Identifying optimal machine learning approach for fish distribution modeling

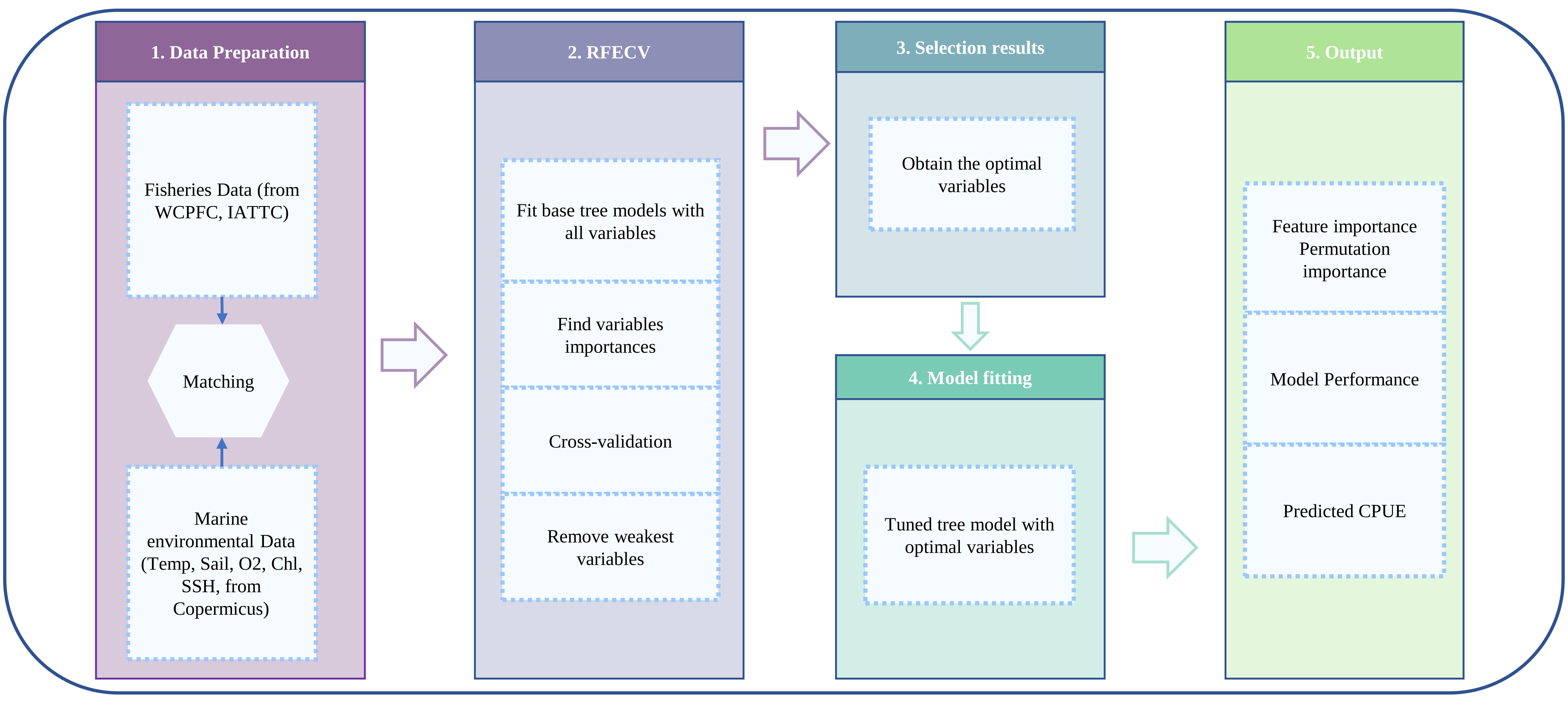
Shaohua Xu, Jintao Wang, Yunkai Li, Xinjun Chen
College of Marine Science, Shanghai Ocean University
Funding: National Natural Science Foundation of China (No. 41876141)
National Key R&D Program of China (2019YFD0901404)

1. Introduction

Background: Yellowfin tuna (*Thunnus obesus*) is an ecological and economical important pelagic species. They are widely distributed in the tropical and subtropical 40° N – 40° S waters of the Pacific, Atlantic, and Indian Oceans with fast swimming speed and high mobility. Fish species distribution models (SDMs) link information on the presence/absence or abundance of species to environmental variables to predict where (and how much of) a species is likely to be present in unsampled locations or time periods. The use of machine learning methods to predict the distribution of species has shown good results. To study the relationship between yellowfin tuna and marine environment and to investigate the distribution of yellow tuna for fishery management and decision-making, it is necessary to select the optimal variables in machine learning.

Objective: We apply feature selection methods to select the suitable predictor variables for constructing tree-based yellowfin tuna distribution models exploring environments-abundance relationships in the Pacific Ocean. The process we manipulated in this study could be as an example of selecting optimal predictor variable sets in machine learning distribution models for pelagic fishes with limited biological cognizance.

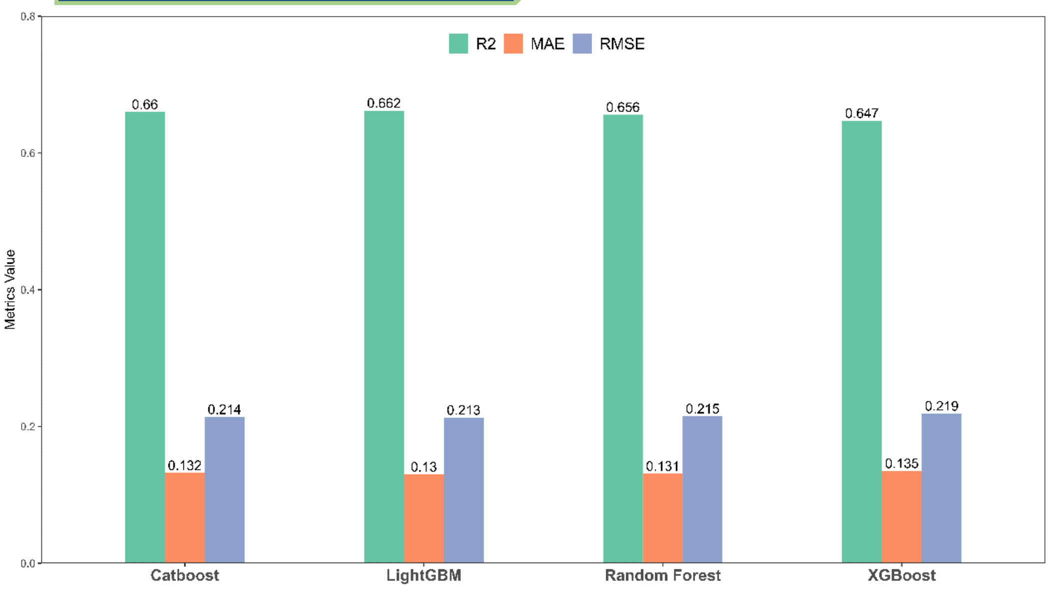
2. Materials and methods



3. Results and analysis

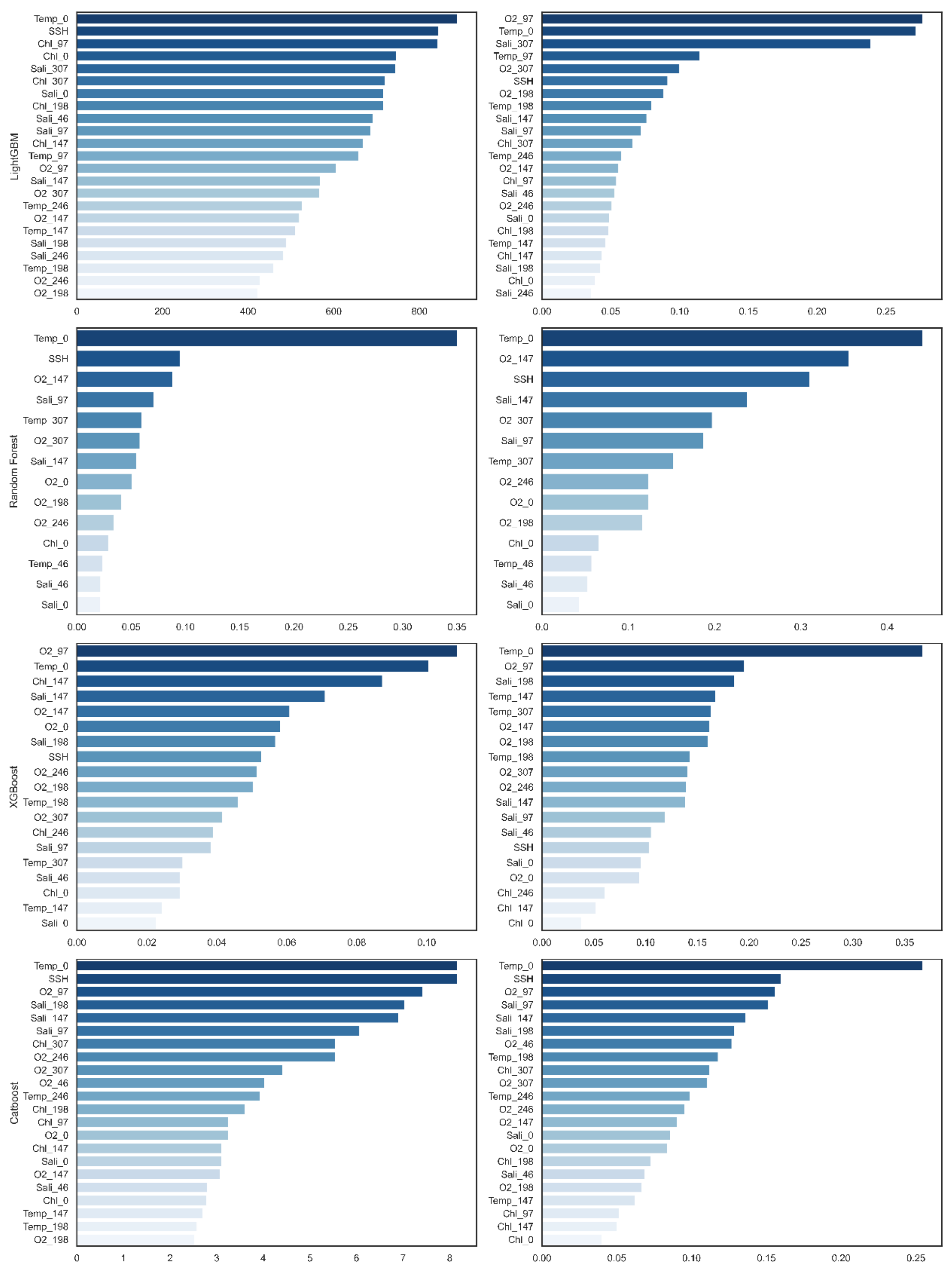
3.1 Performance of tree-based models with significant variables

Performance of the models



Performances of four tree-based models with significant variables. R-squared is an important metric for measuring the model, and every tree-based model has an R-squared of more than 0.6.

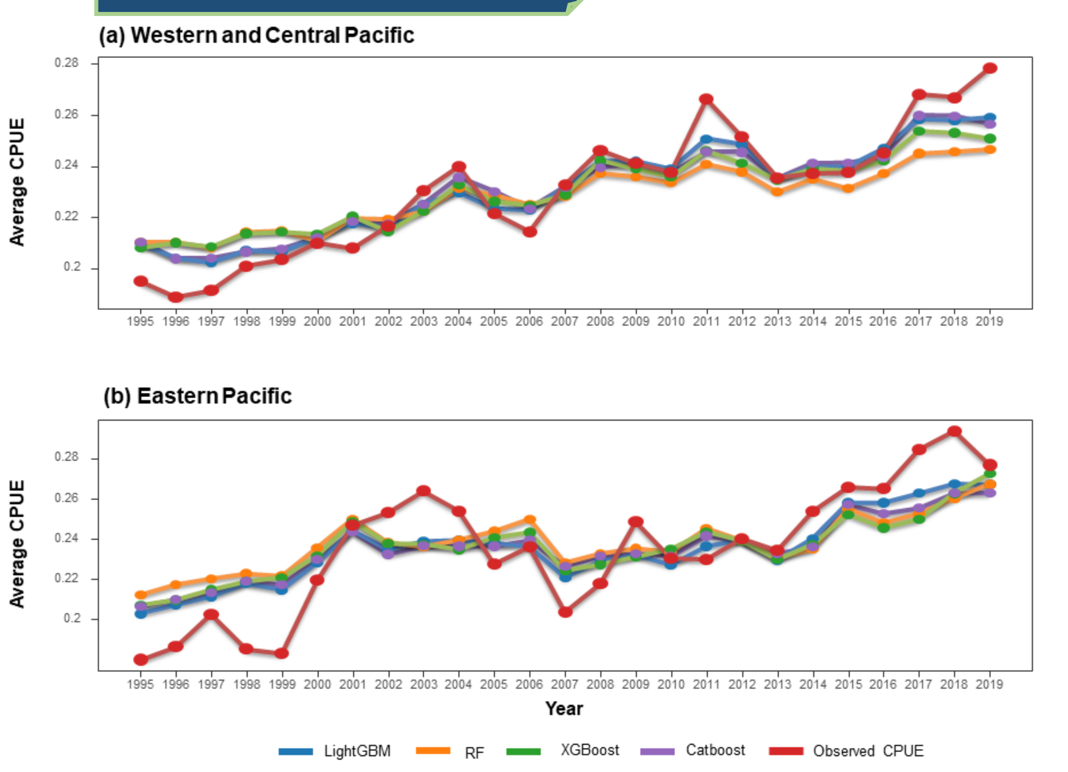
Optimal variables ranking



The importance of optimal variables ranked by the feature importance and the permutation importance for the four tree-based models. Temp_0, O2_97, SSH are all very important variables for yellowfin tuna.

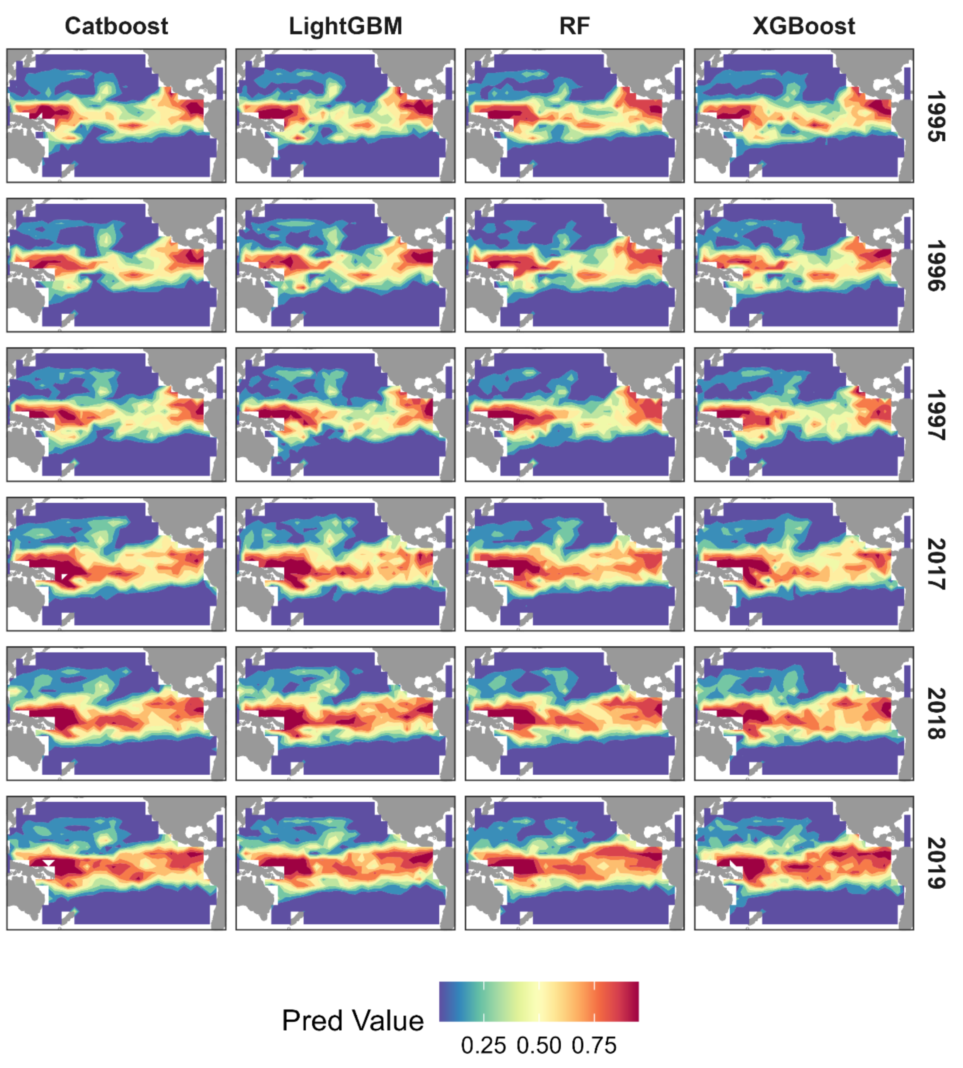
3.2 Prediction of spatiotemporal distribution of yellowfin tuna in the Pacific Ocean

Average CPUE Comparison



The average CPUE for each model was calculated for the years 1995 through 2019 and compared to the average observed CPUE for each year, showing that the projected values were nearly accurate.

Spatiotemporal distribution



During 1995-1997, yellowfin tuna was mainly distributed in the east and west sides of the Pacific Ocean, and less in the central Pacific Ocean. The distribution map reveals an eastward tendency in the fishery's centre of gravity. Each model predicts an almost similar distribution of yellowfin tuna, with no appreciable variations.

4. Conclusions

- Innovatively, we utilized the RFECV method borrowed from feature engineering to systematically and objectively collect the significant environmental variables which shape the distribution of the yellowfin tuna.
- Feature importance and permutation importance provide an intuitive and straightforward way to identify the most influential variables in a model
- Tree-based models trained on the selected predictors have high performance to capture the distribution pattern across the Pacific Ocean.
- Almost all tree-based models ranked Temp_0 as the most important variable, indicating that surface temperature is crucial for yellowfin tuna. All tree-based models consider sea surface height (SSH) an essential environmental variable, both in terms of feature importance and permutation importance.

5. Bibliography

[1] Langley, A., Briand, K., Kirby, D. S., & Murtugudde, R. (2009). Influence of oceanographic variability on recruitment of yellowfin tuna (Thunnus albacares) in the western and central Pacific Ocean. *Canadian Journal of Fisheries and Aquatic Sciences*, *66*(9), 1462–1477.

[2] Schaefer, K. M., Fuller, D. W., & Block, B. A. (2007). Movements, behavior, and habitat utilization of yellowfin tuna (Thunnus albacares) in the northeastern Pacific Ocean, ascertained through archival tag data. *Marine Biology*, *152*(3), 503–525.

[3] Erauskin-Extramiana, M., Arrizabalaga, H., Hobday, A. J., Cabré, A., Ibaibarriaga, L., Arregui, I., Murua, H., & Chust, G. (2019). Large-scale distribution of tuna species in a warming ocean. *Global Change Biology*, *25*(6), 2043–2060. https://doi.org/10.1111/gcb.14630

6. Acknowledgments

We are grateful to Jintao Wang at Shanghai Ocean University (SHOU) for the conception of research, We also thank to the data support from AIFL lab.