Provincial Site Productivity Layer Version 9.0

Technical Report

Prepared by

Ministry of Forests
Forest Analysis and Inventory Branch

May 2026

Version 1.0

Digital copies of this report are available at:

https://www2.gov.bc.ca/gov/content/industry/forestry/managing-our-forest-resources/forest-inventory/site-productivity/provincial-site-productivity-layer

For further information on the British Columbia Forest Inventory Program, visit the Forest Inventory website at:

https://www2.gov.bc.ca/gov/content/industry/forestry/managing-our-forest-resources/forest-inventory

Executive Summary

Site productivity is a primary driver of stand growth in the Timber Supply Review (TSR) process. As such, accurate site productivity estimates are key to achieving greater planning certainty, and a more transparent and reliable fibre supply.

The Provincial Site Productivity Layer (PSPL) v.8.0 currently used in TSR has several known limitations: reliance on a model that assumes linear relationships between environmental conditions and site productivity, a modelling framework that is difficult to update, gaps in species-specific site index estimates across large areas, and — as demonstrated by the Young Stand Monitoring (YSM) ground sampling program — significant differences from field-measured productivity in several TSAs.

To address these limitations, the Forest Analysis and Inventory Branch (FAIB) is modernizing the PSPL through a machine learning-based framework that integrates over 13,000 ground-based site index observations from FAIB ground sampling inventories, Site Index Biogeoclimatic Ecosystem Classification (SIBEC), and Site Index Adjustment (SIA) projects. Random Forest (RF) models were selected for their strong predictive performance and handling of complex datasets. This approach incorporates climatic, topographic, and ecosystem classification predictor variables to better capture the non-linear relationships between environmental conditions and species productivity.

The species-specific RF prediction models were applied at the provincial scale to produce site productivity spatial layers for 14 tree species. When validated against independent YSM ground sampling data, the RF models outperformed the PSPL v.8.0 for all species except black spruce, with particularly strong improvements for key commercial species including Douglas-fir, lodgepole pine, western hemlock, western redcedar, and subalpine fir. The models were applied at the provincial scale to produce version 9.0 of the PSPL, replacing version 8.0.

Beyond the immediate gains in site index prediction accuracy, this update establishes a reproducible and scalable predictive modelling framework that can be readily updated and improved as new ground inventory data or predictor variables become available.

Table of Contents

1. Background

The Provincial Site Productivity Layer (PSPL) v.8.0 currently provides site index (SI) estimates for 22 commercial tree species and has served as a strategic-level input for estimating growth potential of even-aged young stands in the provincial TSR process. However, the current PSPL version suffers from several well-recognized limitations:

- The PSPL v.8.0 integrates ecosystem mapping (PEM/TEM), SIBEC data, and biophysical model outputs. This dependency on a variety of mapping efforts and models managed by multiple programs has resulted in a framework that is difficult to maintain or update.
- The biophysical model (BM) is a regression-based model and was meant to serve as a stop-gap supplemental solution for areas with no PEM/TEM coverage (Nigh, 2012). The BM is based on a limited set of covariates and assumes that the relationship between the predictor variables and SI is linear, which oversimplifies non-linear growth responses to environmental conditions.
- The structure of the PSPL has led to large portions of the province lacking SI values for species that are present on the ground. Processing of the PSPL by Forest Analysis and Inventory Branch (FAIB) for input into managed stand yield tables (MSYT) has previously found that the PSPL only has values for 27% of the required species-specific SI values.[1]
- Data obtained through the FAIB's Young Stand Monitoring (YSM)[2] reports indicate that PSPL SI estimates deviate significantly from field-measured productivity in several management units, with implications on yield projections.

To address these limitations, FAIB is modernizing the PSPL with the objective of establishing a reproducible and scalable framework for predicting site productivity at the provincial scale. This involves integrating ground inventory data with machine learning methods that can model the complex relationships between site environmental conditions and SI and apply the predictions at scale. Machine learning approaches are well-suited to this task because they can incorporate large datasets and capture complex, non-linear patterns in growth response that traditional regression-based methods, such as those used for the PSPL v.8.0 BM, are unable to represent effectively. This approach offers the added benefit that, as new ground data or predictor variables become available, models can be readily updated and improved without the need to rebuild the underlying framework.

[1] FAIB MSYT PSPL processing: https://github.com/bcgov/FAIB_PSPL/blob/main/05_site_index_conversions/01_PSPL_si_conversion_v3.1.md

[2] Provincial Inventory Monitoring reports: https://www2.gov.bc.ca/gov/content/industry/forestry/managing-our-forest-resources/forest-inventory/ground-sample-inventories/provincial-monitoring/reports

2. Approach

2.1. Modelling site productivity with supervised machine learning

Site productivity is determined by the climatic, edaphic and topographic conditions that regulate the availability of water, nutrients and light necessary for tree growth (Mah and Nigh, 2015). These environmental drivers are collinear and exhibit non-linear relationships with site productivity that are difficult to capture with conventional parametric models. Machine learning methods are well-suited to this problem because they can accommodate large numbers of correlated predictors, can handle non-linear predictor-response relationships, and can detect complex interactions among predictor variables. Ensemble machine learning methods in particular, such as random forest, are also robust to outliers, and reduce the risk of overfitting by aggregating across multiple models.

To model and predict site productivity at the provincial scale, a random forest model was trained on ground-sampled SI observations and spatially continuous, province-wide coverage of topographic, climatic, and ecosystem classification predictor variables. After validation against an independent dataset, the model was then applied at the provincial scale to produce species-specific estimates of site productivity at a 1 ha resolution.

2.2. Ground-based SI estimates

The SI ground measurements used for training the model were sourced from FAIB forest inventory ground sample data (YSM and non-YSM), SIBEC, and SIA projects:

- **FAIB YSM:** Tree data from FAIB YSM ground sampling were used as a holdout validation set not used for model training. YSM ground sampling points are established on a systematic grid across young stands (15-50 years of age). The YSM ground sample data were obtained from the latest published version of the forest inventory ground plot data available on the BC Data Catalogue (Version 2, compiled 2025-05-14). YSM sampling is ongoing.
- **FAIB non-YSM:** Tree data from FAIB non-YSM ground plots were used for model training. FAIB non-YSM ground plots are from projects with varying sample designs (systematic, random, subjective). Tree data from the FAIB non-YSM ground plots was obtained from the latest published version of the forest inventory ground sample data available on the BC Data Catalogue (Version 2, compiled 2025-05-14). FAIB non-YSM sampling is ongoing.
- **SIBEC:** SIBEC (EP1215) has been ongoing since 1994 and provides tree species SI estimates by site series for each biogeoclimatic subzone/variant. SIBEC ground plots were subjectively established in targeted young stands. Individual tree data from the SIBEC 2024 database was used. The latest SIBEC sampling occurred in 2023, and at the time of reporting, no additional sampling is planned.
- **SIA**: Site index adjustment projects were designed and implemented by J.S. Thrower & Assoc. / Timberline Natural Resource Group between 1994-2011 and focused on sampling post-harvest regenerated stands in targeted management units. Tree data from the 2012 April 05 version of the SIA database was used for this project. The latest SIA sampling occurred in 2009 and no additional sampling is planned.

All the above SI ground measurements met the SIBEC tree selection sampling standards[3] and included species, lab counted BH age (years), total height (m), and post-processed GPS coordinates. Ground sample data processing steps for all sources are listed in Appendix A. Table 1 shows the total number of ground-based SI observations by species and by source after data processing.

Table 1. Number of ground-based SI observations available for modelling after data preparation.

Species	FAIB YSM	FAIB Non-YSM	SIA	SIBEC	Total
PL	422	108	795	2,262	**3,587**
SW	184	126	378	1,712	**2,400**
FD	136	68	779	1,336	**2,319**
BL	134	143	163	1,264	**1,704**
HW	192	43	649	611	**1,495**
SE	72	42	104	541	**759**
CW	76	17	190	446	**729**
BA	38	10	95	255	**398**
LW	9	16	85	255	**365**
SS	51	2	95	199	**347**
AT	47	31	2	251	**331**
SB	7	64	0	69	**140**
EP	21	12	6	99	**138**
PY	2	1	0	98	**101**
DR	25	4	0	31	**60**
AC	27	13	0	11	**51**
PW	4	0	3	33	**40**
HM	2	10	0	0	**12**
YC	6	1	0	4	**11**
BG	0	0	0	9	**9**
LT	2	4	0	2	**8**
PA	0	2	0	3	**5**
LA	1	1	0	0	**2**
MB	2	0	0	0	**2**
Total	**1,460**	**718**	**3,344**	**9,491**	**15,013**

The R package SindexR v2.0 was used to calculate SI for all observations (except lodgepole pine[4]) based on given tree height and breast height age. Growth intercept (GI) curves were used to calculate SI in instances where AGE_BH ≤ 50 years and where a GI curve exists for the species, otherwise SI curves

[3] SIBEC sampling and data standards: https://www2.gov.bc.ca/assets/gov/environment/plants-animals-and-ecosystems/ecosystems/sibec-documents/standards.pdf

[4] At the time of data preparation, the SIndexR package (SINDEX33.DLL ver 1.52) did not include the Lodgepole Pine Nigh (2017) SI curve that is now available in SINDEX33.DLL v1.54 currently used in SiteTools v4.4. In order to make sure the most current SI curve was used, all lodgepole pine observations were re-calculated using the SiteTools v.4.4 (Sindex dll v1.54) and updated in the model input datasets.

were used. For SI calculation purposes, coastal curves are used where the ground sample is in BEC zones CWH, MH, or CDF.

2.3. Species replacements and conversion equations

Only species with more than 100 ground sampling observations were modelled for this project. The species that were not modelled, with the exception of red alder, are all lacking SI/GI curves. For these species, it is recommended to use the SI estimates of equivalent species as a substitution (Table *2*).

Table 2. Recommended sources of SI substitutes for species not modelled.

Target Species	Reference Species
AC	AT
PW	SS (coast), FD (interior)
HM	HW
YC	CW
BG	BA
LT	LW
PA	PL
LA	LW
MB	DR

In the case of red alder, a direct SI substitution with Douglas-fir was found to produce higher SI values than expected; the following species conversion equation was developed in consultation with Ministry researchers[5]:

- For dry coastal subzones CWHds, CWHxm and CWHdm: red alder SI = 0.55 * Douglas-fir SI
- For all other coastal subzones: red alder SI = 0.73 * Douglas-fir SI

2.4. Predictor variables

2.4.1. Climate grids and bioclimatic variables

Downscaled climate grids for the province were obtained from the climr R package (Daust et al., 2024). Climr employs a change-factor downscaling approach, using BC PRISM as the provincial reference climatology, to spatially refine coarse-resolution data. The provincial TRIM 25m Digital Elevation Model (DEM) was resampled to the 100m HectaresBC grid and used as the elevation input to produce historical observational data for the 2001-2020 period.

Climr generates over 200 monthly, seasonal, and annual climatic variables. Annual and seasonal variables were used directly as model inputs, while monthly temperature and precipitation variables were used to derive the 19 WorldClim bioclimatic variables (Appendix B.), a set of standardized indices representing biologically meaningful aspects of climate, such as temperature seasonality and continentality. Bioclimatic variables were calculated using the R package dismo (Hijmans et al., 2024).

2.4.2. Topographic variables

The TRIM 25m DEM for the province was resampled to the HectaresBC 100m grid and used to calculate topographic indices relevant to site productivity, including relative slope position, slope, aspect,

[5] PSPL Processing 2023 Red Alder conversion equations: https://bcgov.github.io/FAIB_PSPL/#alder

topographic wetness index, among others. A total of 33 topographic variables were calculated using geomorphometric tools in Whitebox R and SAGA GIS (Appendix C.)

Topographic variables that are influenced by neighbourhood size, such as ElevPercentile (EP) and DevFromMeanElev (DEV), were calculated at five scales ranging from 300m to 1.5km windows. Finer resolution windows capture meso-scale features such as ridges and incised valleys, while broader windows are suited to identifying larger landform patterns.

2.4.3. Biogeoclimatic Ecosystem Classification

Biogeoclimatic zones and subzones from Biogeoclimatic Ecosystem Classification (BEC) v.12 (published September 2, 2021) were also used as predictor variables, along with their corresponding natural disturbance regime code (Table 3).

Table 3. Natural disturbance regime codes

Code	**Natural Disturbance Regime**
NDT1	Ecosystems with rare stand-initiating events
NDT2	Ecosystems with infrequent stand-initiating events
NDT3	Ecosystems with frequent stand-initiating events
NDT4	Ecosystems with frequent stand-maintaining fires
NDT5	Alpine Tundra and Subalpine Parkland ecosystems

2.5. Random Forest

Random Forest (Breiman, 2001) was selected as the preferred ensemble algorithm based on ease of parametrization and superior predictive performance relative to XGBoost and Neural Networks in pilot testing. Random Forest (RF) is an ensemble learning method that constructs multiple decision trees using bootstrap resampling of the training data with a random subset of predictor variables at each split and aggregates their outputs to produce a final prediction.

A separate random forest model was developed for each species using the R package ranger (Wright and Ziegler, 2017). Training and test sets were based on a 80/20 random split of the available data. A grid search over mtry, nodesize, sampsize, and replacement confirmed that default ranger hyperparameters produced optimal performance; the only deviation from defaults was the number of trees, set to 200 rather than the ranger default of 500.

RF predictions for continuous variables often display regression to the mean: because predictions are averages of terminal-node outcomes across trees, extreme values are systematically pulled towards the mean, leading in this case to overestimation at low SI values and underestimation at high ones. To mitigate this bias, a best-rotation bias correction (Song, 2015) was applied, which adjusts predictions based on the fitted relationship between residuals and RF predictions.

2.6. Model performance and validation using YSM

Ground SI measurements from the Young Stand Monitoring (YSM) sampling program were used as an external and independent validation dataset to assess how well the model generalizes beyond the training data. One of the explicit goals of YSM is to assess the accuracy of SI estimates in the PSPL. YSM ground samples are collected through statistically unbiased random sampling on a standardized provincial grid, providing broad geographic coverage across a wide range of site conditions. The YSM program will continue to be used for performance evaluation of site productivity estimates derived from the RF model.

Model performance was evaluated using four metrics: coefficient of determination (R^2), root mean squared error (RMSE), Pearson's correlation coefficient, and mean difference (predicted - observed) (MD). These metrics were calculated against ground-sample data from the training, testing and independent validation (YSM) datasets.

A direct comparison of RF and PSPL v.8.0 predictive performance against YSM validation ground sample data was also completed using these same performance metrics.

3. Modelling results and validation

3.1. Random forest predictive performance

Performance results on training and test sets for all species models after bias correction are provided in Appendix D. Model accuracy on the testing set varied by species. Most commercial species models, with a higher number of training and testing sample sizes, had strong predictive performance: Douglas-fir (n = 437, R^2 = 0.78, RMSE = 3.45), Lodgepole pine (n = 631, R^2 = 0.66, RMSE = 2.25), subalpine fir (n = 313, R^2 = 0.63, RMSE = 2.64), and amabilis fir (n = 70, R^2 = 0.80, RMSE = 2.87). The lowest performance occurred generally in species with low training/testing sample sizes: paper birch (n = 23, R^2 = 0.21, RMSE = 3.18), black spruce (n = 27, R^2 = 0.39, RMSE = 2.47), Sitka spruce (n = 48, R^2 = 0.43, RMSE = 4.79), ponderosa pine (n = 20, R^2 = 0.47, RMSE = 2.34), and trembling aspen (n = 57, R^2 = 0.50, RMSE = 2.98).

Prediction error on the testing set was generally higher in coastal species, particularly Sitka spruce (RMSE = 4.79), western hemlock (RMSE = 4.47), and Douglas-fir (RMSE = 3.45). A potential reason for the high error is that site productivity on the coast, where there is often an excess of soil moisture, can vary greatly with subtle variations in slope, bedrock geology or internal soil drainage not captured by a 100m resolution grid (Banner et al., 2005).

3.2. Random forest performance on YSM validation set

Independent validation of RF SI estimates using YSM ground sample data shows variable performance across species (Table 4), with stronger performance generally observed for species with larger training and validation sample sizes. Models for Douglas-fir (n = 136, R^2 = 0.58), lodgepole pine (n = 422, R^2 = 0.44), western larch (n = 9, R^2 = 0.47), western redcedar (n = 76, R^2 = 0.37), and subalpine fir (n = 134, R^2 = 0.34) and western hemlock (n = 192, R^2 = 0.33) and Sitka spruce (n = 51, R^2 = 0.28) have the strongest explanatory performance and correlation metrics.

Minor species, with generally fewer training and validation sample sizes, have weaker model performance: paper birch (n = 21, R^2 = 0.05), trembling aspen (n = 47, R^2 = 0.03), amabilis fir (n = 38, R^2 = −0.07), and black spruce (n = 7, R^2 = −0.83).

An important exception is Engelmann spruce (n = 72, R^2 = −0.18) and white spruce (n = 184, R^2 = −0.13). Despite their relatively large validation sample sizes and significance as a major commercial species, the RF model performs poorly, explaining very little of the observed variation in SI. In the case of the spruce complex, negative R^2 values likely reflect noise stemming from lack of fine-scale microsite (particularly soil moisture and drainage) information and species hybridization. In the interior spruce hybrid zone, hybrid populations form a genetically continuous complex rather than discrete single species (Degner, 2020). Species labels assigned in ground inventory data group these species using coarse assumptions (for example, assuming that spruce in the ESSF is Engelmann spruce and assigning the Engelmann spruce SI curve), whereas these groups likely include a range of species-specific growth responses.

A negative mean difference for most species indicates that the predicted SI estimates remain conservative when compared to the YSM ground sample data.

Table 4. RF prediction validation on YSM ground sample data

Species	n (all)	RF predicted SI				YSM ground sample SI				RF prediction performance			
		Mean	Stdev	Min	Max	Mean	Stdev	Min	Max	R^2	Cor	MD	RMSE
AT	47	19.2	2.3	13.9	23.4	19.7	5.5	7.8	29.4	0.03	0.25	-0.5	5.4
BA	38	24.8	3.1	16.7	29.8	23.9	5.2	11.9	34.4	-0.07	0.27	1.0	5.3
BL	134	17.1	2.6	8.6	21.5	18.0	4.3	4.0	29.4	0.34	0.62	-0.9	3.5
CW	76	20.3	2.8	16.1	27.1	21.2	5.1	8.8	31.1	0.37	0.64	-1.0	4.0
EP	21	19.3	0.8	17.7	20.8	19.6	4.1	9.7	27.2	0.05	0.24	-0.3	3.9
FD	136	26.2	6.6	13.5	37.9	27.7	8.2	9.3	49.1	0.58	0.78	-1.5	5.3
HW	192	24.5	3.5	17.4	31.0	25.1	7.1	8.4	43.7	0.33	0.58	-0.6	5.8
LW	9	21.5	2.5	17.7	24.7	22.5	4.7	12.1	27.9	0.47	0.76	-1.0	3.2
PL	422	18.3	2.6	10.1	23.7	19.7	3.7	7.1	28.7	0.44	0.78	-1.5	2.7
SB	7	10.1	1.7	7.7	13.1	10.3	2.4	6.9	13.4	-0.83	-0.19	-0.2	3.0
SE	72	17.4	2.8	9.9	22.6	20.2	4.5	3.4	27.3	-0.18	0.47	-2.8	4.9
SS	51	28.7	2.9	23.1	35.5	28.1	7.9	10.2	44.8	0.28	0.57	0.7	6.6
SW	184	20.4	2.0	15.1	25.9	22.9	4.1	4.7	33.5	-0.13	0.49	-2.5	4.3

Validation results by species/BEC zones is given in Appendix E. Only BEC units with more than 10 validation pairs are reported.

3.3. Comparison of RF and PSPL v.8.0 performance on YSM validation set

When validated against YSM, RF generally outperformed the existing PSPL v.8.0 across most species (Table 5). Improvements were observed consistently across metrics, indicating that the RF model provides more accurate and reliable estimates compared to the PSPL.

The largest performance gains were for key commercial species with large validation sample sizes. Lodgepole pine, Douglas-fir, western redcedar, western hemlock, Sitka spruce, western larch and subalpine fir all showed substantial increases in R^2 and correlation, and reductions in RMSE compared to the PSPL. For example, lodgepole pine validation R^2 increased from 0.22 in the PSPL v.8.0 to 0.44 with RF, while RMSE decreased from 3.21 m to 2.73 m. Similar improvements were observed for Douglas-fir (R^2 from 0.44 to 0.57; RMSE from 6.19 m to 5.40 m) and western hemlock (R^2 from 0.02 to 0.31; RMSE from 6.93 m to 5.81 m).

For trembling aspen, amabilis fir, Engelmann spruce, white spruce, and Sitka spruce, RF provided marginal improvements when compared to the PSPL v.8.0, but still showed weak overall model fit. Black spruce was the only species for which RF performed worse than the PSPL, though interpretation is constrained by the very small validation sample size (n = 6).

Table 5. Predictive performance of the PSPL and RF when validated against YSM ground sample data.

		PSPL v.8.0				RF			
SPECIES	**n[6]**	**R^2**	**Cor**	**MD**	**RMSE**	**R^2**	**Cor**	**MD**	**RMSE**
AT	28	-0.29	0.31	-2.63	4.87	-0.18	0.18	-1.79	4.67
BA	37	-0.16	0.29	0.92	5.57	-0.09	0.25	0.95	5.39
BL	127	-0.14	0.28	-1.30	4.58	0.32	0.6	-0.87	3.53
CW	75	0.07	0.36	-0.88	4.9	0.38	0.66	-1.03	4.01
EP	20	-0.01	0.09	-0.26	4.15	0.06	0.27	-0.18	3.99
FD	128	0.44	0.7	-1.69	6.19	0.57	0.78	-1.51	5.4
HW	188	0.02	0.28	-1.07	6.93	0.31	0.57	-0.64	5.81
LW	9	-0.18	0.42	-2.19	4.79	0.47	0.76	-1.03	3.21
PL	413	0.22	0.59	-1.22	3.21	0.44	0.78	-1.5	2.73
SB	6	-0.47	-0.26	0.98	2.36	-1.32	-0.28	0.24	2.96
SE	69	-0.81	0.15	-3.63	5.97	-0.13	0.48	-2.61	4.72
SS	49	0.25	0.51	-0.81	6.54	0.33	0.62	0.19	6.16
SW	177	-0.26	0.38	-2.55	4.62	-0.08	0.51	-2.42	4.28

[6] Number of YSM samples with both a RF and a PSPL v.8.0 estimate. Not all YSM samples have a PSPL v.8.0 estimate due to missing SI issues.

4. Provincial Site Productivity Layer v.9.0

The species-specific RF predictions models were applied to the stack of provincial predictor variables to produce 14 provincial SI tiles (Figure 1). These tiles replace the PSPL v.8.0 as the recommended source of best-available data for provincial SI estimates. A SI estimate for any given species is available for every 100 m pixel in the province, regardless of whether the species exists or is suitable for a given site. While this resolves the previous PSPL version's issues with missing SI estimates, any analysis work will need to ensure that SI estimates from this layer are applied only when the species of interest exists or is suitable. The PSPL v.9.0 is available for download through the BC Data Catalogue[7].

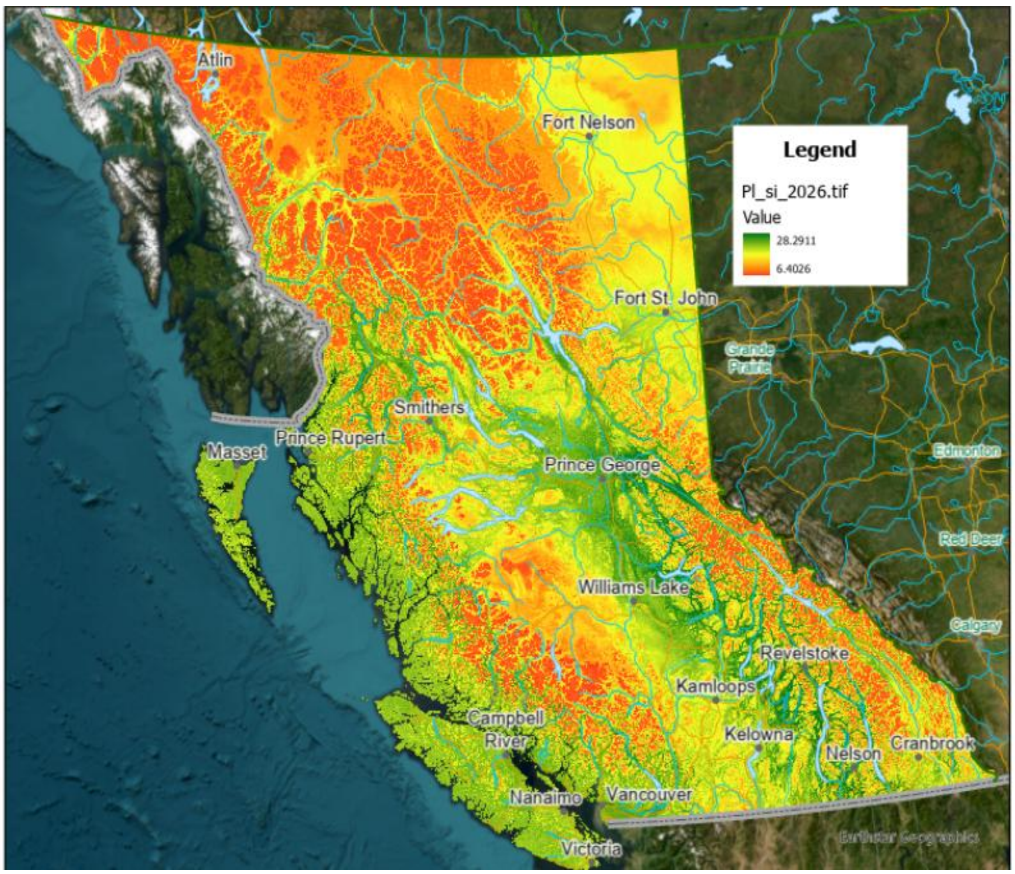


Figure 1. Overview of the Provincial SI tile for lodgepole pine

4.1. Limitations

The performance and applicability of the RF site productivity model and the resulting PSPL v.9.0. are subject to a series of limitations which will inform priorities for future work:

- The PSPL v.9.0 is intended for strategic-level planning and analysis and is not suitable for operational or stand-level decision-making. Where a ground-based measurement of SI is available, it should take precedence over estimates from the PSPL v.9.0.

[7] https://catalogue.data.gov.bc.ca/dataset/04ad45c3-0fdc-4506-bdb4-252c45a63459

- Random Forest algorithms suffer from several limitations, including lack of interpretability, propensity to overfit on small or noisy datasets, and lack of ability to extrapolate beyond the range of its training data.
- Despite bias correction, prediction uncertainty remains highest at the upper and lower extremes of site productivity where ground observations are sparse.
- An estimate of potential site productivity is given for each hectare of the province, which may include non-forested areas. A SI estimate may also be given for a species in areas outside of its actual or potential species distribution range. When using this layer for analysis, analysts must ensure that only estimates for existing or suitable species are used.
- Model performance remains poor for white spruce and Engelmann spruce. More work is needed to determine the cause of poor performance.
- Soil characteristics that drive moisture and nutrients regimes are key drivers of site productivity but are not currently represented in the predictor variables due to the absence of wall-to-wall provincial data.
- The 100 m spatial resolution of predictor variables does not resolve fine-scale topographic gradients, reducing the ability to capture microsite effects in heterogeneous sites.
- RF validation performance is generally weaker for species with fewer training ground sampling points. Additional ground inventory sampling would help improve prediction outcomes.
- Small YSM validation sample sizes limit the ability to evaluate the predictive performance for some species (e.g., ponderosa pine) or BEC zones/subzones.
- SI estimates for species with limited observations (e.g., red alder) rely on species conversion equations or replacement using the estimates from a reference species, introducing additional uncertainty.
- Ground sampling coverage of the entire environmental covariate space was not assessed; additional work is needed to identify areas where the model may be extrapolating outside of the training covariate space, and to identify potential future targeted sampling areas.

4.2. Future work

- Assess how well the training ground samples cover the entire actual covariate space of each species. Results can inform targeted data sampling programs to further improve the model.
- Investigate drivers of poor RF performance in white spruce and Engelmann spruce, including effects of spruce hybridization, covariate gaps in the training data, resolution of predictor variables, or missing predictors.
- Further ground sampling of red alder productivity would make it possible to model SI directly and avoid relying on species conversion.
- Explore opportunities to incorporate additional suitable ground site index observations from existing inventory datasets.
- Integrate high-resolution DEMs (e.g. LiDAR-derived elevation surfaces) and additional predictor variables such as spectral indices.
- As inventory ground data increases, expand the set of species modelled directly to reduce reliance on species conversion and replacement approaches.

5. Conclusion

The modelling framework behind the PSPL v.8.0 was updated to make use of machine learning methods and SI ground sample observations. Results demonstrate that this PSPL v.9.0 RF-based modelling approach provides a clear improvement in provincial site productivity estimation relative to the PSPL v.8.0 for most commercially important species. When validated against independent YSM ground sampling data, the RF models outperformed the PSPL v.8.0 for all species except black spruce, with particularly strong improvements for key commercial species including Douglas-fir, lodgepole pine, western hemlock, western redcedar, and subalpine fir. Predictive performance, while stronger than in the previous PSPL version, remains limited for several minor species and for Engelmann spruce and white spruce; future investigation is recommended to address these limitations. Continued YSM validation will be critical for tracking model performance over time, evaluating improvements from future data expansion, and supporting transparent communication of uncertainty in provincial site productivity estimates.

6. References

Banner, A., P. LePage, J. Moran, and A. de Groot. (compilers). 2005. *The HyP^3 Project: pattern, process, and productivity in hypermaritime forests of coastal British Columbia: a synthesis of 7-year results.* B.C. Min. For., Res. Br., Victoria, B.C. Spec. Rep. 10.

BC Ministry of Forests and Range. 2009. SIBEC sampling and data standards. Prov. B.C.,Victoria

Breiman, L. 2001. Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Daust, K., Mahony, C., Tremblay, B., & Barros, C. 2024. *climr: Downscaling climate data in R* (R package version 0.1.0). https://bcgov.github.io/climr/

Degner, J.C. 2020. Local adaptation in the interior spruce hybrid complex. In: Porth, I.M. and El-Kassaby, A.R. (eds.), *The Spruce Genome*. Springer, Cham. pp. 155–176. https://doi.org/10.1007/978-3-030-21001-4_10

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. 2024. *dismo: Species Distribution Modeling* (R package version 1.3-16). https://rspatial.org/raster/sdm/

Nigh, G. 2012. A biophysical model for estimating site index for the major commercial tree species in British Columbia. Prov. B.C., Victoria, B.C. Tech. Rep. 073.

Mah, S. and G. Nigh. 2015. *SIBEC First and Second-generation Approximation Estimates: backgrounder*. BC Ministry of Forests, Lands, and Natural Resource Operations, Victoria, BC. Available at: https://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/ecosystems/sibec

Song, J. 2015. Bias corrections for random forest in regression using residual rotation. *Journal of the Korean Statistical Society*, *44*(2), 321–326. https://doi.org/10.1016/j.jkss.2015.01.003

Wright M.N., Ziegler A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**(1), 1–17. doi:10.18637/jss.v077.i01.

7. Appendices

Appendix A.

Data preparation of ground-based measurements of SI

YSM	• All SITE_IDENTIFIERs that have had at least one visit in the YSM_MAIN population • SUIT_SI = 'Y' • SUIT_HT = 'Y' • SUIT_TR = 'Y' • Minimum age: >=10 for tree species that have a Growth Intercept (GI) curve, >=20 for tree species with no GI curve available. [8] • Maximum age: < 120 years of age. • If there is more than 1 visit per species with a suitable SI, use only the estimate with the average age at BH closest to 50 years.
Non-YSM	• All SITE_IDENTIFIERs with no visits in the YSM_MAIN population • SUIT_SI = 'Y' • SUIT_HT = 'Y' • SUIT_TR = 'Y' • Minimum age: >=10 for tree species that have a Growth Intercept (GI) curve, >=20 for tree species with no GI curve available.[8] • Maximum age: < 120 years of age • If there is more than one visit per species with a suitable SI, use only the estimate with the average breast height (BH) age closest to 50 years
SIBEC	• All plot locations were converted to NAD83/BC Albers (EPSG:3005, and plots without location were dropped. • SX species codes are reclassified as either SE or SW based on the BEC_ZONE attribute: samples in the ESSF zone are classified as SE; all others as SW
SIA	• Fixed coordinate issues in several projects (see sia_database_2012apr05_fixedcoords.mdb) and converted to NAD83/BC Albers (EPSG:3005). Plots with no locations were dropped. • SX species codes are reclassified as either SE or SW based on the BEC_ZONE attribute: samples in the ESSF zone are classified as SE; all others as SW • Each SIA sample may contain multiple subplots (clusters). The representative BEC zone/subzone per sample is determined based on the Center plot (subplot %in% c('C', 'I', '1', 'A')), or by the most frequent subzone by sample/species when a center plot is not present

[8] The age range used is consistent with SIBEC Sampling and Data Standards.

Appendix B.

Seasonal and annual predictor variables

Code	Variable	Category	Unit
AHM	Annual heat:moisture index	Annual	
bFFP	Beginning of the frost-free period	Annual	day of year
CMD_an	Hargreaves climatic moisture deficit	Annual	mm
CMD_at	Autumn Hargreaves climatic moisture deficit	Seasonal	mm
CMD_sm	Summer Hargreaves climatic moisture deficit	Seasonal	mm
CMD_sp	Spring Hargreaves climatic moisture deficit	Seasonal	mm
CMD_wt	Winter Hargreaves climatic moisture deficit	Seasonal	mm
CMI_an	Climatic Moisture Index	Annual	cm
DDsub0_at	Autumn degree-days below 0°C	Seasonal	
DDsub0_sm	Summer degree-days below 0°C	Seasonal	
DDsub0_sp	Spring degree-days below 0°C	Seasonal	
DDsub0_wt	Winter degree-days below 0°C	Seasonal	
DDsub18_an	Degree-days below 18°C	Annual	
DDsub18_at	Autumn degree-days below 18°C	Seasonal	
DDsub18_sm	Summer degree-days below 18°C	Seasonal	
DDsub18_sp	Spring degree-days below 18°C	Seasonal	
DDsub18_wt	Winter degree-days below 18°C	Seasonal	
DD18_an	Degree-days above 18°C	Annual	
DD18_at	Autumn degree-days above 18°C	Seasonal	
DD18_sm	Summer degree-days above 18°C	Seasonal	
DD18_sp	Spring degree-days above 18°C	Seasonal	
DD18_wt	Winter degree-days above 18°C	Seasonal	
DD5_an	Degree-days above 5°C	Annual	
DD5_at	Autumn degree-days above 5°C	Seasonal	
DD5_sm	Summer degree-days above 5°C	Seasonal	
DD5_sp	Spring degree-days above 5°C	Seasonal	
DD5_wt	Winter degree-days above 5°C	Seasonal	
eFFP	End of the frost-free period	Annual	day of year
EMT	Extreme minimum temperature over 30 years	Annual	°C
Eref_an	Hargreaves reference evaporation	Annual	mm
Eref_at	Autumn Hargreaves reference evaporation	Seasonal	mm
Eref_sm	Summer Hargreaves reference evaporation	Seasonal	mm
Eref_sp	Spring Hargreaves reference evaporation	Seasonal	mm
Eref_wt	Winter Hargreaves reference evaporation	Seasonal	mm

EXT	Extreme maximum temperature over 30 years	Annual	°C
FFP	Frost-free period	Annual	days
MAP	Annual precipitation	Annual	mm
MAT	Mean annual temperature	Annual	°C
MCMT	Mean coldest month temperature	Annual	°C
MSP	Growing season (May to September) precipitation	Seasonal	mm
MWMT	Mean warmest month temperature	Annual	°C
NFFD_an	Number of frost-free days	Annual	
NFFD_at	Autumn number of frost-free days	Seasonal	
NFFD_sm	Summer number of frost-free days	Seasonal	
NFFD_sp	Spring number of frost-free days	Seasonal	
NFFD_wt	Winter number of frost-free days	Seasonal	
PAS_an	Precipitation as snow	Annual	mm
PAS_at	Autumn precipitation as snow	Seasonal	mm
PAS_sm	Summer precipitation as snow	Seasonal	mm
PAS_sp	Spring precipitation as snow	Seasonal	mm
PAS_wt	Winter precipitation as snow	Seasonal	mm
PPT_an	Precipitation	Annual	mm
PPT_at	Autumn precipitation	Seasonal	mm
PPT_sm	Summer precipitation	Seasonal	mm
PPT_sp	Spring precipitation	Seasonal	mm
PPT_wt	Winter precipitation	Seasonal	mm
RH_an	Relative humidity	Annual	%
RH_at	Autumn relative humidity	Seasonal	%
RH_sm	Summer relative humidity	Seasonal	%
RH_sp	Spring relative humidity	Seasonal	%
RH_wt	Winter relative humidity	Seasonal	%
SHM	Summer heat:moisture index	Annual	
Tave_an	Mean temperature	Annual	°C
Tave_at	Autumn mean temperature	Seasonal	°C
Tave_sm	Summer mean temperature	Seasonal	°C
Tave_sp	Spring mean temperature	Seasonal	°C
Tave_wt	Winter mean temperature	Seasonal	°C
TD	Continentality (MWMT-MCMT)	Annual	°C
Tmax_an	Mean daily maximum temperature	Annual	°C
Tmax_at	Autumn mean daily maximum temperature	Seasonal	°C
Tmax_sm	Summer mean daily maximum temperature	Seasonal	°C
Tmax_sp	Spring mean daily maximum temperature	Seasonal	°C

Tmax_wt	Winter mean daily maximum temperature	Seasonal	°C
Tmin_an	Mean daily minimum temperature	Annual	°C
Tmin_at	Autumn mean daily minimum temperature	Seasonal	°C
Tmin_sm	Summer mean daily minimum temperature	Seasonal	°C
Tmin_sp	Spring mean daily minimum temperature	Seasonal	°C
Tmin_wt	Winter mean daily minimum temperature	Seasonal	°C
BIO1	Annual Mean Temperature	Annual	°C
BIO2	Mean Diurnal Range (Mean of monthly (max temp – min temp))	Annual	°C
BIO3	Isothermality (BIO2/BIO7)	Annual	%
BIO4	Temperature seasonality (standard deviation x 100)	Annual	
BIO5	Max temperature of warmest month	Monthly	°C
BIO6	Min temperature of coldest month	Monthly	°C
BIO7	Temperature annual range (BIO5-BIO6)	Annual	°C
BIO8	Mean temperature of wettest quarter	Seasonal	°C
BIO9	Mean temperature of driest quarter	Seasonal	°C
BIO10	Mean temperature of warmest quarter	Seasonal	°C
BIO11	Mean temperature of coldest quarter	Seasonal	°C
BIO12	Annual Precipitation	Annual	mm
BIO13	Precipitation of wettest month	Monthly	mm
BIO14	Precipitation of driest month	Monthly	mm
BIO15	Precipitation seasonality (coefficient of variation)	Annual	%
BIO16	Precipitation of wettest quarter	Seasonal	mm
BIO17	Precipitation of driest quarter	Seasonal	mm
BIO18	Precipitation of warmest quarter	Seasonal	mm
BIO19	Precipitation of coldest quarter	Seasonal	mm

Appendix C.

Topographic predictor variables

Code	Variable	Description	Unit
elev	Elevation	Elevation (m) from 25m DEM resampled to HectaresBC 100m grid by weighted average.	m
elev_min	Elevation	Minimum elevation value in 1ha grid cell.	m
elev_max	Elevation	Maximum elevation value in 1ha grid cell	m
slope	Slope	Slope gradient in degrees calculated from elevation.	deg
slope_min	Slope	Minimum slope value in 1ha grid cell.	deg
slope_max	Slope	Maximum slope value in 1ha grid cell.	deg
northness	Northness	A measure of aspect. Values between 1 (north) and -1 (south), with zero being neither north nor south in aspect (but rather east or west).	n/a
eastnes	Eastness	A measure of aspect. Values between 1 (east) and -1 (west), with zero referring to another aspect that is not east or west (i.e. north or south).	n/a
tri	Ruggedness Index (TRI)	A measure of local topographic relief. The TRI calculates the root-mean-square-deviation (RMSD) for each grid cell in a digital elevation model (DEM), calculating the residuals (i.e. elevation differences) between a grid cell and its eight neighbours.	n/a
ep300	Elevation Percentile (EP) at 300m	The vertical position of a grid cell as the percentile of the elevation distribution within a 300m filter window. Expressed as a percentage from 0% to 100%.	%
ep500	Elevation Percentile (EP) at 500m	The vertical position of a grid cell as the percentile of the elevation distribution within a 500m filter window. Expressed as a percentage from 0% to 100%.	%
ep1500	Elevation Percentile (EP) at 1,500m	The vertical position of a grid cell as the percentile of the elevation distribution within a 1,500m filter window. Expressed as a percentage from 0% to 100%.	%
ep3500	Elevation Percentile (EP) at 3,500m	The vertical position of a grid cell as the percentile of the elevation distribution within a 3,500m filter window. Expressed as a percentage from 0% to 100%.	%
ep15000	Elevation Percentile (EP) at 15,000m	The vertical position of a grid cell as the percentile of the elevation distribution within a 15,000m filter window. Expressed as a percentage from 0% to 100%.	%
dev300	Deviation from Mean Elevation (DEV) at 300m	The difference between the elevation of each grid cell and the mean elevation of the 300m neighbourhood, normalized by standard deviation.	%
dev500	Deviation from Mean Elevation (DEV) at 500m	The difference between the elevation of each grid cell and the mean elevation of the 500m neighbourhood, normalized by standard deviation.	n/a
dev1500	Deviation from Mean Elevation (DEV) at 1,500m	The difference between the elevation of each grid cell and the mean elevation of the 1,500m neighbourhood, normalized by standard deviation.	n/a
dev3500	Deviation from Mean Elevation (DEV) at 3,500m	The difference between the elevation of each grid cell and the mean elevation of the 3,500m neighbourhood, normalized by standard deviation.	n/a

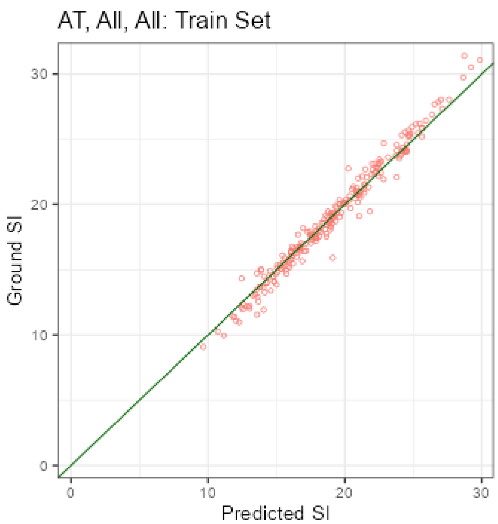
dev15000	Deviation from Mean Elevation (DEV) at 15,000m	The difference between the elevation of each grid cell and the mean elevation of the 15,000m neighbourhood, normalized by standard deviation.	n/a
profcurv	Profile Curvature	Characterizes the profile (or vertical) curvature, or the rate of change in slope along a flow line. Positive values indicate areas of flow acceleration, while negative plan values indicate flow deceleration.	m^{-1}
plancurv	Plan Curvature	Characterizes the contour curvature. Positive plan curvature values indicate areas of flow divergence, while negative plan values indicate flow convergence.	m^{-1}
twi	Topographic Wetness Index (TWI)	Describes the propensity for a site to be saturated to the surface given its contributing area and local slope characteristics.	n/a
twi_min	Topographic Wetness Index (TWI)	Minimum TWI value in 1ha grid cell.	n/a
twi_max	Topographic Wetness Index (TWI)	Maximum TWI value in 1ha grid cell.	n/a
hli	Heat Load Index (HLI)	Characterizes heat load as a function of slope steepness and aspect. HLT folds aspect such that highest heat load values are are in the southwest aspect. Values range from 0 (cold northeast-facing steep slopes) to 1 (warm southwest-facing steep slopes).	n/a
dah	Diurnal Anisotropic Heat Index (DAH)	Diurnal anisotropic heat (DAH) quantifies the combined characteristics of slope aspect and slope angle on the amount of heat a location receives from the sun over a day. It is an indicator of the differential solar heating of a surface. High positive values indicate steep southwest-facing slopes, a value of 0 indicates flat terrain, and low or negative values indicate northeast-facing steep slopes.	n/a
mrvbf	Multiresolution Valley Bottom Flatness (MRBVF)	Characterizes depositional terrain across various scales, based on the observation that that valley bottoms are lower and flatter than surrounding areas, and that the relative flatness increases with the size of the valley. A value of 0 represents an erosional landscape, while values of 1 and above signify areas of deposition.	n/a
mrrtf	Multiresolution Ridge Top Flatness (MRRTF)	Topographic index complementary to MRVBF. MRRTF identifies high, flat areas, such as ridge tops, across various scales. It functions similarly to MRVBF but focuses on elevated, level terrain instead of low-lying, flat areas. A value of 0 indicates a low or steep area, while values of 1 or greater signify progressively larger areas of high, flat ground.	n/a
sp300	Slope position (classified TPI)	Slope position classification according to Topographic Position Index (TPI) and slope, following the Weiss (2001) approach for classification.	n/a
sp500	Slope position (classified TPI)	Slope position classification according to Topographic Position Index (TPI) and slope, following the Weiss (2001) approach for classification.	n/a
sp1500	Slope position (classified TPI)	Slope position classification according to Topographic Position Index (TPI) and slope, following the Weiss (2001) approach for classification.	n/a
sp3500	Slope position (classified TPI)	Slope position classification according to Topographic Position Index (TPI) and slope, following the Weiss (2001) approach for classification.	n/a
sp15500	Slope position (classified TPI)	Slope position classification according to Topographic Position Index (TPI) and slope, following the Weiss (2001) approach for classification.	n/a

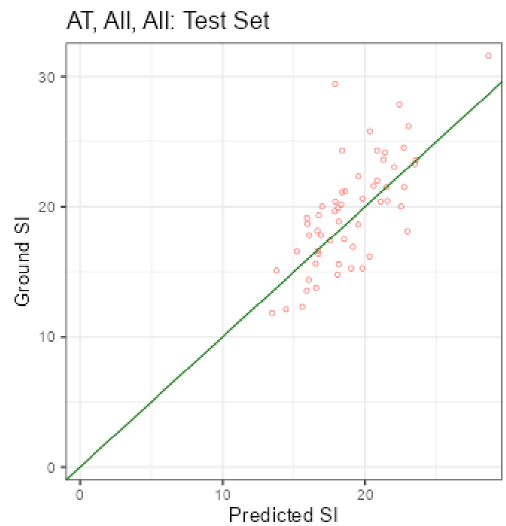
Appendix D.

Random forest prediction performance on training and testing set

Trembling aspen

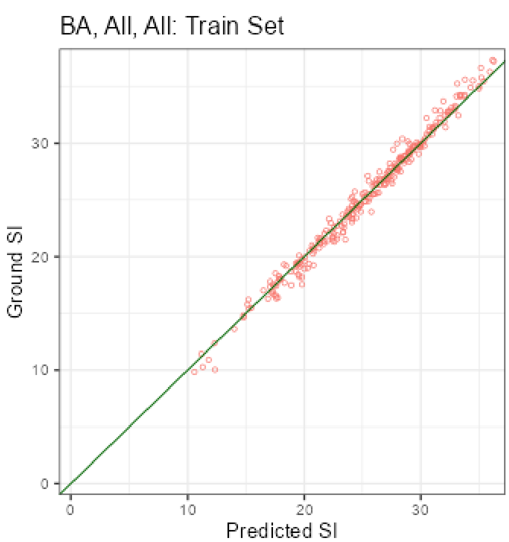
	Train set		Test set
n	226	n	57
R2	0.97	R2	0.50
RMSE	0.74	RMSE	2.98
Corr	0.99	Corr	0.73
MD	0.00	MD	0.63

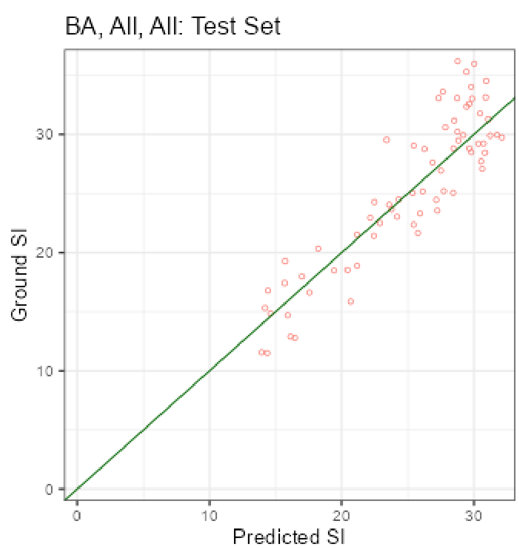




Amabilis fir

	Train set		Test set
n	277	n	70
R2	0.98	R2	0.80
RMSE	0.70	RMSE	2.87
Corr	0.99	Corr	0.90
MD	0.00	MD	0.34

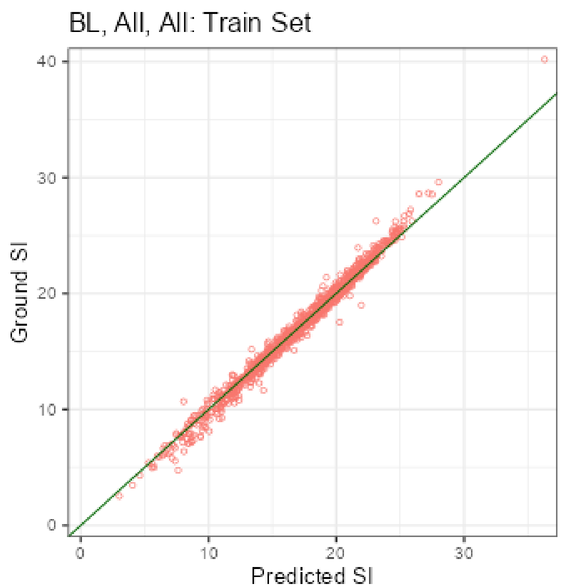


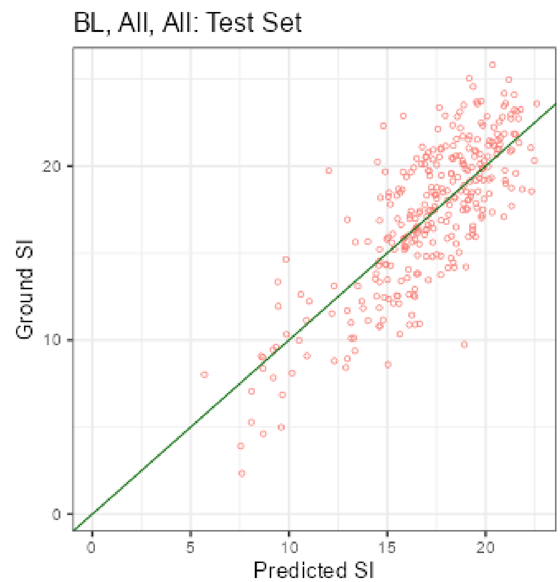


Subalpine fir

Train set	
n	1253
R2	0.99
RMSE	0.54
Corr	0.99
MD	0.00

Test set	
n	313
R2	0.63
RMSE	2.64
Corr	0.79
MD	0.06

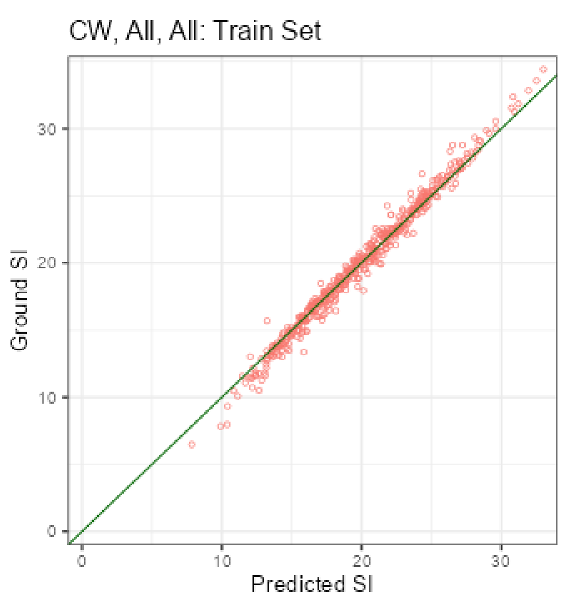




Western redcedar

Train set	
n	504
R2	0.98
RMSE	0.65
Corr	0.99
MD	0.00

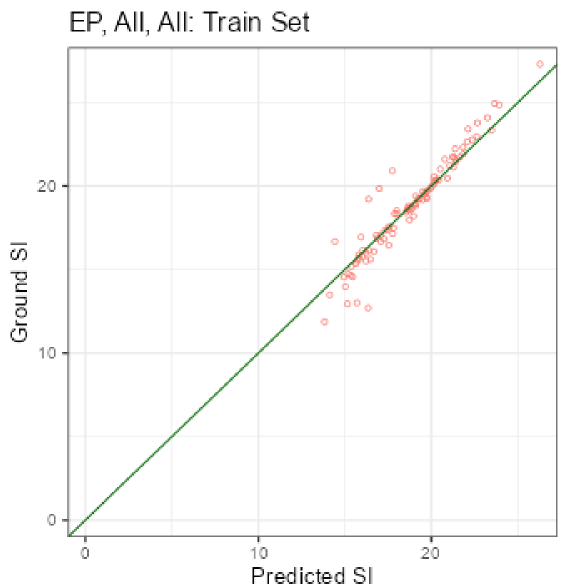
Test set	
n	126
R2	0.58
RMSE	3.32
Corr	0.77
MD	0.2

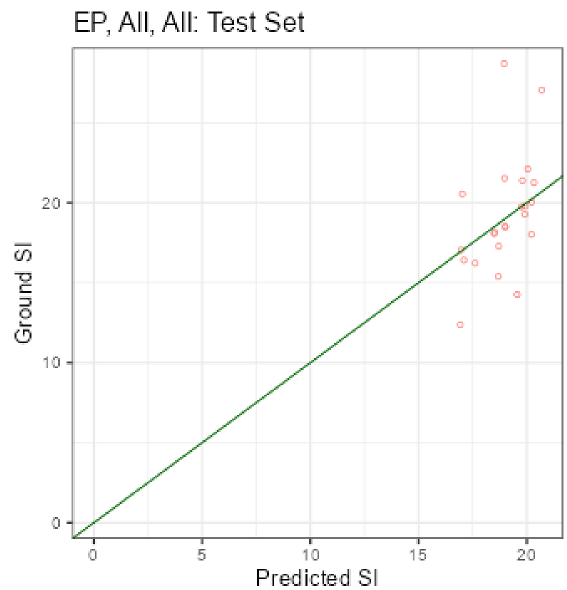


Paper birch

Train set	
n	94
R2	0.90
RMSE	0.94
Corr	0.96
MD	0.01

Test set	
n	23
R2	0.21
RMSE	3.18
Corr	0.49
MD	0.23

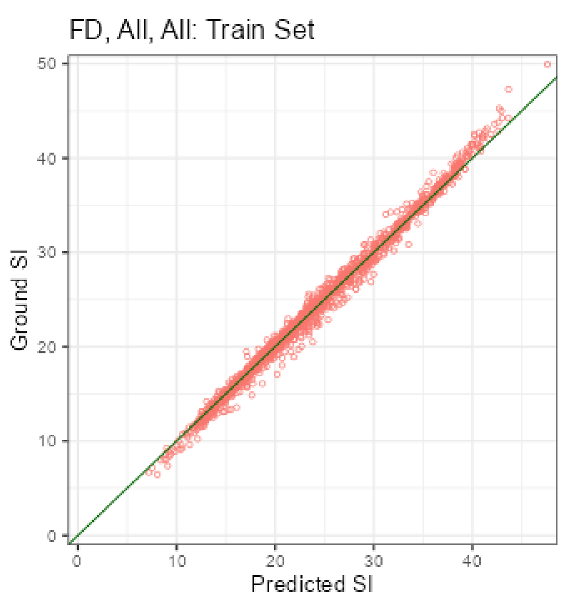




Douglas-fir

Train set	
n	1744
R2	0.99
RMSE	0.67
Corr	1.00
MD	0.01

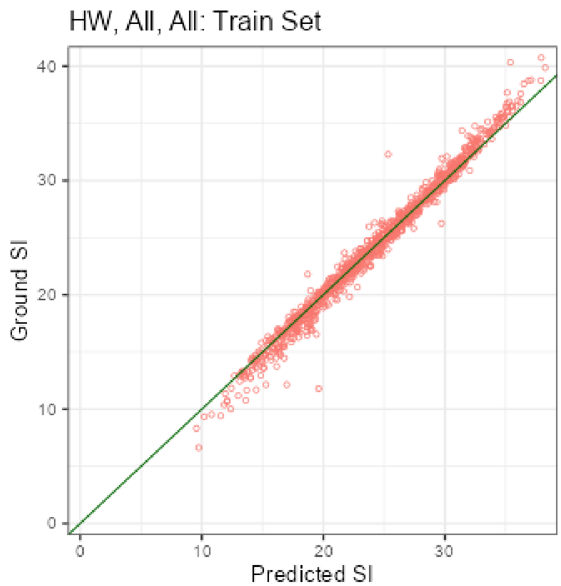
Test set	
n	437
R2	0.78
RMSE	3.45
Corr	0.88
MD	0.02



Western hemlock

Train set	
n	1011
R2	0.98
RMSE	0.84
Corr	0.99
MD	0.00

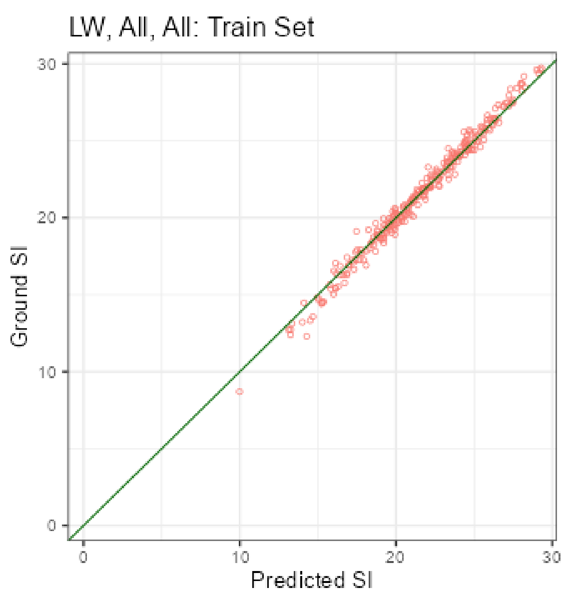
Test set	
n	252
R2	0.55
RMSE	4.47
Corr	0.76
MD	-0.19

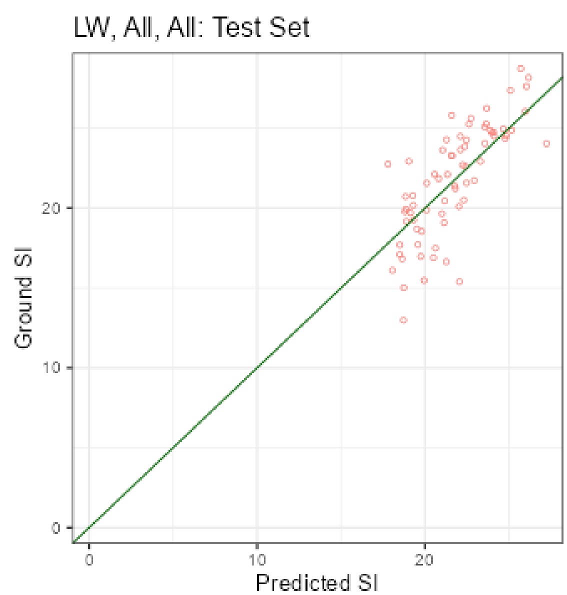


Western larch

Train set	
n	285
R2	0.98
RMSE	0.49
Corr	0.99
MD	0.00

Test set	
n	71
R2	0.58
RMSE	2.26
Corr	0.77
MD	0.08

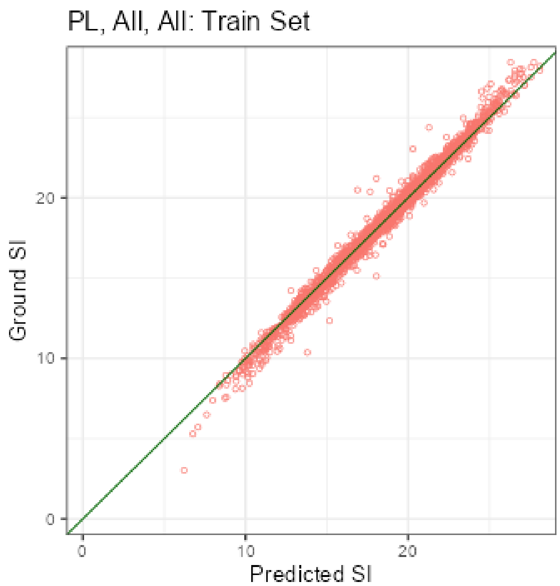


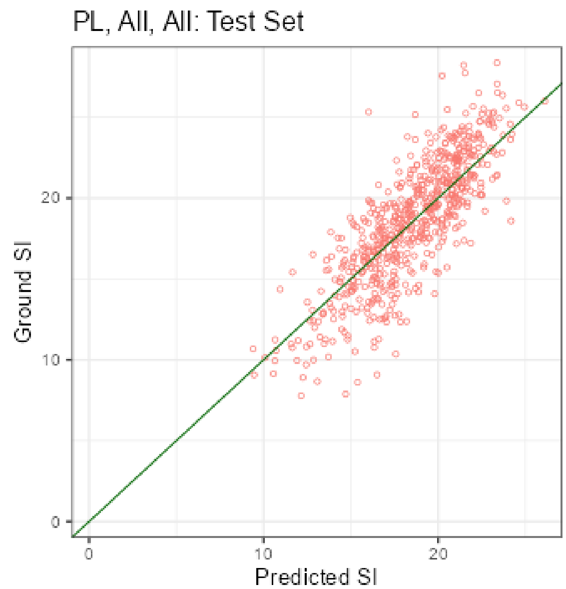


Lodgepole pine

Train set	
n	2532
R2	0.99
RMSE	0.44
Corr	0.99
MD	0.01

Test set	
n	631
R2	0.66
RMSE	2.25
Corr	0.82
MD	0.09

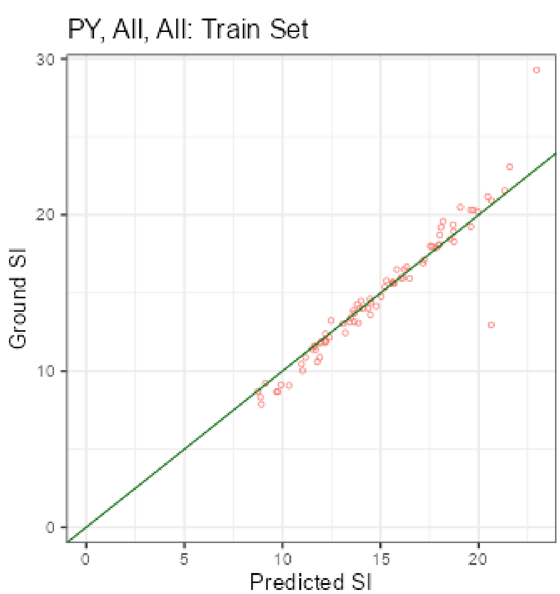


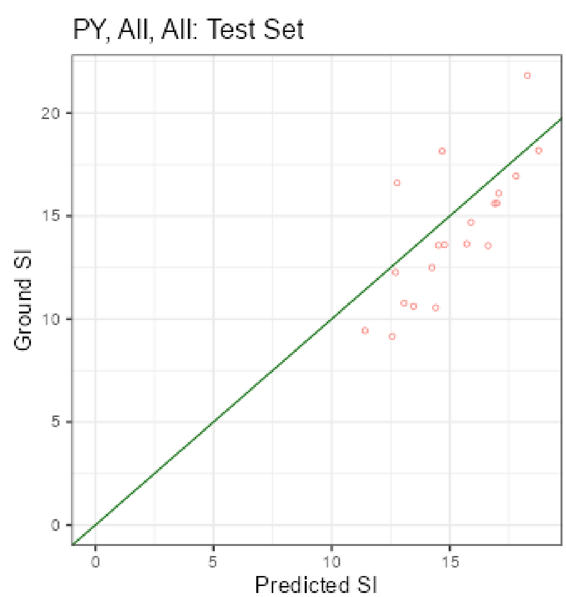


Ponderosa pine

Train set	
n	79
R2	0.90
RMSE	1.26
Corr	0.95
MD	-0.06

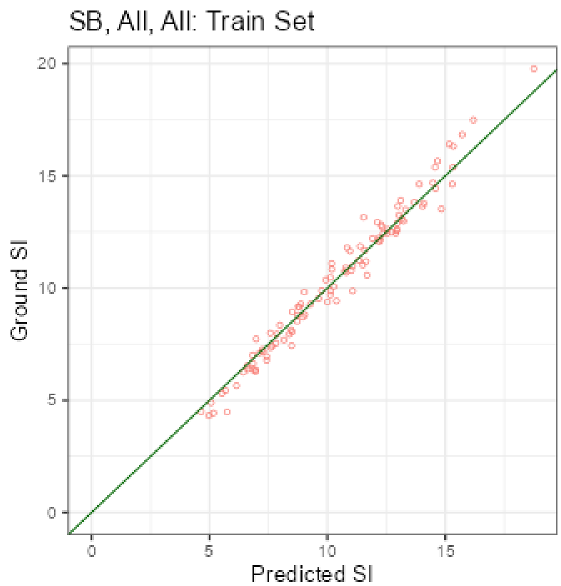
Test set	
n	20
R2	0.47
RMSE	2.34
Corr	0.75
MD	-0.96

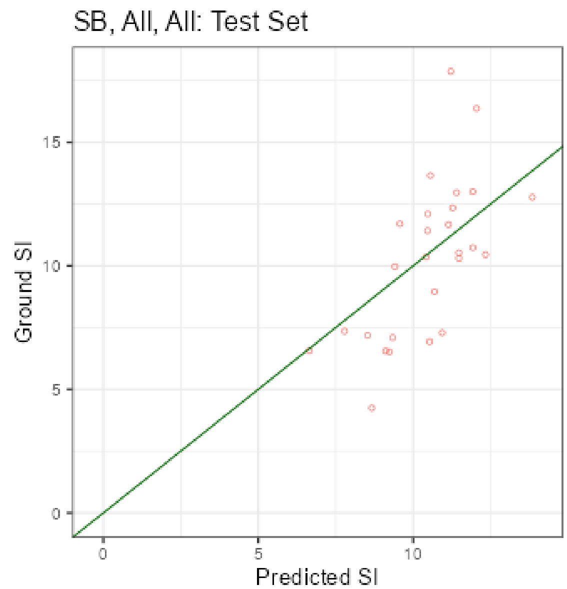




Black spruce

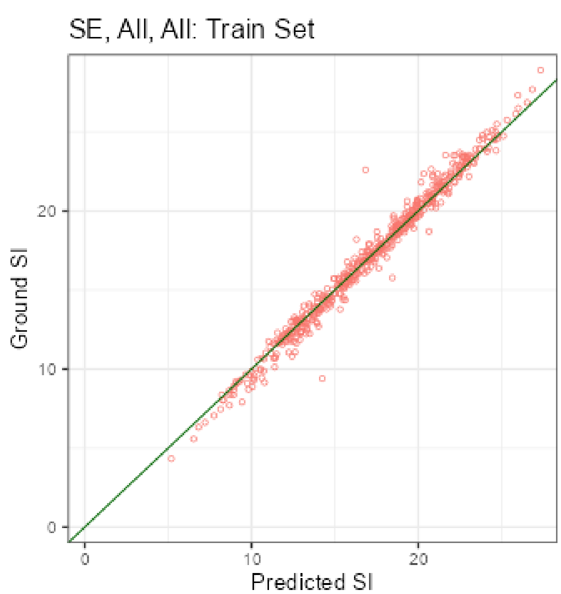
	Train set		Test set
n	106	n	27
R2	0.97	R2	0.39
RMSE	0.57	RMSE	2.47
Corr	0.99	Corr	0.65
MD	0.00	MD	-0.21

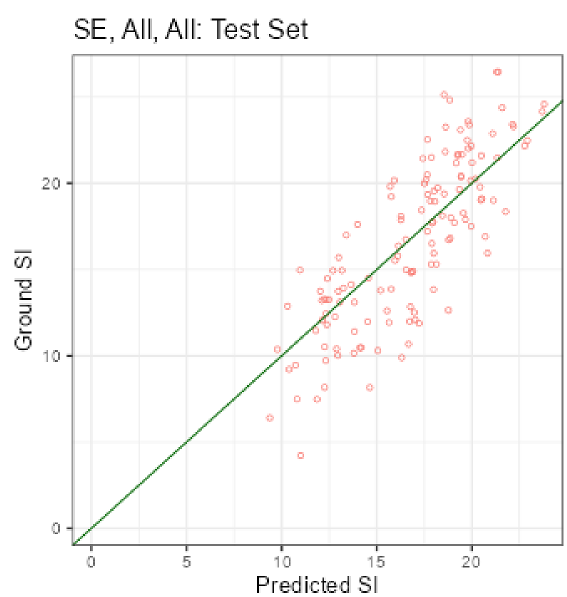




Engelmann spruce

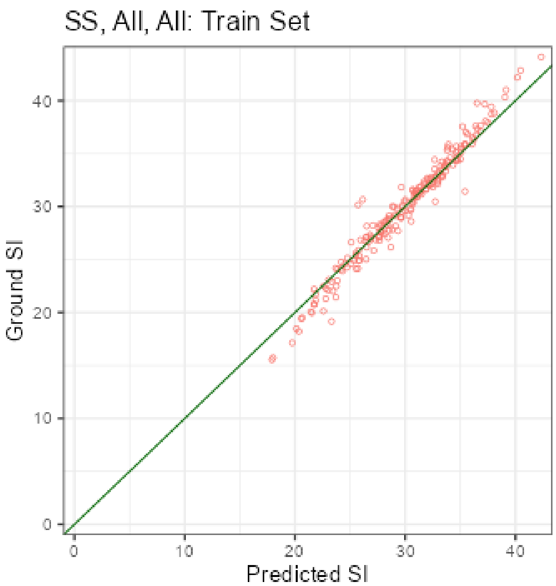
	Train set		Test set
n	550	n	136
R2	0.98	R2	0.64
RMSE	0.64	RMSE	2.86
Corr	0.99	Corr	0.81
MD	0.01	MD	-0.16

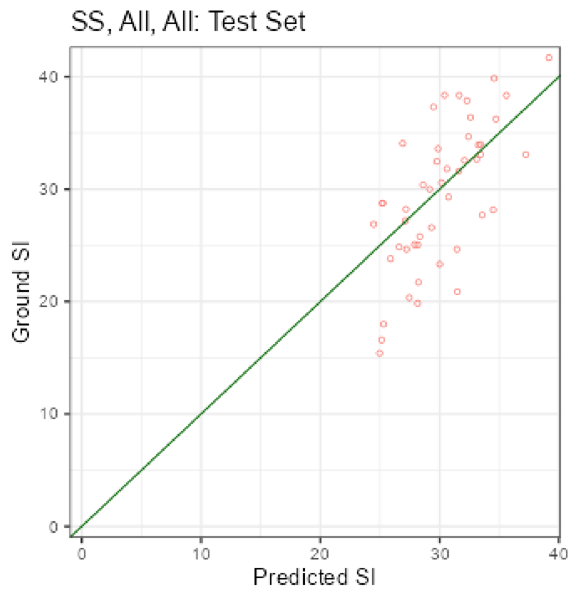




Sitka spruce

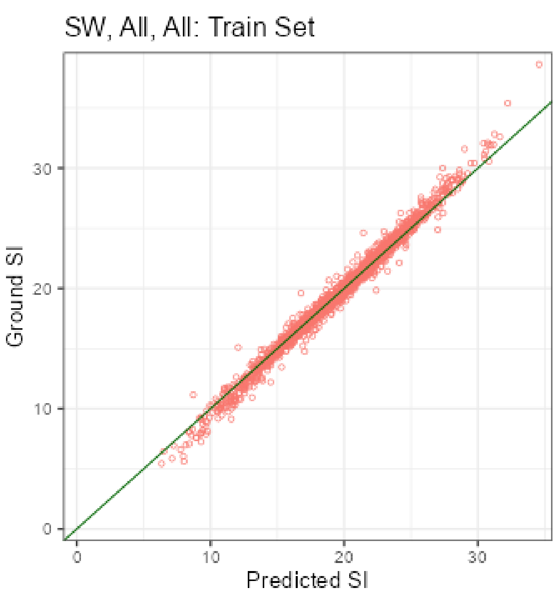
	Train set	Test set
n	206	48
R2	0.96	0.43
RMSE	1.13	4.79
Corr	0.98	0.68
MD	0.02	-0.65





White spruce

	Train set	Test set
n	1772	443
R2	0.98	0.56
RMSE	0.56	3.21
Corr	0.99	0.77
MD	0.00	-0.04



Appendix E.

Random forest prediction performance on YSM validation by species and BEC zone.

Species	BEC	n (all)	R^2	Pearson's correlation	MD	RMSE
AT	**all**	**47**	**0.03**	**0.25**	**-0.5**	**5.4**
	SBS	19	0.03	0.25	-0.5	5.4
BA	**all**	**38**	**-0.07**	**0.27**	**1.0**	**5.3**
	CWH	36	-0.11	0.23	0.8	5.4
BL	**all**	**134**	**0.34**	**0.62**	**-0.9**	**3.5**
	ESSF	63	0.29	0.57	-0.8	3.6
	ICH	18	0.05	0.39	-1.2	3.6
	MS	10	-0.49	-0.20	1.0	2.7
	SBS	43	0.18	0.56	-1.3	3.4
CW	**all**	**76**	**0.37**	**0.64**	**-1.0**	**4.0**
	CWH	50	0.37	0.64	-0.5	4.2
	ICH	24	0.20	0.71	-1.6	3.5
EP	**all**	**21**	**0.05**	**0.24**	**-0.3**	**3.9**
FD	**all**	**136**	**0.58**	**0.78**	**-1.5**	**5.3**
	CWH	60	0.24	0.52	-1.3	6.4
	ICH	40	0.02	0.42	-1.8	4.5
	IDF	15	0.35	0.60	-0.3	3.8
HW	**all**	**192**	**0.33**	**0.58**	**-0.6**	**5.8**
	CWH	154	0.29	0.57	-0.5	6.2
	ICH	36	-0.13	0.31	-1.5	3.6
PL	**all**	**422**	**0.44**	**0.78**	**-1.5**	**2.7**
	BWBS	12	-0.09	0.70	-2.7	3.8
	ESSF	59	0.12	0.68	-1.6	2.6
	ICH	38	0.07	0.55	-1.4	2.7
	IDF	29	0.62	0.84	-0.9	2.1
	MS	65	0.37	0.76	-1.5	2.8
	SBPS	53	0.53	0.74	-0.3	2.6
	SBS	163	-0.15	0.60	-1.9	2.8
SE	**all**	**72**	**-0.18**	**0.47**	**-2.8**	**4.9**
	ESSF	72	-0.18	0.47	-2.8	4.9
SS	**all**	**51**	**0.28**	**0.57**	**0.7**	**6.6**
	CWH	51	0.28	0.57	0.7	6.6

SW	**all**	**184**	**-0.13**	**0.49**	**-2.5**	**4.3**
	BWBS	16	-1.91	0.32	-4.5	5.5
	ICH	54	-0.38	0.37	-2.4	4.0
	MS	13	0.08	0.53	-1.4	3.0
	SBS	94	-0.24	0.46	-2.7	4.5