Natural Science Foundation of China (2017YFA0604302), National Natural Science Foundation of China (42175074), Natural Science Foundation of Yunnan Province (202302AN360006), Technological Infrastructure project “Earth System Science Numerical Simulator Facility”

3 National Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Being 100029
4 International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Being 100029

Abstract Soybean is one of the world's four major grain crops and the most important source of vegetable oil and protein. China, as the world's fourth-largest soybean producer and largest consumer, relies heavily on Northeast China for domestic production, which accounts for approximately half of the national output. The interannual variability of soybean yield in Northeast China is primarily driven by meteorological factors. Its accurate prediction is crucial for food security and market stability. Previous statistical prediction studies have been limited to local areas or single provinces and have only provided fitting skills or short-term (≤5 years) prediction skills. To address these limitations, this study developed prediction models for interannual soybean yield variations at the provincial scale in Northeast China during 1981–2018 using six machine learning methods (ridge regression, Lasso regression, support vector regression, K-nearest neighbors regression, decision tree regression, and random forest) based on meteorological factors, with their cross-validation performance evaluated and compared. The main findings are: (1) Ridge regression showed the best overall performance among the six methods, with cross-validation correlation coefficients reaching 0.48 ($P<0.01$), 0.58 ($P<0.001$), and 0.72 ($P<0.001$) in Heilongjiang, Jilin, and Liaoning provinces, respectively; (2) Compared to stepwise linear regression, ridge regression demonstrated superior performance in a correlation coefficient (R) and root mean square error (RMSE), with slightly lower accuracy only in amplitude prediction for Jilin and Liaoning provinces; (3) Predictor selection and sample augmentation generally improved the cross-validation prediction skills of machine learning models; (4) The critical meteorological impact window concentrated in the flowering and pod-setting period (July-August), during which the positive effects of temperature, precipitation, and sunshine duration significantly enhanced final yields through promoting pod formation, grain development, and photosynthesis. These findings provide scientific

support for soybean yield prediction and agricultural risk management in Northeast China.

Key Words Soybean yield prediction, Machine learning, Meteorological factors, Northeast China, Interannual variability

1 引言

大豆是全球四大粮食作物（水稻、小麦、玉米、大豆）之一，也是世界上最大宗的油料作物和最重要的植物蛋白来源（Hasegawa et al., 2022; FAO, 2023）。中国是全球排名第四的大豆生产国（年产量约 1500~2000 万吨）和最大的大豆消费国（年消费量约 1~1.2 亿吨），中国的大豆消费高度依赖进口的现状长期威胁我国粮食安全（江涛等, 2012）。在此背景下，国家层面将提升大豆自给率列为战略重点，而准确的大豆产量预测可为农业生产管理、市场调控及政策制定提供科学依据。

中国东北地区（黑龙江省、吉林省和辽宁省）作为全球三大黑土带之一（Wang et al., 2024），该区域土壤肥沃、分布集中连片，气候适宜，是全国最大的大豆产区，产量约占全国大豆产量的 50%（Xin et al., 2010; 中国国家统计局, 2024）。得益于品种改良、农艺优化和政策支持（Liu et al., 2008; He et al., 2020; Chandio et al., 2022; Guo et al., 2023），东北地区大豆产量长期呈稳定增长趋势（于贵瑞等, 2023），可通过外推法获得较为准确的预测趋势变化，但对天气气候变化引起的年际产量波动（即气象产量）的准确预测仍是当前的主要难点（Ray et al., 2015）。大量研究揭示了可能影响东北大豆产量的气象因子，如气温（包括均温、积温、最值、以及高于某一域值的温度）（Tao et al., 2008; 刘景利等, 2013; 崔明元, 2014; 吕金莹等, 2019; Zhang et al., 2022; 王德明, 2022; Guo et al., 2022）、降水量与相对湿度（于晓秋和郭玉, 2002; Tao et al., 2008; 刘景利等, 2013;王德明, 2022 ; Zhang et al., 2022; Guo et al., 2022）以及日照时长（闫平等, 2005; Tao et al., 2008; 刘景利等, 2013）等。

目前，东北大豆产量年际变化预测研究主要基于两类方法：传统统计方法和机器学习方法。于晓秋等（2007）和韩文革等（2009）基于 Logistic 时间序列模

型与气象因子的逐步线性回归方法，拟合了 1970~2006 年黑龙江省九三地区的大豆产量。赵放等（2024）基于正交多项式和积分回归方法，使用旬平均的气温、降水与日照时长拟合了 1980~2021 年东北地区大豆的趋势产量并评估了产量受各气象要素的影响。王贺然等（2018）以 5 日平均的气温、降水和日照时长为预测因子，使用多元线性回归方法预测了辽宁省 2014~2016 年的大豆产量。王贺然等（2018）、邱美娟等（2018）和陈雪等（2023）使用基于气候适宜度指数（温度、降水和日照时长的非线性函数）的一元线性回归模型分别估计了辽宁省 2014~2016 年、吉林 2015 和 2016 年，以及黑龙江省 2017~2019 年的大豆产量。然而，以上传统统计方法多关注历史数据的拟合优度，或只提供 2~3 年的预测技巧，更长验证集的东北大豆产量预测方案有待进一步的讨论和研究。

在国际大豆主产区，气象因子驱动的大豆产量预测研究同样取得了显著进展。如 Petersen（2019）通过 1970~2015 年美国县级数据的线性回归分析证明极端高温显著降低大豆产量，并使用该回归模型预测未来情景下的产量变化。Luan et al.（2021）和 Luan et al.（2023）基于美国雨养农业区 1970~2010 年县级气象与产量数据，采用含交互项的多项式回归模型，分别揭示了大豆产量与降水、温度，以及降水、蒸发之间的非线性协同效应。然而，这些研究仍然依赖全时段数据拟合，缺乏交叉验证以评估模型的预测能力。与拟合研究不同，Monteiro et al.（2022）通过机器学习中的随机森林和支持向量机算法，结合留一法交叉验证，实现了收获前 30 天的巴西大豆产量预测（相对误差 9.2%~41.5%），且证明了随机森林和支持向量机模型对比传统多元线性回归的优越性。

近年来，中国东北地区也开始探索机器学习方法在大豆产量预测中的应用。经典的机器学习算法中，线性模型（如岭回归和 Lasso 回归）与非线性模型（如支持向量机回归、随机森林和 K 近邻回归）各具特色。线性模型通过正则化技术有效处理自变量间的多重共线性问题，提供更稳定的参数估计，而非线性模型则善于捕捉作物产量与气象因素间的复杂的非线性关系。支持向量机回归利用核函数处理变量间的交互效应，随机森林能识别关键气象驱动因子，K 近邻回归则适合处理具有空间相关性的预测问题。Li et al.（2023）基于生长期温度、降水和日照时长，集成上述 5 种算法构建模型预测 2009~2013 年中国东北部分县的大豆产量，发现随机森林的解释方差比起线性模型有明显提升。此外，Lu et al.（2024）

使用随机森林、卷积神经网络、极致梯度提升以及 CNN-LSTM-Attention 模型，对 2020 年东北三省的大豆产量进行了预测研究。Song et al.（2024）使用支持向量机回归、随机森林、遗传算法优化反向传播神经网络和卷积神经网络预测大豆产量，但仅对吉林省进行了预测。尽管这些研究展示了机器学习在东北地区的应用潜力，但其空间尺度和验证周期仍普遍存在不足。

现有东北大豆产量年际变化预测研究存在两个主要局限：一是研究的空间尺度局限于县域、单个省份或其局部区域；二是模型评估多局限于拟合效果或短期（不超过 5 年）预测技巧，缺乏长期预测效果的系统评估。综上所述，目前尚未有覆盖东北三省省级尺度的大豆产量预测及其长期预测技巧评估研究。

本研究以气温、降水、日照时长和相对湿度等气象要素为预测因子，采用岭回归、Lasso 回归、支持向量机回归、K 近邻回归、决策树回归和随机森林六种机器学习方法，构建了 1981~2018 年东北地区省级大豆产量年际变化预测模型。我们采用“留一法”交叉验证框架，该框架可在小样本情形下最大化训练数据的规模，有助于提升模型的稳定性和准确性，同时为评估模型的长期预测技巧提供充分样本。

论文结构如下：第 2 节详细介绍研究区域、数据预处理、预测方法及评估指标；第 3 节分析六种机器学习方法在 1981-2018 年的交叉验证预测技巧；第 4 节讨论机器学习中降维和扩样的重要性，比较岭回归与传统逐步线性回归的预测效果，并探讨气象因子影响大豆产量的潜在生理机制；第 5 节总结研究结论并展望未来可能的研究方向。

2 数据与方法

2.1 研究区域

中国东北地区（118°–135°E，38°–53°N）涵盖黑龙江、吉林和辽宁三省，总面积约 80 万平方公里，属温带季风气候区。区域内降水量呈东南（>800 mm）向西北（400~500 mm）递减，年日照 2400~2900 小时，夏季降水占全年 60%~70%，水热同步特性显著（路亚洲等, 2012; 吴佳和高学杰, 2013），为大豆开花和结荚鼓粒期提供了适宜的气候条件。作为全球三大黑土带之一，其土壤以腐殖质丰富、

保水性强、肥力高著称（FAO, 2015）。平坦的地形及三江、松嫩、辽河三大平原的连片耕地共同构成了我国机械化程度最高的大豆主产区（中华人民共和国农业农村部, 2021）。

1980 年至 2018 年间，黑龙江、吉林和辽宁三省的大豆种植面积分别为 280 万公顷、40 万公顷和 30 万公顷，分别占全国大豆种植总面积的 33.9%、4.9%和 3.1%（中国国家统计局, 2024）。大豆种植主要集中在黑龙江西部地区及吉林、辽宁中部（图 1）。在此期间，三省的年均大豆产量分别为 460 万吨、80 万吨和 40 万吨，单产分别为 1650 千克/公顷、2061 千克/公顷和 1810 千克/公顷。三省的大豆产量分别占全国总产的 35.0%、6.2%和 3.0%。

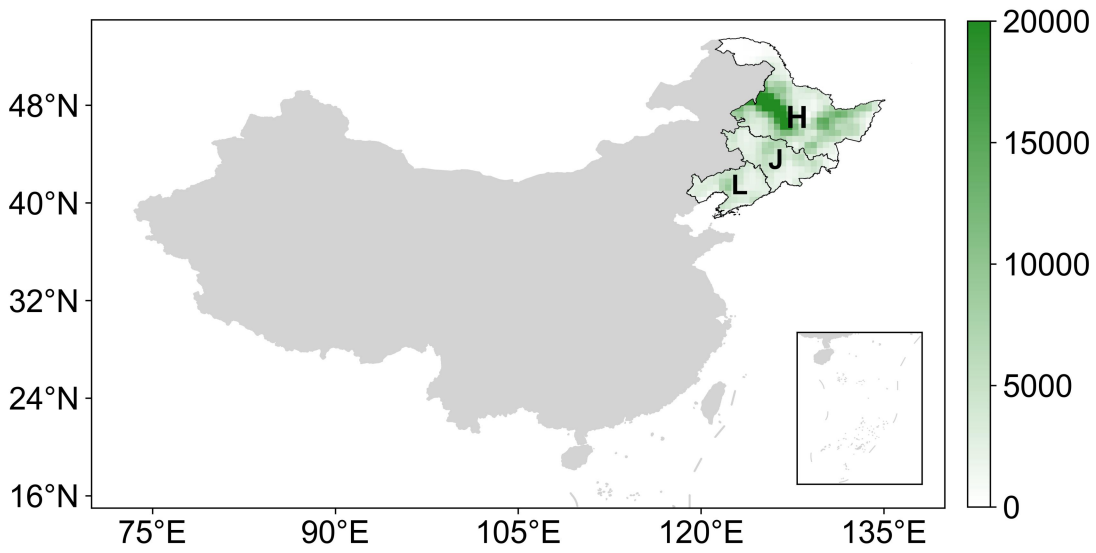

图 1　东北地区大豆播种面积的空间分布（单位：公顷），其中 H、J、L 分别代表黑龙江省、吉林省和辽宁省（资料来源：MIRCA2000 数据集, Portmann et al., 2010）。

Figure 1　Spatial distribution of soybean planting area in Northeast China (Unit: hectares (ha)). H, J, and L denote Heilongjiang, Jilin, and Liaoning Provinces, respectively. This spatial distribution data is derived from the MIRCA2000 dataset (Portmann et al., 2010).

2.2 数据与预处理

2.2.1 数据

本研究中使用的数据包括农业观测数据和气象数据。

在农业观测数据方面，1980~2018 年黑龙江、吉林和辽宁三省的省级大豆单产数据来源于国家统计局发布的《中国统计年鉴》以及中国农业部出版的《中国农业年鉴》（中国农业年鉴编辑委员会, 2021; 中国国家统计局, 2024）。大豆的播种面积数据采用 MIRCA2000 数据集提供的 2000 年前后作物播种面积空间分布数据（Portmann et al., 2010）。

在气象数据方面，我们采用 CN05.1 数据集（吴佳和高学杰, 2013）。其空间分辨率为 0.25°×0.25°，基于中国境内 2400 多个气象站的观测资料插值生成。我们从中提取了 1980~2018 年的逐日气象数据，包括降水量（Pr）、相对湿度（RH）、日照时长（SD）、平均气温（Tavg）、最低气温（Tmin）和最高气温（Tmax）六个变量。

2.2.2 预处理

气象数据预处理分为以下三步：

生长季界定：黑龙江省播种期设定为 5 月 15 日，吉林省和辽宁省为 5 月 5 日，三省统一收获期为 9 月 24 日（刘景利等, 2007; Gong et al., 2021）。我们仅采用生长季（播种至收获）内的气象数据用于分析。

滑动平均计算：以往研究表明，在多个长度的滑动平均窗口中，5 日滑动平均对黑龙江省（产量最高省）的预测效果最好（胡岳, 2023），因此我们对日值气象数据计算 5 日滑动平均值。

省级空间平均：基于作物播种空间分布数据集 MIRCA2000，以格点内大豆播种面积比例≥1%的区域为大豆播种区，在播种区内计算各气象因子的省级空间平均值。

由此，我们得到了对应年份的气象输入因子：每年生长季内、省级尺度大豆播种区内经 5 日滑动平均所获得的每个气象要素 129 个（吉林、辽宁为 139 个）值，六个气象要素共计 774 个（吉林、辽宁为 834 个）值。

2.3 方法

2.3.1 年际增量法

本研究采用范可等（2007）提出的年际增量法提取大豆产量的年际变化，计算大豆的年际增量（年际变化）为当前年份与前一年的产量之差：

$$\Delta Yield(t) = Yield(t) - Yield(t-1) \quad (1)$$

其中 $Yield(t)$ 为第 t 年的大豆产量，$\Delta Yield(t)$ 为第 t 年的大豆产量年际增量。

同样的方法也用于计算上文经预处理后的 5 日滑动平均气象输入因子的年际变化。这种去趋势的方法能够放大年际波动，有效地过滤长期趋势和年代际变化，有助于捕捉预测信号，已成功应用气候和天气预测（例如，Fan and Wang, 2009; Fan et al, 2012; 范可和田宝强, 2013; 范可等, 2016)。

2.3.2 “留一法”交叉验证

本研究采用“留一法”交叉验证的思想构建和评估预测模型（Stone, 1974; 李芳, 2008; Yates et al., 2022）。具体而言，对于时间序列中的每个观测年份 i（i=1, 2, ..., N），我们始终将目标年份 i 作为待预报的测试集，而使用其余 N−1 个年份的完整数据建立预测模型，以预测目标年份。由于机器学习需要进行参数的选择和调整，故在原方法的框架下进行进一步修改，将其余 N−1 个年份的样本进一步随机划分为 80%的训练集和 20%的验证集，根据验证集上观测与预测间最大相关系数为标准选择最优参数组合。在每次迭代中，使用训练集的数据以及验证集得到的最优参数组合建模以预报测试集（目标年份产量年际变化）。通过在整个样本长度上迭代执行这一过程，最终获得所有 N 个年份的独立预测结果。模型预测性能通过系统比较全部年份的预测值与实际观测值来量化，该评估指标被定义为交叉验证技巧（cross-validation skill）。

“留一法”交叉验证的思想适用于小样本时间序列分析，具有以下优势：首先，每次预测均将 N−1 个样本用于模型的训练与调参，以最大限度地利用有限的数据集；其次，严格保持训练集、验证集与测试集三者之间的独立性，确保验证过程的客观性，避免过拟合；此外，N 个样本用于评估预测技巧，为评估模型的长期预测技巧提供充分样本。

2.3.3 评价指标

本研究采用皮尔逊相关系数（R）、标准偏差比值（std ratio）和均方根误差（RMSE）三项指标系统评估模型的交叉验证技巧。

我们采用了皮尔逊相关系数（R）评估预测值的时间变化模态，该指标为预测值与观测值之间的相关系数，其公式为：

$$R = \frac{\sum_{i=1}^{N}\left(y_{obs,i} - \overline{y_{obs}}\right)\left(y_{pred,i} - \overline{y_{pred}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{obs,i} - \overline{y_{obs}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{pred,i} - \overline{y_{pred}}\right)^2}} \quad (2)$$

式中，N 为时间序列长度，i（i=1, 2, ..., N）为时间序列中的每个样本年份。$y_{obs,i}$ 和 $y_{pred,i}$ 分别为第 i 年的观测值和预测值，$\overline{y_{obs}}$ 和 $\overline{y_{pred}}$ 为其均值。我们采用学生 t 检验（Student's t-test）来评估相关系数的统计学意义，当观测值和预测值之间的相关系数在 0.05 水平上显著（即优于 95%置信度下的随机预测）时，认为模型能有效捕捉年际变化的模态特征。

我们采用了预测与观测的标准偏差比值（std_{pred}/std_{obs}，后文记为 r_{std}）评估预测值的时间变化幅度，其公式为：

$$r_{std} = std_{pred}/std_{obs} = \sum_{i=1}^{N}\left(y_{pred,i} - \overline{y_{pred}}\right)\Big/\sum_{i=1}^{N}\left(y_{obs,i} - \overline{y_{obs}}\right) \quad (3)$$

公式中的变量意义同式（2）。当 r_{std} 大于 1 时，反映预测高估了幅度，小于 1 则反映低估。当 r_{std} 趋近于 1 时，表明预测模型能准确复现观测序列年际变化的幅度。

我们采用均方根误差（RMSE）对预测值与观测值的一致性进行综合评估：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{pred,i} - y_{obs,i}\right)^2} \quad (4)$$

公式中的变量意义同式（2）。该指标通过整合预测值与观测值在模态匹配度和幅度一致性的偏差，量化模型对预测值的整体复现能力。RMSE 越小，表明模型的预测能力越强。

2.3.4 机器学习方法

本研究使用了六种经典的机器学习方法，包括岭回归、Lasso 回归、支持向

量机回归、K 近邻回归、决策树回归与随机森林。其中，岭回归、Lasso 回归是在线性回归基础上进行改进的正则化方法，支持向量机能够自由选择核函数类型以达到线性或非线性的回归效果，K 近邻回归、决策树和随机森林则能够识别数据中的非线性关系。以下对这六种方法进行分别介绍，并在最后介绍算法应用的过程中对数据进行的因子选择和样本叠加处理。

表 1 机器学习的关键参数及物理意义。

Table 1 Key parameters and physical interpretations of machine learning models.

方法	参数名称	参数代号	取值或范围
岭回归	L_2 正则化项	λ	10^{-10}~10^{10}
Lasso 回归	L_1 正则化项	λ	10^{-10}~10^{10}
支持向量机回归	核函数	kernel	“linear”、“poly”、“rbf”
	惩罚系数	C	10^{-4}~10^{4}
	间隔带宽度	ε	“scale”、“auto”
	多项式核的次方	degree	2、3（只在多项式核函数中）
K 近邻回归	最近邻居个数	k	1~10
	距离度量	distance_metrics	“euclidean”, “minkowski”, “manhattan”, “chebyshev”
	权重函数	weights	“uniform”, “distance”
决策树回归	最大深度	max_depth	1~30
	最小样本分割数	min_samples_split	2~10
	最小样本叶数	min_samples_leaf	1~10
随机森林	树的数量	n_estimators	10, 20, 50, 100, 200
	最大深度	max_depth	None, 10, 20, 30
	最小样本分割数	min_samples_split	2, 5, 10
	最小样本叶数	min_samples_leaf	1, 2, 5, 10

2.3.4.1 **岭回归**

岭回归是针对多重共线性问题的线性回归改进方法，其通过在目标函数中引入 L_2 正则化项（系数平方和惩罚项）提升模型稳定性。目标函数定义为：

$$\min_{\beta_0,\beta_j}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_0 x_{ij}^2\right)+\lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad (5)$$

其中，y_i 表示第 i 个样本的大豆产量观测值（i=1, 2, ..., n），β_0 为模型截距项，β_j 为第 j 个气象因子的回归系数（j=1, 2, ..., p），x_{ij} 表示第 i 个样本的第 j 个气象因子值，n 为样本量，p 为气象因子数。公式（5）中的第一项为均方误差项，

第二项为 L_2 正则化项，λ为正则化强度参数，λ≥0（表 1）。本文中λ的取值通过交叉验证在 10^{-10}~10^{10} 范围内以 $10^{0.1}$ 为间隔对数搜索确定，以平衡欠拟合（λ过大导致系数过度压缩）与过拟合（λ过小削弱正则化效果）风险。该方法的核心优势在于对全部系数进行均衡压缩而非变量剔除（Hoerl and Kennard, 1970），既能抑制多重共线性干扰，又可保留所有气象因子的信息贡献（如不同生育期温湿度均可能对作物产量存在潜在影响），避免关键农艺特征丢失。

2.3.4.2 Lasso 回归

Lasso（Least Absolute Shrinkage and Selection Operator）回归通过引入 L_1 正则化项（系数绝对值之和惩罚项）实现特征选择，其目标函数为：

$$\min_{\beta_0,\beta_j}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_0 x_{ij}^2\right)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (6)$$

其中变量的定义与公式（5）中的变量定义一致。公式（6）中的第一项为均方误差项，第二项为 L_1 正则化项，其正则化强度参数λ（λ≥0）通过交叉验证确定（表 1，参数取值范围及间隔与岭回归一致）以控制模型稀疏性强度：λ增大时更多系数被压缩至零生成稀疏模型，λ减小时则趋近普通线性回归。相较于岭回归的均衡压缩，Lasso 回归通过几何约束将不重要变量的系数精确压缩至零（Tibshirani, 1996），特别适用于处理高维气象数据（如滑动平均处理的气象变量）中的冗余特征筛选，在降低模型复杂度的同时增强可解释性。

2.3.4.3 支持向量机回归

支持向量机回归基于结构风险最小化原则，通过核函数将数据映射到高维空间以捕捉非线性关系。其目标函数为：

$$\min\left\{\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^{n}\left(\xi_i+\xi_i^*\right)\right\} \quad (7)$$

其中第一项控制模型复杂度，第二项为 ε-不敏感带外的误差惩罚项（宽度 2ε 的间隔带内不计误差，记 ξ_i、ξ_i^* 为超出上、下界的误差）。支持向量机算法通过引入 ε-不敏感带，仅对落在间隔带外的样本施加惩罚，增强了模型的稳定性。核

函数类型（线性“linear”、多项式“poly”或径向基“rbf”）决定特征空间的映射方式，多项式核通过调整阶数（degree）可有效解析气象变量与产量的高阶交互效应。惩罚系数 C 控制模型对训练误差的敏感度：C 值较大时倾向于精确拟合但易过拟合，C 值较小时则增强泛化性但可能忽略关键特征。本研究采用交叉验证优化核函数类型、C 及 ε 的组合（具体参数范围见表 1），通过边界支持向量构建最优超平面，在抑制观测噪声干扰的同时保留气候驱动的核心信号，尤其适用于处理温度、降水等气象因子的非线性累积效应。

2.3.4.4 K 近邻回归

K 近邻回归是一种基于局部相似性的非参数方法，通过搜索训练集中与目标样本最邻近的 K 个数据点，以均值或距离加权方式预测输出值，其核心公式为：

$$\hat{y} = \frac{1}{\mathrm{K}} \sum_{i \in N_{\mathrm{K}}(x)} y_i \tag{8}$$

其中，$N_{\mathrm{K}}(x)$表示目标样本的 K 个最近邻。关键参数包括最近邻居个数（K）、距离度量（distance_metrics，取值为欧氏距离“euclidean”、闵可夫斯基距离“minkowski”、曼哈顿距离“manhattan”或切比雪夫距离“chebyshev”）和权重函数（weights，设为均值“uniform”或反比加权“distance”），这些参数的最优取值组合通过交叉验证确定（表 1）。最近邻居个数 K 的选择对模型性能影响显著：K 过大易平滑局部特征（欠拟合），K 过小则对噪声敏感（过拟合）。距离度量和权重函数共同定义“相似性”计算规则，进一步影响模型的预测精度。相较于参数化模型，K 近邻回归无需显式建模变量间关系，而是根据数据本身的特点来进行预测，这适用于小样本和局部特征显著的气象建模问题，但对数据规模和噪声敏感（Fix and Hodges, 1951）。

2.3.4.5 决策树回归

决策树回归通过递归特征分割构建树状结构，将特征空间划分为多个子区域（叶节点），并以区域内样本均值作为预测值。其目标函数最小化子区域内的样本方差：

$$\min\left\{\sum_{i\in R_L}\left(y_i-\overline{y_L}\right)^2+\sum_{j\in R_R}\left(y_j-\overline{y_R}\right)^2\right\} \quad (9)$$

其中 R_L 和 R_R 为分割后的左右子区域，$\overline{y_L}$ 和 $\overline{y_R}$ 为区域均值。关键参数包括树的最大深度（max_depth）、最小样本分割数（min_samples_split）和最小样本叶数（min_samples_leaf）（表 1），并通过交叉验证进行参数组合的优化。相较于传统线性模型，决策树无需数据标准化且具有直观解释性，适用于解析气象因子（如温度、降水阈值）与产量之间的分段非线性关系，但需谨慎控制分割深度以防止过度细分所带来的过拟合（Breiman et al., 1984）。

2.3.4.6 随机森林

随机森林回归通过集成多棵决策树的预测结果（取均值）提升模型稳定性，其随机性体现在两方面：一是每棵树基于训练集的自助采样（bootstrap）构建，二是在特征分裂时仅从随机子集中选择局部最优特征而非全局最优。这种双重随机机制有效降低了模型方差（Breiman, 2001）。随机森林的关键参数包括树的数量（n_estimators）、最大深度（max_depth）、最小样本分割数（min_samples_split）和最小样本叶数（min_samples_leaf），这些参数共同影响模型的性能和泛化能力。树的数量决定了集成的规模，较多的树通常能提高模型稳定性，但会增加计算成本；最大深度控制单棵树的复杂度，过深可能导致过拟合，过浅则无法捕捉关键特征，导致欠拟合；最小样本分割数和最小样本叶数则用于限制节点的分裂条件，防止模型对训练数据的过度拟合。相较于单棵决策树，随机森林通过集成策略显著提升模型稳定性，适用于高维气象数据（如多生育期气候变量）与产量的非线性关系建模（Breiman, 2001）。

2.3.4.7 因子选择与样本叠加

为了更精确地找出显著影响产量的气象因子时段，在 2.2.2 节中我们使用 5 日滑动平均的方法来预处理气象因子数据，这会得到远多于样本数量的预测因子，且预测因子互相之间存在一定的共线性。为了提高预测的准确性，我们需要在训练集上进行因子选择（即仅保留与产量存在显著相关性的气象因子，显著性水平设为 $p<0.05$），以适当减少因子数量；并进行三省样本的叠加，通过这种方

式来相对增加样本数量，进一步提高获得更高预测技巧的可能性。

在最终的模型中，我们合并三个省份（每省 38 年）获得一个 114 样本的集合。在“留一法”交叉验证框架下，对每个样本（测试集=1），使用剩下 113 个样本所建模型对其进行预测，以充分利用数据集。将这 113 个样本再按 8:2 随机划分，即取其中 90 个样本作为训练集，23 个样本作为验证集，使用不同的关键参数组合，以训练集的样本建模来预测验证集数据，选定验证集的预测值与实际值相关系数最高的关键参数组合。最后，使用训练集数据及对应最优参数组合建模，对测试样本进行预测。重复以上操作，获得每个样本的预测值，按照省份评估预测技巧。

3 预测结果

表 2　六种机器学习方法预测 1981~2018 年东北大豆产量年际变化的交叉验证技巧。其中*，，***分别代表观测和预测间相关系数在 0.05，0.01，0.001 水平上显著。**

Table 2　Cross-validation skills of six machine learning methods in predicting interannual variability of soybean yield in Northeast China from 1981 to 2018 (*: P<0.05, **: P<0.01, ***, P<0.001).

省份	机器学习方法	R	r_{std}	RMSE
黑龙江	岭回归	0.48**	0.9	251
	Lasso 回归	0.48**	0.94	254
	支持向量机回归	0.32	0.79	274
	K 近邻回归	−0.06	0.63	312
	决策树回归	0.57***	1.17	265
	随机森林	0.21	0.65	276
吉林	岭回归	0.58***	0.74	339
	Lasso 回归	0.46**	0.62	369
	支持向量机回归	0.51***	0.7	362
	K 近邻回归	0.26	0.3	397
	决策树回归	0.38*	0.6	392
	随机森林	0.63***	0.35	339
辽宁	岭回归	0.72***	0.79	264
	Lasso 回归	0.74***	0.72	256
	支持向量机回归	0.7***	0.71	267
	K 近邻回归	0.46**	0.84	369
	决策树回归	0.24	0.88	444
	随机森林	0.65***	0.43	298

岭回归和 Lasso 回归在三省都能有技巧地预测大豆产量的年际变化（表 2）。岭回归是三省整体最优的模型，在黑龙江省的交叉验证相关系数达 0.48，通过 0.01 的显著性检验，而在吉林和辽宁，该相关系数可分别达 0.58 和 0.72，均通过 0.001 的显著性检验（图 2）。Lasso 回归在黑龙江省、吉林省的交叉验证相关系数分别为 0.48 和 0.46（$P<0.01$）；而在辽宁相关系数达 0.74（$P<0.001$）、RMSE 降至 256 $kg \cdot ha^{-1}$，二者均超过岭回归成为辽宁省预测效果最好的模型，尽管幅度预测略低于岭回归。作为正则化方法，岭回归和 Lasso 回归是唯二在三省都取得显著预测技巧的模型，其中岭回归在吉林省和辽宁省都取得了显著的技巧，而 Lasso 回归在辽宁省取得最佳预测效果，展现了正则化方法的优越性。这表明，仅使用简单的正则化对线性回归进行系数上的约束和收缩即可获得良好的交叉验证预报技巧。

随机森林和支持向量机回归在吉林省和辽宁省大豆产量年际变化预测中表现出有效性，但其在黑龙江省的预测精度显著降低（表 2）。具体而言，随机森林在吉林省和辽宁省的交叉验证相关系数分别为 0.63 和 0.65（$P<0.001$），且为吉林省模态捕捉效果最好的模型，RMSE 也低至 339 $kg \cdot ha^{-1}$，与岭回归在该省的 RMSE 持平。然而，随机森林在吉林和辽宁的预测幅度明显偏小，r_{std} 仅分别为 0.35 和 0.43，显著低于理论理想值 1，表明其对产量时间序列的变幅存在明显的系统性低估，这可能源于随机森林集成过程中过多决策树的均值化效应导致的局部特征平滑化。支持向量机回归在吉林省和辽宁省的预测表现同样显著（交叉验证相关系数分别为 0.51 和 0.70，$P<0.001$），但其在黑龙江省的相关系数未能达到统计显著（$R=0.32, P\geq0.05$）。值得注意的是，模型参数训练与组合优化的结果显示，全部 114 个留一法交叉验证子模型均选择线性核函数，暗示东北三省大豆产量与气象因子的关联模式仍以线性响应为主导。这一现象或可解释为何引入正则化约束的线性回归方法（岭回归与 Lasso 回归）在整体预测精度和稳健性上优于其他致力于挖掘非线性关系的机器学习模型。

决策树回归能够有技巧地预测黑龙江省和吉林省的产量年际变化，而在辽宁省缺乏技巧（表 2）。具体而言，该模型在黑龙江省的交叉验证相关系数达 0.57（$P<0.001$），为三省最优表现，然而其 RMSE（265 $kg \cdot ha^{-1}$）高于岭回归（251 $kg \cdot ha^{-1}$），表明其对产量序列的整体捕捉能力仍存在局限性。值得注意的是，黑

龙江省的决策树回归算法也是唯一预测标准偏差高于观测标准偏差的案例，预测幅度高于观测幅度 17%，而其他所有省份、所有算法均表现出对观测幅度的系统性低估，包括整体最优的岭回归和在三省都显著的 Lasso 回归算法，其预测幅度分别低于观测幅度 10~26%、6~38%。这种现象可能与算法特性有关：决策树基于节点分割的算法机制可能保留更多局部变异特征，而岭回归和 Lasso 回归通过 L_2 或 L_1 正则化约束对回归系数进行全局压缩或稀疏化处理，在抑制模型对极端值和噪声敏感的同时，也倾向于系统性削弱模型对变幅的表征能力。

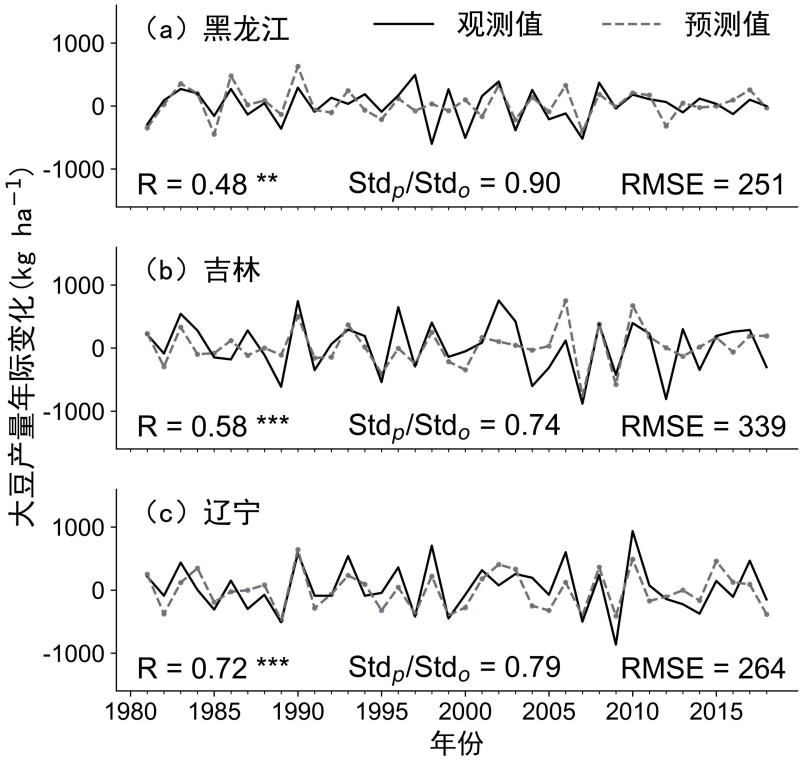

图 2 使用岭回归预测得到的 1981~2018 年东北三省大豆产量年际变化时间序列。其中*, **, ***分别代表观测和预测间相关系数在 0.05，0.01，0.001 水平上显著。

Figure 2. Time series of interannual variability of soybean yield predicted by ridge regression in Northeast China from 1981 to 2018 (*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$).

岭回归模型成功预测到东北三省大豆产量的多个典型的极端年份：如吉林省 1990 年、2008 年的高产和 2007 年、2009 年的低产，辽宁省 1990 年、2008 年的高产和 1989 年、2007 年的低产。但该模型仍存在一定局限性：首先，预测的变化幅度系统性偏低，黑龙江、吉林和辽宁分别低估了 10%、26%和 21%（图 2）；其次，部分极端值未能准确捕捉，如未能反映黑龙江 1997 年的高产与 1998 年的

低产，吉林 2002 年、2003 年的高产与 1989 年、2004 年、2012 年和 2018 年的低产。在辽宁省，虽然预测出 1998 年、2006 年、2010 年的高产及 2009 年的低产，但预测幅度显著偏低。这些局限可能源于两个方面：一是部分极端产量年份可能受到非气象因素的显著影响，政策调整和农艺管理等人为因素可能是重要影响因素。二是岭回归算法本身的特性，即通过压缩回归系数来实现模型正则化，这可能导致预测值整体变化幅度偏低。

4 讨论

4.1 气象因子最优滑动平均窗口的选择

表 3　不同时间尺度窗口下岭回归模型的交叉验证预测效果比较。其中*，，***分别代表观测和预测间相关系数在 0.05，0.01，0.001 水平上显著。**

Table 3　Comparison of cross-validated prediction performance of ridge regression models with different temporal windows (*: P<0.05, **: P<0.01, ***, P<0.001).

省份	窗口	R	r_{std}	RMSE
黑龙江	5 日滑动平均	0.48**	0.9	251
	7 日滑动平均	0.29	0.78	277
	10 日滑动平均	0.14	0.64	286
	月平均	−0.04	0.55	299
	生长季平均	0.06	0.26	263
吉林	5 日滑动平均	0.58***	0.74	339
	7 日滑动平均	0.55***	0.54	344
	10 日滑动平均	0.49**	0.69	368
	月平均	0.37*	0.34	382
	生长季平均	−0.04	0.13	416
辽宁	5 日滑动平均	0.72***	0.79	264
	7 日滑动平均	0.63***	0.9	314
	10 日滑动平均	0.62***	1.15	357
	月平均	0.58***	0.41	312
	生长季平均	0.59***	0.4	313

在 2.2.2 气象数据的预处理中，我们采用 5 日滑动平均窗口处理气象因子，然而，该窗口的不同选择可能影响预测效果。因此，我们对岭回归模型对比测试了其他时间尺度窗口：7 日滑动平均、10 日的滑动平均、月平均以及整个生长季的平均。表 3 展示，对于所有省份而言，5 日滑动平均窗口都对应预测结果的最高相关系数（R）和最低均方根误差（RMSE），且标准偏差比值（r_{std}）较合理，表明对于基于岭回归的大豆产量预测模型而言，5 日滑动平均是气象因子预处理的最优选择。

4.2 因子选择与样本叠加的必要性

在 2.3.4.7 中，我们对数据集进行了两步处理：其一，在训练集上进行因子选择（即仅筛选与产量之间相关系数达 0.05 显著性水平的气象因子），以减少特征数量、降低维度；其二，叠加三省样本，以增加样本量。选择显著因子能够有效降低模型复杂度，减少冗余信息的干扰。叠加三省样本不仅增加了样本量，还引入更多空间异质性数据，增强了模型对不同区域气候条件的适应能力。此外，我们还尝试了主成分分析（Principal Component Analysis, PCA）作为另一种降低维度的方式，即通过线性变换将原始气象因子转换为互不相关的成分，并保留累计贡献率≥90%的主成分。

图 3~5 分别比较了黑龙江省、吉林省和辽宁省六种处理方法（原始不处理、进行 PCA 处理、进行显著因子选择、进行样本叠加、进行 PCA 与样本叠加、进行显著因子选择与样本叠加）的相关系数（R）、幅度（r_{std}）与均方根误差（RMSE）三项交叉验证技巧。对比未作任何处理的原始结果而言，当同时选择显著因子和叠加样本时，三省的岭回归、Lasso 回归和随机森林三种机器学习对相关系数（R）的交叉验证预测技巧均有提升；支持向量机在黑龙江基本持平，在吉林和辽宁提升。六种机器学习方法中，仅黑龙江和吉林的 K 近邻回归和辽宁的决策树回归在处理后比起处理前对相关系数（R）的预测技巧明显降低。对幅度（r_{std}）的预测技巧可能降低也可能提升，例如，算法为岭回归时，因子选择与样本叠加的组合处理使得辽宁省的幅度预测略微变差，但对黑龙江省和吉林省而言效果变好。该处理使得对 RMSE 的预测普遍降低或基本持平，仅在辽宁省使用决策树算法

时明显升高。在本研究中，使用主成分分析（PCA）进行降维的效果不佳。

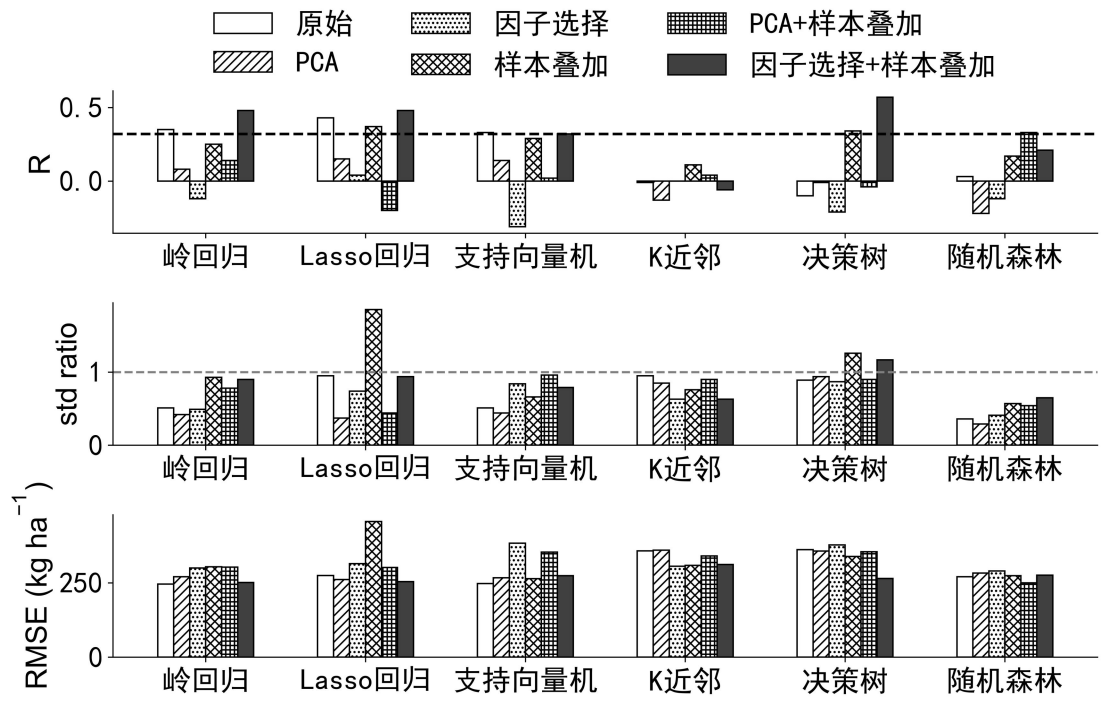

图 3　黑龙江省不同处理方法所得到的交叉验证预测技巧，从上往下分别为预测值和观测值之间的皮尔逊相关系数（R）、标准偏差比值（std ratio）、均方根误差（RMSE）。黑色虚线为相关系数的 0.05 显著性水平线，灰色虚线为 std=1 线。

Figure 3　Cross-validation prediction skills for different methods for Heilongjiang Province. From top to bottom are the correlation coefficient (R), standard deviation ratio (std ratio), and root mean square error (RMSE) between predicted and observed values. The black dashed line indicates the 0.05 significance level for correlation coefficients, and the gray dashed line represents the reference line for std ratio = 1.

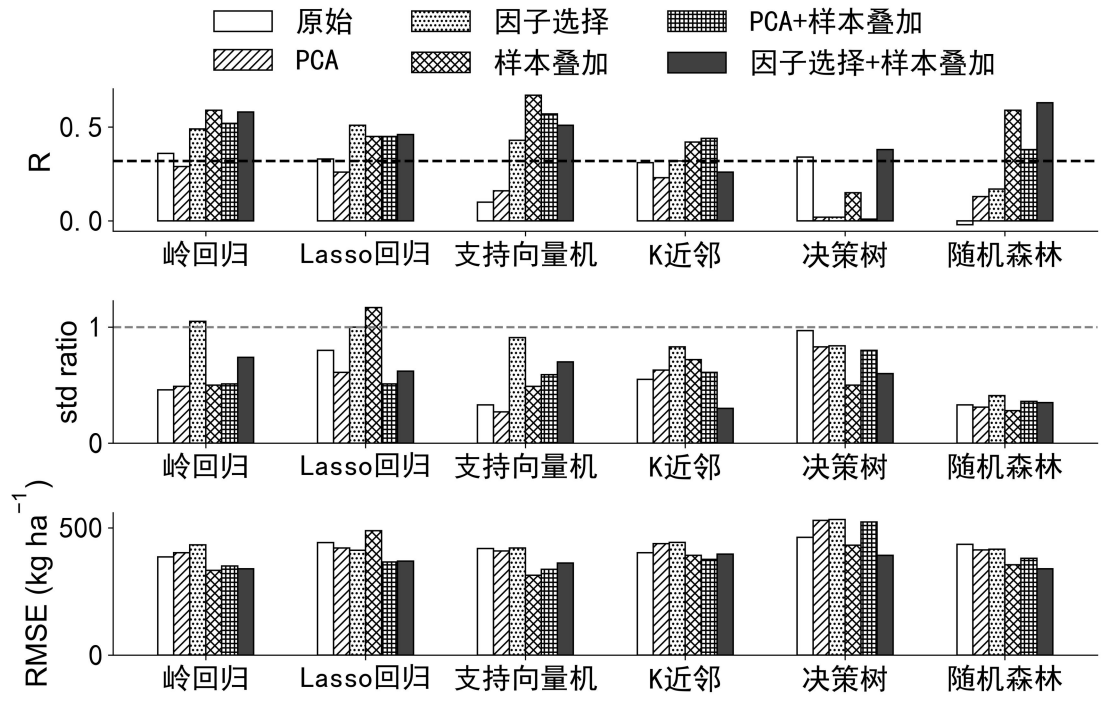

图 4　同图 3，但省份为吉林省。

Figure 4　Same as Figure 3 but for Jilin Province.

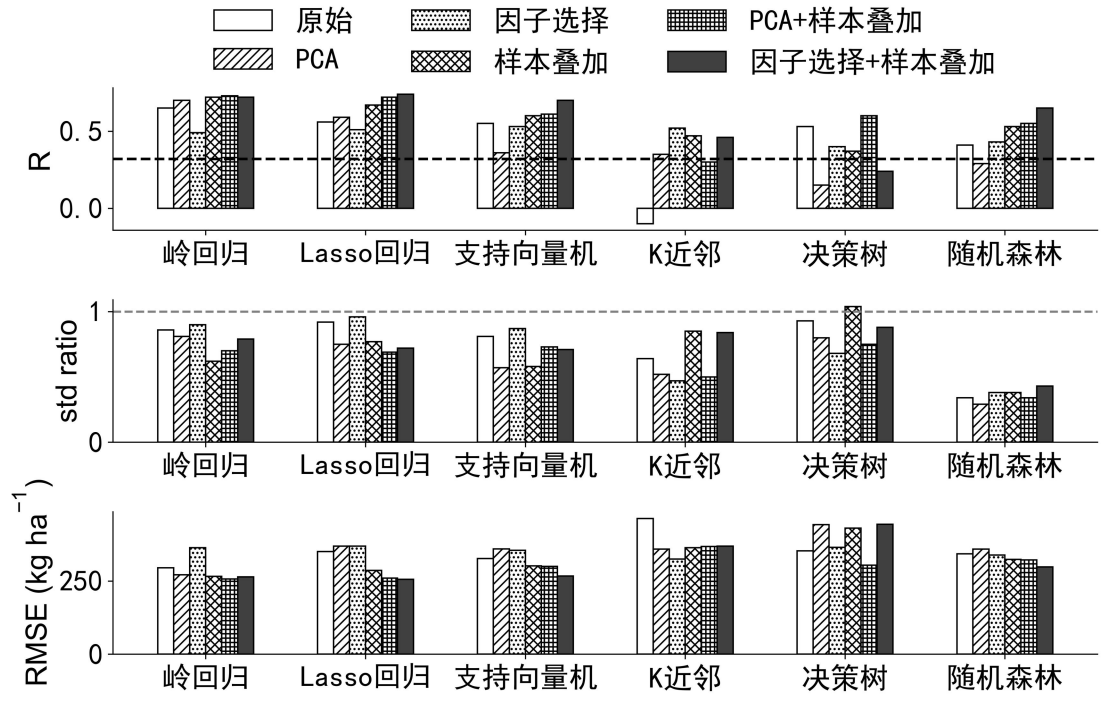

图 5　同图 3，但省份为辽宁省。

Figure 5　Same as Figure 3 but for Liaoning Province.

4.3 岭回归与逐步线性回归结果的比较

在“留一法”交叉验证的框架下，我们评估了六种机器学习方法的交叉验证技巧，其中岭回归的结果整体最优。我们将岭回归的结果与同一框架下使用逐步线性回归这一传统统计方法的预测结果作对比评估，由表 4 可见，岭回归在三省的交叉验证预测技巧相比逐步线性回归而言均有显著的提升，RMSE 也均大幅度降低，表明岭回归在复现东北大豆产量年际变化模态方面具有比逐步线性回归更优秀的性能，展现了正则化方法相对于传统统计回归的优越性。然而，在幅度的预测中，逐步线性回归在吉林省和辽宁省较优。

表 4　岭回归与逐步线性回归的交叉验证技巧对比。其中*，，***分别代表观测和预测间相关系数在 0.05，0.01，0.001 水平上显著。**

Table 4　Comparison of cross-validation skills between ridge regression and stepwise linear regression (*: P<0.05, **: P<0.01, ***, P<0.001).

省份	方法	R	r_{std}	RMSE
黑龙江	岭回归	0.48**	0.9	251
	逐步线性回归	0.34*	1.14	319
吉林	岭回归	0.58***	0.74	339
	逐步线性回归	0.25	0.91	479
辽宁	岭回归	0.72***	0.79	264
	逐步线性回归	0.6***	1	337

4.4 岭回归中的关键气象因子及其系数解释

我们分析了岭回归中对产量影响较大的关键气象因子（图 6）。7 月 7 至 11 日的相对湿度、8 月 16 至 20 日的日照时长、8 月 5 至 9 日的相对湿度、8 月 22 至 26 日的降水量以及 7 月 31 日至 8 月 4 日的最低温度为影响大豆产量最重要的气象因子。因子系数在绝大多数情况下大于零，且气象因子与产量之间的相关系数均值均大于等于 0.25，表明岭回归模型筛选出的关键气象因子与大豆产量之间普遍呈现稳定的正向关联。关键气象因子均处于 7 至 8 月的开花结荚和鼓粒时段，这一阶段是大豆生长发育最旺盛的时期（张波等, 2008; 刘金宇等, 2013），充足的养分和水分对最终产量形成具有决定性作用。该阶段更高的温度、更充足的水

分、日照条件以及水热的协同作用均有利于大豆的光合作用和生殖生长过程，确保花、荚和籽粒的正常生长发育，提高单株荚数和籽粒灌浆速率，有利于高产的形成（李炜等, 2008; 张波等, 2008）。

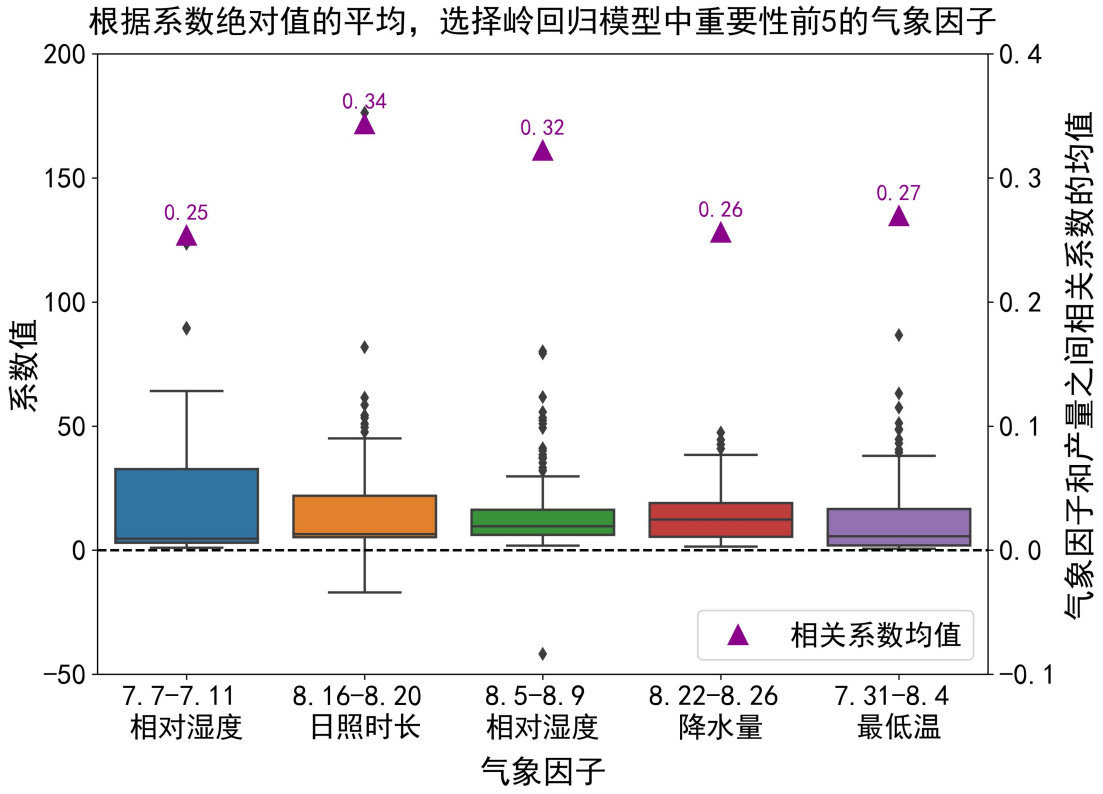

图 6 岭回归模型中重要程度前 5 的气象因子系数值分布（箱线图，左 y 轴）和气象因子与产量之间相关系数的均值（紫色三角，右 y 轴）。由于在每个“留一法”交叉验证子模型中分别进行显著因子的筛选和模型的建立，故重要因子首先选择为频次最高的因子，其次选择为系数绝对值的平均值最大的因子。因前 34 个因子在所有留一模型中均被全部选入（114 次），故选择系数绝对值平均最大的前 5 个因子，即图中所示。

Figure 6 Coefficient distributions (box plots, left y-axis) and mean correlation coefficients (purple triangles, right y-axis) of the top 5 meteorological factors in ridge regression models. The top factors were selected based on (1) highest selection frequency and (2) largest mean absolute coefficients across all leave-one-out cross-validation (LOOCV) sub-models. The first 34 factors were selected in all 114 LOOCV models; among them, the 5 factors with the highest mean absolute coefficients are shown.

5 总结与展望

本文采用六种机器学习方法，基于气象因子建立并评估了“留一法”交叉验证框架下的 1981~2018 年东北地区省级大豆产量年际变化的机器学习预测模型。结果表明：

（1）六种机器学习模型中，岭回归和 Lasso 回归在三省都能有技巧地预测大豆产量的年际变化，表明正则化方法的优越性。岭回归为三省整体最为显著的模型，在黑龙江、吉林和辽宁的交叉验证相关系数分别达到 0.48（P＜0.01）、0.58（P＜0.001）和 0.72（P＜0.001）。这表明，使用 L_2 正则化对线性回归进行约束即可在东北三省大豆产量预测中获得极为良好的性能。正如 van Klompenburg et al.（2020）的综述研究所示，在农业产量预测中，对于不同尺度、地理位置和作物类型，最佳模型有所不同，更为复杂的模型并不总是为产量预测提供最优性能。要寻求最优性能，应对具有不同复杂度的模型及不同特征组合进行系统的测试和比较。

（2）在岭回归模型中，5 日滑动平均法被证明为气象因子预处理的最优方法，在三省（黑龙江、吉林、辽宁）的表现均优于 7 日滑动平均、10 日滑动平均、月平均及生长季平均，具体表现为具有最高的相关系数（R）和最低的均方根误差（RMSE）。然而，以往研究（胡岳, 2023）显示，5 日滑动平均并非在所有情况下都最优：当采用逐步线性回归（而非本文的岭回归）并结合留一法交叉验证进行分省预报时，该窗口仅在黑龙江省表现出最优预测性能。值得注意的是，两项研究在黑龙江省的结论高度一致——5 日滑动窗口是唯一通过显著性检验的预处理方法（其他窗口均不显著）。而在吉林省，多个窗口能够通过 0.05 水平的显著性检验；在辽宁省，所有时间窗口均能通过 99%的显著性检验（在岭回归算法下为 99.9%）。这种区域差异可能与各省大豆产区的气候和地形特征有关：黑龙江省大豆种植区经纬度跨度大，省内气候条件和地形复杂多样；相比之下，吉林和辽宁两省的气象条件相对均一，因此其产量预测的稳定性更高，产量与气象因子之间的关联性更好。

（3）在同一框架下与逐步线性回归进行比较时，岭回归的交叉验证相关系数（R）在三省均显著提升，且均方根误差（RMSE）均大幅降低，这表明了岭回归在捕捉模态以及预测观测一致性方面相对于传统逐步线性回归的巨大优势。

然而，可能由于岭回归对模型系数的压缩作用，岭回归对吉林省和辽宁省的幅度预测不如逐步线性回归。

（4）多数情况下，对数据集进行因子选择和样本叠加的处理能够一定程度上提升机器学习方法的交叉验证预测技巧。该处理使得对相关系数的预测普遍提高或持平，仅在黑龙江和吉林使用 K 近邻算法和辽宁使用决策树算法时明显降低。对 RMSE 的预测普遍降低或基本持平，仅在辽宁省使用决策树算法时明显升高；对幅度的预测技巧可能降低也可能提高。

（5）在三省总体层面上，气象因子对产量形成的主要作用体现在 7 至 8 月的开花结荚期，期间温度、水分和日照时长对产量的形成均有正向作用。该阶段充足的水热条件有利于植株花、荚的形成和荚果、籽粒的发育，促进光合作用，提高碳同化量，提高最终产量。

本研究构建了东北大豆产量年际变化的机器学习预测框架，通过系统比较六类机器学习算法，为区域农业气象统计建模中的算法选择、时间尺度选择、数据优化策略选择提供了实证依据。本研究推动了农业气象预测方法的发展，为气候变化背景下的农业风险管理提供了科学依据，为其他作物和区域的产量预测研究提供了有价值的见解和参考。然而，研究仍存在以下局限性，未来有待进一步研究和解决：

（1）机器学习方法有待拓展。本研究主要采用六种经典的机器学习方法，尚未充分尝试深度学习等更加复杂的算法，如长短期记忆神经网络（LSTM）用于捕捉气象因子的时序特征，卷积神经网络（CNN）用于提取空间分布特征，以及残差神经网络（ResNet）用于探索深度潜在关系，这些方法的应用有望进一步提升东北大豆产量年际变化的预测精度。

（2）多源数据融合有待深化。本研究主要依赖气象数据，未来可整合多源遥感数据（如 MODIS 等）和地面观测数据，特别是生长季 NDVI 动态变化、叶面积指数（LAI）等植被参数，以及土壤水分、养分等环境因子，构建更全面的预测指标体系。未来的研究应探索更多算法和数据来源，以进一步提高东北地区大豆产量预测的准确性和可靠性。

（3）在实际进行预测时，通常需要提前给出预测结果。例如，对当年进行预测，常需提前 1 月（而并非采用生长季的全部气象因子）；对未来进行预测，

需先使用动力模式预报出未来气象要素，再使用预报得到的气象要素预测未来产量。因此，真实预测中的技巧将会有所降低。可以借鉴引言中 Monteiro et al.（2022）对巴西大豆的预测方法，使用生长季内更早的数据进行产量的动态预报，以提前获取当年的预测结果。

对以上问题的进一步研究将有助于建立更精准、更可靠的东北大豆产量预测系统，为区域农业生产管理和粮食安全保障提供更强大的科学支撑。

此外，本研究在东北大豆产量年际变化的预测上积累的经验可为实际应用提供参考：

（1）东北大豆产量年际变化与气象因子间的关系以线性特征为主。岭回归和 Lasso 回归两种正则化方法能够在三省都取得显著的预测技巧，表明仅通过简单的正则化约束即可显著提升线性回归的预测性能；且支持向量机回归对核函数类型的自动选择时均选择了线性核函数（而非多项式、径向基等非线性核函数）。因此，在东北地区大豆产量预测中，可优先考虑线性模型，而非复杂的非线性方法。

（2）可采用因子选择与样本叠加的方法来提升机器学习模型的预测能力。

参考文献（References）

Breiman L, Friedman J, Olshen R A, et al. 1984. Classification and Regression Trees [M]. New York: Chapman and Hall/CRC, 368pp.

Breiman L. 2001. Random Forests [J]. Machine Learning, 45(1): 5–32.

Chandio A A, Akram W, Sargani G R, et al. 2022. Assessing the impacts of meteorological factors on soybean production in China: What role can agricultural subsidy play? [J]. Ecological Informatics, 71: 101778. doi: 10.1016/j.ecoinf.2022.101778

陈雪, 高梦竹, 赵晶, 等. 2023. 高寒地区大豆产量动态预报研究 [J]. 农学学报, 13(6): 91–96. Chen Xue, Gao Mengzhu, Zhao Jing, et al. 2023. Study on dynamic forecast of soybean yield in alpine region [J]. Journal of Agriculture, 13(6), 91–96. doi: 10.11923/j.issn.2095-4050.cjas2022-0065

崔明元. 2014. 气象因子对通化地区大豆产量构成因素影响的分析 [D]. 延边大学硕士学位论文, 28pp. Cui Mingyuan. 2014. Analysis of meteorological factor on the soybean yield components in Tonghua [D]. Master's thesis (in Chinese), Yanbian University, 28pp. doi: 10.27010/d.cnki.gdbnu.2022.000558

范可, 王会军, Choi Y J. 2007. 一个长江中下游夏季降水的物理统计预测模型 [J]. 科学通报, 52(24), 2900–2905. Fan Ke, Wang Huijun, Choi Young-Jean, 2007. A physically-based statisitical forecast model for the middle-lower reaches of the Yangtze River Valley summer rainfall [J]. Chinese Science Bulletin (in Chinese), 52(24), 2900–2905. doi: 10.1360/csb2007-52-24-2900

Fan K, Wang H J. 2009. A new approach to forecasting typhoon frequency over the western North Pacific [J]. Weather and Forecasting, 24(4), 974–986. doi: 10.1175/2009WAF2222194.1

Fan K, Liu Y, Chen H P, et al. 2012. Improving the Prediction of the East Asian Summer Monsoon: New Approaches [J]. Weather and Forecasting, 27(4): 1017–1030. doi: 10.1175/WAF-D-11-00092.1

范可, 田宝强. 2013. 东北地区冬半年大雪-暴雪日数气候预测 [J]. 科学通报, 58(8): 699–706. Fan Ke, Tian Baoqiang. 2013. Prediction of wintertime heavy snow activity in Northeast China [J]. Chinese Science Bulletin, 58(8): 1420–1426. doi: 10.1007/s11434-012-5502-7

范可, 田宝强, 刘颖. 2016. 2015/2016 年极强厄尔尼诺事件下我国动力和统计结合实时气候

预测研究 [J]. 大气科学学报, 39(6): 744–755. Fan Ke, Tian Baoqiang, Liu Ying. 2016. Hybrid dynamical and statistical climate prediction in China during the extremely strong El Niño of 2015 and 2016 [J]. Transactions of Atmospheric Sciences (in Chinese), 39(6), 744–755. doi: 10.13878/j.cnki.dqkxxb.20160814003

FAO. 2015. World Reference Base for Soil Resources 2014 [M]. Rome: FAO.

FAO. 2023. World food and agriculture — statistical yearbook 2023 [M]. Rome: FAO. doi: 10.4060/cc8166en

Fix E, Hodges J L. 1951. Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties [R]. Technical Report 4. Randolph Field, TX: USAF School of Aviation Medicine.

Gong L J, Tian B X, Li Y G, et al. 2021. Phenological Changes of Soybean in Response to Climate Conditions in Frigid Region in China over the Past Decades [J]. International Journal of Plant Production, 15: 363–375, doi: 10.1007/s42106-021-00145-5

Guo S B, Guo E J, Zhang Z T, et al. 2022. Impacts of mean climate and extreme climate indices on soybean yield and yield components in Northeast China [J]. Science of The Total Environment, 838: 156284. doi: 10.1016/j.scitotenv.2022.156284

Guo S B, Zhang Z T, Zhang F L, et al. 2023. Optimizing cultivars and agricultural management practices can enhance soybean yield in Northeast China [J]. Science of The Total Environment, 857: 159456. doi: 10.1016/j.scitotenv.2022.159456

韩文革, 葛家麒, 于晓秋, 等. 2009. 多目标预测模型在黑龙江垦区大豆产量预测中的应用 [J]. 数学的实践与认识, 39(13): 51–55. Han Wenge, Ge Jiaqi, Yu Xiaoqiu. 2009. The applications of the multi-objectives forecasting model of soybean yield in Heilongjiang [J]. Mathematics in practice and theory (in Chinese), 39(13): 51–55.

Hasegawa T, Wakatsuki H, Ju H, et al. 2022. A global dataset for the projected impacts of climate change on four major crops [J]. Scientific Data, 9: 58. doi: 10.1038/s41597-022-01150-7

He L, Jin N, Yu Q. 2020. Soybean phenological changes in response to climate warming in three northeastern provinces of China [J]. Science of The Total Environment, 707: 135638. doi: 10.1016/j.scitotenv.2019.135638

Hoerl A E, Kennard R W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems [J]. Technometrics, 12(1): 55–67.

胡岳. 2023. 东北大豆产量年际变化的预测方法研究 [R]. 北京: 中国科学院大气物理研究

所. Hu Yue. 2023. Prediction of interannual variability of soybean yield in Northeast China (in Chinese) [R]. Beijing: Institute of Atmospheric Physics, Chinese Academy of Sciences.

江涛, 姜荣春, 王军. 2012. 从大豆产业开放及其产业格局演变看粮食安全 [J]. 国际贸易, (2): 45–49+53. Jiang Tao, Jiang Rongchun, Wang Jun. 2012. Food Security through the Lens of Soybean Industry Liberalization and Structural Evolution [J]. Intertrade (in Chinese), (2): 45–49+53. doi: 10.14114/j.cnki.itrade.2012.02.005

李芳. 2008. 季度预测理论方法的研究及其在东亚季风区夏季降水预测中的应用[D]. 中国科学院大气物理研究所博士学位论文, 95pp. Li Fang. 2008. Research on quarterly prediction theory methods and their application in summer precipitation prediction in the East Asian Monsoon Region [D]. Ph. D. dissertation (in Chinese), Institute of Atmospheric Physics, Chinese Academy of Sciences, 95pp.

Li Q C, Xu S W, Zhuang J Y, et al. 2023. Ensemble learning prediction of soybean yields in China based on meteorological data [J]. Journal of Integrative Agriculture, 22(6): 1909–1927. doi: 10.1016/j.jia.2023.02.011

李炜. 2008. 主要气象因子对大豆生长发育及产量的关联分析[J]. 黑龙江农业科学, (2): 41–43. Li Wei. 2008. Grey Relational Analysis of Meteorological Factors on Soybean Growth and Yield [J]. Heilongjiang Agricultural Sciences (in Chinese), (2): 41–43. doi: 10.3969/j.issn.1002-2767.2008.02.014

刘景利, 杨扬, 史奎桥, 等. 2007. 1985—2005 年锦州地区大豆物候期变化及气候响应[J]. 气象与环境学报, 23(4): 29–32. Liu Jingli, Yang Yang, Shi Kuiqiao, et al. 2007. Soybean phenology and its responses to climatic factors from 1985 to 2005 in Jinzhou region [J]. Journal of Meteorology and Environment (in Chinese), 23(4): 29–32.

刘景利, 杨扬, 梁涛, 等. 2013. 气象要素对大豆产量的影响分析 [J]. 气象与环境学报, 29(5): 136–139. Liu Jingli, Yang Yang, Liang Tao, et al. 2013. The effect of meteorological factors on soybean yield [J]. Journal of Meteorology and Environment (in Chinese), 29(5): 136–139. doi: 10.3969/j.issn.1673-503X.2013.05.022

刘金宇, 王兴环, 刘洋宸, 等. 2022. 气象因子对宝清县大豆和玉米产量的影响及预测[J]. 农业灾害研究, 12 (2): 107–109. Liu Jinyu, Wang Xinghuan, Liu Yangchen, et al. 2022. Influence and prediction of meteorological factors on meteorological yield of soybean and maize in Baoqing County. Journal of Agricultural Catastrophology (in Chinese), 12(2):

107–109. doi: 10.3969/j.issn.2095-3305.2022.02.035

Liu X B, Jin J, Wang G H, et al. 2008. Soybean yield physiology and development of high-yielding practices in Northeast China [J]. Field Crops Research, 105(3): 157–171. doi: 10.1016/j.fcr.2007.09.003

Lu J, Fu H K, Tang X H, et al. 2024. Deep learning for multi-source data-driven crop yield prediction in Northeast China [J]. Agriculture, 14(6): 794. doi: 10.3390/agriculture14060794

路亚洲, 宋广树, 孙蕾, 等. 2012. 东北地区主要气象要素分布特征分析 [J]. 中国农学通报, 28 (14): 290–294. Lu Yazhou, Song Guangshu, Sun Lei, et al. 2012. The Analysis of Distribution Characteristics of Main Meteorological Factors in the Northeast China [J]. Chinese Agricultural Science Bulletin (in Chinese), 28 (14): 290–294. doi: 10.11924/j.issn.1000-6850.2011-3557

Luan X Y, Bommarco R, Vico G. 2022. Coordinated evaporative demand and precipitation maximize rainfed maize and soybean crop yields in the USA [J]. Ecohydrology, 16(1): e2500. doi: 10.1002/eco.2500

Luan X Y, Bommarco R, Scaini A, et al. 2021. Combined heat and drought suppress rainfed maize and soybean yields and modify irrigation benefits in the USA [J]. Environmental Research Letters, 16(6): 064023. doi: 10.1088/1748-9326/abfc76

吕金莹, 闫超, 贾天宇, 等. 2019. 松嫩平原活动积温变化及其对作物产量的影响 [J]. 生态学杂志, 38(11): 3349–3356. Lü Jingyin, Yan Chao, Jia Tianyu, et al. 2019. The variation of accumulated temperature in Songnen Plain and its impact on crop yield [J]. Chinese Journal of Ecology (in Chinese), 38(11): 3349–3356. doi: 10.13292/j.1000-4890.201911.025

Monteiro L A, Ramos R M, Battisti R, et al. 2022. Potential Use of Data-Driven Models to Estimate and Predict Soybean Yields at National Scale in Brazil [J]. International Journal of Plant Production, 16(4): 691–703. doi: 10.1007/s42106-022-00209-0

Petersen L K. 2019. Impact of Climate Change on Twenty-First Century Crop Yields in the U.S. [J]. Climate, 7(3): 40. doi: 10.3390/cli7030040

Portmann F T, Siebert S, Döll P. 2010. MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high‐resolution data set for agricultural and hydrological modeling [J]. Global Biogeochemical Cycles, 24(1): GB1011. doi: 10.1029/2008GB003435

邱美娟, 郭春明, 王冬妮, 等. 2018. 基于气候适宜度指数的吉林省大豆单产动态预报研究

[J]. 大豆科学, 37(3): 445–451. Qiu Meijuan, Guo Chunming, Wang Dongni, et al. 2018. Study of soybean yield forecast in Jilin Province based on climate suitability index method [J]. Soybean Science (in Chinese), 37(3): 445–451. doi: 10.11861/j.issn.1000-9841.2018.03.0445

Ray D K, Gerber J S, MacDonald G K, et al. 2015. Climate variation explains a third of global crop yield variability [J]. Nature Communications, 6: 5989. doi: 10.1038/ncomms6989

Song L F, Song R X, Wang C C. 2024. Prediction of soybean yield in Jilin based on diverse machine learning algorithms and meteorological disaster indices[J]. The International Journal of Multiphysics, 18(2): 278–287. URL: https://themultiphysicsjournal.com/index.php/ijm/article/view/1272

Stone M. 1974. Cross-validation and multinomial prediction [J]. Biometrika, 61(3): 509–515. doi: 10.1093/biomet/61.3.509

Tao F L, Yokozawa M, Liu J Y, et al. 2008. Climate–crop yield relationships at provincial scales in China and the impacts of recent climate trends [J]. Climate Research, 38(1): 83–94. doi: 10.3354/cr00771

Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society: Series B (Methodological), 58(1): 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

van Klompenburg T, Kassahun A, Catal C. 2020. Crop yield prediction using machine learning: A systematic literature review [J]. Computers and Electronics in Agriculture, 177: 105709. doi: 10.1016/j.compag.2020.105709

王德明. 2022. 60 年来绥化市热量和降水变化特征及对粮食产量的影响 [D]. 东北农业大学硕士学位论文, 67pp. Wang Deming. 2022. Variation characteristics of heat and precipitation in Suihua City over the past 60 years and effects on grain yield [D]. Master's thesis (in Chinese), Northeast Agricultural University, 67pp.

王贺然, 张慧, 王莹, 等. 2018. 基于两种方法建立辽宁大豆产量丰歉预报模型对比 [J]. 中国农业气象, 39(11): 725–738. Wang Heran, Zhang Hui, Wang Ying, et al. 2018. A comparative study on forecast model for soybean yield by using different statistic methods in Liaoning Province [J]. Chinese Journal of Agrometeorology (in Chinese), 39(11): 725–738. doi: 10.3969/j.issn.1000-6362.2018.11.004

Wang W X, Deng X Z, Yue H X. 2024. Black soil conservation will boost China's grain supply and reduce agricultural greenhouse gas emissions in the future [J]. Environmental Impact Assessment Review, 106: 107482. doi: 10.1016/j.eiar.2024.107482

吴佳, 高学杰. 2013. 一套格点化的中国区域逐日观测资料及与其它资料的对比 [J]. 地球物理学报, 56(4): 1102–1111. Wu Jia, Gao Xuejie. 2013. A gridded daily observation dataset over China region and comparison with the other datasets. Chinese Journal of Geophysics, 56(4): 1102–1111. doi: 10.6038/cjg20130406

Xin M H, Zhang Z G, Han Y C, et al. 2023. Soybean phenological changes in response to climate warming in three northeastern provinces of China [J]. Field Crops Research, 302: 109082, doi: 10.1016/j.fcr.2023.109082

闫平, 姜丽霞, 王萍, 等. 2005. 5~7 月光照对大豆产量的影响分析 [J]. 黑龙江气象, 22(4): 3–4. Yan Ping, Jiang Lixia, Wang Ping, et al. 2005. Analysis of the illumination effect on the soy output from May to July [J]. Heilongjiang Meteorology (in Chinese), 22(4): 22–43, doi: 10.3969/j.issn.1002-252X.2005.04.010

Yates L A, Aandahl Z, Richards S A. 2022. Cross validation for model selection: A review with examples from ecology [J]. Ecological Monographs, 93(1): e1557. doi: 10.1002/ecm.1557

于贵瑞等. 2023. 全球变化对生态脆弱区资源环境承载力的影响研究 [M]. 北京: 科学出版社, pp.401–402. Yu Guirui, et al. 2023. The effects of global changes on the resource and environmental carrying capacity of the ecologically fragile areas (in Chinese) [M]. Beijing: Science Press, pp.401–402.

于晓秋, 郭玉. 2002. 气象因子对大豆产量的影响 [J]. 黑龙江气象, 19(2): 3–4. Yu Xiaoqiu, Guo Yu. The Effect of Meteorological Factors on the Soybean Output [J]. Heilongjiang Meteorology (in Chinese), 19(2): 3–4. doi: 10.3969/j.issn.1002-252X.2002.02.002

于晓秋, 葛家麒, 刘长海. 2007. 组合预测方法在黑龙江垦区大豆产量预测中的应用 [J]. 数学的实践与认识, 37(24): 27–32. Yu Xiaoqiu, Ge Jiaqi, Liu Changhai. 2007. The applications of the method of the combination forecasting in the forecasting of soybean yield in Heilongjiang Reclamation Area [J]. Mathematics in practice and theory (in Chinese), 37(24): 27–32.

张波, 王宝河, 薛艳萍, 等. 2008. 影响 2007 年宁安农场大豆产量的气象因子分析[J]. 现代农业科技, (5): 183–185. Zhang Bo, Wang Baohe, Xue Yanping, et al. 2008. The analysis of

meteorological factors affecting soybean yield at Ning'an Farm in 2007 [J]. Modern Agricultural Science and Technology (in Chinese), (5): 183–185. doi: 10.3969/j.issn.1007-5739.2008.05.126

Zhang H Q, Chandio A A, Yang F, et al. 2022. Modeling the Impact of Climatological Factors and Technological Revolution on Soybean Yield: Evidence from 13-Major Provinces of China [J]. International Journal of Environmental Research and Public Health, 19(9): 5708. doi: 10.3390/ijerph19095708

赵放, 林伟楠, 赵春亮, 等. 2024. 气候因子对东北地区大豆生产的影响效应分析 [J]. 大豆科学, 4(6): 758–767. Zhao Fang, Lin Weinan, Zhao Chunliang, et al. 2024.Analysis of the impact of climatic factors on soybean production in Northeast China [J]. Soybean Science (in Chinese), 4(6): 758–767. doi: 10.11861/j.issn.1000-9841.2024.06.0758

中华人民共和国国家统计局. 2024. 中国统计年鉴 2024 [M]. 北京: 中国统计出版社. National Bureau of Statistics of China. 2024. China Statistical Yearbook (in Chinese) [M]. Beijing: China Statistics Press.

中国农业年鉴编辑委员会. 2021. 中国农业年鉴 2021 [M]. 北京: 中国农业出版社. Editorial Committee of China Agricultural Yearbook. 2021. China Agricultural Yearbook 2021 (in Chinese) [M]. Beijing: China Agriculture Press.

中华人民共和国农业农村部, 2021. 国家黑土地保护工程实施方案（2021—2025 年）[R/OL]. Ministry of Agriculture and Rural Affairs of the People's Republic of China, 2021. Implementation Plan for the National Black Soil Conservation Project (2021–2025) (in Chinese) [R/OL]. (2021-07) [2025-02-20]. Available at http://www.ntjss.moa.gov.cn/zcfb/202107/P020210729662220053137.pdf