# Predicting Missing Values in Biodiversity Datasets Using Phylogenetics and Spatial Mapping

Jay Jain, Quentin Read, Andrew Finley, Phoebe Zarnetske

# Motivation

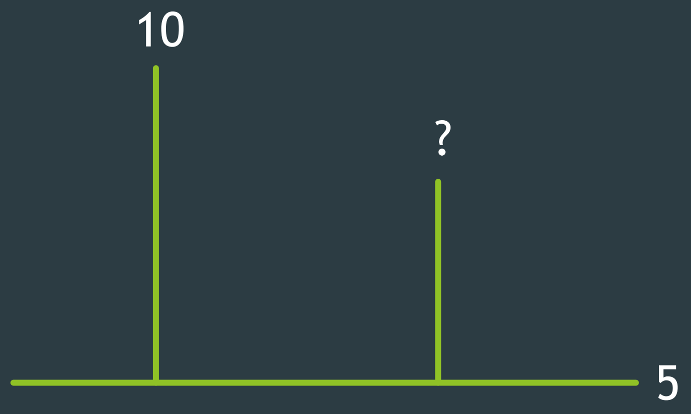
- Complete datasets are important for ecologists to understand biodiversity in age of climate change.
- What causes variation in diversity across space and time?
- Ecological datasets with many missing values have become the norm.
- Methods to impute missing values exist,

  but do not provide optimal estimates.
- Imputing trait values in tree species from

  West coast of USA
  - Dataset: Averages of 6 traits from 64 species

# Methods

- Phylogenetic imputation



- Multivariate imputation

| | Trait 1 | Trait 2 |
|---|---|---|
| **Species 1** | **5** | **?** |
| **Species 2** | **10** | **20** |

- Experiment: Removing 25% of known values at random from dataset and trying to impute missing values for each of three methods.

| | **Trait 1** | **Trait 2** | **Trait 3** | **Trait 4** | **Trait 5** |
|---|---|---|---|---|---|
| **Species 1** | 24 | **?** | 5 | 85 | 56 |
| **Species 2** | 46 | 26 | **?** | 23 | 34 |
| **Species 3** | 25 | **?** | 8 | 2 | 762 |
| **Species 4** | 28 | 37 | 23 | 657 | 37 |
| **Species 5** | 2 | **?** | 567 | 46 | **?** |
| **Species 6** | **?** | 37 | 26 | **?** | 234 |
| **Species 7** | 7 | **?** | 4 | 35 | 83 |
| **Species 8** | 73 | 72 | **?** | **?** | 26 |

# Computing Tools

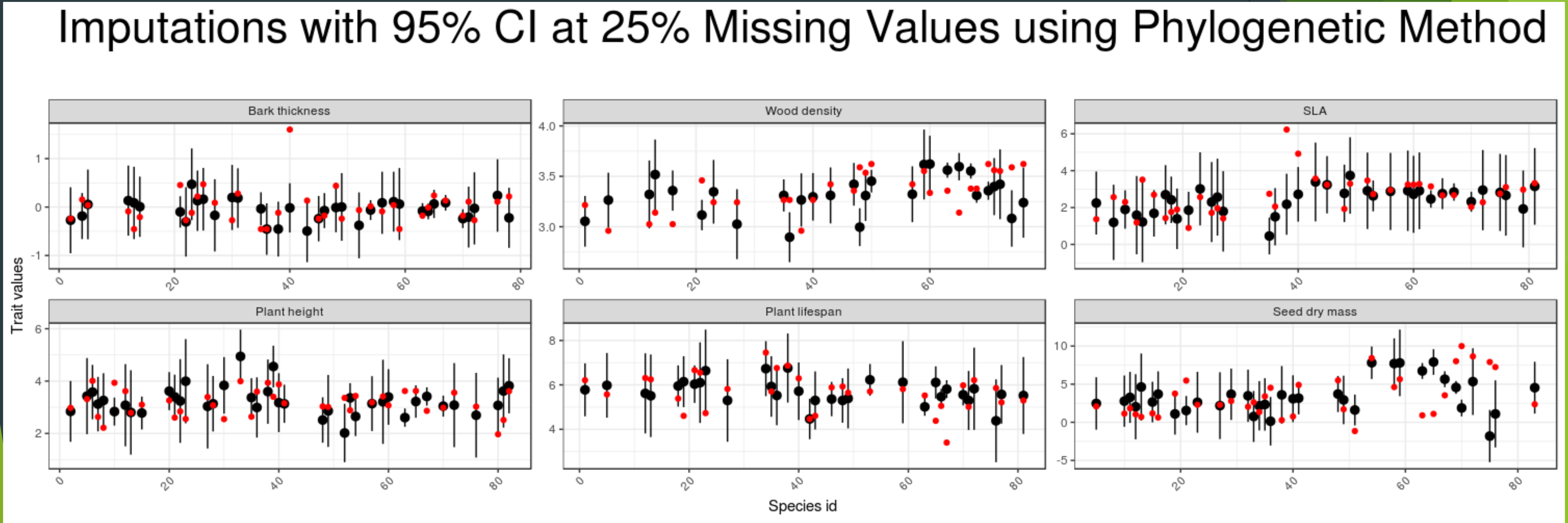
HPCC
High Performance Computing Center

**R**
r-project.org

**STAN**
mc-stan.org

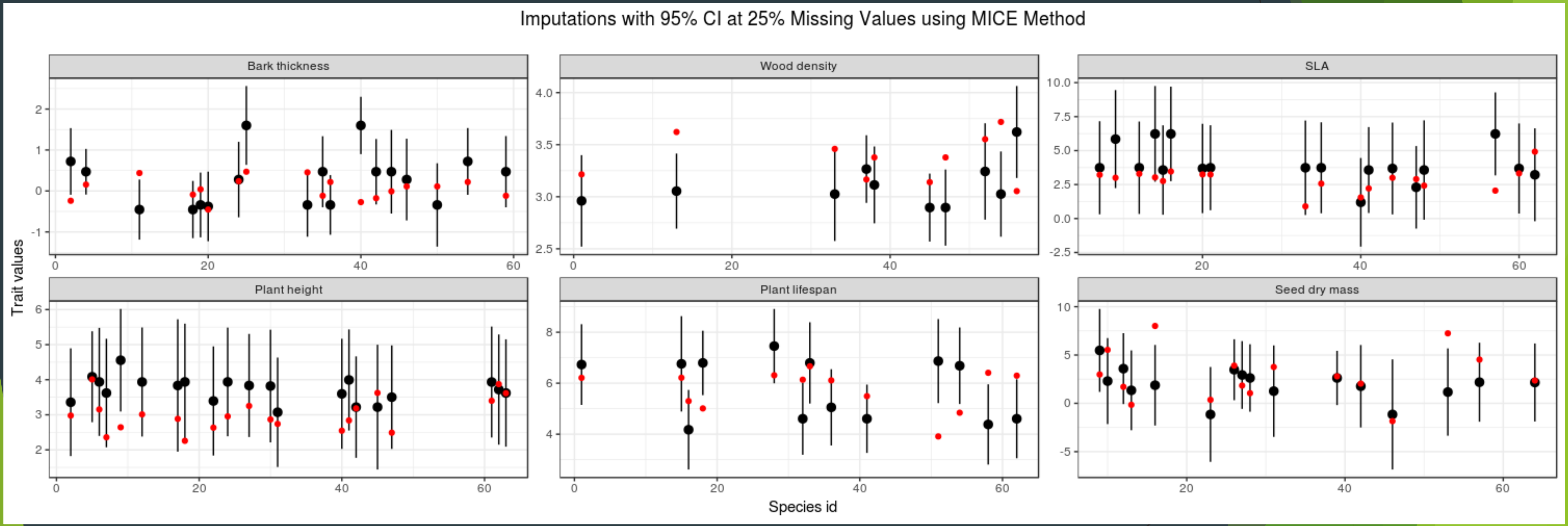
**HPC**
icer.msu.edu

# Phylogenetic Imputation



Imputations with 95% CI at 25% Missing Values using Phylogenetic Method

# Multivariate Imputation



# Combined Model

$$y \sim X\beta + Z\alpha + \varepsilon$$

y = imputed traits

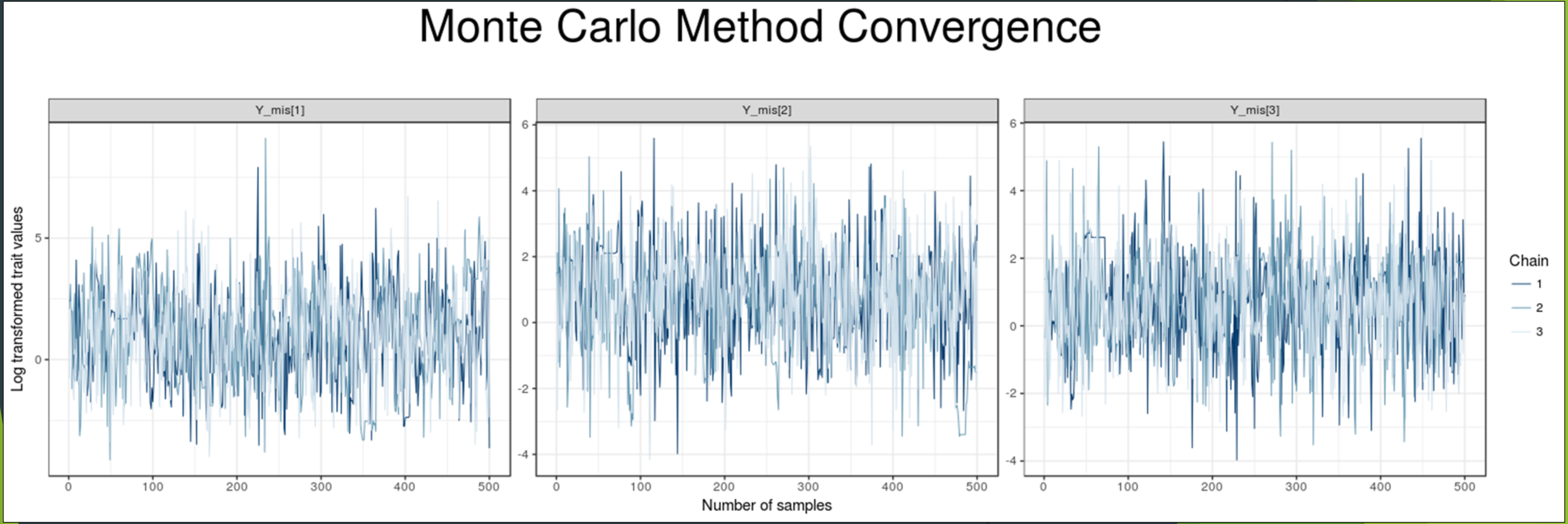
X = predictor matrix

β = fixed effects/slope

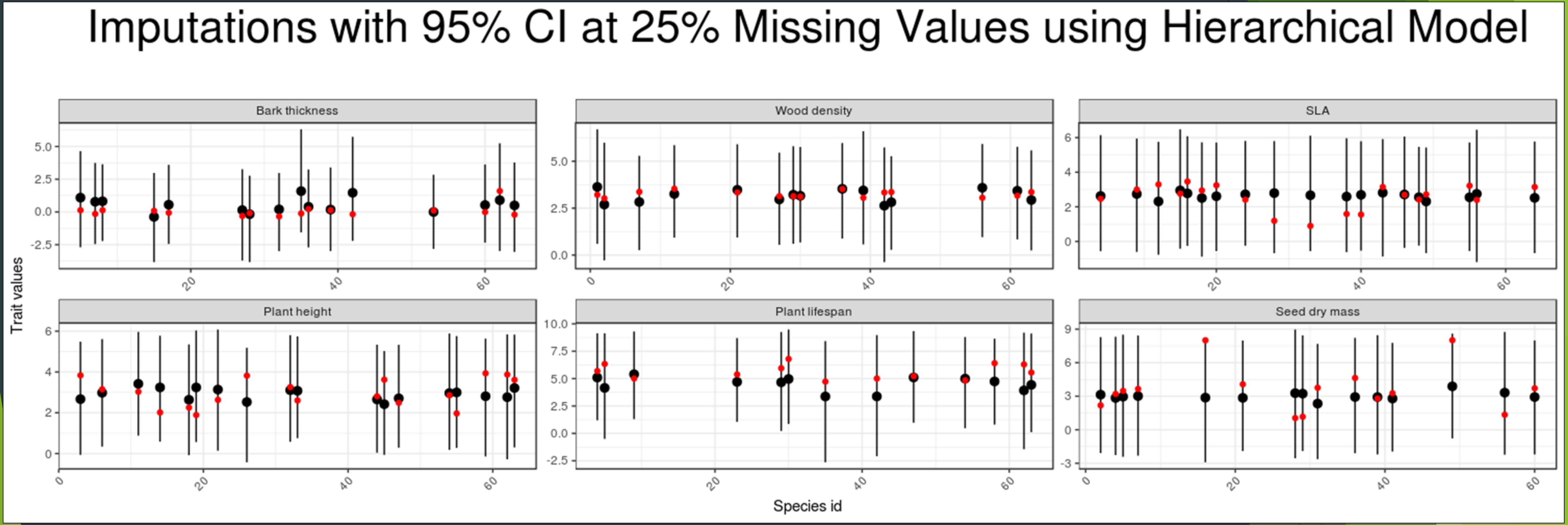
Z = identity matrix

α = phylogenetic random effects

ε =  residuals

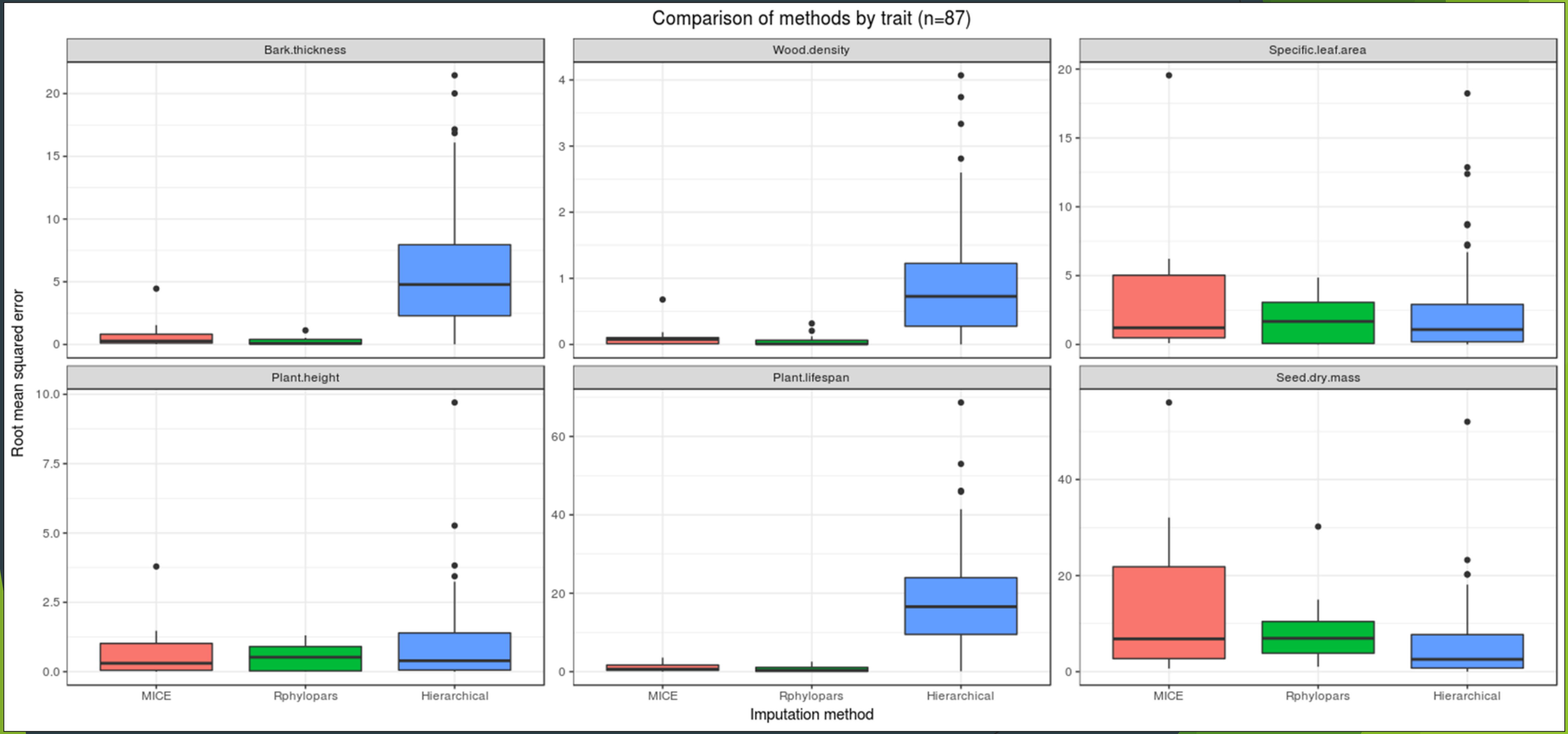


# Hierarchical Model



- Model accounts for phylogenetic relatedness, trait covariance, and environmental predictors.
- Bayesian method fully characterizes uncertainty.

# Comparison of Methods



# Conclusion

- The model worked better for imputing some, but not all traits compared to phylogenetic and multivariate imputation methods implemented in isolation.
  - Used RMSE values as metric, where RMSE is average deviation of imputed value from true value.
  - RMSE = 0 is optimal
- May be because plant lifespan, bark thickness, and wood density have a large range of possible values.

# Future Work

- Test different environmental and climate predictors
- Include spatial random effects
- Increase number of iterations in Monte Carlo method
- Test data sets with different traits and species

# QUESTIONS?