Supplemental Information

steppe

Figure 1 - Site Info

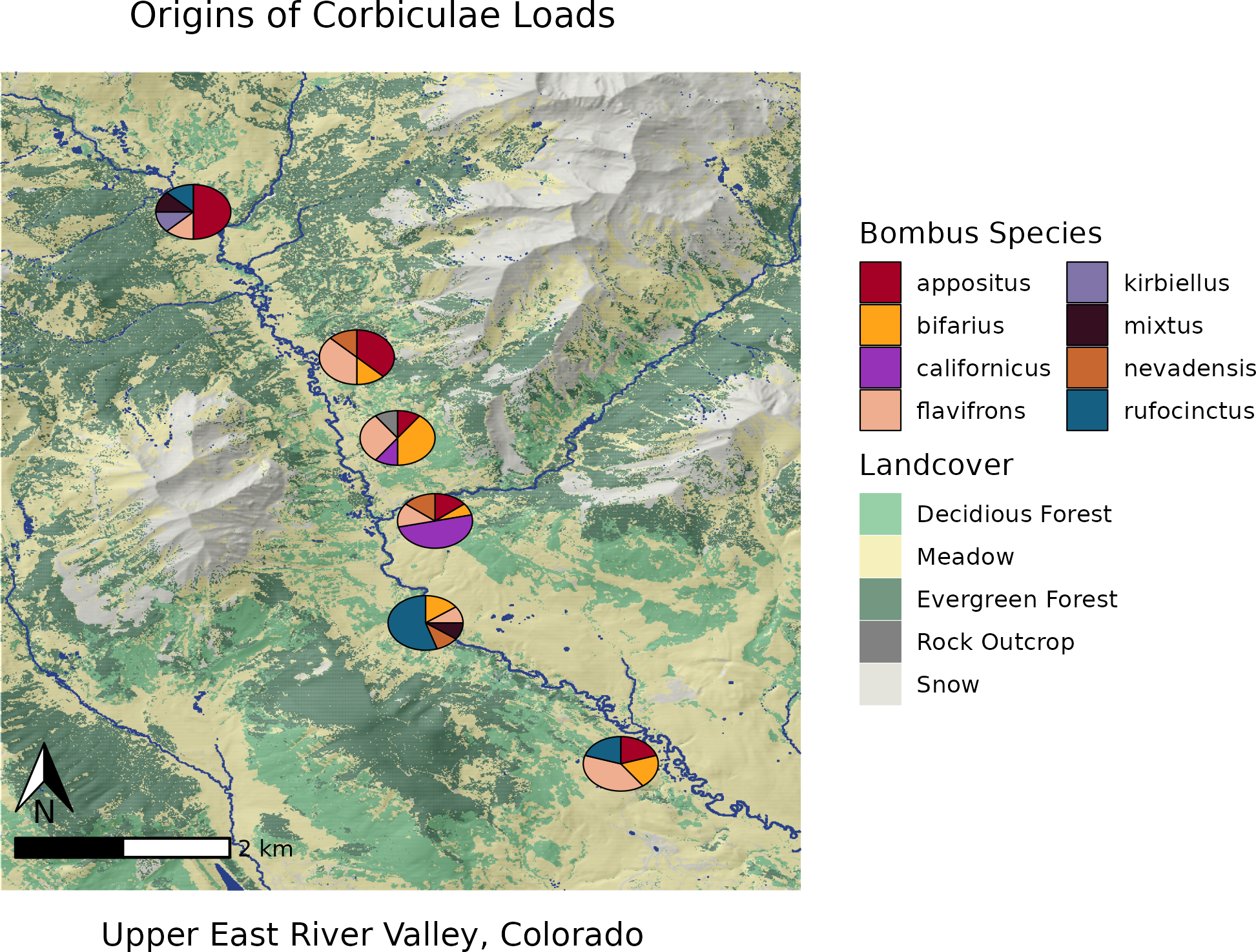


Table 2 - Models used for Species Distribution Model Ensembles

The two machine learning models utilize Ensemble learning.

**Ensemble learning** utilizes many sets of trees, each tree being composed of many binary decisions, to create a single model. Each independent variable ( - or *feature*) may become a node on the tree - i.e. a location on the tree where a binary decision will move towards a predicted outcome. Each of the decision tree models which ensemble learning utilizes is a weak model, each of which may suffer due to high variance or bias, but which produce better outcomes than would be expected via chance. When ensembled these models generate a strong model, a model which should have more appropriately balanced variance and bias and predicts outcomes which are more strongly correlated with the expected values than the individual weak models.

***Random Forest (RF)*** the training data are continually bootstrap re-sampled, in combination with random subsets of features, to create nodes which attempt to optimally predict a known outcome. A large number of trees are then aggregated, via the most common predictions, to generate a final classification prediction tree. Each individual prediction tree is generated independently of the others.

***Boosted Regression Tree (BRT)*** (or Gradient Boosted tree) An initial tree is grown, and all other trees are derived sequentially from it, as each new tree is grown the errors in responses from the last tree are weighed more heavily so that the model focuses on selecting dependent variables which refine predictions. All response data and predictor variables are kept available to all trees.

***Bias*** predictions from an algorithm are systematically in error due to being prejudiced for or against certain results, due to assumptions during learning.

***Variance*** errors in models due to an over-reliance and sensitivity of training to outliers in training data.

In general, Random Forest models have high bias and low variance, where boosted regressions trees have lower bias and higher variance. Theoretically, the weaknesses and strengths of bootstrap aggregation (bagging) as implemented by Random Forests are supplemented by the boosting.

Table 3 - Species Distribution Models Predictors

| Layer | LM | Description | Name | Source |
| --- | --- | --- | --- | --- |
| 1. | N | Mean annual cloudiness - MODIS | Cloud Cover (EarthEnv) | Wilson et al. 2016 |
| 2. | Y | Cloudiness seasonality 1 - MODIS | Cloud Cover (EarthEnv) | Wilson et al. 2016 |
| 3. | N | Cloudiness seasonality 2 - MODIS | Cloud Cover (EarthEnv) | Wilson et al. 2016 |
| 4. | Y | Cloudiness seasonality 3 - MODIS | Cloud Cover (EarthEnv) | Wilson et al. 2016 |
| 5. | N | Beginning of the frost-free period | ClimateNA | Wang et al. 2016 |
| 6. | N | Climatic moisture deficit | ClimateNA | Wang et al. 2016 |
| 7. | N | Degree-days above 5C | ClimateNA | Wang et al. 2016 |
| 8. | N | Mean annual precipitation | ClimateNA | Wang et al. 2016 |
| 9. | Y | Mean annual precipitation as snow | ClimateNA | Wang et al. 2016 |
| 10. | Y | Temperature seasonality | ClimateNA | Wang et al. 2016 |
| 11. | Y | 2015 Percent Grass/Herbaceous cover - MODIS | (MOD44B) |  |
| 12. | Y | 2015 Percent Tree cover from Landsat 7/8 | (GLCC) |  |
| 13. | Y | Soil probability of bedrock (R Horizon) | SoilGrids | Hengl et al. 2017 |
| 14. | N | Soil organic carbon (Tonnes / ha) | SoilGrids | Hengl et al. 2017 |
| 15. | N | Surface soil pH in H2O | SoilGrids | Hengl et al. 2017 |
| 16. | Y | Surface soil percent sand | SoilGrids | Hengl et al. 2017 |
| 17. | Y | Soil USDA class | SoilGrids | Hengl et al. 2017 |
| 18. | N | Topographic elevation | Topography (EarthEnv) | Amatulli et al. 2018 |
| 19. | Y | Topographic elevation, moving window. | Topography (EarthEnv) | Amatulli et al. 2018 |
| 20. | Y | Topographic percent slope | Topography (EarthEnv) | Amatulli et al. 2018 |
| 21. | Y | Topographic wetness index | Topography (EarthEnv) | Amatulli et al. 2018 |
| 22. | Y | Topographic aspect | Topography (EarthEnv) | Amatulli et al. 2018 |
| 23. | Y | Annual potential solar radiation computed | r.sun |  |
| 24. | N | Estimated actual (w/-cloud) solar radiation | r.sun | Wilson et al. 2016 |
| 25. | Y | Log-transformed distance to surface water | Global Surface Water Explorer | Pekel et al. 2016 |
| 26. | Y | Percent surface water | Global Surface Water Explorer | Pekel et al. 2016 |

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data volume 5, Article number: 180040. DOI: <doi:10.1038/sdata.2018.40>.

Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, et al. (2017) SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12(2): e0169748. <https://doi.org/10.1371/journal.pone.0169748>

Pekel, JF., Cottam, A., Gorelick, N. et al. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422 (2016). <https://doi.org/10.1038/nature20584>

Wang T, Hamann A, Spittlehouse D, Carroll C (2016) Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. PLoS ONE 11(6): e0156720. <https://doi.org/10.1371/journal.pone.0156720>

Wilson AM, Jetz W (2016) Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions. PLoS Biol 14(3): e1002415. <https://doi.org/10.1371/journal.pbio.1002415>

Table 4 - Molecular Reference Specimen Table

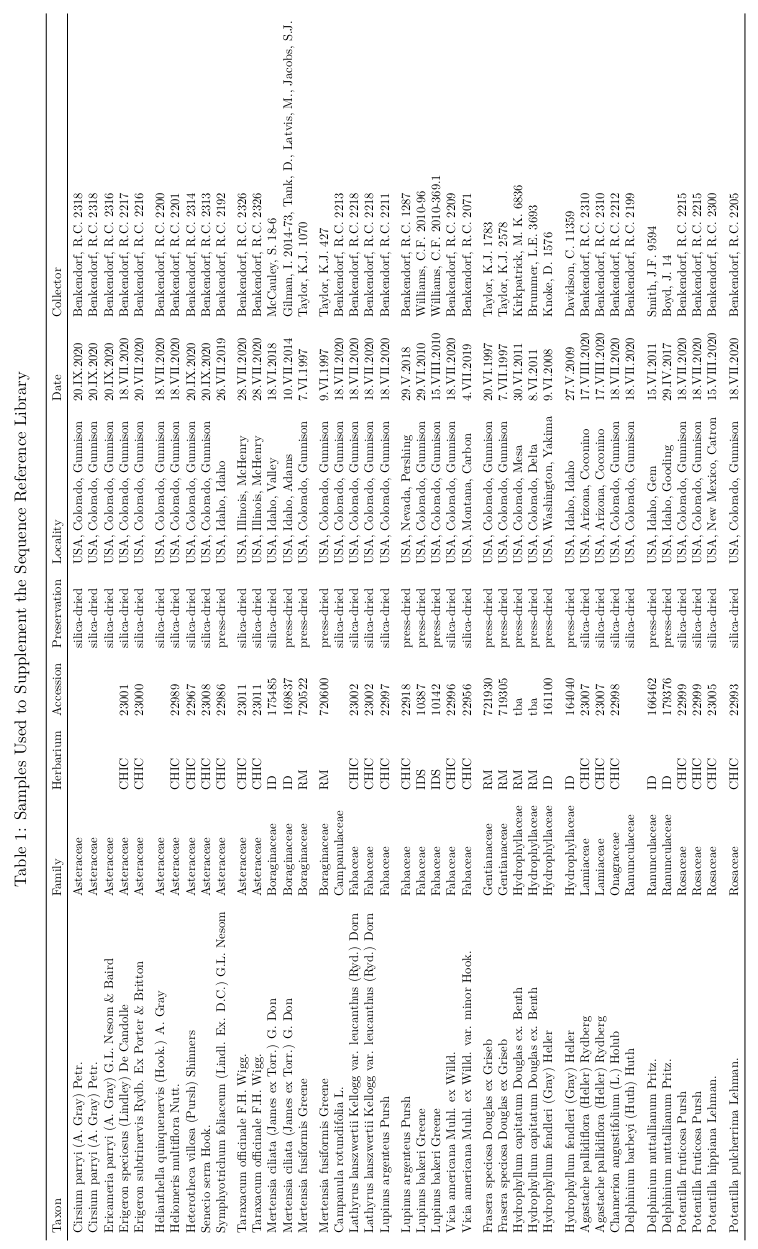


Figure 2 - Pollen Key

![](data:application/pdf;base64,)

![](data:application/pdf;base64,)

Table 5 - All Species in the Sequence Databases

![](data:application/pdf;base64,)

![](data:application/pdf;base64,)

![](data:application/pdf;base64,)

![](data:application/pdf;base64,)

Figure 3 - Reads Per Loci



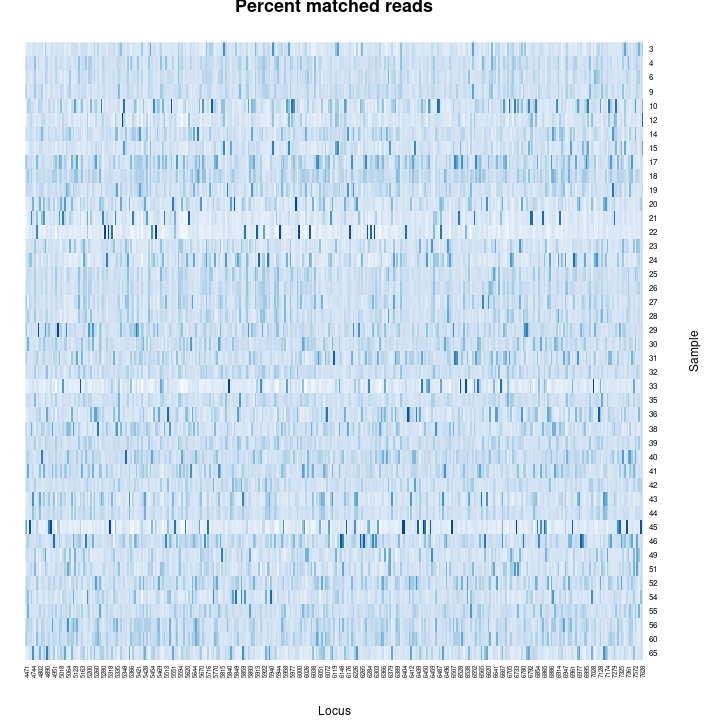


Figure 4 - Comparison of Kraken2, Bracken, and BLAST

![](data:application/pdf;base64,)