AutoML-based Almond Yield Prediction and Projection in California

Shiheng Duan
Lawrence Livermore National Laboratory
Livermore, CA 94550
duan5@llnl.gov

Shuaiqi Wu
University of California, Davis
Davis, CA 95616
shqwu@ucdavis.edu

Erwan Monier
University of California, Davis
Davis, CA 95616
emonier@ucdavis.edu

Paul Ullrich
University of California, Davis
Davis, CA 95616
paullrich@ucdavis.edu

Abstract

Almonds are one of the most lucrative products of California, but are also among the most sensitive to climate change. In order to better understand the relationship between climatic factors and almond yield, an automated machine learning framework is used to build a collection of machine learning models. The prediction skill is assessed using historical records. Future projections are derived using 17 downscaled climate outputs. The ensemble mean projection displays almond yield changes under two different climate scenarios, along with two technology development scenarios, where the role of technology development is highlighted. The mean projections and distributions provide insightful results to stakeholders and can be utilized by policymakers for climate adaptation.

1 Introduction

California is the most significant almond-producing region in the world. Eighty percent of the almonds consumed globally and all commercially sold in the United States are produced in California. Almonds were California's most profitable agricultural export product and the second-largest commodity overall in 2020 [1]. Previous studies have shown that climate factors can influence almond growth [2, 3], and when these climate factors change due to global warming, the yield of almonds will inevitably suffer. Thus, a precise projection is necessary for both scientific research and climate adaptations.

Because of the difficulty in simulating the growth process of perennial crops, very few process-based crop models can be utilized to investigate the relationship between climatic variables and almond

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

yields. Data-driven approaches, including machine learning (ML) and deep learning (DL) models, have gained popularity in Earth system modeling since the emergence of artificial intelligence due to their capacity to fit nonlinear functions. Many studies have applied ML and DL models for various tasks, such as hydrology projection, air quality analysis and severe weather forecasting [4, 5, 6]. The ML and DL models can be designed in an end-to-end manner to examine the relationship between input and output variables. While this strategy is often effective, a proper model architecture is required, which often involves specialized knowledge. Automated machine learning (AutoML) frameworks are designed to save these efforts and to facilitate the use of ML models by researchers without a background in computer science, as is frequently the case for climate studies.

The goal of this study is to build an AutoML framework that creates pipelines from California's climatic variables to estimate almond yields. Section 2 introduces the data sources and AutoML model. In section 3, the model is trained and tested with observational records, and future projections are in section 4.

2 Data and Machine Learning Model

2.1 GridMET and Almond Yields

Climate variables were derived from the GridMET climate dataset, which provides high-resolution (1/24th degree) daily meteorological data for the contiguous United States from 1979 to the present [7]. Instead of simply using monthly or seasonal averaged climate variables, we extracted climate data from GridMET based on the phenology of almond, such as the specific humidity of the bloom period and chill hours of dormancy. Gridded climate variables were averaged or summed based on the historical location of almond orchards from the CropScape geospatial Cropland Data Layer (CDL) product developed by the United States Department of Agriculture, which maintains the consistency between climate variables and almond yield and improves the accuracy of our analysis [8]. Figure 1 depicts the location of almond orchards in California over the past 15 years. Climate data inputs were normalized by removing their mean values and dividing them by their standard deviation. Previous studies have revealed quadratic relationships between climate variables and perennial crops yields in California [2]. Therefore, we include a total of 13 phenology-climate variables and their squares as features to the ML model. The county-level almond yields (ton/acre) and planted areas (acre) were reported by the California County Agricultural Commissioners and available on the website of USDA (United States Department of Agriculture) for the period of 1980 to 2020 [9]. To account for advancements in farming practices and technological improvements, we include a linear trend variable for each county. In addition, the county names (16 counties) were used as a categorical predictor in our analysis to represent the differences in static factors such as soil properties and agricultural policies among counties.

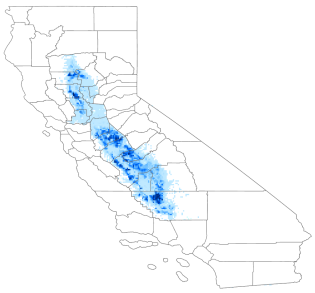

Figure 1: Map of Almond orchards. Darker color represents denser plant area.

2.2 MACA

The Multivariate Adaptive Constructed Analogs (MACA) dataset is used for projection purposes. It includes downscaled products from 20 global climate models from the Coupled Model Intercomparison Project Phase 5 (CMIP5). We subselected the 17 climate models without missing values in

our study area. These models have been developed at different modeling centers across the globe, and can account for structural uncertainties such as model design and parameterization choice. The selected models can be found in the supplement. Following the CMIP5 experimental design, the historical period is from 1850 to 2006 [10]. For the period from 2006 to 2020, which overlaps with our observational period and gridMET forcings, both the RCP4.5 and RCP8.5 scenarios are used and compared against observational records.

2.3 AutoGluon

Automated machine learning (AutoML) is a type of framework that automates the construction of machine learning pipelines from input data to targets [11]. It integrates several necessary processes in machine learning applications, including but not limited to: data preprocessing, data augmentation, feature engineering, and hyperparameter tuning. AutoGluon is an AutoML framework that automates ML pipelines on tabular, text and image datasets [12]. It has been used for various problems, such as landslide hazards [13] and drought forecasts [14]. In contrast to other AutoML frameworks, AutoGluon places less emphasis on hyperparameter optimization. Instead, it makes use of multi-layer stack ensembling and repeated k-fold bagging, which has been demonstrated to outperform other prominent AutoML frameworks in benchmark tasks [12].

3 Prediction Performance

The coefficient of determination (R^2 score) and root mean squared error (RMSE) are used to quantify the model predictability, and AutoGluon is compared against random forest and linear regression. A 5-fold cross-validation approach is used together with a train-test split method (70%-30%) to benchmark the ML models. There are various presets in AutoGluon that specify the hyperparameters. In general, high accuracy corresponds to longer training time and larger models for disk consumption. The choice of presets may affect the model performance, but to maintain simplicity and automation, we only select one preset ('high-quality') in this study.

As listed in Table 1, both the AutoGluon and random forest have relatively lower skill with the train-test split, with the exception of linear regression. This is probably due to the decrease in the number of training samples, since 80% samples are used as the training set in cross-validation, whereas only 70% are used for the train-test split. Notably, the AutoGluon model achieves the best performance for both the cross-validation and train-test settings.

Table 1: Machine learning model performance. Values in bold represent the best performance.

Model	Cross-Validation R^2	Cross-Validation RMSE	Train-Test R^2	Train-Test RMSE
AutoGluon	**0.754**	**0.132**	**0.740**	**0.135**
Linear Regression	0.715	0.143	0.724	0.139
Random Forest	0.611	0.167	0.599	0.168

4 Projection and Climate Adaptation

Projections under various climate change scenarios are produced using the trained AutoGluon model and MACA forcing data. Figure 2 shows the comparison of the historical simulation with observations. The results from 17 MACA models are generally consistent with one another, with a small inter-quantile range, and are effective at capturing the trend in almond yield. Comparing the ensemble mean simulations, RCP4.5 is slightly closer to observations ($R^2 = 0.5205$) than RCP8.5 ($R^2 = 0.5199$). Despite the fact that these values are lower than those in Table 1, it is to be expected given that these climate models are only forced by aerosol and greenhouse-gas concentrations and so are only climatologically consistent with observed history. Shorter-term phenomena that can affect yield (i.e., heat waves or wildfires) are averaged out.

AutoGluon projections under RCP4.5 and RCP8.5 scenarios are shown in Figure 3. In addition to climate change scenarios, we also include two technology scenarios: 'WOTech' denotes agricultural technology remaining at its current level as of 2020, whereas 'WTech' denotes agricultural technology continuing improving through 2100. Compared with historical simulations, the future projections display higher uncertainty, with larger inter-quantile ranges, likely due to the discrepancy among

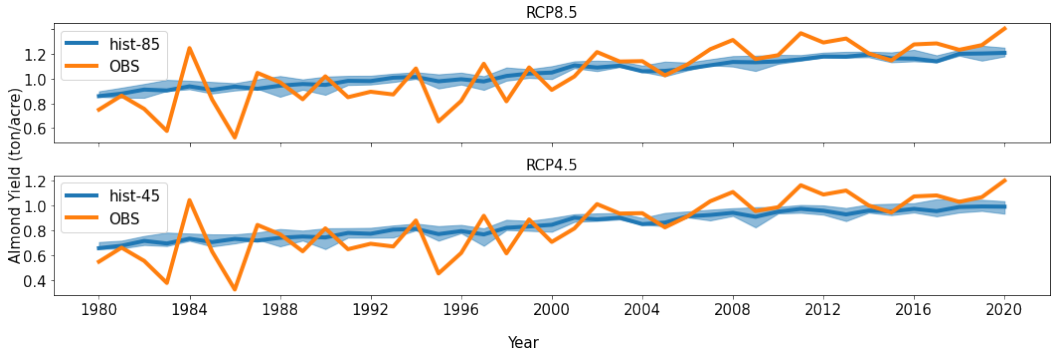

Figure 2: Almond yield observations and historical simulations. The blue line depicts the mean simulations from the selected 17 MACA models, and the blue shading represents the inter-quantile range. Observation is abbreviated to 'OBS'. 'RCP' projection data is used over the period from 2005 to 2020.

climate models. It is evident by comparing the various technology scenarios within the same climatic scenario that technological advancement is necessary to maintain historical trends in annual almond yield growth. This finding highlights the important role of technological advancement in climate adaptation. On the other hand, almond yield will decline in both the RCP4.5 and RCP8.5 scenarios after peak technological advancement. As for the climate change scenarios, the RCP8.5 shows a slightly greater decline in yields compared with RCP4.5, likely due to high temperatures under this scenario.

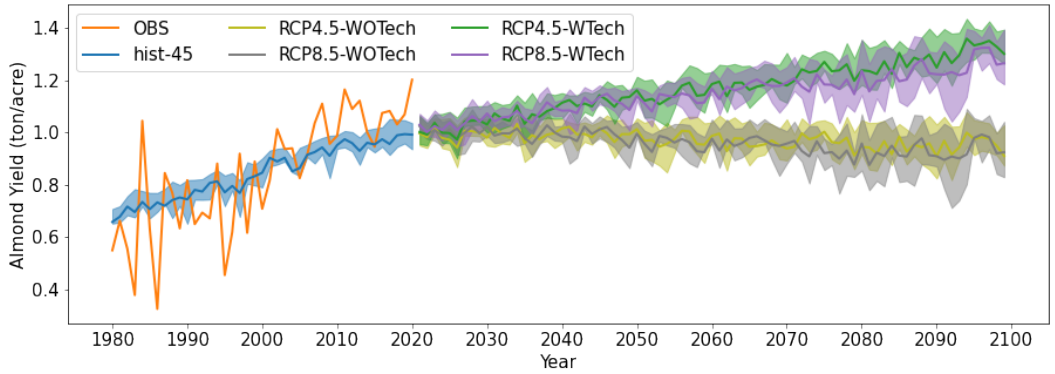

Figure 3: AutoGluon projections under RCP4.5 and RCP8.5 scenarios. Solid lines represent the ensemble mean from the selected MACA simulations, whereas shadings show the inter-quantile range. Observation is abbreviated to 'OBS'.

5 Conclusions and Future Work

The use of deep learning and machine learning techniques to answer climate change-related problems is growing in popularity. Not all tasks, however, can offer sufficient data samples for complex neural networks. In this study, we demonstrate how an AutoML framework can predict and project California's almond yield. As an ensemble ML model, AutoGluon achieves better performance compared with single ML models, without requiring more data samples or computational resources. It also enables us to analyze the relative importance of input variables using, i.e., a permutation-based method, from which the total precipitation during the bloom period (February-Mid March) is shown as the most important variable. These results are not shown here, as they are unrelated to our projection analysis.

Our projection results highlight the critical role that technology plays in climate adaptation. Increases in almond yield require further technological developments; in the absence of further development,

climatological factors will lead to a decrease in yields. However, the cost of developing new techniques is not taken into consideration in our analysis. From the perspective of stakeholders, it could be promising to incorporate our projection results within a socioeconomic model to gain insight into the balance between investment cost and expected output.

6 Acknowledgment

We would like to thank computational resources funded by the Walter and Margaret Milton Graduate Atmospheric Science Award at University of California, Davis. The workshop presentation is funded by the Lawrence Livermore National Laboratory Postdoc Development Program. Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] CDFA - Statistics — cdfa.ca.gov. `https://www.cdfa.ca.gov/Statistics`. [Accessed 14-Aug-2022].

[2] David B Lobell, Christopher B Field, Kimberly Nicholas Cahill, and Celine Bonfils. Impacts of future climate change on california perennial crop yields: Model projections with climate and crop uncertainties. *Agricultural and Forest Meteorology*, 141(2-4):208–218, 2006.

[3] Chaopeng Hong, Nathaniel D Mueller, Jennifer A Burney, Yang Zhang, Amir AghaKouchak, Frances C Moore, Yue Qin, Dan Tong, and Steven J Davis. Impacts of ozone and climate change on yields of perennial crops in california. *Nature Food*, 1(3):166–172, 2020.

[4] Shiheng Duan, Paul Ullrich, and Lele Shu. Using convolutional neural networks for streamflow projection in california. *Frontiers in Water*, page 28, 2020.

[5] Yunqian Lv, Hezhong Tian, Lining Luo, Shuhan Liu, Xiaoxuan Bai, Hongyan Zhao, Shumin Lin, Shuang Zhao, Zhihui Guo, Yifei Xiao, et al. Meteorology-normalized variations of air quality during the covid-19 lockdown in three chinese megacities. *Atmospheric Pollution Research*, page 101452, 2022.

[6] Aaron J Hill, Russ S Schumacher, and Israel Jirak. A new paradigm for medium-range severe weather forecasts: probabilistic random forest-based predictions. *arXiv preprint arXiv:2208.02383*, 2022.

[7] John T Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013.

[8] CDFA - Statistics — cdfa.ca.gov. `https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php`, . [Accessed 14-Aug-2022].

[9] CDFA - Statistics — cdfa.ca.gov. `https://www.nass.usda.gov/Statistics_by_State/California/Publications/AgComm/`, . [Accessed 14-Aug-2022].

[10] Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American meteorological Society*, 93(4):485–498, 2012.

[11] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[12] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[13] Wenwen Qi, Chong Xu, and Xiwei Xu. Autogluon: A revolutionary framework for landslide hazard analysis. *Natural Hazards Research*, 1(3):103–108, 2021.

[14] Shiheng Duan. Automl-based drought forecast with meteorological variables. *arXiv preprint arXiv:2207.07012*, 2022.