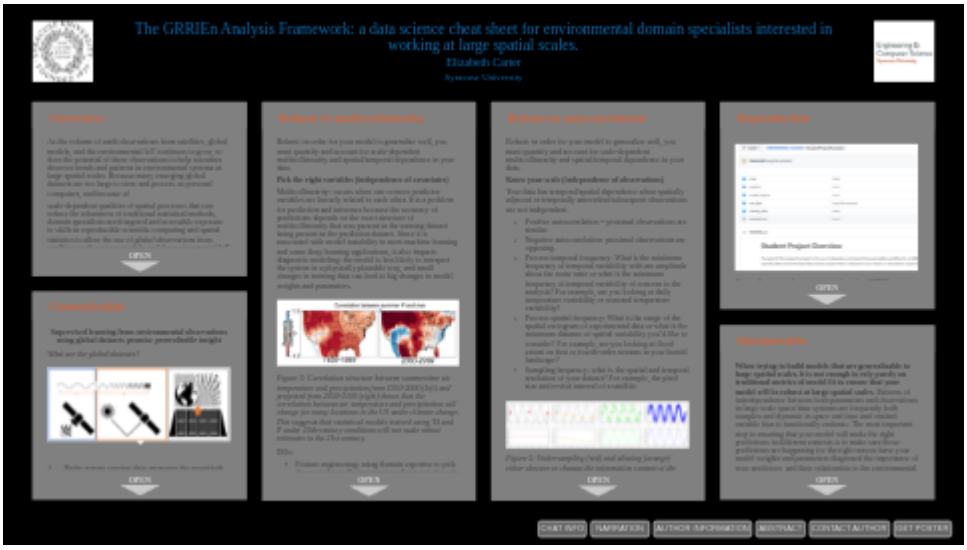
The GRRIEn Analysis Framework: a data science cheat sheet for environmental domain specialists interested in working at large spatial scales.



Elizabeth Carter

Syracuse University

Engineering & Computer Science
Syracuse University

PRESENTED AT:

OVERVIEW:

As the volume of earth observations from satellites, global models, and the environmental IoT continues to grow, so does the potential of these observations to help scientists discover trends and patterns in environmental systems at large spatial scales. Because many emerging global datasets are too large to store and process on personal computers, and because of

scale-dependent qualities of spatial processes that can reduce the robustness of traditional statistical methods, domain specialists need targeted and accessible exposure to skills in reproducible scientific computing and spatial statistics to allow the use of *global observations* from satellites, earth systems models, and the environmental IoT to generalize insights from *in-situ* field observations across unsampled times and locations. The GRRIEn framework, which stands for **generalizable, robust, reproducible, and interpretable environmental** analytic framework was developed for this purpose.

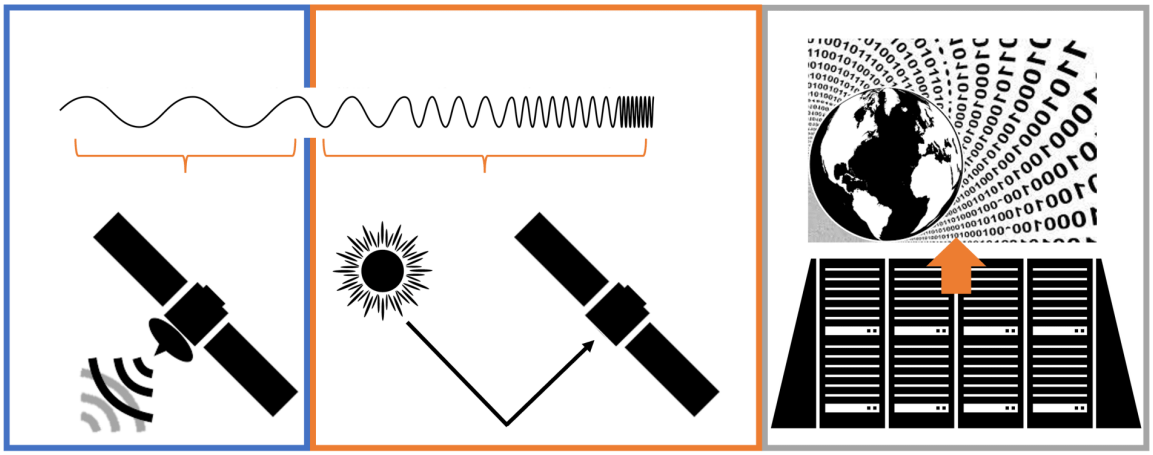
- **G**eneralizable: how well do your experimental results from a sample extend to the population as a whole?
- **R**obust: do your statistics show good performance on data drawn from a wide range of probability and joint probability distributions?
- **R**eproducible: can other scientists understand and replicate your analysis and yield the same results?
- **I**nterpretable: do your model parameters reflect a physically plausible diagnosis of the system?

Environmental analysis

GENERALIZABLE

Supervised learning from environmental observations using global datasets promise *generalizable* insight

What are the global datasets?



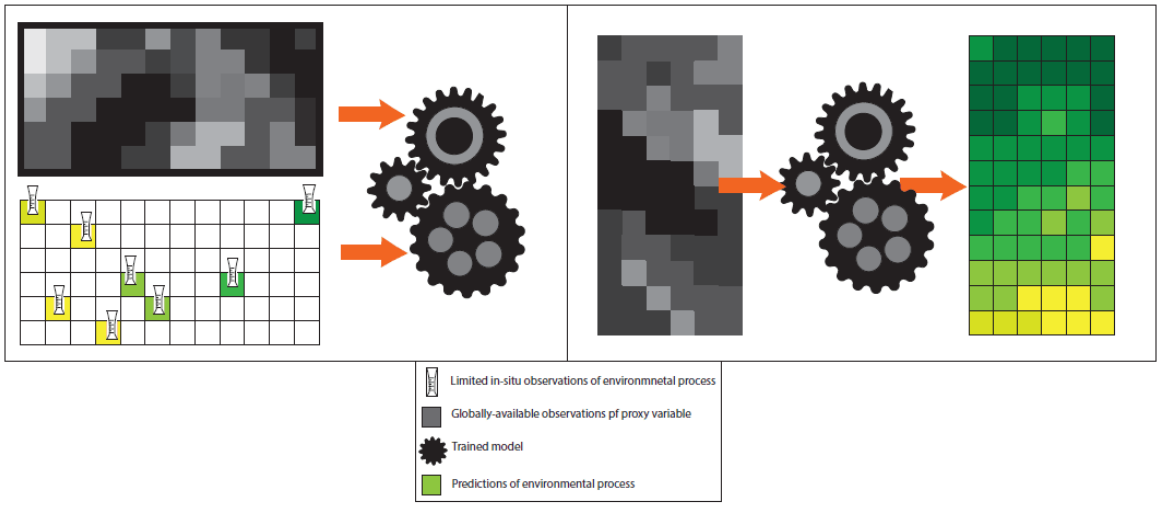
1. Radar remote sensing data: measures the magnitude of microwave radiation reflected from a pulse of a specific wavelength transmitted by the satellite, tells us about structural and dielectric properties of earth surface features.

2. Optical remote sensing data: measures the magnitude of electromagnetic radiation reflected from sunlight at select wavelengths between 300-3000μm, tells us about the chemical characteristics of earth surface features.

3. Coupled earth system model output: coupled process models describing terrestrial, oceanic, and atmospheric fluxes of matter or energy in environmental systems, run at standard time step and grid scale for the globe, provide spatially continuous estimates of thousands of difficult and impossible to measure parameters for all locations at various frequencies.

Other notable datasets: Paleoclimate reconstructions, LiDAR data, SfM photogrammetric models, socioeconomic data, gravity altimetry, passive microwave.

What is supervised learning?



Supervised learning lets us use global datasets to generalize the results of our local experiments through interpolation, prediction, and diagnostic analysis.

1. **Interpolate:** “fill in gaps” in spatial/temporal record of environmental process represented by in-situ observations using globally-available proxy signal.

2. **Predict:** infer values in yet-to be explored spaces/times using globally-available proxy signal.

3. **Diagnose:** explore values of model weights/parameters to understand how globally available proxy signa indicates physical drivers of variability in the environmental process.

ROBUST TO MULTICOLLINEARITY

Robust: in order for your model to generalize well, you must quantify and account for scale-dependent multicollinearity and spatial/temporal dependence in your data.

Pick the right variables (independence of covariates)

Multicollinearity: occurs when one or more predictor variables are linearly related to each other. It is a problem for prediction and inference because the accuracy of predictions depends on the exact structure of multicollinearity that was present in the training dataset being present in the prediction dataset. Since it is associated with model instability in most machine learning and some deep learning applications, it also impacts diagnostic modeling: the model is less likely to interpret the system in a physically plausible way, and small changes in training data can lead to big changes in model weights and parameters.

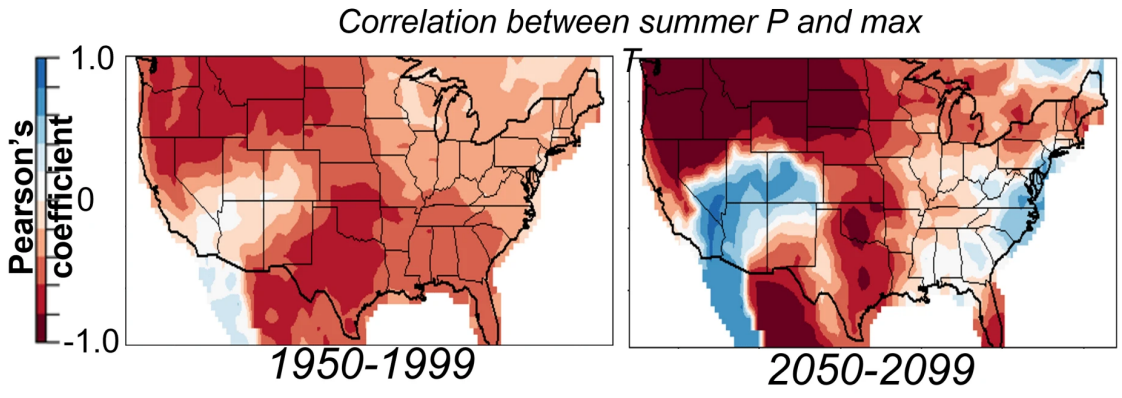


Figure 3: Correlation structure between summertime air temperature and precipitation from 1950-2000 (left) and projected from 2050-2100 (right) shows that the correlation between air temperature and precipitation will change for many locations in the US under climate change. This suggests that statistical models trained using TA and P under 20th-century conditions will not make robust estimates in the 21st century.

DOs:

- Feature engineering: using domain expertise to pick the most physically important predictor out of a suite of correlated variables.
- Data-driven feature reduction: Use data compression (aka PCA)
- Choose to minimize variance, not bias, in your loss function (i.e. elastic net regression).

DON'Ts:

- Ignore the issue
- Stepwise selection

ROBUST TO AUTOCORRELATION

Robust: in order for your model to generalize well, you must quantify and account for scale-dependent multicollinearity and spatial/temporal dependence in your data.

Know your scale (independence of observations)

Your data has temporal/spatial dependence when spatially adjacent or temporally antecedent/subsequent observations are not independent.

- Positive autocorrelation = proximal observations are similar.
- Negative autocorrelation: proximal observations are opposing.
- Process temporal frequency: What is the minimum frequency of temporal variability with am amplitude about the noise ratio or what is the minimum frequency of temporal variability of concern to the analysis? For example, are you looking at daily temperature variability or seasonal temperature variability?
- Process spatial frequency: What is the range of the spatial variogram of experimental data or what is the minimum distance of spatial variability you'd like to consider? For example, are you looking at flood extent on first or fourth-order streams in your humid landscape?
- Sampling frequency: what is the spatial and temporal resolution of your dataset? For example, the pixel size and revisit interval of a satellite.

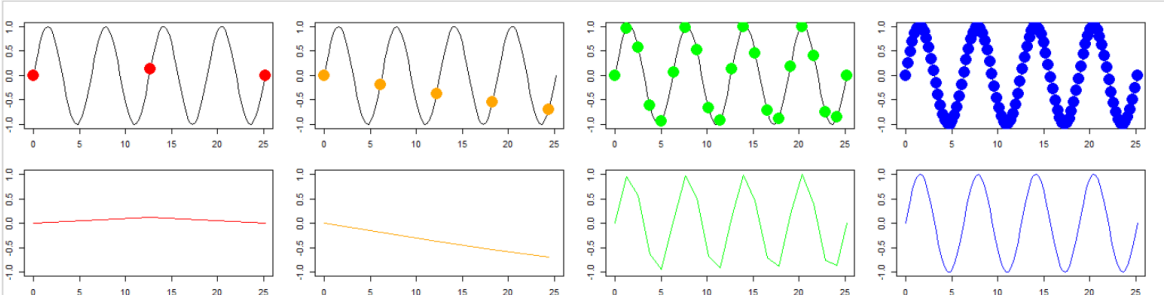


Figure 5: Undersampling (red) and aliasing (orange) either obscure or change the information content of the signal; sampling at twice the signal frequency (green) allows us to recreate the signal with the fewest possible observations. Oversampling the signal (blue) retains the information content of the signal but will cause problems for inference: you have induced a violation of the assumptions of independence of observations, and if you don't account for this autocorrelation you will overestimate the importance of certain predictors in diagnostic modeling.

DO's:

- Resample or high-pass filter your data to 2-4 times the highest frequency of interest or set a spatial chunk size of 2-4 times the highest frequency of interest (for convolutional neural networks or batch processing).
- Include spatial/temporal covariates that are predictive at above/below target frequencies.
- Use autocorrelation tolerant models (e.g. include autocovariance functions, autoregressive functions)

DON'Ts:

- Ignore the problem! Always check lag correlation and model residuals.
- Overstate the significance of your diagnostic results with autocorrelation is present.

REPRODUCIBLE

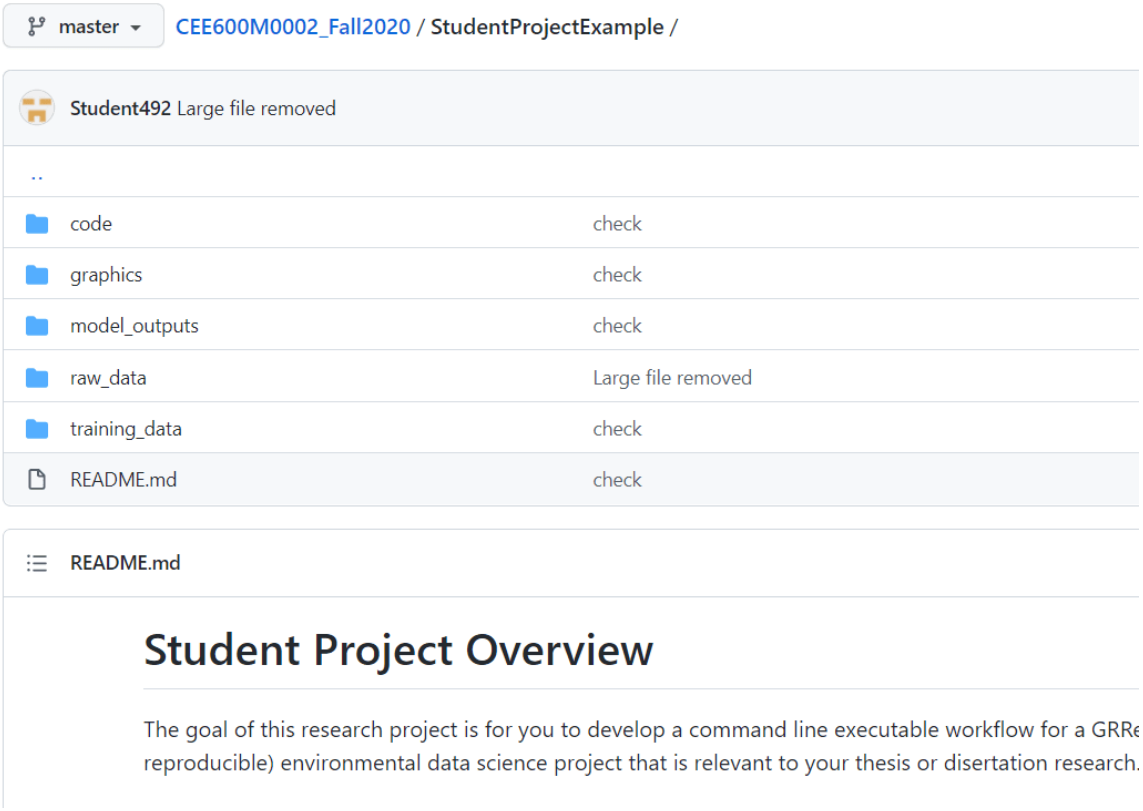


Figure 5: an sample project repository for GRRIEn analysis

Your project repository should include at a minimum:

- Raw data folder: for your in-situ data.
- Processed data folder: Contains analysis-ready datasets merging your in-situ data and global proxy variables.
- Model outputs folder: Contains trained model objects.
- Figures and tables: Contains graphics and organized tables presented in publications.
- Code: contains all scripts used in the analysis, at a bare minimum:
    1. Data download code: utilizing online geodatabase API to programmatically access data. *Reads from an online geodatabase, saves to Processed_data*
    2. Data preprocess and merge code: functions required to colocate in-situ signal and globally available proxy data in space and time, including CRS conversions, resampling functions, and zonal statistics functions. Convert merged data into an analysis ready format (i.e. numpy array, pandas dataframe, R DataFrame). Preprocess analysis ready data for model specification including any feature reduction, variable transformations, and processing of missing data. *Reads from raw_data, processed_data; writes to processed_data*
    3. Model train and validate code: splits data into training and testing datasets; trains statistical model; calculates model fit and summary statistics. *Reads from processed_data, writes to model_outputs.*
    4. Figure and table generating code: code required to generate figures and tables. *Reads from processed_data, model_ouputs, writes to figures_and_tables*
- README file: provides motivation for study, data citations, and explicit instructions on how to generate a complete analysis using public repository, including hardware and software specifications, and citations for the finished publication.
- License, software environment or container, and gitignore file.

INTERPRETABLE

When trying to build models that are generalizable to large spatial scales, it is not enough to rely purely on traditional metrics of model fit to ensure that your model will be robust at large spatial scales. Patterns of interdependence between both parameters and observations in large-scale space/time systems are frequently both complex and dynamic in space and time and omitted variable bias is functionally endemic. The most important step to ensuring that your model will make the right predictions in different contexts is to make sure those predictions are happening for the right reason: have your model weights and parameters diagnosed the importance of your predictors, and their relationship to the environmental process that serves as your predictand, in a physically plausible way?

- For simple regression/classification: do your parameters make physical sense? Are your parameter variances low and stable across model training runs?
- For deep learning: use feature importance metrics that are robust under multicollinearity and autocorrelation.

AUTHOR INFORMATION

https://ekcarter.expressions.syr.edu/ (https://ekcarter.expressions.syr.edu/)

ekcarter@syr.edu

ABSTRACT

Globally available, georeferenced data from earth observing satellites and coupled earth systems models provide new opportunities to use limited field observations to infer trends in environmental processes across unsampled locations and over time. Building statistical models that are intended generalize earth surface processes across large spatial scales represents a new frontier in supervised statistical inference. For example, care must be taken to collect unbiased samples given the complexity of the earth system: surface features can appear vastly different from the perspective of multispectral and SAR imagery in different atmospheric/landscape contexts. Environmental processes occur at variable spatial/temporal scales, and sampling resolution can drastically alter the appearance of patterns, and lead to spatial and temporal autocorrelation which can bias model weights and/or parameter estimates. Multicollinearity in multivariate datasets, which is also scale dependent, can inflate variance in parameter/weight estimates. All these in tandem can undermine the robustness of models in predicting in out-of-sample contexts. To overcome this, the GRRIEn (Generalizable, Reproducible, Robust, and Interpretable Environmental) analysis framework is introduced as a standard method of training and validating supervised data-driven models at large spatial scales. The method is explained, and demonstrated with a case study detecting surface water at CONUS scale using SAR and multispectral imagery.