# AutoML-based Almond Yield Prediction and Projection in California

Presentation at NeurIPS 2022 Workshop
Tackling Climate Change with Machine Learning

Shiheng Duan[1], Shuaiqi Wu[2], Erwan Monier[2], Paul Ullrich[2]
[1]Lawrence Livermore National Laboratory, [2]University of California, Davis

Dec 9, 2022



Lawrence Livermore National Laboratory

# Why Almond? Why AutoML?

- Almond is an important product in California.
- Climatic variables and technologies affect almond yields.
- Process-based models are more suitable for annual crops. Data-driven models are more common for perennial plants.
- Hyperparameters can affect model performance. Task of Combined Algorithm Selection and Hyperparameter optimization, or CASH.

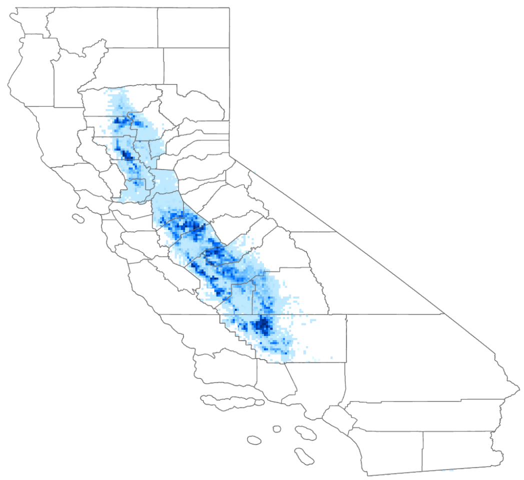

Fig. Map of Almond orchards.

# AutoGluon

- AutoGluon: AutoML for Text, Image, and Tabular Data.
- Benchmarks: Compared with 5 other AutoML frameworks on OpenML AutoML benchmarks and Kaggle benchmarks (50 tasks in total). AutoGluon generates the best score in 30 tasks.
- Idea: Instead of the CASH task, focus on the ensembles.
  - Multi-Layer Stack Ensemble. “Model of models”
  - Repeated k-fold Bagging. K models with k folds of data.
- Preset of hyperparameters.
  - Hyperparameters for individual model
  - Complexity of stacking and bagging.

# Data

- For prediction (1980-2020):
  - GridMET. It is extracted by the phenology of almond and averaged by the location of almond orchards. 13 climatic variables and their squares.
  - USDA. Almond yield in ton/acre for each county.
  - Technology. Linear trend for each county.
  - County names. Categorical variable representing the soil and technology differences.
- For projection (1980-2100):
  - MACA. Same climatic variables as GridMET, but with downscaled CMIP5 simulations.
  - Historical period from 1980-2005 and RCP4.5/8.5 for 2005-2100.
  - County names. Same as the prediction period.
  - Technology. Continue the linear trend or stopped at 2020.

# Prediction Results

- County-level metrics

| Model | 5-Fold CV $R^2$ | 5-Fold CV 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 |

- 5-Fold cross validation is stratified by counties. No spatial extrapolation. Train-test with a ratio of 70%-30%.

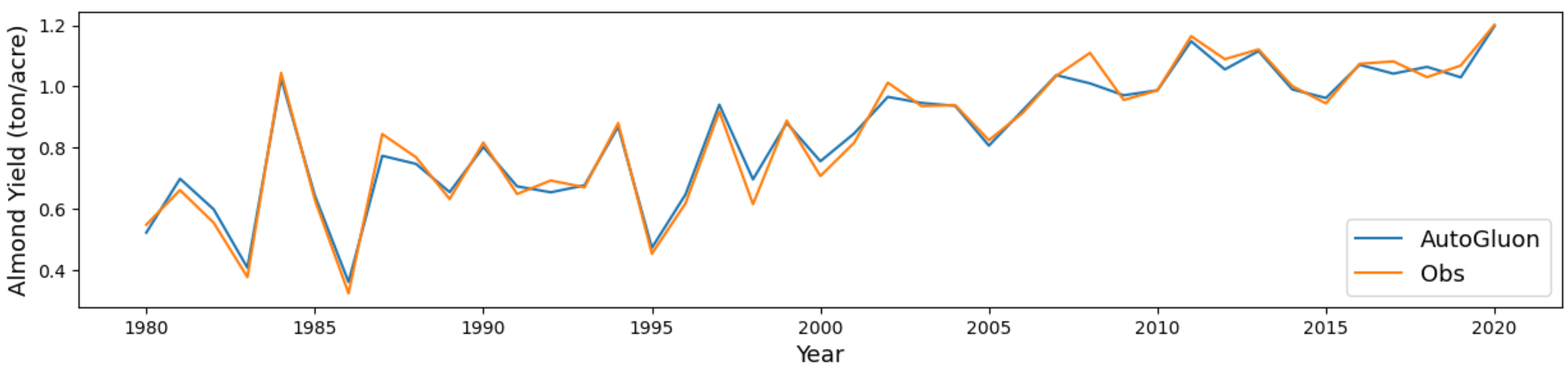


Fig. Prediction of State-level Almond yields.

# Projection with MACA

- Ensemble projections with 17 CMIP5 models. Historical, RCP4.5 and RCP8.5 scenarios.

- Two technology scenarios.

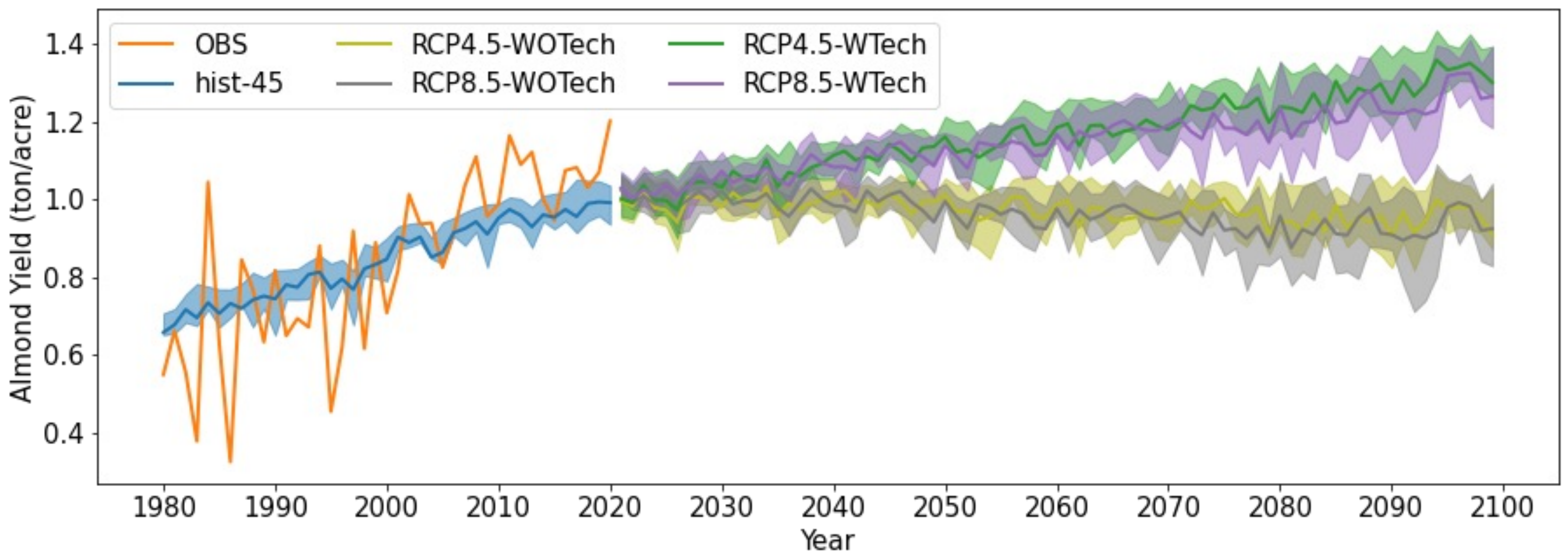


Fig. Projections with MACA forcings.

# Conclusions and Future work

- An easy-to-use AutoML framework:AutoGluon. It can be served as a baseline for any complex deep learning models.
- Permutation-based feature importance score. Either valid the machine learning model or discover new physical laws.
- Sensitivity test. Decompose changes in almond yields.


- Questions or suggestions: duan5@llnl.gov or shqwu@ucdavis.edu



The computational resources are funded by the Walter and Margaret Milton Graduate Atmospheric Science Award at University of California, Davis. The presentation is funded by the Lawrence Livermore National Laboratory Postdoc Development Program.

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC