Geospatial Data for Modeling Soil Carbon Stocks across Pacific Northwest Watersheds

Get Data

Documentation Revision Date: 2026-02-12

Dataset Version: 1

Summary

This dataset provides predicted soil organic carbon (SOC) for 2021-2022 (nominal) as well as the predictor data for spatial models in four watersheds of the Pacific Northwest (PNW). These data support the study of wetland carbon storage within this landscape. Field sample collection for soil carbon stocks at 114 locations provide observations for modeling and were collected through 2021-2022. The raster and vector predictor layers are sourced from lidar and satellite imagery, which span dates from 2012-2022. The four study watersheds include the Heen Latinee Experimental Forest (HLEF) located in southeast Alaska near Juneau and three watersheds in Washington state: the Hoh River Watershed (HRW) located on the west coast of the Olympic Peninsula, the Mashel River Watershed (MRW) located on the western side of the Cascade Mountain Range near Tahoma (Mt. Rainier), and the Colville Watershed (CVW) located in northeastern Washington. The Wetland Intrinsic Potential (WIP) tool was implemented for each study watershed to model the gridded land surface as a continuous probability of wetland presence, with each grid pixel containing a value from 0-100%. Geospatial datasets related to vegetation, climate, lithology and geology, and topography were gathered to determine predictors for SOC stocks. Google Earth Engine was used to obtain satellite imagery for calculation of vegetation spectral indices from the five year median of Sentinel-2 reflectance. Two model types were used in the research to model SOC stock and SOC percent: a linear mixed effects model (LMM) and a quantile random forest (RFM). The LMM was used to test specific hypotheses about important predictors and examine predictor coefficients. The RFM was used to accommodate potential non-linear relationships in the data as well as incorporate a more flexible approach to predictor selection. Mapped predictions of SOC stock across all four study watersheds were generated using the model with the best fit to the test dataset. Additional geospatial masks were used to remove surface water areas and urban zones. The data are provided in comma separated values (CSV), GeoTIFF, and GeoPackage formats.

There are 22 data files in this dataset that includes two files in comma-separated values (CSV) format, 12 GeoTIFFs, and eight GeoPackage files.

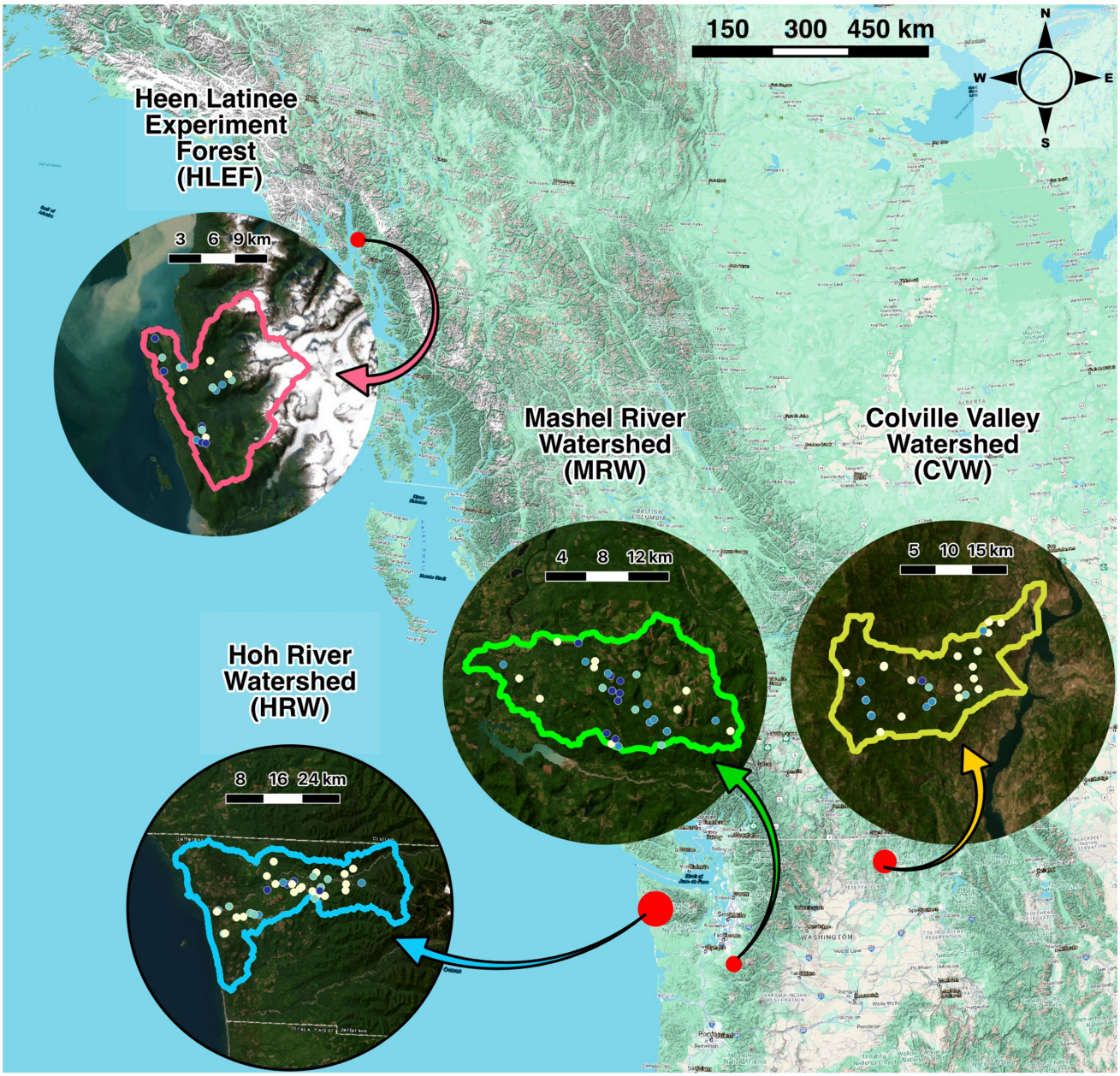

Figure 1. Map of study watersheds in the Pacific Northwest of North America. The Heen Latinee Experimental Forest (HLEF) is located in southeast Alaska. The Hoh River Watershed (HRW), Mashel River Watershed (MRW), and Colville Valley Watershed (CVW) are located in Washington state. The size of the red dots is proportional to the size of the inset circle for each study area. Basemap imagery provided by OpenStreetMap. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

Citation

Stewart, A., M. Halabisky, D.V. D'amore, D. Spinola, C. Babcock, L.M. Moskal, and D. Butman. 2026. Geospatial Data for Modeling Soil Carbon Stocks across Pacific Northwest Watersheds. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/2449

Table of Contents

1. Dataset Overview
2. Data Characteristics
3. Application and Derivation
4. Quality Assessment
5. Data Acquisition, Materials, and Methods
6. Data Access
7. References

1. Dataset Overview

This dataset provides predicted soil organic carbon (SOC) for 2021-2022 (nominal) as well as the predictor data for spatial models in four watersheds of the Pacific Northwest (PNW). These data support the study of wetland carbon storage within this landscape.

Field sample collection for soil carbon stocks at 114 locations provide observations for modeling and were collected through 2021-2022. The raster and vector predictor layers are sourced from lidar and satellite imagery, which span dates from 2012-2022. The four study watersheds include the Heen Latinee Experimental Forest (HLEF) located in southeast Alaska near Juneau and three watersheds in Washington state: the Hoh River Watershed (HRW)

located on the west coast of the Olympic Peninsula, the Mashel River Watershed (MRW) located on the western side of the Cascade Mountain Range near Tahoma (Mt. Rainier), and the Colville Watershed (CVW) located in northeastern Washington.

The Wetland Intrinsic Potential (WIP) tool was implemented for each study watershed to model the gridded land surface as a continuous probability of wetland presence, with each grid pixel containing a value from 0-100% (Halabisky et al., 2023). Geospatial datasets related to vegetation, climate, lithology and geology, and topography were gathered to determine predictors for SOC stocks. Google Earth Engine (Gorelick et al., 2017) was used to obtain satellite imagery for calculation of vegetation spectral indices from the five year median of Sentinel-2 reflectance. Two model types were used in the research to model SOC stock and SOC percent: a linear mixed effects model (LMM) and a quantile random forest (RFM). The LMM was used to test specific hypotheses about important predictors and examine predictor coefficients. The RFM was used to accommodate potential non-linear relationships in the data as well as incorporate a more flexible approach to predictor selection. Mapped predictions of SOC stock across all four study watersheds were generated using the model with the best fit to the test dataset. Additional geospatial masks were used to remove surface water areas and urban zones.

Project: Carbon Monitoring System

The NASA Carbon Monitoring System (CMS) program is designed to make significant contributions in characterizing, quantifying, understanding, and predicting the evolution of global carbon sources and sinks through improved monitoring of carbon stocks and fluxes. The System uses NASA satellite observations and modeling/analysis capabilities to establish the accuracy, quantitative uncertainties, and utility of products for supporting national and international policy, regulatory, and management activities. CMS data products are designed to inform near-term policy development and planning.

Related Publication

Stewart, A. Improving soil organic carbon spatial distribution and interpretation of cross-scale drivers with probabilistic wetland representation. 2026. In preparation.

2. Data Characteristics

Spatial Coverage: Watersheds in southeast Alaska and Washington, U.S.

Spatial Resolution: 4-5 m

Temporal Coverage: 2021 to 2022 (nominal period for predictions using data from 2012 to 2022)

Temporal Resolution: One-time estimate

Study Area: Latitude and longitude are given in decimal degrees.

Site	Westernmost Longitude	Easternmost Longitude	Northernmost Latitude	Southernmost Latitude
Alaska and Washington, U.S.	-134.9827	-118.0855	58.7242	46.7724

Data File Information

There are 22 data files in this dataset which includes twelve files in cloud-optimized GeoTIFF (.tif) format, eight files in geopackage format (.gpkg), and two files in comma-separated values (.csv) format.

Site abbreviations (<site>) **used in the data file names.** Abbreviations may be lower case or capitalized.

- HLEF: The Heen Latinee Experimental Forest located near Juneau, Alaska
- Hoh: The Hoh River Watershed (HRW) located on the west coast of the Olympic Peninsula, Washington
- MAS: The Mashel River Watershed (MRW), located on the western side of the Cascade Mountain Range near Tahoma (Mt. Rainier), Washington
- Col: The Colville Valley Watershed (CVW), located on the northeastern portion of Washington

Table 1. File names and descriptions.

Naming Conventions	Example File Name	Description
<site>_PredictorStack_Class.tif	hoh_PredictorStack_Class.tif	Four GeoTIFF files, one per site, used to generate soil carbon predictions containing classified layers (20 bands)
PNW_<site>_SOC_RFM_1m.tif	PNW_mas_SOC_RFM_1m.tif	Four GeoTIFF files, one per site, with SOC predictions at 1-m depth from the Random Forest Model
PNW_<site>_prediction_interval_SOC_RFM_1m.tif	PNW_hlef_025_975_SOC_RFM_1m.tif	Four GeoTiff files, one per site. Each file has 2 bands: SOC stock prediction interval of 2.5th quantile and 97.5th, for SOC using the Random Forest Model
<site>_lab_pts_<EPSG>.gpkg	col_lab_pts_2855.gpkg	Four files: geopackage vector files with point locations associated with lab measurements, such as soil SOC, soil type, and pH. The EPSG codes for the projected coordinate systems are included in the file names: EPSG:2855 for "col" and "hoh", EPSG:6394 for "hlef", and EPSG:2856 for "mas".
<site>_pts_<EPSG>.gpkg	col_pts_2855.gpkg	Four files: geopackage vector files that contain the study area point locations These files contain SOC and other data similar to above, but not pH, soil type, etc.
pnw_lab_data.csv	pnw_lab_data.csv	PNW laboratory data
pnw_stocks_data.csv	pnw_stocks_data.csv	PNW carbon stocks data

Table 2. Variables in the files *<site>_PredictorStack_Class.tif*.

Variable	Units	Description
CHM	m	Canopy Height Model
MAP	mm	Mean Annual Precipitation
MAT	degrees C	Mean Annual Temperature
PET	mm	Potential Evapotranspiration
DTM	m	Digital Terrain Model or Elevation
geomorphons	-	Geomorphons categories for geomorphology
HLI	1	Heat Load Index
LITH	-	Lithology
nlcd_reclass	-	Reclassified names for the National Land Cover Database
NDVI_median	1	Median Normalized Difference Vegetation Index
MNDWI_median	1	Median Modified Normalized Water/Wetness Index
EVI_median	1	Median Enhanced Vegetation Index
NDYI_median	1	Median Normalized Difference Yellow Index
dev_1000	m	Deviation from mean elevation at 1000-m scale
dev_300	m	Deviation from mean elevation at 300-m scale
dev_50	m	Deviation from mean elevation at 50-m scale
grad_1000	1	Gradient at 1000-m scale (slope gradient = rise / run)
grad_300	1	Gradient at 300-m scale
grad_50	1	Gradient at 50-m scale
WIP	percent	Wetland Intrinsic Potential

Table 3. Variables in the files *PNW_<site>_SOC_RFM_1m.tif*.

Variable	Units	Description
sum	Mg ha^{-1}	sum of all layers for 1-m SOC stocks

Table 4. Variables in the files PNW_<site>_prediction_interval_SOC_RFM_1m.tif.

Variable	Units	Description
soc_025	g cm^{-2}	SOC stock prediction interval of 2.5th quantile
soc_975	g cm^{-2}	SOC stock prediction interval of 97.5th quantile

Table 5. Variables in the files *<site>_lab_pts_<EPSG>.gpkg*.

Variable	Units	Description
sample_ID	-	Unique sample ID
lower_depth	cm	Depth of bottom interval/horizon
carbon_stock_g_cm2	g cm^{-2}	Soil carbon stock
carbon_perc	percent	Soil carbon percentage
pH	1	pH
Sand	percent	Soil sand percentage
Silt	percent	Soil silt percentage
Clay	percent	Soil clay percentage
SiltClay	percent	Soil silt + clay percentage
site	-	Study Area
CHM	m	Canopy Height Model
MAP	mm	Mean Annual Precipitation
MAT	degrees C	Mean Annual Temperature

PET	mm	Potential Evapotranspiration
DTM	m	Digital terrain model elevation
geomorphons	-	Geomorphons categories for geomorphology
HLI	1	Heat Load Index
LITH	-	Lithology
nlcd_reclass	-	Reclassified names for the National Land Cover Database
NDVI_median	1	Median Normalized Difference Vegetation Index
MNDWI_median	1	Median Modified Normalized Water/Wetness Index
EVI_median	1	Median Enhanced Vegetation Index
NDYI_median	1	Median Normalized Difference Yellow Index
dev_1000	m	Deviation from mean elevation at 1000-m scale
dev_300	m	Deviation from mean elevation at 300-m scale
dev_50	m	Deviation from mean elevation at 50-m scale
grad_1000	1	Gradient at 1000-m scale
grad_300	1	Gradient at 300-m scale
grad_50	1	Gradient at 50-m scale
WIP	percent	Wetland Intrinsic Potential

Table 6. Variables in the files *<site>_pts_<EPSG>.gpkg*.

Variable	Units	Description
sample_ID	-	Unique sample ID
lower_depth	cm	Depth of bottom interval/horizon
SOC_stock_spline	g cm^{-2}	Splined SOC stock data for 25cm intervals
site	-	Study area name
CHM	m	Canopy Height Model
MAP	mm	Mean Annual Precipitation
MAT	degrees C	Mean Annual Temperature
PET	mm	Potential Evapotranspiration
DTM	m	Digital Terrain Model or Elevation
GEO	-	Surficial Geology categories
geomorphons	-	Geomorphons categories for geomorphology
HLI	1	Heat Load Index
names	-	Names for land cover from the National Land Cover Database
LITH	-	Lithology
nlcd_reclass	-	Reclassified names for the National Land Cover Database
NDVI_median	1	Median Normalized Difference Vegetation Index
MNDWI_median	1	Median Modified Normalized Water/Wetness Index
EVI_median	1	Median Enhanced Vegetation Index
NDYI_median	1	Median Normalized Difference Yellow Index
dev_1000	m	Deviation from mean elevation at 1000-m scale
dev_300	m	Deviation from mean elevation at 300-m scale
dev_50	m	Deviation from mean elevation at 50-m scale
grad_1000	1	Gradient at 1000-m scale
grad_300	1	Gradient at 300-m scale
grad_50	1	Gradient at 50-m scale
WIP	percent	Wetland Intrinsic Potential

Table 7. Variables in *pnw_lab_data.csv*.

Variable	Units	Description
sample_ID	-	Unique sample ID
top_sample_soil_horizon	cm	Top of sample soil horizon
center_sample_soil_horizon	cm	Center of sample soil horizon
bottom_sample_soil_horizon	cm	Bottom of sample soil horizon
soil_rock	percent	Rock fragment percentage
soil_carbon	percent	Soil carbon percentage
carbon_stock	g cm^{-2}	Soil carbon stock
latitude	degrees north	Latitude in decimal degrees
longitude	degrees east	Longitude in decimal degrees
pH	1	pH
soil_sand	percent	Soil sand percentage
soil_silt	percent	Soil silt percentage
soil_clay	percent	Soil clay percentage
silt_clay	percent	Soil silt + clay percentage

Table 8. Variables in *pnw_stocks_data.csv*.

Variable	Units	Description
sample_ID	-	Unique sample ID
top_sample_soil_horizon	cm	Top of sample soil horizon
center_sample_soil_horizon	cm	Center of sample soil horizon
bottom_sample_soil_horizon	cm	Bottom of sample soil horizon
soil_rock	percent	Rock fragment percentage
latitude	degrees north	Latitude in decimal degrees
longitude	degrees east	Longitude in decimal degrees
soil_carbon	percent	Soil carbon percentage
carbon_stock	g cm^{-2}	Soil carbon stock
bulk_density	g cm^{-3}	Soil bulk density

3. Application and Derivation

These data could be useful to climate change studies and policies.

4. Quality Assessment

Mapped predictions of SOC stock across all four study watersheds were generated using the model with the best fit to the test dataset (based on R^2 and RMSE). This model was used to generate SOC predictions for each depth interval (25 cm) and then combined to calculate a 1-m SOC stock as well as the 97.5% and 2.5% quantiles of SOC stock to show uncertainty in the range.

Additional geospatial masks were used to remove surface water areas and urban zones. The Modified Normalized Difference Water Index (MNDWI) was used to remove surface water with a threshold of 0.3 (Xu, 2006). The National Land Cover Database (NLCD) was used to mask urban areas such as roads (Dewitz, 2021). Geospatial masks using the WIP threshold of 0.50 or 50% were used to estimate wetland and upland 1-m SOC stocks and were compared with 1-m SOC stock data from the National Wetland Condition Assessment (NWCA), which was upscaled using the National Land Cover Database (NLCD) in (Uhran et al., 2021).

5. Data Acquisition, Materials, and Methods

Study areas

Field surveys, sample collection, and geospatial modeling were conducted across four study watersheds in the Pacific Northwest that span a climatic gradient (Figure 1). The Heen Latinee Experimental Forest (HLEF) is located in southeast Alaska near Juneau, AK with a mean annual precipitation (MAP) of approximately 2284 mm and a mean annual temperature (MAT) of 4.3 degrees C. The Hoh River Watershed (HRW) is located on the west coast of the Olympic Peninsula in Washington state where MAP is approximately 3061 mm and MAT is 10.0 degrees C. The Mashel River Watershed (MRW) is located on the western side of the Cascade Mountain Range near Tahoma, Washington, (Mt. Rainier) and has a MAT of 8.4 degrees C and a MAP of approximately 1716 mm. The Colville Watershed (CVW) is located on the northeastern portion of Washington state and has a MAT of 6.5 degrees C and a MAP of 652 mm.

The landscapes of the four study areas contain a range of landforms including mountains, hillslopes, alluvial floodplains, and glacial moraines and outwash plains. Geologic ages range from late Quaternary deposits from retreating glaciers in the HLEF and HRW, Oligocene-Eocene volcanic and sedimentary rocks in the MRW to Pre-Tertiary metamorphic rocks in the CRW. Vegetation is predominantly forested with the HLEF containing a mix of Western Hemlock (*Tsuga heterophylla*), and Sitka Spruce (*Picea sitchensis*) with small amounts of Yellow Cedar (*Cupressus nootkatensis*). The HRW forests contain Sitka Spruce (*Picea sitchensis*), Western Hemlock (*Tsuga heterophylla*), Western Red Cedar (*Thuja plicata*), and Bigleaf Maple (Acer macrophyllum). The MRW forests contain predominantly Douglas fir (*Pseudotsuga menziesii*), Western Hemlock (*Tsuga heterophylla*), and Red Alder (*Alnus rubra*). The CVW forests contain a mix of Ponderosa Pine (*Pinus ponderosa*), Douglas fir (*Pseudotsuga menziesii*), Lodgepole pine (*Pinus contorta*), Western Larch (*Larix occidentalis*), Western Hemlock (*Tsuga heterophylla*), and Western Red Cedar (*Thuja plicata*).

Wetland Identification with the Wetland Intrinsic Potential (WIP) Tool

The Wetland Intrinsic Potential (WIP) tool was implemented for each study watershed to model the gridded land surface in each study area as a continuous probability of wetland presence, with each grid pixel containing a value from 0-100% (Halabisky et al., 2023). The WIP tool relies on multi-scale terrain metrics derived from digital terrain models (DTMs) as predictors of wetland presence on the landscape, which has been shown to improve wetland classification (Maxwell et al., 2016; Maxwell and Warner, 2019). These terrain metrics along with additional remote sensing metrics representing wetland identifiers and formation factors were used in a Random Forest model (Breiman, 2001) with training data derived from the National Wetland Inventory and field-visit observations. The Random Forest model prediction assigns grid pixels a probability value (0-1.0, revised to 0-100% for interpretability) based on a “wetland” or “upland” designation. We follow Halabisky et al. (2023) and others in using a probability threshold of ≥50% to define the cutoff for defining a wetland (≥50%) or upland (<50%). Both the data and framework of the HRW WIP implementation were utilized to derive similar WIP models for the HLEF, MRW, and CRW study watersheds.

Field sampling and laboratory analysis

Soil samples were collected over the course of two summer field seasons in 2021-2022. Sample locations were pre-selected using a stratified random selection of points across the WIP probability for each study watershed in order to evaluate the full spread of the wetland-upland gradient. The WIP probability from 0-100% was stratified into 30 equal size bins, and sample points were generated for those bins at random locations within each study watershed. Sample locations were then visited in the field as close to the sample point coordinates as possible and then updated with new GPS/GNSS location data. Overall, 114 samples were collected from locations with 30 pedons each in the MRW and CVW, 36 pedons in the HRW, and 18 pedons in the HLEF. The HLEF also contained three legacy soil pedons sampled in 2019. Soil pedons were dug to at least 1-m depth unless there was a restrictive layer. Soil pedons were characterized according to NRCS field sampling guides (Soil Survey Staff, 2022) and soil samples for bulk density and bulk chemistry were extracted by horizon and at the horizon center. Laboratory analysis started with drying samples to a constant weight at 70 degrees C and sieving to 2 mm to remove coarse fragments. The mass of the <2-mm fraction was then used to calculate bulk density and analyze carbon content, particle size distribution, and pH. Carbon content was measured using a Shimadzu Elemental Analyzer. Particle size distribution was measured using the hydrometer method. Soil pH was measured using 1:1 soil to water ratio, although a small number of soils required 1:1.5 or 1:2 soil to water ratios.

Geospatial data

Geospatial datasets were gathered to determine predictors for SOC stocks across the three study areas. Geospatial factors were related to vegetation, climate, lithology and geology, and topography.

Google Earth Engine (Gorelick et al., 2017) was used to obtain satellite imagery for calculation of vegetation spectral indices from the five year median of Sentinel-2 reflectance. The enhanced vegetation index (EVI) (Huete et al., 2002) and normalized difference vegetation index (NDVI) were calculated for each study area. A lidar derived canopy height model from the Washington Department of Natural Resources (WA DNR) was used which measured the elevation difference between the lidar DTM and the lidar digital surface model.

For climate predictors, 800-m^2 gridded 30 year climate normals for mean annual precipitation (MAP) and mean annual temperature (MAT) were downloaded from the Northwest Alliance for Computation Science and Engineering (PRISM Climate Group, 2014.) for the three Washington study watersheds. Gridded historical climate data (~1 km) were used from WorldClim 2.1 to extract MAP and MAT for the HLEF. Potential Evapotranspiration (PET) from the Version 3 of the Global Aridity Index and Potential Evapotranspiration Database for all four study areas (Zomer et al., 2022) were also downloaded. PET was divided by MAP to create an Aridity Index estimate of water availability. A heat load index (HLI) from the lidar DTM (McCune and Keon, 2002) was calculated to capture finer scale effects of solar radiation and heat.

Geology and lithology maps were downloaded from the Washington Department of Natural Resources at a 1:100,000 scale and from the USGS (Wilson et al., 2015). These maps were consolidated into six broad categories that represent soil Parent Material (PM) and time since soil inception: Glacial Till and Drift, Glacial Outwash, Igneous, Metamorphic, Sedimentary, and Unconsolidated sediments.

Modeling approach

Linear Mixed Modeling

Two model types were used in the research to model SOC stock and SOC percent: a linear mixed effects model (LMM) and a quantile random forest (RFM). An LMM approach was used to test specific hypotheses about important predictors and examine predictor coefficients. To compare models, both datasets for SOC stock and SOC % were split into training and testing datasets using a 85:15 ratio. The same training and testing datasets were used for all models to compare performance and accuracy. Models were also compared with and without the WIP predictor.

In the LMMs, the sample location was chosen as a random effect and sample depth was specified as a random slope within the sample location. Several models were chosen that represented a priori hypotheses with varying interactions between the predictors as well as eliminating predictors that were not significant down to a null model with no predictors. The final model was chosen using Akaike’s information criterion (AIC) and parsimony then evaluated for violations of homogeneity of variance and linearity. Although some observations deviated slightly from the homoscedasticity assumption, it was deemed that the model assumptions were sufficiently met and therefore used to evaluate drivers and predict SOC stock and SOC %.

All data used to build models of SOC stock and SOC % were centered and scaled if they were continuous predictors. Correlated variables were removed if they had an r >0.6. However, at this scale MAP and PET:MAP are highly correlated, so separate models were used to assess the strength between the two predictors in our mixed model selection. The initial model selection process using maximum likelihood showed that models using PET:MAP, instead of MAP, were better fits to the data according to AIC.

For SOC stock, LMM selection using AIC comparisons removed models containing interactions among fixed effects, resulting in SOC stock modeled as a function of PET:MAP, WIP, Canopy Height (CHM), Heat Load Index (HLI), Parent Material (PM), and Depth as fixed effects. After this LMM was found, the dredge function from the MuMIn R package was used to further identify if incorporated interactions outperformed the current LMM structure determined from earlier (Barton, 2023). The final LMM was chosen with the lowest AIC and/or fewest parameters for parsimony.

The final model was evaluated on the test dataset by measuring the R^2 and root mean square error (RMSE). Predictor variable importance was estimated using the 95% confidence intervals on the scaled coefficient estimates with important predictors having a confidence interval that did not include 0.

SOC stock data and model residuals were also examined for spatial autocorrelation by examining semivariograms and the spatial autocorrelation within each study watershed. The expectation was for there to be minimal autocorrelation.

Random Forest Modeling - SOC stock prediction and uncertainty

The Quantile Random Forest (RFM) was used to accommodate for potential non-linear relationships as well as incorporate a more flexible approach to predictor selection. The *mlr3* package in R was used to implement the RFM from the ranger package (Lang et al., 2019; Wright and Ziegler, 2017). The num.trees, mtry, and max.depth parameters were tuned using repeated 10-fold cross validation on the training dataset. Predictor variable importance was determined using SHAP (SHapley Additive exPlanations) values using a 200-observation subset of the training dataset as the rows to be explained and the full training dataset as the background (Lundberg and Lee, 2017). Interactions between predictors were investigated using the interpretable machine learning *iml* package and the Interaction function (Molnar et al., 2018).

Mapped predictions of SOC stock across all four study watersheds were generated using the model with the best fit to the test dataset (R^2 and RMSE). This model was used to generate SOC predictions for each depth interval (25 cm) and then combined to calculate a 1-m SOC stock as well as the 97.5% and 2.5% quantiles of SOC stock to show uncertainty.

Estimates of SOC stock were also extracted from open data sources that have been used to model the spatial distribution of SOC at regional to global scales. For all study areas, an estimated mean 1-m SOC stock was used from an ensemble combination of global and regional models compiled and calculated by Jones and D'Amore (2024) for Alaska, British Columbia, Washington, Oregon, California, and Hawai'i. Additionally, a published map of 1-m SOC stocks was created specifically for the Pacific Coastal Temperate Rainforest by McNicol et al. (2019) which encompasses the HLEF.

6. Data Access

These data are available through the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC).

Geospatial Data for Modeling Soil Carbon Stocks across Pacific Northwest Watersheds

Contact for Data Center Access Information:

- E-mail: uso@daac.ornl.gov
- Telephone: +1 (865) 241-3952

7. References

Barton, K. 2023. *MuMIn: Multi-model inference* manual. https://CRAN.R-project.org/package=MuMIn

Breiman, L. 2001. Random Forests. Machine Learning 451:5–32. https://doi.org/10.1023/A:1010933404324

Dewitz, J. 2021. National Land Cover Database NLCD 2019 products. Version 3.0. February 2024. U.S. Geological Survey. https://doi.org/10.5066/P9KZCM54

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202:18–27. https://doi.org/10.1016/j.rse.2017.06.031

Halabisky, M., D. Miller, A.J. Stewart, A. Yahnke, D. Lorigan, T. Brasel, and L.M. Moskal. 2023. The wetland intrinsic potential tool: Mapping wetland intrinsic potential through machine learning of multi-scale remote sensing proxies of wetland indicators. Hydrology and Earth System Sciences 2720:3687–3699. https://doi.org/10.5194/hess-27-3687-2023

Huete, A., K. Didan, T. Miura, E.P. Rodriguez, X. Gao, and L.G. Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 831:195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jones, D., and D.V. D'Amore. 2024. Ensemble model and input model rasters for soil organic carbon stock mean and uncertainties for Alaska, British Columbia, Washington, Oregon, California, and Hawai'i. Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2024-0009

Lang, M., M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, and B. Bischl. 2019. mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software 4:1903. https://doi.org/10.21105/joss.01903

Lundberg, S.M., and S.-I. Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems.* Volume 0. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maxwell, A.E. and T.A. Warner. 2019. Is high spatial resolution DEM data necessary for mapping palustrine wetlands? International Journal of Remote Sensing 40:118-137. https://doi.org/10.1080/01431161.2018.1506184

Maxwell, A.E., T.A. Warner, and M.P. Strager. 2016. Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables. Photogrammetric Engineering and Remote Sensing 826:437–447. https://doi.org/10.14358/PERS.82.6.437

Mayer, M., and D. Watson. 2024. *kernelshap: Kernel SHAP* manual. https://github.com/ModelOriented/kernelshap

McCune, B., and D. Keon. 2002. Equations for potential annual direct incident radiation and heat load. Journal of Vegetation Science 134:603–606. https://doi.org/10.1111/j.1654-1103.2002.tb02087.x

McNicol, G., C. Bulmer, D. D'Amore, P. Sanborn, S. Saunders, I. Giesbrecht, S.G. Arriola, A. Bidlack, D. Butman, and B. Buma. 2019. Large, climate-sensitive soil carbon stocks mapped with pedology-informed machine learning in the North Pacific coastal temperate rainforest. Environmental Research Letters 14:014004. https://doi.org/10.1088/1748-9326/aaed52

Molnar, C., G. Casalicchio, and B. Bischl. 2018. iml: An R package for interpretable machine learning. Journal of Open Source Software 326:786. https://doi.org/10.21105/joss.00786

PRISM Climate Group. 2014. 30 Year Normals. Oregon State University. https://prism.oregonstate.edu

Soil Survey Staff. 2022. *Keys to soil taxonomy* 13th ed. USDA Natural Resources Conservation Service. https://www.nrcs.usda.gov/sites/default/files/2022-09/Keys-to-Soil-Taxonomy.pdf

Stewart, A. Improving soil organic carbon spatial distribution and interpretation of cross-scale drivers with probabilistic wetland representation. 2026. In preparation.

Uhran, B., L. Windham-Myers, N. Bliss, A.M. Nahlik, E.T. Sundquist, and C.L. Stagg. 2021. Improved wetland soil organic carbon stocks of the conterminous U.S. through data harmonization. Frontiers in Soil Science Volume 1 - 2021. https://www.frontiersin.org/articles/10.3389/fsoil.2021.706701

Wilson, F.H., C.P. Hults, C.G. Mull, and S.M. Karl. 2015. *Geologic map of Alaska: U.S. Geological Survey Scientific Investigations Map 3340.* Scientific

Investigations Map No. pamphlet 196 p. 2. https://pubs.usgs.gov/publication/sim3340

Wright, M.N. and A. Ziegler. 2017. ranger: A fast implementation of Random Forests for high dimensional data in C++ and R. Journal of Statistical Software 77:1-17. https://doi.org/10.18637/jss.v077.i01

Xu, H. 2006. Modification of normalized difference water index NDWI to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing 27:3025–3033. https://doi.org/10.1080/01431160600589179

Zomer, R.J., J. Xu, and A. Trabucco. 2022. Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. Scientific Data 91:409. https://doi.org/10.1038/s41597-022-01493-1