User Guide: Northeast Alpine Zone (35 sites) Landsat NDVI Time Series Analysis Code

Overview

This R workflow processes multi-decadal Landsat NDVI time series data from 35 montane study sites across northeastern US and eastern Canada to analyze long-term vegetation trends and phenological changes in alpine ecosystems. The workflow consists of 12 sequential processing steps that transform raw Google Earth Engine-exported CSV files into analyses of long-term trends and phenological change.

The workflow starts with time series NDVI data exported by Google Earth Engine as CSV text files. Each study site contains two zones: the Alpine Zone (AZ) above treeline with canopy heights <2m, and the Upper Subalpine Zone (USAZ) defined as a 200m buffer around each alpine zone. The analysis spans the growing season (June 1 - September 30) from 1984-2024, using data from the Landsat-5, -7, -8, and -9 satellites. The workflow produces a large number of output files with information about NDVI dynamics at all 35 sites and 7 subregions. A slightly revised set of R code files does similar processing steps for sites where data are available on specific ecological communities within the alpine zone.

Key outputs include deseasonalized annual NDVI trends (seasonal maximum and seasonal mean); phenological change (timing of maximum NDVI); various robustness assessments, comparisons, and spatial autocorrelation analyses; and graphics.

Step 01: Data Import and Integration

This step reads and combines individual CSV files exported from Google Earth Engine with site metadata to create a unified analysis dataset. The script handles the hierarchical directory structure (state/province → site → data files) and resolves the (rare) duplicate observations that may occur at sites located in the endlap between consecutive pairs of Landsat images along the same WRS-2 orbit path.

Statistical Methods: n/a.

(i) Input Files:

- 35sites_info.csv: Site metadata including coordinates, zone areas, and subregion classifications
- Individual CSV files in “input/[SP]/[Site]_GEE_v4/” directories containing NDVI time series data ending in “_pc50.csv”

(ii) Output Files:

- **Intermediate**: step01_analysis_file.csv - Combined dataset for subsequent processing
- **Results**: n/a

(iii) Important Settings:

- Directory paths can be modified at the beginning of the script
- File naming patterns (_pc50.csv, _GEE_v4) define which files are processed

(iv) Required R Packages:

- tidyverse: Data manipulation and file I/O
- lubridate: Date handling

(v) Important Notes:

- Handles duplicate observations by selecting the observation with higher percent valid pixels or averaging when percent valid is equal (these cases are not common but do occur, only for endlap between successive Landsat images along a single orbit path)
- Creates standardized site_id codes (SP_Site) for consistent identification across the workflow
- Logs the number of sites and observations successfully loaded

---

Step 02: Data Preparation and Quality Assessment

This step generates data coverage summaries by decade and subregion.

Statistical Methods: Descriptive statistics and data coverage assessment across temporal and geographic dimensions.

(i) Input Files:

- step01_analysis_file.csv: Combined dataset from Step 01

(ii) Output Files:

- **Intermediate**: step02_analysis_data_prepared.csv - Analysis-ready dataset with derived variables
- **Results**: step02_coverage_summary.csv - Coverage by subregion/decade; step02_site_summary.csv - Site-level coverage statistics; coverage_by_site.csv and coverage_by_subregion.csv in tables subdirectory

(iii) Important Settings:

- Quality threshold: pct_valid >= 0.5 (50% valid pixels minimum)
- Alpine zone size classes: Small (<100 ha), Medium (100-1000 ha), Large (>1000 ha)

(iv) Required R Packages:

- tidyverse: Data manipulation and summarization

(v) Important Notes:

- Generates coverage statistics for observation density across space and time

---

Step 03: Robustness Checks - Data Quality

This step assesses potential data quality issues including the relationship between percent valid pixels and NDVI values, seasonal patterns in data availability, and systematic differences between Landsat paths that may result from east-viewing vs west-viewing geometry.

Statistical Methods: Correlation analysis, t-tests for path effect comparison, seasonal pattern analysis.

(i) Input Files:

- step02_analysis_data_prepared.csv: Prepared dataset from Step 02

(ii) Output Files:

- **Results**: step03_pctvalid_within_site_correlations.csv - Correlations between percent valid and NDVI by site; step03_low_pctvalid_sites.csv - Sites ranked by data quality; step03_seasonal_pctvalid.csv - Monthly patterns; step03_ndvi_by_pctvalid_ranges.csv - NDVI by quality ranges; step03_multi_path_sites.csv - Sites imaged from multiple paths; step03_path_effects_analysis.csv - Statistical comparison of path effects; step03_path_coverage_by_time.csv - Temporal distribution of path coverage

(iii) Important Settings:

- Correlation thresholds for flagging potential bias (|r| > 0.3)
- Minimum observation requirements for statistical tests

(iv) Required R Packages:

- tidyverse: Data manipulation and analysis

- lubridate: Date processing

(v) Important Notes:

- n/a

Step 04: Phenological Modeling (Time-varying Seasonal Cycles)

This step implements the phenological modeling using Generalized Additive Models (GAM) with tensor product smoothing to capture time-varying seasonal cycles. The approach models long-term underlying phenology while accounting for interannual variation, and handles deseasonalization of NDVI observations.

Statistical Methods: GAM with tensor product smoothing, Leave-One-Year-Out cross-validation, bootstrap analysis for parameter sensitivity.

(i) Input Files:

- step02_analysis_data_prepared.csv: Prepared dataset from Step 02

(ii) Output Files:

- **Intermediate**: step04_deseasonalized_data.csv - NDVI observations with seasonal adjustments for mean and max
- **Results**: step04_phenological_model_summary.csv - Model performance statistics; step04_loyo_cv_results.csv - Cross-validation results; step04_nh_pre_az_k_iteration_test.csv - Parameter sensitivity analysis for sample site

(iii) Important Settings:

- Tensor smoothing parameters: k_doy_final = 4, k_year_final = 8 (users may adjust for different smoothing levels)
- Growing season definition: season_start = 152 (June 1), season_end = 273 (September 30)
- Minimum observations for model fitting: 20 per site-zone combination (not really relevant for this study; there are generally over 100 per site-zone combination)

(iv) Required R Packages:

- tidyverse: Data manipulation
- mgcv: GAM modeling

- purrr: Functional programming

(v) Important Notes:

- Generates two sets of adjusted NDVI values with each value being an estimate of (a) seasonal max NDVI, or (b) seasonal mean NDVI
- Conservative smoothing parameters prevent overfitting due to sparse data in most years
- Cross-validation results provide model validation metrics
- The tensor product approach allows seasonal cycles to vary smoothly across years
- While the volume of data at individual sites across the 41-year period is lower than would be preferred for this modeling, it is better than the obvious alternatives (no seasonal adjustment of observations; seasonal adjustment using only a static model of phenology)

Step 05: De-seasonalization and Annual Summaries

This step aggregates the deseasonalized NDVI observations to annual values, applying Gaussian weighting based on distance from modeled peak growing season timing (for NDVI max only). The approach produces less-noisy annual estimates while accounting for uneven temporal sampling within seasons. It mitigates observation frequency bias and reduces noise in the trends that will be calculated in the next step.

Statistical Methods: Gaussian-weighted annual aggregation, standard error calculation for uncertainty quantification.

(i) Input Files:

- step04_deseasonalized_data.csv: Deseasonalized observations from Step 04

(ii) Output Files:

- **Intermediate**: step05_annual_for_trends.csv - Annual summaries filtered for trend analysis (≥8 years)
- **Results**: step05_annual_summaries.csv - Complete annual summaries; step05_sites_for_trends.csv - Site filtering criteria; step05_data_completeness_summary.csv - Data completeness assessment

(iii) Important Settings:

- Gaussian weighting parameter: sigma = 60 days (controls how observations are weighted relative to peak timing; this large value ensures a large “window” such that even early/late observations will contribute to the seasonal estimate)
- Minimum years for trend analysis: 8 years per site-zone combination (this has no practical effect with this study’s data - all sites have close to 41 years of data - but it is left here in case the code gets repurposed for other uses)

(iv) Required R Packages:

- tidyverse: Data manipulation and summarization

(v) Important Notes:

- Weighting function gives highest weight to observations near the modeled seasonal peak
- Produces both raw NDVI and deseasonalized annual summaries for comparison

---

Step 06: Long-Term Trend Analysis

This step quantifies long-term NDVI trends using LOESS regression and Mann-Kendall tests. LOESS models are fitted to annual data with consistent start (1984) and end (2024) points across all sites, enabling direct comparison of trend magnitudes. The approach includes statistical significance testing and trend classification. Significance is based on a fairly conservative estimate of degrees of freedom (# of years of data at a given site, minus 4). Note that subsequent analysis of the output from this step shows that the LOESS-based estimates of seasonal mean and peak NDVI are very close to those from the GAM in Step 7.

Statistical Methods: LOESS regression with extrapolation capability, Mann-Kendall trend tests, weighted regression using annual observation counts.

(i) Input Files:

- step05_annual_for_trends.csv: Annual summaries from Step 05

(ii) Output Files:

- **Results**: step06_trend_analysis_results_[mean|max].csv - LOESS trend results for mean and max NDVI; step06_trend_summary_by_region_[mean|max].csv - Regional trend summaries; step06_mk_test_results_[mean|max].csv - Mann-Kendall test results

(iii) Important Settings:

- LOESS span: span_val = 0.75 (controls smoothing level; smaller values = less smoothing)
- Trend period: start_year = 1984, end_year = 2024 (standardized across all sites)
- Degrees of freedom: # of years minus 4
- Significance threshold: $p < 0.05$ for trend classification

(iv) Required R Packages:

- tidyverse: Data manipulation
- Kendall: Mann-Kendall trend tests

(v) Important Notes:

- LOESS extrapolation by 1 time-step allows consistent trend period even for sites missing first year of data (this is via the surface = "direct" option in the LOESS function)
- Provides both LOESS and Mann-Kendall non-parametric trend assessments

---

Step 07: Phenological Change Analysis

This step quantifies changes in phenological metrics (peak timing, peak NDVI amplitude, seasonal mean NDVI) between study period endpoints using bootstrap resampling for uncertainty estimation. The analysis employs GAM models fitted to individual sites to extract seasonal characteristics, then estimates 95% confidence intervals through bootstrap sampling.

Statistical Methods: Bootstrap resampling (100 iterations default), GAM modeling, confidence interval estimation for phenological metrics.

(i) Input Files:

- step02_analysis_data_prepared.csv: Raw NDVI observations from Step 02
- step04_phenological_model_summary.csv: Model success indicators from Step 04

(ii) Output Files:

- **Results**: step07_phenological_metrics.csv - Phenological metrics with confidence intervals; step07_phenological_changes_start_end.csv - Change magnitudes and significance; step07_bootstrapped_predictions_[site_id]_[zone].csv - Detailed predictions for single site analysis (optional)

(iii) Important Settings:

- Bootstrap iterations: n_bootstrap = 100 (can be increased for higher precision)
- Analysis period: start_year = 1984, end_year = 2024
- Growing season: start_doy = 152, end_doy = 273
- Single site option: run_single_site = TRUE/FALSE (enables detailed output for one site)
- GAM parameters: k_doy_final = 4, k_year_final = 8 (must match Step 04 settings)

(iv) Required R Packages:

- tidyverse: Data manipulation
- mgcv: GAM modeling

(v) Important Notes:

- This is the most computationally intensive step (runtime: several minutes vs. seconds for other steps)
- Bootstrap confidence intervals represent 95% CI around phenological metrics
- Significance assessment based on non-overlapping confidence intervals between time periods
- Single site analysis option provides detailed bootstrap predictions for method validation

---

Step 08: Robustness Checks - Observation Frequency Bias

This step evaluates potential biases arising from systematic changes in observation frequency over time and tests the sensitivity of trend estimates to different analytical approaches. The analysis includes balanced subsampling, early vs. late period comparisons, and LOESS span sensitivity testing.

Statistical Methods: Linear regression for balanced vs. full dataset comparison, correlation analysis for observation frequency effects, LOESS sensitivity analysis across multiple span parameters.

(i) Input Files:

- step05_annual_for_trends.csv: Annual summaries from Step 05

(ii) Output Files:

- **Results**: step08_observation_frequency_bias.csv - Assessment of observation frequency effects; step08_full_vs_balanced_comparison.csv - Comparison of trend estimates using different data subsets; step08_early_vs_late_period_comparison.csv - Temporal period comparisons; step08_loess_sensitivity_analysis.csv - Sensitivity to LOESS parameters

(iii) Important Settings:

- Balanced sampling criteria: 3-6 observations per year, ≥5 years in each half of study period
- Correlation thresholds for bias detection: $|r| > 0.4$ for NDVI-observation relationship
- LOESS span range: 0.5 to 0.9 for sensitivity testing

(iv) Required R Packages:

- tidyverse: Data manipulation
- broom: Model output tidying

(v) Important Notes:

- Balanced subsampling tests whether increasing observation density over time affects trend estimates

---

Step 09: Zone Comparisons (Alpine vs. Upper Subalpine)

This step conducts paired comparisons between Alpine Zone (AZ) and Upper Subalpine Zone (USAZ) measurements to quantify systematic differences in NDVI values and trends. The analysis employs mixed-effects modeling to account for site-level correlation and examines regional patterns in zone differences. (The elevation component of this is not particularly meaningful, but included for completeness).

Statistical Methods: Paired t-tests, mixed-effects modeling (lme), temporal difference analysis, correlation assessment for trend comparisons.

(i) Input Files:

- step05_annual_for_trends.csv: Annual summaries from Step 05
- step06_trend_analysis_results_[mean|max].csv: Trend results from Step 06

(ii) Output Files:

- **Results**: step09_paired_sites_annual_[mean|max].csv - Paired annual comparisons; step09_site_specific_comparisons_[mean|max].csv - Site-level zone differences; step09_trend_analysis_[mean|max].csv - Trend comparisons between zones; step09_regional_zone_summary_[mean|max].csv - Regional patterns; step09_elevation_zone_effects_[mean|max].csv - Elevation effects on zone differences; step09_zone_mixed_model_results_[mean|max].csv - Mixed-effects model results

(iii) Important Settings:

- Minimum years for site-specific comparisons: 8 years (ensures robust statistical testing; not actually relevant here, because all sites have close to 41 years of data)
- Mixed-effects model includes zone, year, subregion, and relative elevation as fixed effects

(iv) Required R Packages:

- tidyverse: Data manipulation
- nlme: Mixed-effects modeling

(v) Important Notes: -

- Pairs sites that have both AZ and USAZ data for direct comparison

---

Step 10: Regional and Multi-factor Analysis

This step synthesizes results across multiple factors including subregional patterns, elevation effects, alpine zone size influences, and phenological changes. The analysis identifies the strongest and most consistent trends across the study domain and examines interactions between environmental factors. Neither alpine zone size nor relative elevation is emphasized in the current study, but they are included here for completeness.

Statistical Methods: Multi-factor summary statistics, correlation analysis for elevation effects, interaction analysis between elevation and alpine zone size.

(i) Input Files:

- step06_trend_analysis_results_[mean|max].csv: Trend results from Step 06
- step07_phenological_changes_start_end.csv: Phenological changes from Step 07

(ii) Output Files:

- **Results**: step10_regional_multifactor_trends_[mean|max].csv - Multi-factor trend summaries; step10_elevation_effects_within_subregions_[mean|max].csv - Elevation-trend relationships; step10_az_size_effects_[mean|max].csv - Alpine zone size effects; step10_geographic_gradient_analysis_[mean|max].csv - Geographic patterns; step10_phenological_multifactor_summary.csv - Phenological change patterns; step10_strongest_most_consistent_trends_[mean|max].csv - Identification of robust patterns; step10_elevation_size_interaction_[mean|max].csv - Factor interactions

(iii) Important Settings:

- Trend strength classification: Strong (≥0.02/decade), Moderate (0.01-0.02/decade), Weak (<0.01/decade)

(iv) Required R Packages:

- tidyverse: Data manipulation and analysis

(v) Important Notes:

- n/a

Step 11: Spatial Autocorrelation Analysis

This step evaluates spatial autocorrelation in NDVI trends and phenological changes using Moran's I statistics. The analysis tests whether nearby sites exhibit more similar trends than distant sites.

Statistical Methods: Moran's I spatial autocorrelation tests, distance-based correlation analysis, neighbor identification using geographic distance criteria.

(i) Input Files:

- 35sites_info.csv: Site coordinates
- step06_trend_analysis_results_[mean|max].csv: Trend results from Step 06
- step07_phenological_changes_start_end.csv: Phenological changes from Step 07

(ii) Output Files:

- **Results**: step11_spatial_autocorrelation_results_trends.csv - Moran's I results for trend data; step11_spatial_autocorrelation_results_phenology.csv - Spatial autocorrelation in phenological changes; step11_distance_trend_similarity.csv - Distance-similarity relationships; step11_spatial_analysis_summary.csv - Summary of spatial patterns

(iii) Important Settings:

- Maximum distance for neighbor identification: 500 km (can be adjusted based on study region size)
- Distance categories for similarity analysis: <50 km, 50-150 km, 150-500 km, >500 km
- Significance threshold: $p < 0.05$ for spatial autocorrelation

(iv) Required R Packages:

- tidyverse: Data manipulation
- spdep: Spatial dependence analysis
- sp: Spatial data classes
- nlme: Mixed-effects modeling

(v) Important Notes:

- Analyzes both AZ and USAZ zones separately for each NDVI metric

---

Step 12: Data Visualization (In Development)

This step creates various graphics for the analysis results including time series plots, phenological change graphics, and others. Note that the code specifying the output file can be edited to save the results as a .png or .svg file. There are other settings that can also be modified; examine closely before using code.

- step12a_graphics_step06a_lineplot: This code creates a line plot for the LOESS models from step06.
- step12b_graphics_step06b_boxplot: Not used in this study.
- step12c_graphics_step06c_bubbleplot: This code creates a bubble plot, where each site-zone combination has a circle symbol whose XY position is based on its initial NDVI value (X) and change in NDVI (Y), whose area is based on its zone polygon area, and whose color is based on zone (alpine zone in tan, USAZ in green).
- step12d_graphics_step07_phenology_early_vs_late: This code makes plots comparing the early (1^{st} five years) vs late (last 5 yrs) phenological models for 2 or 3 sites, with 2 to 7 zones per site. It uses bootstrapping to generate confidence intervals for the plots. It includes observations (which include annual weather noise) plus models (which include only underlying phenology based on long term change, not short term variability)

- step12f_graphics_step04_pheno_curves: This generates a “small multiples” figure showing many years of phenological models and observations for a single site and zone.

EcoComm (Ecological Communities) code

The code files in this subfolder are for doing the same processing steps (generally speaking – not all steps are included) for sites where detailed data on ecological communities have been mapped. The step numbering matches the steps in the main code, and the descriptions of most steps are similar.