Random forest regression models in ecology:

Accounting for messy biological data and producing predictions with uncertainty

Caitlin I. Allen Akselrud

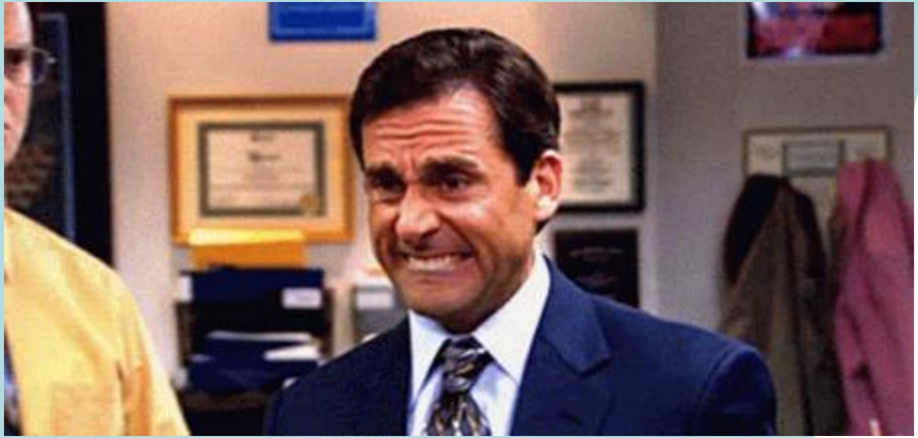
NOAA Southwest Fisheries Science Center,
Fisheries Stock Assessment
&
University of Washington,
PhD Candidate

NOAA FISHERIES

Machine learning was not designed for ecological data sets.

Ecological data has:

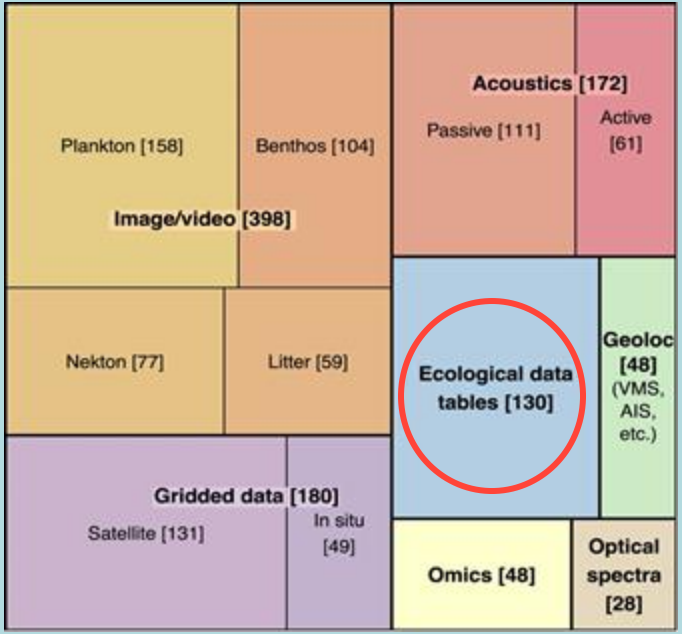
- Missing points or blocks of data
- Short data sets (few observations)
- Autocorrelation



NOAA FISHERIES

And yet, machine learning can be incredibly useful.

- Make predictions where traditional statistical models struggle
- Pick up non-linear relationships



Rubbens, P. et al.
"Machine learning in marine ecology: an overview of techniques and applications." ICES Journal of Marine Science 80.7 (2023): 1829-1853.

NOAA FISHERIES

Is machine learning for forecasting usable with problematic data sets?

- With caution and an eye for detail
- This talk covers adaptations for forecasting using ecological data with machine learning that are broadly applicable.
- Example using random forest methodology to make catch forecasts for the California market squid fishery (*Doryteuthis opalescens*)
- Additional details are published: Allen Akselrud (2024) in *Fisheries Research*.

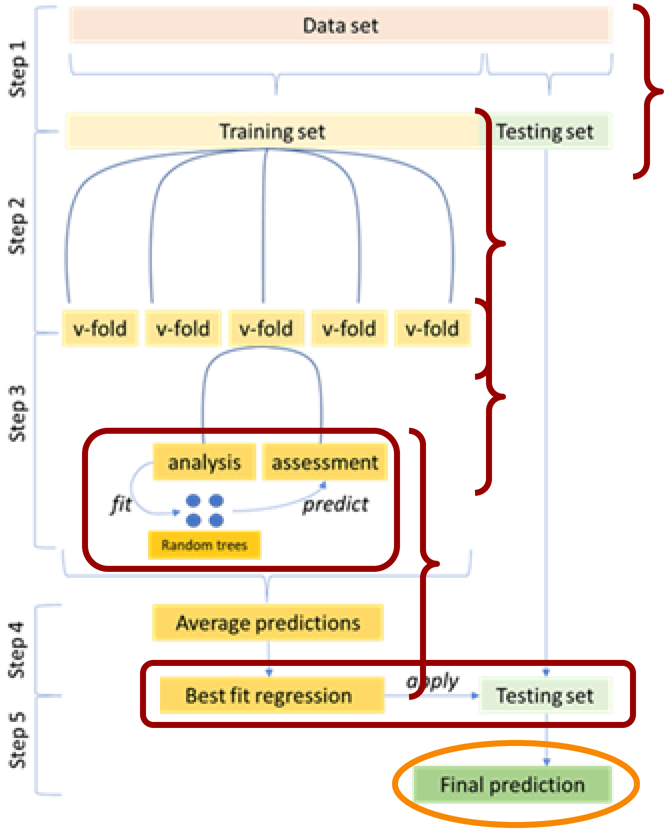
NOAA FISHERIES

Roadmap

- Start your engines: basics of random forest methodology
- Wrong turn: thoughtful failures
- Course correction: fixing problems from ecological data
    - Detour into the forest: understanding how random forests grow
- Pitstop: apply to real data
- Cruise control: how methods changes improve results
- Destination: takeaways

NOAA FISHERIES

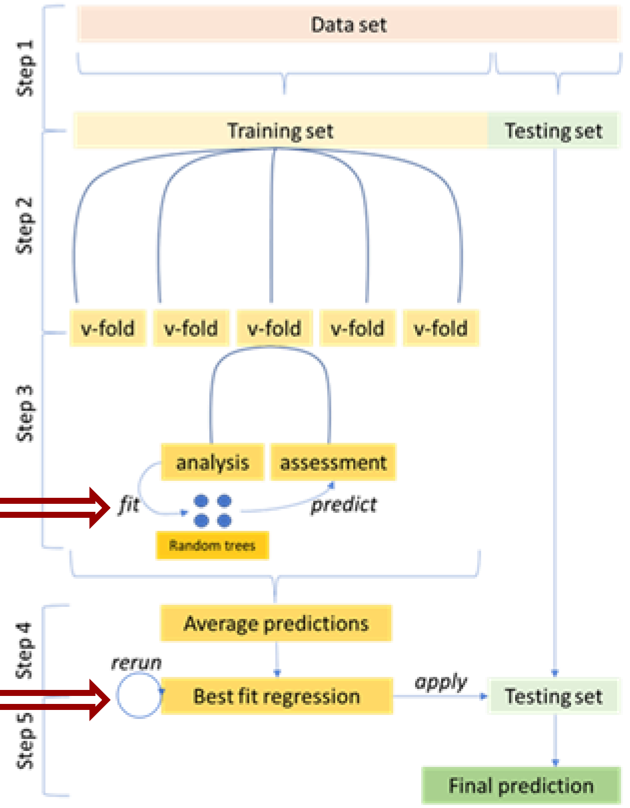
Basics of random forest methodology



Thoughtful failures: what went wrong?

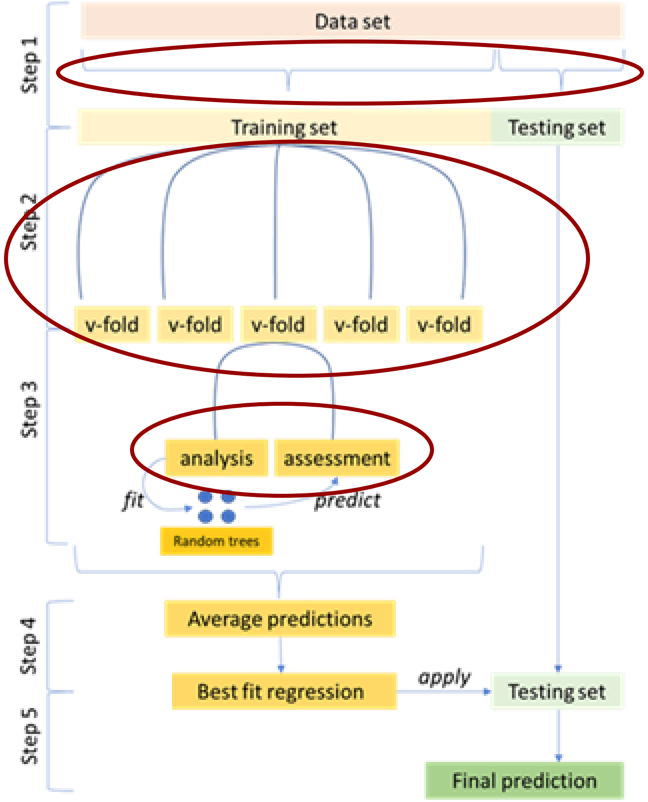
② *Getting a different answer every time: sparse data*

Epistemic uncertainty



① *Overfitting*

Thoughtful failures: what went wrong? Overfitting



Data must be split temporally, rather than randomly.

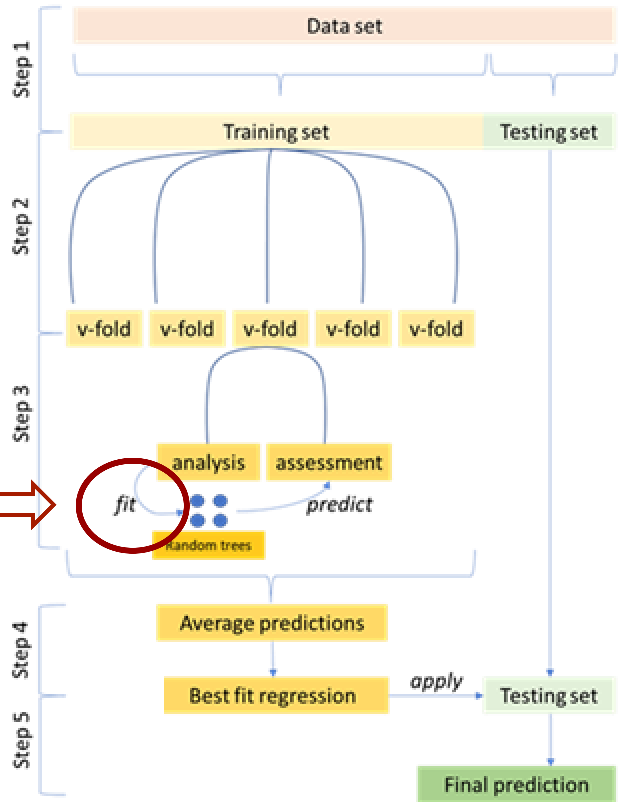
Improving the methods: Overfitting

Year	1	2	3	4	5	6	7	8	9	10	11
Fold 1	Analysis		Assess								
Fold 2											
Fold 3											
Fold 4											
Fold 5											
Fold 6											
Fold 7											
Fold 8											
Fold 9											

NOAA FISHERIES

Thoughtful failures: what went wrong? Sparse data

hyperparameter tuning



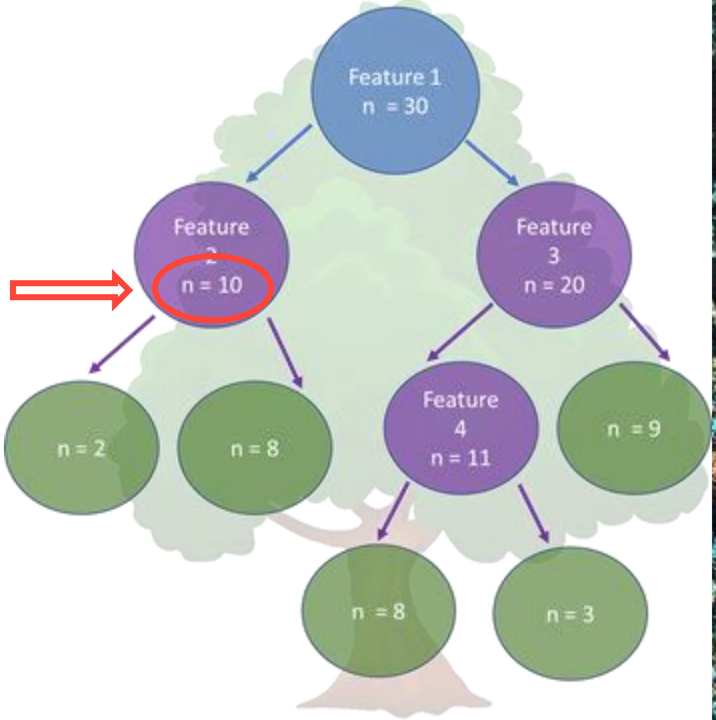
What is a random forest regression?

An ensemble of regression trees with 3 hyperparameters:

- ➢ How many data points per branch?
- ➢ How many features per tree?
- ➢ How many trees?

How do you grow a tree?

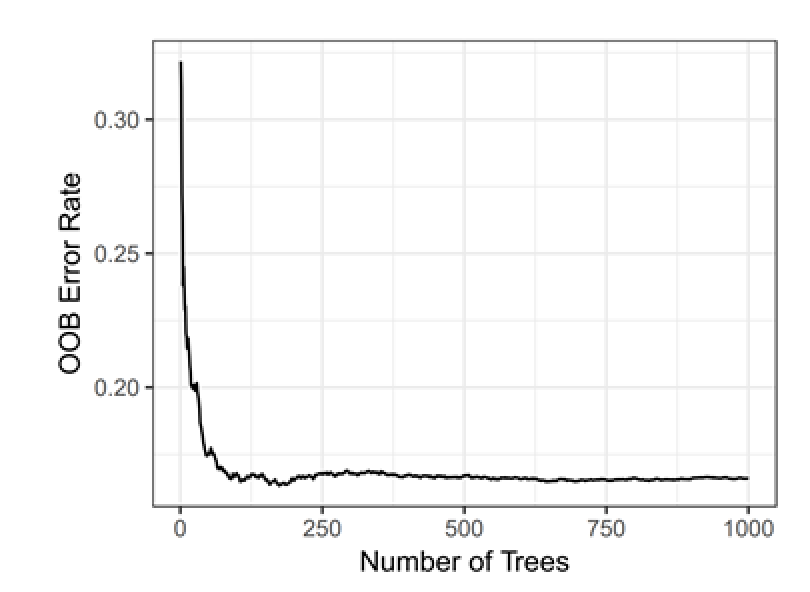
There is some minimum number of data points, below which you no longer create branches



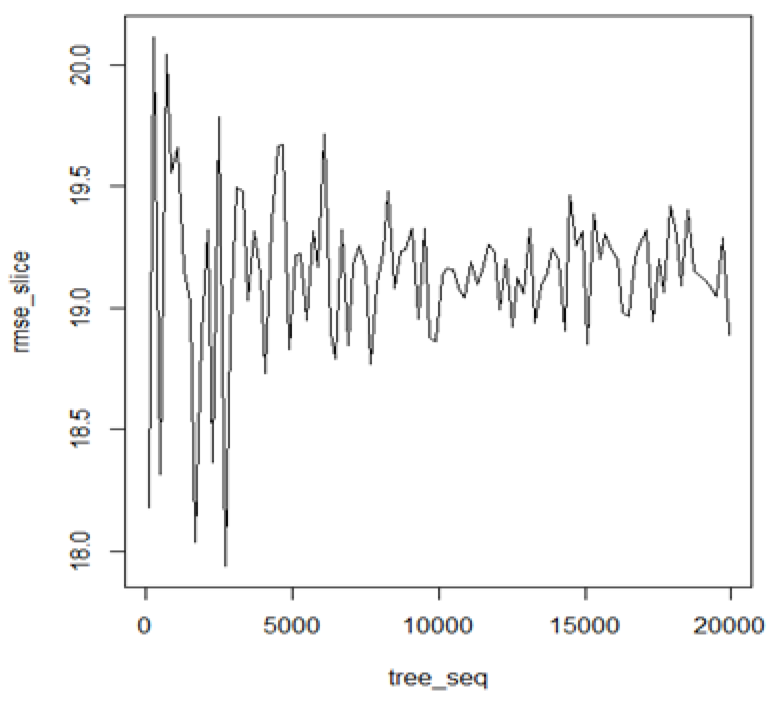
How many features?
A random selection with a minimum

All your data	Tree 1	Tree 2	Tree 3	Tree 4	Tree 5	Tree 6
Feature 1						
Feature 2						
Feature 3						
Feature 4						
Feature 5						
Feature 6						
Feature 7						
Feature 8						

How many trees?



Ehrlinger, J., 2015. ggrandomforests: Visually exploring a random forest for regression. arXiv preprint arXiv:1501.07196.



Sparse data and non-optimal tuning

Hyperparameter tuning

Minimum number of data points per branch
Minimum number of features per tree
Number of trees

Tuning:

- Over every possible combination
- Over an optimal set of combinations (without repeats) – maximum entropy

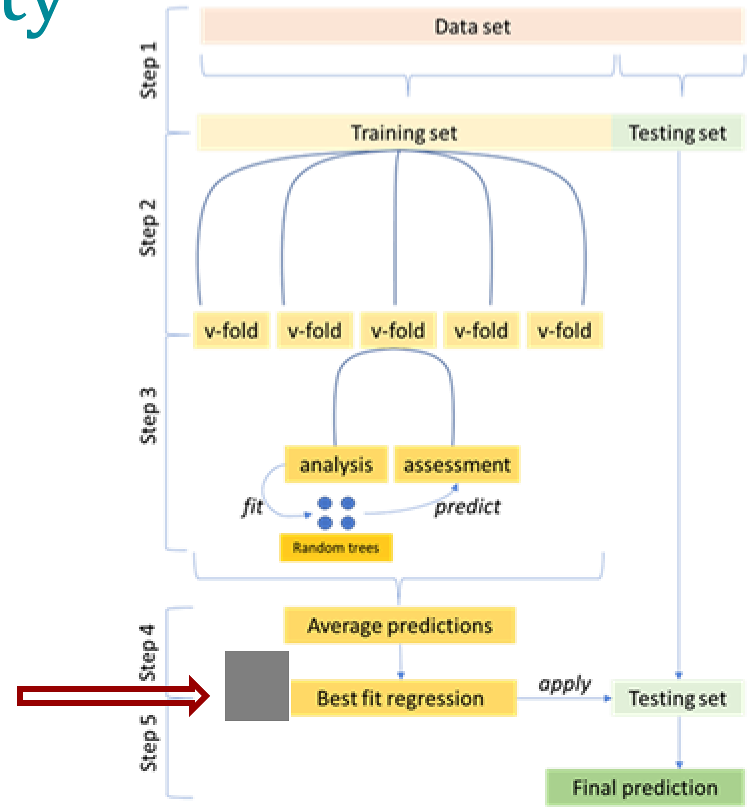
Improving the methods: Model selection

What metric do we use to select for the best set of hyperparameters?

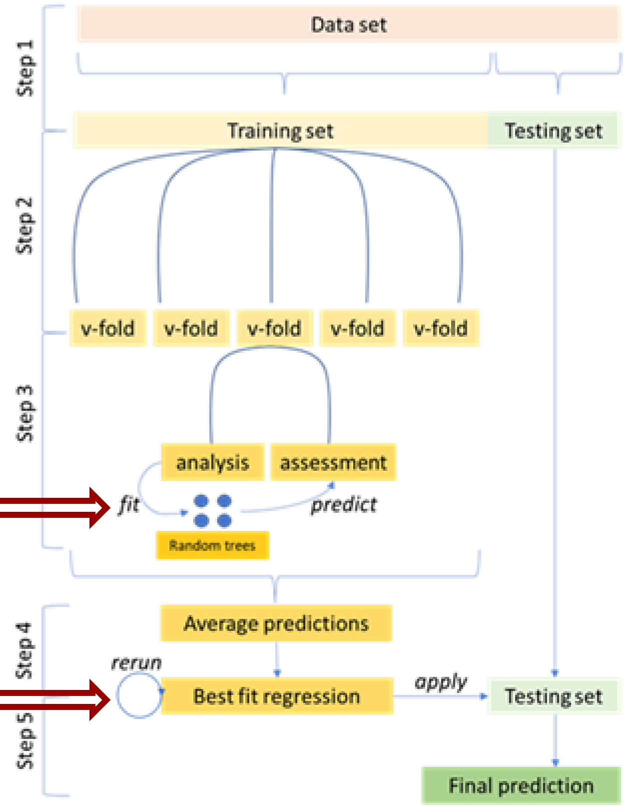
- ➢ Depends on your question
- ➢ In forecasting, we want the most precise prediction, so we may use:
    - Root mean squared error (RMSE)
    - Mean squared error (MSE)
    - Mean absolute error (MAE)
    - Ratio of performance to deviation (RPD)
- ➢ For model fit, R^2 is typically used

Improving the methods: Epistemic uncertainty

Re-run the tuned model multiple times to get your predictive range



Improving the methods

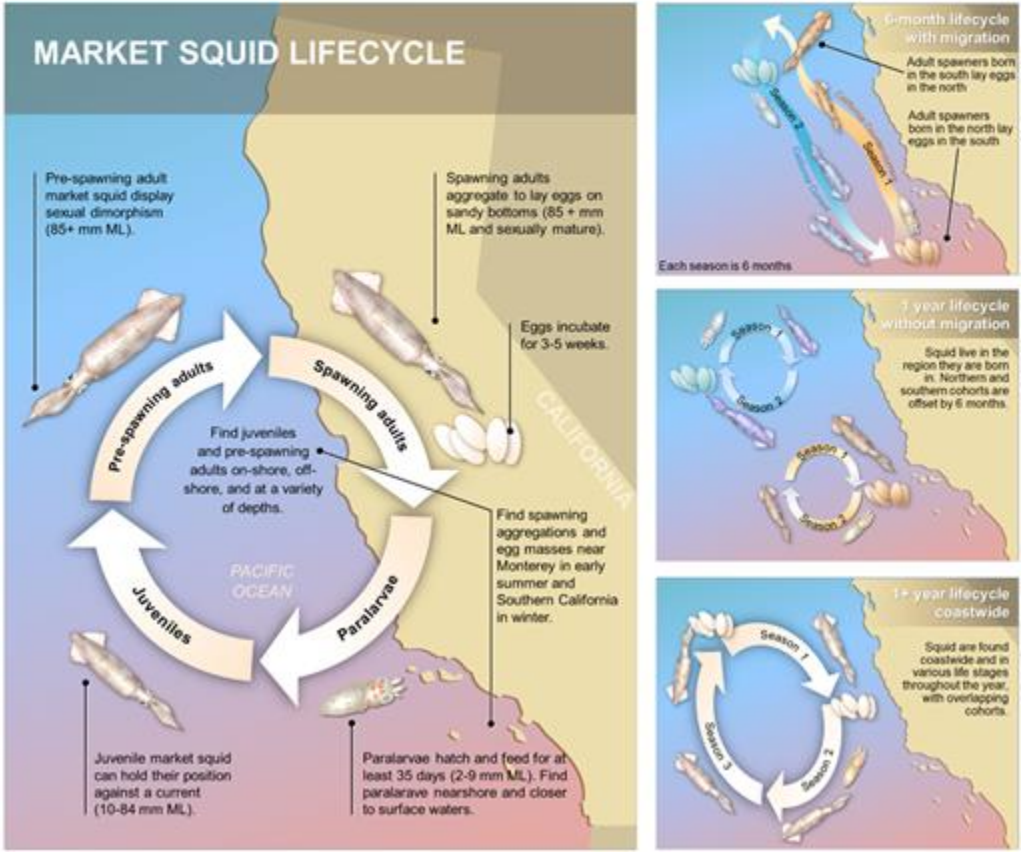


Data structuring

Hyperparameter tuning

Uncertainty

Application to fisheries: California market squid



- ➢ 3 plausible life history strategies
- ➢ Observational data
- ➢ Environmental data
- ➢ 6 model configurations (3x2): life histories with or without environment

NOAA FISHERIES

Results of methodological improvements

Random forest prediction skill Percent of predictions within a category				
Life history	Environment included	Good	Fine	Poor
Short	Yes	44	33	22
Short	No	48	33	19
Medium	Yes	41	37	22
Medium	No	41	33	26
Long	Yes	33	44	22
Long	No	30	41	30

Results of methodological improvements

Epistemic uncertainty for each model configuration				
Life History	Environment included	Minimum RMSE	Maximum RMSE	RMSE difference
Short	Yes	24.01	25.41	1.39
Short	No	23.89	27.28	3.39
Medium	Yes	24.01	26.39	2.38
Medium	No	27.00	30.09	3.09
Long	Yes	23.88	26.04	2.16
Long	No	26.61	28.63	2.02

NOAA FISHERIES

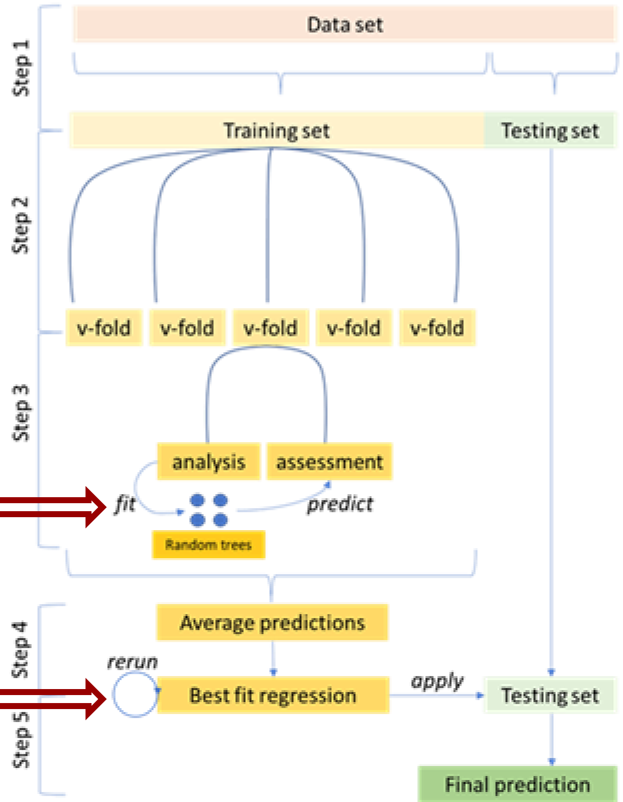
Future work

- Simulation study over more problems with ecological data
- The implications of applying better practices (or failing to) for each type of data problem
- Come chat with me if you have suggestions…

Takeaways

"The increased flexibility and accessibility of random forest methods does not mean that they can be blindly applied to any kind of data without caution" (Boulesteix et al., 2012).

Takeaways



Data structuring

Hyperparameter tuning

Uncertainty

Thank you for your time and attention!

Please feel free to get in touch with me:

Caitlin Allen Akselrud

caitlin.allen_akselrud@noaa.gov
text/call: 858-546-5613

For more details, please see:

Akselrud, C.I.A., 2024. Random forest regression models in ecology: Accounting for messy biological data and producing predictions with uncertainty. Fisheries Research, 280, p.107161.

NOAA FISHERIES