GRRIEn analysis: a data science cheat sheet for earth scientists learning from global earth observations

Elizabeth Carter,[a] Carolynne Hultquist,[b] Tao Wen[c]

[a] *Assistant Professor, Civil and Environmental Engineering, Syracuse University, Syracuse NY*

[b] *Lecturer, School of Earth and Environment, University of Canterbury, Christchurch, NZ*

[c] *Assistant Professor, Earth and Environmental Sciences, Syracuse University, Syracuse NY*

Corresponding author: Elizabeth Carter, ekcarter@syr.edu

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-22-0065.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Globally available environmental observations (EOs), specifically from satellites and coupled earth systems models, represent some of the largest datasets of the digital age. As the volume of global EOs continues to grow, so does the potential of this data to help earth scientists discover trends and patterns in earth systems at large spatial scales. To leverage global EOs for scientific insight, earth scientists need targeted and accessible exposure to skills in reproducible scientific computing and spatiotemporal data science, and to be empowered to apply their domain understanding to interpret data-driven models for knowledge discovery. The GRRIEn (**G**eneralizable, **R**eproducible, **R**obust, and **I**nterpreted **En**vironmental) analysis framework was developed to prepare earth scientists with an introductory statistics background and limited/no understanding of programming and computational methods to use global EOs to successfully generalize insights from local/regional field measurements across unsampled times and locations. GRRIEn analysis is **g**eneralizable, meaning results from a sample are translated to landscape scales by combining direct environmental measurements with global EOs using supervised machine learning; **r**obust, meaning that model shows good performance on data with scale-dependent feature and observation dependence; **r**eproducible, based on a standard repository structure so that other scientists can quickly and easily replicate the analysis with a few computational tools; and **i**nterpreted, meaning that earth scientists apply domain expertise to ensure that model parameters reflect a physically plausible diagnosis of the **en**vironmental system. This tutorial presents standard steps for achieving GRRIEn analysis by combining conventions of rigor in traditional experimental design with the open-science movement.

SIGNIFICANCE STATEMENT

Earth science researchers in the digital age are often tasked with pioneering big data analysis, yet have limited formal training in statistics and computational methods such as databasing or computer programming. Earth science researchers often spend tremendous amounts of time learning core computational skills, and making core analytical mistakes, in the process of bridging this training gap, at risk to the reputability of observational geostatistical research. The GRRIEn analytical framework is a practical guide introducing community standards for each phase of the computational research pipeline: dataset engineering, model training, and model diagnostics, to promote rigorous, accessible use of global EOs in earth systems research.

File generated with AMS Word template 2.0

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

1. Introduction

The past fifty years have ushered in exponential growth in the volume of earth observations. As of writing, there are over 300 earth observing satellites in operation by national space agencies globally, with another 79 platforms approved or in development for the next decade (Committee on Earth Observing Satellites 2022). Increased availability of telecommunications bandwidth and lowered costs of electronic components have led to a proliferation of *in-situ* automatic earth monitoring networks (Stephens et al. 2020; Balsamo et al. 2018). Scaffolding the storage and processing of all this data, we've seen a more than a trillion fold increase in global computer power in the last 50 years (Tredinnick and Laybats 2018). This shift in information content of environmental systems has led to a shift in research methods. To quote from the Stanford *Earth Matter's* magazine: "The satellite and supercomputer are the rock, hammer, and compass of modern geoscientists." Even though modern earth scientists are often tasked with pioneering big-data analysis, of the top-ten ranked undergraduate programs in earth science (US News and World Report 2022), the majority require one or fewer semesters of coursework in probability and statistics, and none have required coursework in computational methods such as databasing or computer programming.

In the past several decades, the research community has proposed many guidelines to promote scientific rigor in observational computational research. One such framework that is widely accepted by the research community is Open Science by Design, as proposed by the National Academies of Sciences, Engineering, and Medicine (National Academies of Sciences, Engineering, and Medicine 2018). Open Science by Design guides stakeholders to practice "open science," a movement to make scientific research reproducible and accessible, throughout the entire research life cycle (ideation, knowledge generation, validation, dissemination, preservation). The same concept was later adopted by the Earth Science community to form the ICON principles (Integrates, Coordinates, Openly, Networks) (Goldman et al. 2021) to promote open science approaches in Earth Science with a particular emphasis on the need for growing open interdisciplinary collaboration. Both concepts emphasize that it is important for researchers to make research results FAIR (Findability, Accessibility, Interoperability, Reusability; Wilkinson et al., 2016), which facilitates peer review of the entire computational research pipeline for improved research quality.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

GRRIEn (**G**eneralizable, **R**obust, **R**eproducible, and **I**nterpreted **En**vironmental, pronounced ‘grēn’ like the color) supervised learning using global earth observations (EOs) analysis builds on FAIR standards to empower earth scientists to generate high-quality, easily reproducible geostatistical workflows utilizing openly available global geospatial data. We do this by outlining best-management-practices for each step in the computational analysis pipeline: dataset engineering, model training, and model diagnostics (Fig. 1).

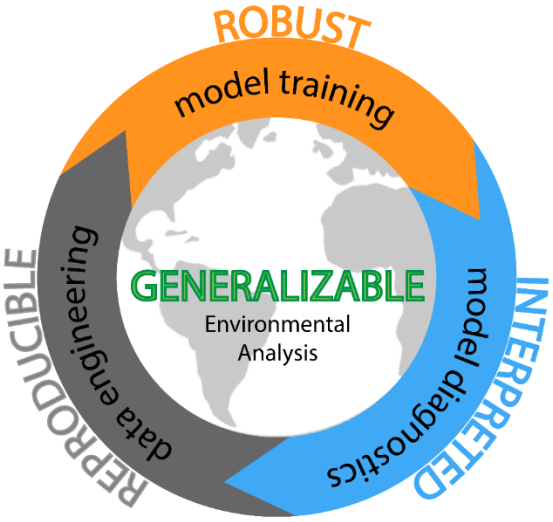


Fig. 1. GRRIEn analysis introduces best-management practices to extract *generalizable* insight on environmental systems using supervised learning of global EOs, including standards for *reproducible* data engineering, *robust* model training, and domain-specialist *interpreted* model diagnostics. Figure adapted from (National Academies of Sciences, Engineering, and Medicine 2018).

GRRIEn analysis applies to context in which earth scientists use global EOs to **generalize** insights from limited earth system measurements to unsampled times and locations. We define major types of global EOs, classify the primary objectives when using global EOs in supervised learning, and briefly describe how global EOs can be replicably converted into analysis-ready data using geodatabase APIs (Application Programming Interfaces) and open-source programming languages (Section 2). We describe the two spatiotemporal data pitfalls, scale-dependent observation and feature dependency, detail how they impact the **robustness** of supervised modelling frameworks for the different modelling objectives, and present checklists for diagnosis and model-agnostic management of these pitfalls in dataset engineering and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

model training (Section 3, Supplementary Appendix C). Drawing from a suite of most-used tools in reproducible computational research, we propose a standard software repository structure to facilitate a highly adaptive, highly **reproducible** data sourcing, engineering, and modelling pipeline for replication of diverse supervised GRRIEn analysis workflows (Section 4). Finally, we describe how experimental design principles translate into the era of global EOs, give an overview of explainable machine learning and AI, and explain the critical role of the modern earth scientists in **interpreting** the physical plausibility of trained data-driven models (Section 5). We conclude with how to incorporate GRRIEn into the experimental design and manuscript outline process and describe limitations of the method and future work (Section 6).

Supplemental Appendix A, "Required (Computational) Tools," is a detailed outline of the assumed background knowledge for GRRIEn analysis. For each topic, we include resources for where to find more information and learn new skills. We strongly suggest reading this supplemental resource, and reviewing any background concepts as necessary, prior to learning the GRRIEn method.

2. Generalizable

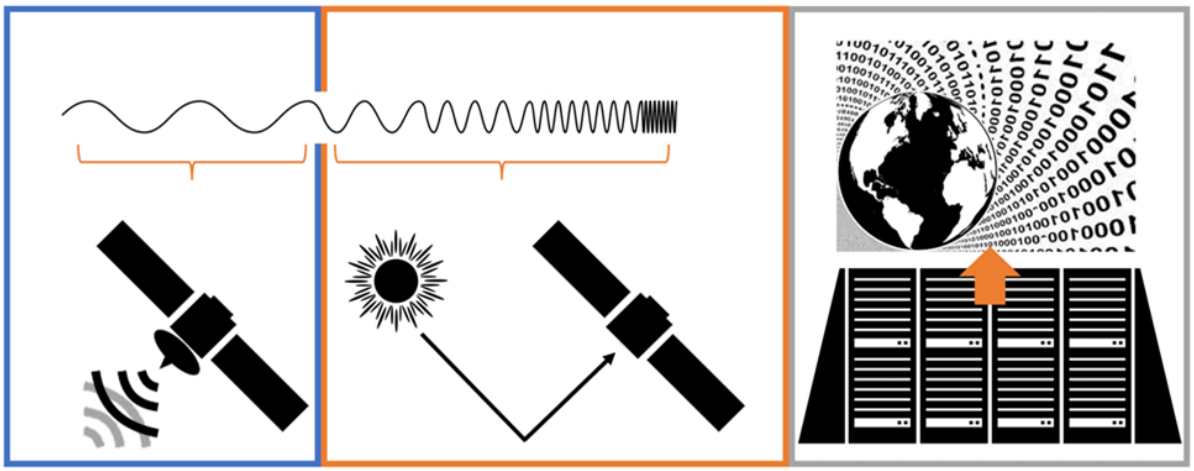
Earth science disciplines have evolved in a historically data-poor environment. Early to mid-20th century geoscientists relied primarily on data from short-term site or laboratory based experiments (Clark and Gelfand 2006), and then used process models to infer modes of spatiotemporal variability across unsampled times and locations (Hilborn and Mangel 2013). A process model functions like an algorithmic narrative, merging theory and available data to quantitatively articulate and compare plausible hypotheses concerning drivers of spatiotemporal environmental variability (Hilborn and Mangel 2013). As one cannot properly quantify the importance of unparameterized, under-parameterized, or inaccurately parameterized phenomenon using process-based models, this method has strong potential for perpetuating confirmation bias in earth systems theories (Nickerson 1998; Bond et al. 2007; Shi et al. 2019). Global earth observations (global EOs), defined here as spatially continuous, temporally repeating data with continental to global coverage, are outcomes of an international movement to improve characterization of earth systems (Nativi et al. 2015). As global EOs are approximately spatiotemporally continuous, they provide an observation-

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

based framework to translate insight from limited field measurements across unsampled times and locations to validate theories of spatiotemporal drivers of environmental processes. For example, characterizing spatiotemporal variability in precipitation using point data from rain gages is a highly socially relevant field of study that has been a major thread in hydrological study for decades (Kidd et al. 2017). In many gridded precipitation datasets, output from numerical weather models are used to spatially interpolate precipitation between gages (e.g. Hersbach et al. 2020). Recent studies show that satellite observations, while containing independent sources of error, can help elucidate the nature of specific biases in precipitation estimation from numerical weather models (Xu et al. 2022). This is one of many examples of how global EOs have facilitated evolution of theories of environmental systems that are described in process models. The task ahead is straightforward: how can we accurately use these global EOs as proxies of environmental processes to augment our understanding of how and why diverse earth systems vary across space and time?

a. Getting started with global earth observations.

The diversity and scope of earth observations are staggering and ever expanding (Balsamo et al. 2018), but to get started, we recommend a working knowledge of the following three data types that provide global coverage at a regular spatial and temporal scale: active satellite remote sensing data, passive satellite remote sensing data, and coupled earth systems model (CESM) output (Fig. 2.) More detail on these data types can be found in Supplemental Appendix A Section 1.



Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Fig. 2. Three common sources of global gridded earth observations (EOs) include active satellite remote sensing data, such as synthetic aperture radar imagery (left), passive satellite remote sensing data, including optical and passive microwave imagery (center), and gridded outputs from global coupled earth system models (right).

b. Objectives of using global EOs in environmental systems analysis.

In supervised learning, models are trained on coupled records of both input variables (predictors), and output variable(s) [label(s) or predictand(s)]. The GRRIEn framework deals specifically with supervised learning. It is meant to be applied to contexts where some measurements of an environmental process (predictand) are available, but the experimental questions cover a larger spatial extent, a different period of time, and/or are at a different sampling frequency than the measured data. As global EOs are (approximately) spatiotemporally continuous observations, the overall objective of GRRIEn analysis is to train a supervised learning algorithm that can predict an environmental process using globally available EOs as input data. This algorithm can be used to *generalize* insights from direct observations of an environmental process sampled experimentally or *in-situ* (hereafter called the label or output variable in a trained algorithm) across unlabeled times or locations where global EOs are available (see examples in Table 1). The prediction space of the generalizing algorithm will be bounded by an area of interest (AOI) and period of interest (POI) that may or may not overlap with the labelled sample. Global EOs serving as predictors (i.e., input data to the predictive model) provide full coverage of the POI and AOI, whether or not it is possible to directly observe the environmental process at that scale (Fig. 3).

Objective 1:	**Example:**	**Example predictor:**	**Example predictand/label:**	**Modeling framework:**
Interpolation	**Mapping atmospheric NO2 concentration** (Young et al. 2016)	**418 gridded geographic covariates and Aura Satellite Ozone Monitoring Instrument imagery**	**U.S. EPA Air Quality System site NO2 time series**	**Universal kriging**
Description: Interpolation is “filling in gaps” in the spatial/temporal record of an environmental process. Researchers use supervised models to gap-fill missing data in a time series, or statistically upsample or downsample the spatial density of observations to create a uniform grid of data across an AOI. With interpolation, our AOI and POI intersect with the sample space of direct observations.				

File generated with AMS Word template 2.0

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Objective 2:	**Example:**	**Example predictor:**	**Example predictand/ label:**	**Modeling framework:**
Extrapolation	Predicting crop yields under future climate scenarios (Challinor et al. 2014)	PRISM precipitation, vapor pressure deficit, and temperature (training). CMIP5 composite precipitation, vapor pressure deficit, and air temperature (prediction)	County-level USDA Risk Management Agency yields	Generalized Additive Model (GAM)
Description: Extrapolation is the prediction of variability in the environmental process in yet-to-be explored spaces/times. Since our scope of interest in an environmental process often exceeds our ability to sample it directly in space and/or time, extrapolation is a common objective motivating the training of supervised models.				
Objective 3:	**Example:**	**Example predictor:**	**Example predictand/label:**	**Modeling framework:**
Diagnosis	Characterizing drivers of extreme precipitation (Carter et al. 2021)	NOAA-CPC Soil Moisture data, NCAR-NCEP Reanalysis data, EN4.2.1 quality controlled subsurface ocean temperature and salinity objective analysis	Gage-based precipitation data	Guided regularized random forest regression with feature importance
Description: In diagnostic modeling, we use the trained algorithm as its own kind of data to gain insight on the nature, causes, and associations of space/time variability in the environmental process Diagnostic modeling is often conducted to validate predictive or interpolative models, and we argue that domain specialists must play a fundamental role for it to be effective.				

Table 1. Examples of supervised learning for interpolation, extrapolation, and diagnostic modelling using globally available EOs as input.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

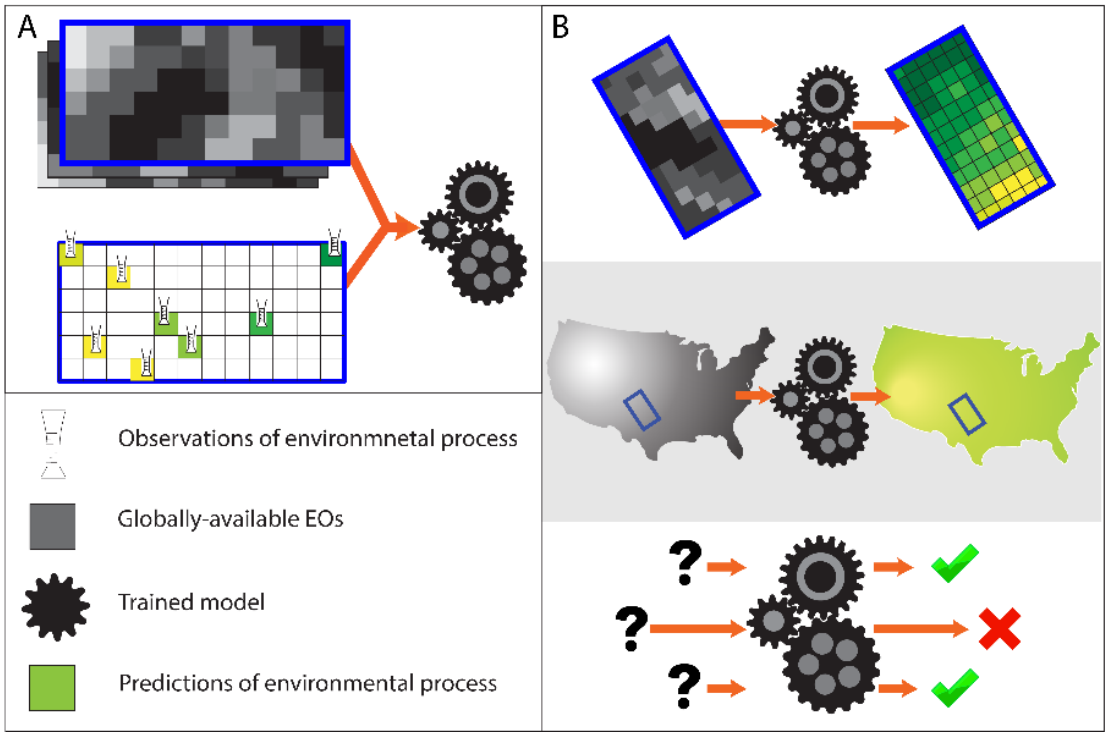


Fig. 3. A) In supervised learning, a statistical model is trained that maps variability in input data to variability in a limited number of labels (in-situ or experimental measurements) which are contemporaneous in space and/or time. B) The trained model can then be used either for interpolation (filling in the gaps in spatiotemporal variability in the output for an AOI and POI that overlaps with the training data, top); extrapolation (predicting in an AOI and/or POI that exceeds the bounds of the training data, middle); and/or to diagnose drivers and modes of spatial coverability between inputs and labels (bottom).

There are three primary objectives when using global EOs in supervised learning: interpolation (prediction in an AOI or POI intersecting with sample space), extrapolation (prediction in an AOI or POI that does not intersect with the labelled sample space), or diagnostic modelling (using the trained algorithm as data to gain insight on the nature, causes, and associations of space/time variability in the environmental process) (Fig. 3). Table 1 provides detailed explanations and sample studies using global EOs as predictors in supervised analysis for interpolation, extrapolation, and diagnostic modelling.

3. Robust

In experimental research, the goal is to collect a representative sample to allow for the statistical evaluation of specific hypothesis(es) about independent drivers of variability in a

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

system. Experimental researchers start with a highly controlled environment, such as a laboratory or field space, which will hold all physical variables deemed important by domain scientists constant, except the variables that will be experimentally varied (the predictors). Good experimental design ensures that data will be collected across the full range of possible values for each predictor variable. In addition, in multivariate experiments, independent representation of predictors is ensured by collecting data across all potential combinations of values of predictor variables. For example, in an experiment to determine how radiation and soil moisture (the input variables or predictors) impact plant growth (the output variable or predictand), low radiation high soil moisture, low radiation low soil moisture, high radiation high soil moisture, and high radiation high soil moisture treatments should be included. The number of samples to be collected across treatments is defined *a priori* based on the desired confidence in results [*e.g.*(Campbell et al. 1969)].

The field of statistics was developed to quantify certainty of results from controlled experimental trials, and machine learning and deep learning are subfields of statistics (Runge et al. 2019; Blei and Smyth 2017). Because of this, all observational environmental data, especially spatial and temporal observational data collected at large scales, have characteristic divergences from data collected in controlled experimental trials, some of which must be addressed in order for data science algorithms to generalize well. Here, we discuss three sources of such divergences: inability to parameterize all drivers of variability in an environmental system; limitations to ensuring independence of observations sampled in time and space; and limitations to sampling for independence of multivariate predictors.

First, in observational experiments, we lack controlled environments. Many landscape phenomena will change at similar spatial and temporal scales as the environmental process of interest. If any feature driving variability in the environmental process is not parameterized (i.e., included as a predictor) in the model, that model is subject to *omitted variable bias*: something important is happening to the environmental process, and the impact is observable in our response variable, but it is not parameterized in the model.

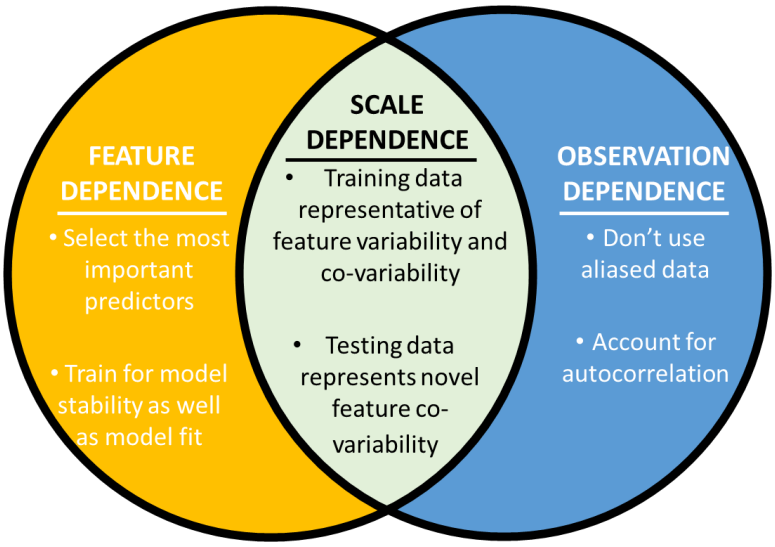
Second, in observational data science, we often lack the ability to collect balanced, cross-replicated samples of multivariate predictors. This is because in environmental systems multivariate predictors tend to be intrinsically related to, or dependent, on each other. For example, since rain comes from cloudy skies, there will be more observations of low

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

radiation high soil moisture conditions, and few to no observations of high radiation high soil moisture conditions, in observational plant growth data (Carter et al. 2018b). This is called *feature dependence,* or a condition where we have some structure of covariability between individual predictors in our training data.

Third, especially when relying on global EOs which are collected at standard temporal and spatial intervals, we do not get to define *a-priori* how many observations are collected per unit space or time to confirm or deny a pre-defined hypothesis. Since observations which are nearer to each other in space or in time tend to be related, our observations which are sampled at regular intervals in space and time will not be independent. Implicit *observation dependence* in spatiotemporal data means that our sampling frequency may either obscure or overrepresent the modes of spatiotemporal variability that we are trying to evaluate in our model.

In large scale spatiotemporal systems, unaddressed scale-dependent feature and observation dependence can severely impact interpolation, extrapolation, and interpretation of supervised models. When working with global EOs, it is therefore important that we take certain steps to characterize feature and observation dependence in our datasets, and address it in dataset engineering and model training, to avoid misinterpretation of our results (Fig. 4) Please note that a checklist for robust spatiotemporal models can be viewed at the end of this section (Fig. 10), as well as a flowchart for diagnosing model performance issues using the spatial distribution of model residuals (Fig. 12). These figures are referenced throughout the section.



Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Fig. 4. The robust rules for GRRIEn analysis.

a. Robust to dependent features.

There are two main sources of feature dependence in observational datasets. *Intrinsic feature dependence* occurs when measurements are imperfect representations of an underlying latent process that cannot be measured directly. For example, we cannot directly measure soil texture, but we can measure percent of sand, silt, and clay in a soil sample. All these variables will be intrinsically correlated (a higher ratio of sand implies a lower ratio of silt/clay), yet each give different information on the unmeasurable quality of soil texture. *Incidental feature dependence* is caused by our inability to capture a representative sample. With incidental correlation, not all combinations of predictor variables exist in a study area (e.g., the lack of cross replication between radiation and soil moisture mentioned in the introduction to Section 3).

Multicollinearity is a special case of feature dependence that occurs when one or more predictor variables (features) are linearly related. To understand how multicollinearity impacts inference and prediction in machine learning algorithms, we should start by reviewing how multicollinearity impacts ordinary least squares (OLS) linear regression. In OLS regression, the standard error around coefficient estimates on collinear predictors are increased. The stronger the linear relationship between the variables, the larger the inflation of standard error on their coefficient estimates. This means that any single realization of a coefficient estimate, derived from a sample, is likely to be further from the unknown "true" value that represents the linear dependence between the predictor and predictand across the whole population. Multicollinearity therefore leads to model instability, or a situation where small changes in training data can lead to large changes in model coefficient estimates.

Model instability is a problem for several reasons. To start, when two collinear predictor variables are included in a model, we cannot statistically identify which variable has a direct association or a causal impact with the predictand (Dormann et al. 2013). So given two predictors, one of which has a direct association with the predictand, and one of which has no direct physical association with the predictand but is incidentally correlated with the other predictor, no statistical model can determine which is the causal driver. This issue impacts all data-driven models, not just linear regression (Dormann et al. 2013; Alin 2010; Feng et al.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

2019; Kim 2019). The performance of the model no longer depends on the value of individual independent predictor variables, but on the joint distribution of dependent predictor variables. If we enter into a prediction space where the two predictors are no longer incidentally correlated, the model may make inaccurate predictions, specifically if the model attributed substantial weight to the second, incidentally correlated predictor. In other words, with substantial feature dependency in input data, the model performance is contingent on the structure of multicollinearity that was present in the training dataset being conserved across the multivariate population as a whole. This means that even if a trained model shows good in-sample model fit statistics (i.e., our predictions are similar to our available data), if the model is being used to make predictions with a dataset where the correlation structure between predictors is different than it was during model training, such as we'd expect with incidental feature dependence, these predictions are likely to be inaccurate. This is ultimately because parameters in an unstable model are substantially less likely to represent physically plausible characterizations of the relationship between individual predictor variables and the predictand (Section 3.c. "Checking your work;" Fig. 12a).

This can be problematic in environmental systems analysis, as the structure of collinearity between predictor variables is often dynamic in both space and time (Fig 5). For example, all meteorological variables will exhibit some degree of collinearity, and the strength and sign (positive or negative) of collinearity between meteorological variables will change over space (as a signature of local climate) as well as over time (as a signature of specific seasonal or weather patterns) (Dormann et al. 2013; Thornton et al. 1997; Carter et al. 2018a). Since we have limited ability to directly sample environmental processes across landscape scales where we might capture this variability in feature dependence, in observational environmental systems analysis, incidental multicollinearity in training data is a common problem that cannot always be avoided. In Fig 5a, this is demonstrated in the spatial variability in local correlation between summertime air temperature and monthly precipitation from 1950-2000 across the United States. This covariance has a physical cause: as it gets drier, incident solar radiation that would have been used to evaporate soil moisture is partitioned to sensible heat flux, increasing air temperatures. The correlation coefficient changes spatially because the specific covariance structure of meteorological variables is a signature of local climate (Supplemental Appendix 2, Figure B1). Under climate change, we expect the exact structure of local covariability of meteorological variables to shift as well, as is demonstrated in Fig.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

5b, which shows the local correlation coefficient between ensemble mean summertime and temperature and monthly precipitation projected for 2050-2099 under a moderate emissions scenario in the CMIP5 multi-model ensemble.

Substantial changes in local correlation between summertime air temperature and precipitation under climate change also has a physical explanation, as we expect complex, non-linear changes in the dependency structure between meteorological fields as air temperatures increase. For example, local summertime air temperatures will be increasing because of increased long wave absorptivity of the atmosphere (a different physical driver than meteorological drought), and precipitation will shift due to shifting atmospheric circulation and hydrologic intensification. We cannot collect a sample of meteorological data that is representative of the structure of covariability under climate change because these conditions do not yet exist. As the specific correlation structure between meteorological variables changes over space and time, and since the prediction skill of data-driven algorithms depends on the conservation of the correlation structure of predictor variables in the prediction space, we are likely to see spatiotemporal structure in the error of observational algorithms trained with meteorological variables as predictors, specifically in extrapolation (Dormann et al. 2013; Carter et al. 2018a). Since the structure of collinearity changes over space and time in many environmental processes, even within the training data, multicollinearity must be carefully assessed during routine exploratory data analysis and cannot safely be ignored during model training (Dormann et al. 2013). Because of this, we often talk about two metrics of model performance in a multicollinear system: **model accuracy** (how well do predictions match labels) and **model stability** (how consistently variations in specific input variables map to changes in predictions between training samples and model parameterizations) (Graham 2003).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

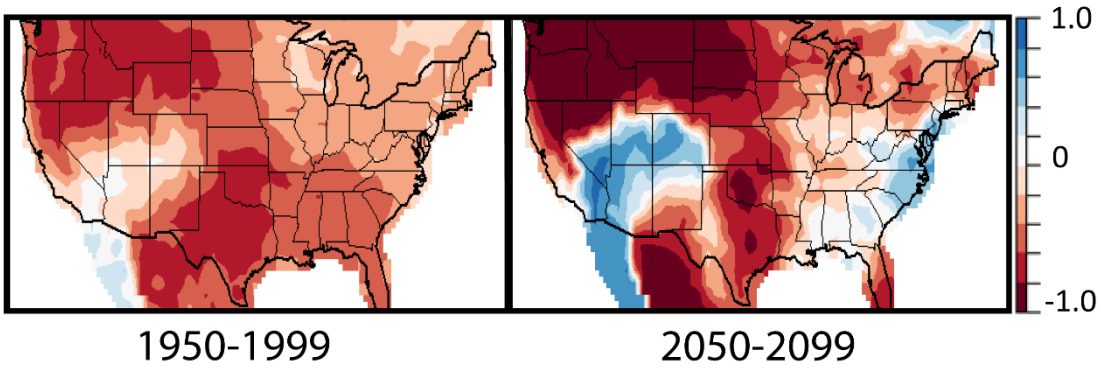


Fig. 5. Pearson's correlation coefficient [scaled between -1 and 1, color bar (Benesty et al. 2009)] between bias-corrected statistically downscaled Climate Model Intercomparison Project 5 ensemble mean monthly precipitation and daily max temperature. Historical observations 1950-1999 (left) and moderate emissions forecast (RCP 4.5) for 2050-2099 (right) both indicate spatiotemporal variability collinearity between summertime maximum temperature and precipitation. Covariance of meteorological variables is a signature of local climate. As local climates shift due to global warming, so will the local covariability of meteorological variables (right). This generates complexity for predicting environmental process response to meteorological variables under climate change (Taylor et al. 2012).

1) Diagnosing and data engineering with feature dependence

(i) Quantify multicollinearity globally and locally

With multivariate spatial datasets, multicollinearity must be evaluated both globally (assuming all observations represent a single population) and locally (treating different regions and/or time periods as unique populations). Global multicollinearity can be visualized by looking at a scatterplot matrix (pairwise correlation coefficients), or quantified by way of the dataset's variance inflation factors (VIFs), condition numbers (CN) (Alin 2010), or variance decomposition proportions (VDs) (Brauner and Shacham 1998). Multicollinearity can also be diagnosed locally by calculating geographically weighted VIF, CN, or VD at different spatial bandwidths (Kalogirou 2013; Wheeler 2009; Lu et al. 2014). Researchers should note substantial variability in locally-calibrated metrics of multicollinearity and be transparent that these multicollinearities could affect model interpolation, prediction, as well as interpretation/explanation.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

(ii) Select representative training/testing data

To avoid overfitting when training algorithms, it is important to divide the entire pool of matched label/predictand observations into training data (data which the model parameters are calibrated to), and testing data (data which the model predictions are evaluated against). For iteratively trained algorithms or while hyperparameter tuning, validation data subsets are sampled from the training data to evaluate skill of evolving model formulations (Berrar 2019). To avoid incidental multicollinearity, training/validation or/and testing data should contain a range of values between the minimum and maximum expected value for each input and output variable, with probability distribution of the sample reflecting the population as a whole. When there is spatial or temporal variability in the structure of multicollinearity in the dataset, training data should also be strategically sampled, and when appropriate training/validation/testing subsetting should be strategically designed to be representative of the range of covariability in input variables. This often means stratifying your sampling area to select diverse examples representing different regional or temporal representations of input variable collinearity (positive or negative, weak or strong) (Tamura et al. 2017). This way, we can quantify the models' mean performance across the training or training/validation sample, as well as on anomalous examples that may be present in the population as a whole.

In observational analysis, researchers often do not have control over training sample collection. When it isn't feasible to collect representative training/validation data, testing data, which is data withheld from training and used to evaluate model stability, should always reflect anomalous (from global) spatiotemporal collinearity (Salazar et al. 2022). For example, the global Pearson's correlation coefficient of two variables is 0.7 in a 1000 square km study area, but the local correlation with a bandwidth of 5km can range from -0.8 to 0.95. You would want to include regions with strong negative, neutral, and strong positive correlation in your testing data.

When thinking about your intended extrapolation contexts, it is important to apply domain expertise to evaluate whether there might be shifts in spatiotemporal collinearity of your predictors that is not represented in your training data. Are you planning to make predictions in regions or times where other factors, like climate change, might change the relationship between your input variables? Shifts in the correlation structure of input

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

variables between your training data and your AOI or POI will likely impact the fidelity of extrapolations (Fig. 12b, (Dormann et al. 2013).

(iii) Reduce the number of input variables

Unnecessary intrinsic multicollinearity, where several collinear predictors are included in the model with no known relationship to the predictand, can undermine model accuracy, stability, and interpretability, and should be avoided as a rule. In observational analysis, it is rare to have access to perfectly cross-replicated measurements of known or suspected drivers of variability in an environmental process. Instead, we have to compromise and use what observations we have available as "proxies" for the things we want to measure. When making use of proxy data, there's a difference between making scientifically informed choices about which proxies are important, and just throwing all available data into your model and letting it decide for you (a process commonly known as "data mining"). At this point in the big data revolution, a well-trained domain scientist will do a better job picking important variables than a supercomputer. Step zero in addressing multicollinearity in your dataset is "feature engineering," or using your domain expertise to select only the most physically important variables. The inclusion of any predictor variable in a data science model should be motivated by established theory and literature in your field. Think of each of your predictor variables as a hypothesis that you would like to test regarding what physical drivers or measurements are most relevant to your environmental system. If your meaningful predictors are intrinsically collinear, consider factor reduction [a review of methods can be found in (Chan et al. 2022)], or use elastic net regularization (see below). Stepwise selection, where variables are chosen by evaluating change in prediction error when variables are either included or omitted from the suite of predictors, is strongly discouraged in collinear datasets, as it is likely that model instability will lead to rejection of important variables (Smith 2018).

2) TRAINING AND VALIDATING MODELS WITH FEATURE DEPENDENCE

(i) Train for model stability, as well as model fit

Prediction skill (and its opposite, error) can be decomposed into two components: accuracy (variance) and precision (bias). The sum of variance and bias in predictions is the total error of the model (Fig. 6). The default objective of most model fitting protocols, like OLS regression, is to minimize total error in model predictions of an unbiased model. When

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

multicollinearity is present, models trained to minimize bias will tend towards being overfit. An overfit model has poor generalization, i.e., model parameters describe noise in the training data, not patterns in the population as a whole. Thus, overfit models yield inaccurate out-of-sample predictions (See Fig. 12a).

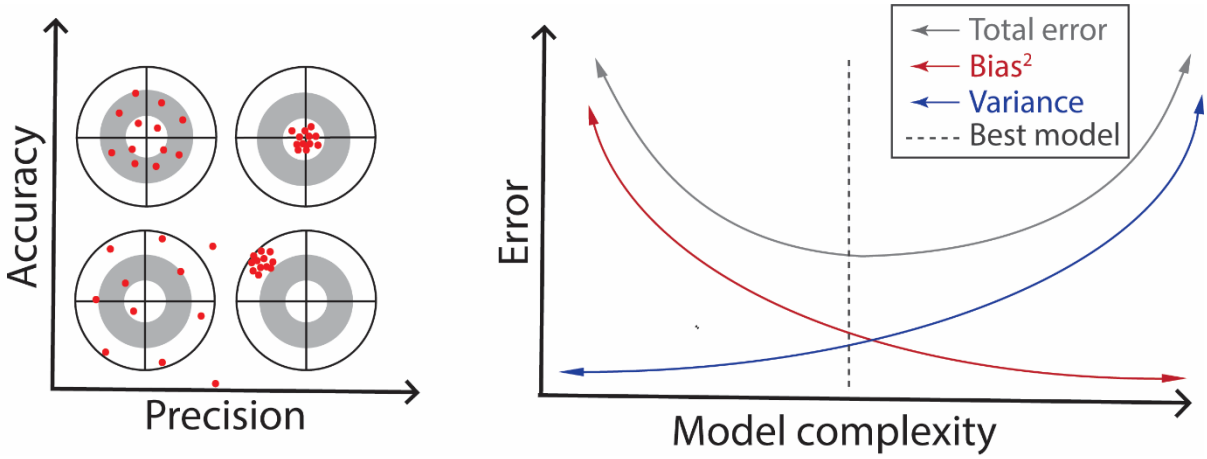


Fig. 6. Total model skill (and its opposite, error) can be decomposed into accuracy (variance) and precision (bias). To avoid overfitting and minimize total error in models with multicollinear data, we intentionally add bias to our model parameters using regularization parameters. Figure adapted from Bigbossfarin, CC0, via Wikimedia Commons.

Regularization parameters, like ridge (Hoerl and Kennard 1970), lasso (Tibshirani 1996), and elastic net [a linear combination of lasso and ridge parameters, (Zou and Hastie 2005)], are integrated into statistical (Dormann et al. 2013), machine learning (Li et al. 2021; Carter et al. 2021; Mudereri and Dube 2019), and deep learning models (Murugan and Durairaj 2017; Versloot 2020) to shrink model estimators. Regularization is therefore a systematic bias that we add to models to stabilize parameter estimates and reduce overfitting. We do this because we know that the lowest total error of a model will occur somewhere between the lowest variance model and the lowest bias model (Fig. 6, right). A biased model might miss some of the patterns in the data, but since it will not fit noise, it often generates a more physically plausible representation of the system, and therefore yields more accurate interpolations and extrapolations (Dormann et al. 2013). The ridge and lasso regularization parameters contain coefficients which can be tuned. Increasing or decreasing the coefficient on your regularization parameters increases or decreases the amount of regularization, which increases and decreases the bias in your parameter estimates associated with collinearity, in order to increase model stability and minimize total error.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Different regularization techniques behave in different ways with multicollinearity. The ridge parameter achieves a "grouping effect:" if there is a group of highly collinear predictors, it will effectively partition the magnitude of response variance more equally between each collinear predictor variable, instead of arbitrarily all to one, improving model stability and reducing estimator bias. With a ridge penalty, unimportant coefficients will be reduced in magnitude, but no parameters will be shrunk to zero. As such, it is not directly suitable for feature selection. The lasso parameter, on the other hand, effectively eliminates uninfluential predictors by shrinking their associated coefficients to zero. Unlike ridge, however, it does not achieve the "grouping effect," and therefore the magnitude of coefficients on collinear predictors is likely to be subject to volatility. Elastic net regularization, which is a linear combination of ridge and lasso, achieves "group selection effect," producing physically plausible, stable coefficients that partition response variance within important groups of collinear predictors, while removing unimportant predictors from the model completely (achieving all in one factor reduction and model training). Elastic net shows consistent skill over other regularization and in-situ factor reduction methods in model fit and stability in high-dimensional, collinear datasets when added to loss functions of diverse machine learning and deep learning algorithms (Srisa-An 2021; Dormann et al. 2013). Supplemental Appendix C provides an example of using regularization to manage multicollinearity.

(ii) Utilize cross-validation and ensemble learning

Models which are trained on independent, randomly permuted subsets of training/validation data in order to minimize out-of-sample prediction error produce more stable and robust models under multicollinearity. Among these, models which use *ensemble learning* to integrate insights from random subsampling [e.g. bagged decision trees (Breiman 1996a), random forest (Breiman 2001), stacked models (Breiman 1996b), AdaBoost (Freund and Schapire 1999), Gradient Boosting Machines (Friedman 2001)] perform better than those that use *iterative learning* (Wang et al. 2021; Hembram et al. 2021; Adnan et al. 2020; Smith et al. 2013). For example, random forest regression, a bootstrapped ensemble learning method, is consistently shown to have higher stability and prediction accuracy than other machine learning/deep learning algorithms, such as naive bayes, boosted regression trees (Hembram et al. 2021), support vector machines (Adnan et al. 2020) and artificial neural

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

networks (Smith et al. 2013) that train iteratively, when used on collinear observational datasets.

(iii) Account for space-time variability in feature dependence

Models which calculate local (to space or time) parameters or weights are more stable and robust with dynamic patterns of multicollinearity than models which rely on global parameter estimates (Mahadi et al. 2022; Carter et al. 2018b; Wen et al. 2018). Some examples of locally calibrated models include geographically weighted and time varying regression, in which both regression parameters and regularization parameters can be locally calibrated (Murakami et al. 2021; Li and Lam 2018; Wheeler 2009; Kalogirou 2013; Bárcena et al. 2014; Mahadi et al. 2022); and convolutional neural networks, which are common in computer vision and can learn complex spatial as well as spectral patterns in multidimensionally gridded datasets (such as remotely sensed imagery or multivariate meteorological data) so long as these complex patterns are consistent throughout the AOI (Chen et al. 2014; Audebert et al. 2019). For example, CNNs have been shown to produce more accurate predictions than models which rely only on 1-dimensional (spectral) variability, such as generalized linear models, random forest regression, and artificial neural networks (Saha et al. 2022, 2021). For datasets where spatiotemporal collinearity is dynamic in time, CNNs which are locally calibrated, such as those which incorporate recurrent networks, show promising results (Guo et al. 2022; Zang et al. 2020; Chen et al. 2021).

b. Robust to dependent observations

Environmental processes change in space and in time, often because of different drivers. For example, average annual temperatures across a continent will vary from place to place as a function of latitude, topography, and global circulation (spatial), while hourly temperatures for a given location will change as a function of seasonality, weather, and the diurnal cycle (temporal) (Fig. 7). Since environmental processes change over space and time, and since space and time are both continuous fields, observations that are nearer to each other in space and time will be related (Anselin and Li 2020). Depending on what spatiotemporal driver you wish to characterize, your observations need to be sampled at some interval in space and/or time to capture important patterns, trends, and drivers of variability, without undersampling (sampling too far apart in space and/or time such that the information content of the signal is lost or changed) or oversampling (sampling too close together in space and/or time, which is

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

akin to repeating a measurement) (Section 3.b.1.i., Figure 8). In experimental analysis, defining the sampling interval (how much time will pass between when subsequent samples will be taken), sampling frequency (the inverse of sampling interval), and sample size (how many observations will be collected in total, the product of sampling interval and study duration) are important considerations for experimental design.

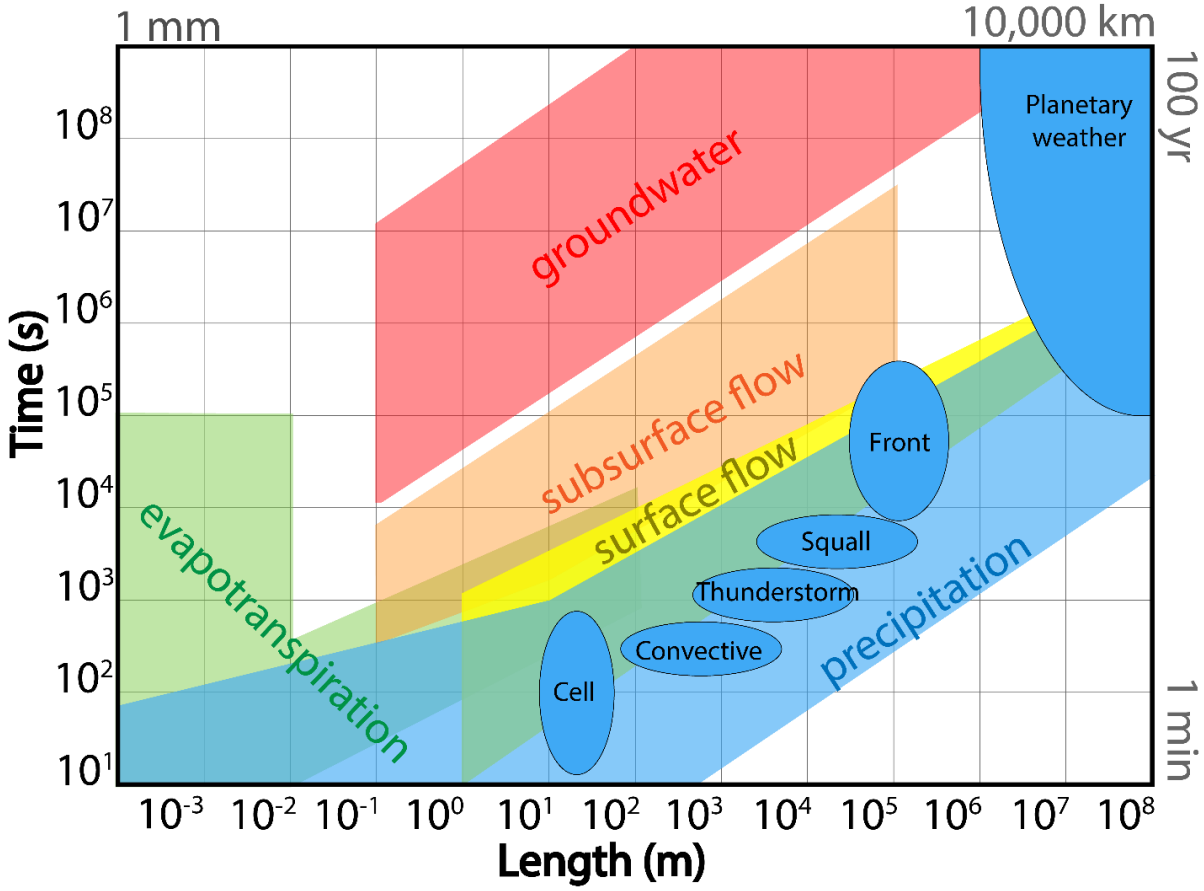


Fig. 7. Schematic of the spatial and temporal scales ($f_{process}$ = 1/time or 1/length) of terrestrial water budget components modelled in hydrologic and hydrometeorological subroutines of coupled earth systems models. Adapted from (Cristiano and Veldhuis 2017)

1) DIAGNOSING AND DATA ENGINEERING WITH OBSERVATION DEPENDENCE

(i) Define the process frequency

Most environmental systems will change over space and time due to different drivers, so the process frequency ($f_{process}$) of interest in an analysis will have a temporal component and a spatial component. The temporal process frequency is the minimum frequency (roughly the difference between high and low values, such as the maximum rainfall intensity and zero

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

rainfall intensity during a storm) of temporal variability of concern to the analysis (e.g. daily $f_{process}$= 1/day; or seasonal $f_{process}$ = 1/90 days). Similarly, the process spatial frequency is the minimum linear distance of spatial variability of interest in the analysis. For example, if the goal of an analysis is to build a model that can interpolate convective precipitation between gage stations, using Fig. 7 as a guide, the spatial $f_{process}$ is approximately 1/100m, and the temporal $f_{process}$ is approximately 1/120 seconds. Defining the spatial and temporal $f_{process}$ is a critical part of observational experimental design, as variability in many environmental processes is associated with different physical forcings at different temporal/spatial scales. For example, a point-based precipitation time series integrates variability from mesoscale, synoptic scale, and planetary scale precipitation drivers, each representing a unique set of atmospheric forcings (Fig. 7).

(ii) Define the sampling frequency

The sampling frequency (f_{sample}) is the spatial and temporal resolution of your dataset. For example, the spatial f_{sample} of a satellite image is the inverse of the pixel size, and the temporal f_{sample} is the inverse of the time between subsequent image captures. The magnitude of f_{sample} relative to $f_{process}$ has stark implications for your ability to accurately characterize an environmental process (de Knegt et al. 2010). Fig. 8 shows a theoretical continuous environmental process as a *f(t)*. We explore three different potential $f_{samples}$ as points (top row). We then attempt to recreate our process signal by linearly interpolating between these sample points (bottom row). Undersampling (Fig. 8, left) changes the information content the environmental process signal, a phenomenon commonly known as *aliasing*. Undersampling is associated with information loss, we cannot resolve drivers of environmental process variability that act below our sampling frequency. Undersampling cannot be resolved computationally, it is a sampling problem which motivates the design of higher-resolution satellites and finer scale coupled earth system model simulations (Luvall et al. 2017). We should consider any data that is sampled at a lower spatial or temporal frequency than the process frequency of interest to be *aliased.*

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

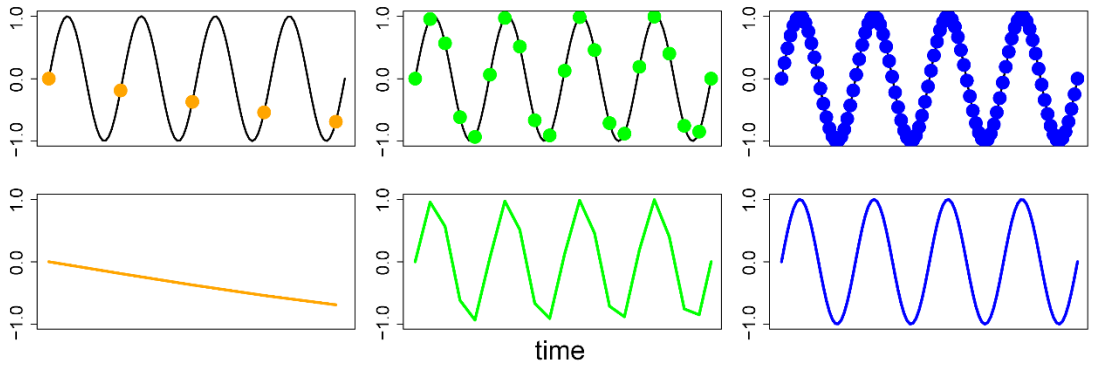


Fig. 8. (top) mismatched sampling frequency (f_{sample}, points) and process frequency (line) can induce aliasing (left, bottom) or autocorrelation (right, bottom) in recreated signals. The Nyquist frequency ($f_{sample} = 2 \times f_{process}$, middle column) allows for recreation of a continuous signal from a discretized (sampled) signal with the fewest possible observations.

Fig. 8 (middle column) shows a sampling frequency of exactly twice the signal frequency ($f_{sample} = 2 \times f_{process}$), also known as the Nyquist frequency (Shannon 1949). The Nyquist frequency allows us to recreate a continuous signal from a discretized (sampled) signal with the fewest possible observations. Oversampling the signal (Fig. 8, right) retains the information content of the signal, but since we are now sampling too close together in time, our observations are no longer considered independent, which causes substantial problems in inference, prediction, and diagnostic modeling (Dubin 1998; F. Dormann et al. 2007).

In observational environmental analysis, just as we cannot specify *a-priori* what we measure, we also typically cannot specify where and when the measurements are made. Global EOs discussed here are produced at standard spatial and temporal $f_{samples}$. For example, a reanalysis dataset that contains hourly data on a 1km grid, or a collection of satellite imagery with 7-day repeat coverage on a 10m grid. In observational experimental design, decisions must be made about which spatiotemporal gridded EO datasets are most appropriate as inputs for an analysis. Likewise, we often need to combine observations at different sampling frequencies. In these cases, we will statistically upsample/upscale (use interpolation to model the dataset at higher temporal or spatial resolution) or downsample/downscale (use statistical aggregation to model the dataset at lower temporal or spatial resolution) one or more observations to represent the data as concomitant in space and/or time. When dealing with multi-phase, multi-frequency observational signals, we

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

recommend use of the "engineer's Nyquist" to estimate the ideal f_{sample}, which is at least 2.5 times the process frequency (Srinivasan et al. 1998), eq 1):

$$f_{sample} > 2.5 \times f_{process}, \qquad \text{(Eq 1)}$$

The engineer's Nyquist provides some additional safeguard against aliasing environmental processes with heterogenous or non-stationary phase or frequency.

For example, using the above example analyzing convective precipitation with a spatial $f_{process}$ of 1/100m and a temporal $f_{process}$ of 1/120 seconds (Fig. 7), we would to use precipitation data with a nominal spatial resolution less than 40m ($f_{sample} > 2.5 \times \frac{1}{100m}$ or $\frac{1}{40m}$) and a nominal temporal resolution less than 48 seconds ($f_{sample} > 2.5 \times \frac{1}{120s}$ or $\frac{1}{48s}$). The spatial scale of convective precipitation systems is often below the spatial sampling frequency of precipitation gauges and weather satellites. Because of this, gridded precipitation datasets, even those with spatial resolution below the size of convective storms, tend to have negative bias in total precipitation estimates associated with the spatial aliasing of convective precipitation. In some places, like the Midwestern United States, the majority of total precipitation is delivered by localized convective storms, leading to substantial negative bias in gridded precipitation estimates (Risser et al. 2019).

(iii) Do not use potentially aliased signals

In reality, *a-pri*ori definition of your spatial and temporal $f_{process}$ may not be a possibility. First, when dealing with complex, dynamic, unknown, or multiple process frequencies, it is unlikely that any single, heterogeneous "engineers Nyquist" exists that will uniformly protect against aliasing and inducing autocorrelation across the entire AOI and POI [Figure 7, (Subba Rao and Terdik 2017)]. Second, because of historical limitations in environmental sampling, the spatiotemporal scales of variability for many environmental processes remain unknown. But even if you are unsure of your target $f_{process}$, it is important to know the minimum $f_{process}$ that can be successfully evaluated from your data so that you do not incorrectly interpret aliased results. The minimum analyzable temporal $f_{process}$ for a dataset is minimum frequency with a signal to noise ratio above 0 decibles (dB) of your experimental data time series (Subba Rao and Terdik 2017). The minimum analyzable spatial $f_{process}$ of your spatial time series can be deduced by corresponding to the range of the spatial variogram of experimental

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

data (Garrigues et al. 2006; Lark 2002; Bogaert and Russo 1999). A dataset cannot be used to assess the importance of, or make predictions using proxies of, physical drivers acting below the minimum analyzable $f_{process}$.

(iv) Quantify autocorrelation

Spatiotemporal processes vary as a function of time (*f(t)*) and of two-dimensional space (*f(x,y)*; where x and y may represent latitude and longitude). Autocorrelation describes a situation where a variable is correlated with itself at some distance, either in time or in space, called a lag. At a lag of zero, we expect a signal to be perfectly correlated with itself. As we increase the lag, we expect the absolute value of the correlation to decrease. Positive autocorrelation occurs when subsequent observations have similar values, and negative autocorrelation occurs when subsequent observations have opposing values (Fig. 9).

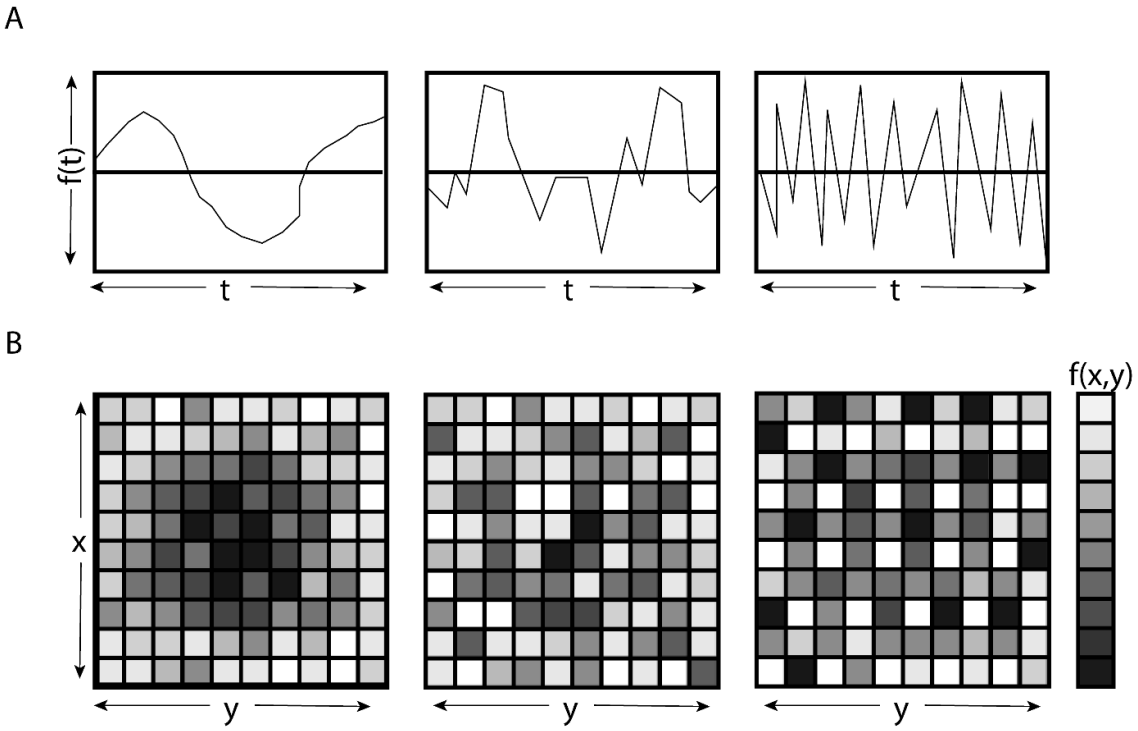


Fig. 9. A. Time series showing positive (left) neutral (middle) and negative (right) autocorrelation. B. Spatial process showing positive autocorrelation or clustering (left), no autocorrelation (middle), and negative autocorrelation (dispersal, right). Adapted from (Fortin and Dale 2009).

Autocorrelation can be a feature of the environmental process itself, or a feature of error in how the process has been observed or sampled. Examples of spatial processes that lead to

File generated with AMS Word template 2.0

autocorrelation include spatial diffusion (a process that spreads over space and time from an origin), spillover (a process that spreads across real or perceived boundaries, such as air pollution crossing national boundaries or diseases jumping between species), spatial interaction (a process characterized by movement in conscious response to certain characteristics of the environment, such as animal migration), and dispersal (a process that seeks distance from itself, like would be observed in the population distribution of territorial mammals such as panthers). Examples of spatial error that lead to autocorrelation include oversampling ($f_{process} >> f_{sample}$), measurement error (for example, antenna patterns in a radar image, a stream sensor that experiences calibration drift over time), and model misspecification (for example, omitted variable bias or bias in a parameter estimate) (Dubin 1998; F. Dormann et al. 2007; McMillen 2003).

Like multicollinearity, autocorrelation in spatiotemporal data can cause substantial problems in statistical inference, and must be quantified both in time and in space prior to modelling. Most methods for evaluating autocorrelation will calculate some metric of covariance (such as linear or rank correlation, covariance, or semi-covariance) of a signal that has been offset from itself at increasing temporal or spatial distances, or lags. Our two primary goals in evaluating autocorrelation in exploratory analysis are to ascertain whether significant autocorrelation is present at a given f_{sample}, and at what lag this autocorrelation becomes insignificant (referred to as the range of the data). For the purpose of GRRIEn analysis, we recommend using Moran's I statistic (Getis 2010) to calculate whether significant spatial autocorrelation is present globally in the data (Fig 11, Supplemental Appendix C), and the generation of spatial variograms (Garrigues et al. 2006; Oliver and Webster 1986) to estimate the range of this autocorrelation, or the distance at which autocorrelation becomes negligible (Fig 12, Supplemental Appendix C). For temporal autocorrelation, we recommend using an autocovariance function (ACF) to calculate the both the significance and range of temporal autocorrelation (Ma and Genton 2000; McLeod 1975). The diagnosis and management of autocorrelation in spatiotemporal data is the subject of extensive research, so these recommendations are necessarily simplistic. Excellent introductory information on management of autocorrelation in data can be found in the literature (F. Dormann et al. 2007; Dubin 1998; Ramezan et al. 2019).

2) MODEL TRAINING AND VALIDATION WITH OBSERVATION DEPENDENCE

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

To understand how autocorrelation impacts statistical inference, we'll start with the example of oversampling. Sampling too close together in time and/or space ($f_{sample} \gg 3 \times f_{process}$) is somewhat analogous to double counting ballots in an election. If we double count at random, it will inflate our sampling variance, and add instability to our results (Neville et al. 2004). If we double-count systematically, it will bias our results (Jensen and Neville 2002). Whether we're double counting at random or using some sort of structure, we'll think we have more votes than we do, and will therefore have false confidence in our results (Ferraciolli et al. 2019). Like multicollinearity, autocorrelation is associated with model instability (Neville et al. 2004), but since it can artificially inflate confidence in predictor estimates (i.e. a negative bias in standard error of model parameters), it can make erroneous or biased estimators more difficult to identify during model validation (Fig. 12a, Supplemental Appendix C).

Because aliasing represents information loss, it is important to err on the side of oversampling your data. Failure to diagnose and model autocorrelation, however, has been associated with inflated estimates of model skill, bias in parameter estimates, and bias in feature selection in statistical, machine learning, and deep learning models (Kattenborn et al. 2022; Segurado et al. 2006; Sergeev et al. 2019; Brenning 2005; Ferraciolli et al. 2019), including convolutional neural networks (CNNs) (Kattenborn et al. 2022). Because of this, addressing autocorrelation is critical for models to generalize well in observational spatiotemporal systems.

The methods for addressing autocorrelation in supervised learning fall into two major categories: strategically downsampling/downscaling the input data (e.g., resampling to the range of the spatial variogram) prior to model training, and modelling autocorrelation during model training. The two main approaches to modelling autocorrelation include spatial lag models, in which autocorrelation is parameterized directly in the model, or spatial error models, which force autocorrelation to the structure of model residuals. Some autocorrelation-tolerant modelling methods include auto covariate regression, autoregressive models, spatial eigenvector mapping, generalized least squares regression (Dormann et al. 2007), and latent variable grouping (Neville and Jensen 2005; Carter et al. 2016). Spatial lag models are appropriate when modelling spatial processes, i.e., autocorrelation that is a generating feature of the spatial process serving as the predictand. Spatial error models are

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

appropriate when autocorrelation is suspected to be an artifact of sampling, measurement, or model specification error. If autocovariance associated spatial processes, measurement error (Fig. 12d) or dependence with spatiotemporal features (Fig. 12c) are inaccurately parameterized in the model, this will be apparent when looking at model residuals (Section 3.c, "Checking your work;" Fig. 12).

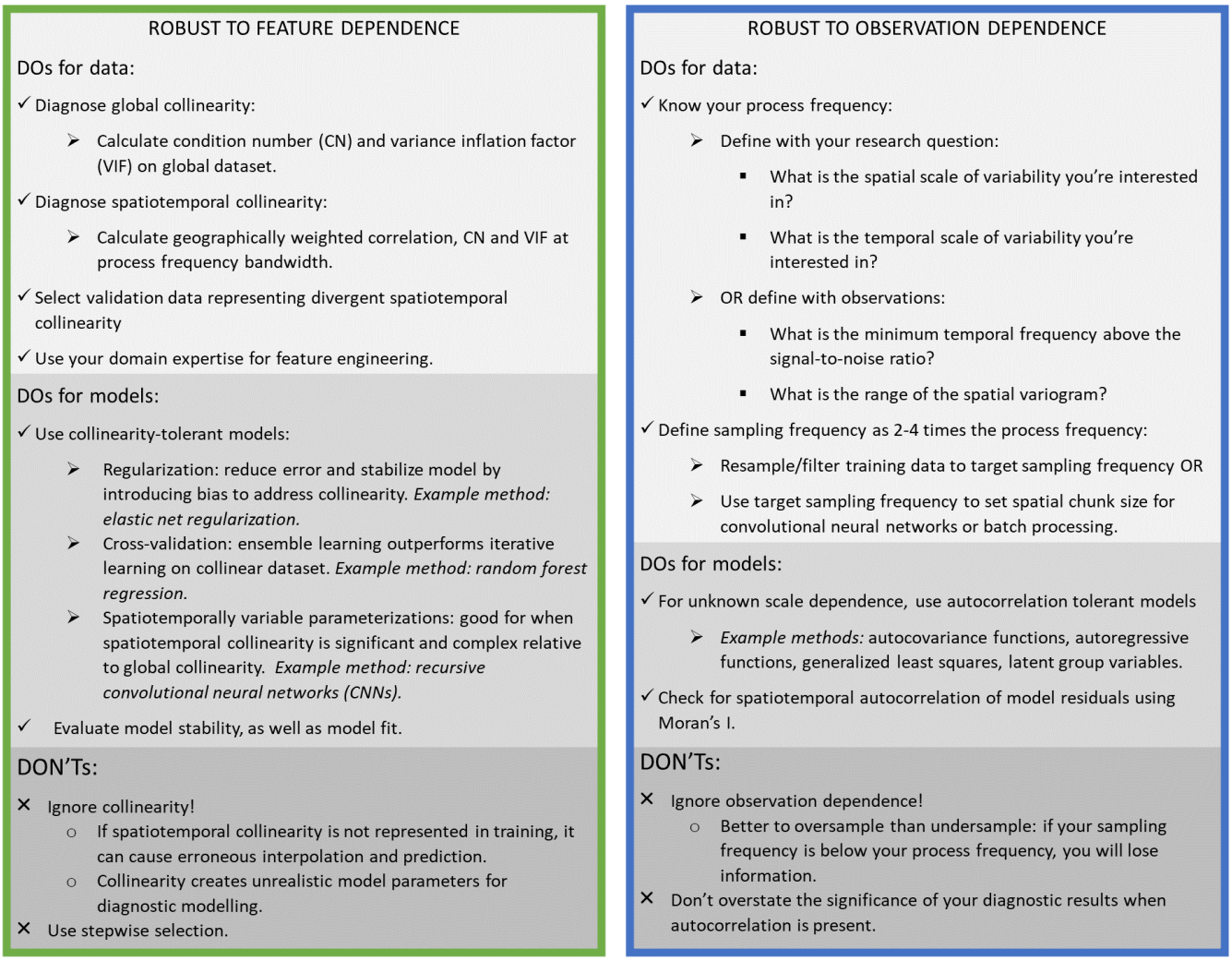


Fig. 10. Checklist for robust data engineering and model development with dependent features (left) and observations (right) in spatiotemporal systems.

c. Checking your work

In GRRIEn analysis, it is essential to example the residuals of your model (difference between predictions and observations of your predictand or error, Fig. 11) for spatial and temporal patterns. Spatial and temporal patterns in model residuals indicate that the model

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

will not generalize well, and offer clues as to why it is not robust (Dubin 1998; F. Dormann et al. 2007). Spatial or temporal autocorrelation of model residuals can be indicative of model instability induced by parameter and observation dependence, or undiagnosed modes of intrinsic autocovariance in our environmental process. It can also indicate omitted variable bias: a condition where some important driver of spatiotemporal variability has not been parameterized in the model. Carefully examining the presence of spatial autocorrelation of model residuals (Moran's I, Supplemental Appendix C, Fig. 11), the range of spatial autocorrelation of model residuals (spatial variogram), and the spatial distribution of model residuals (visual interpretation of spatial plots) can assist in diagnosing model misspecifications (Fig. 12).

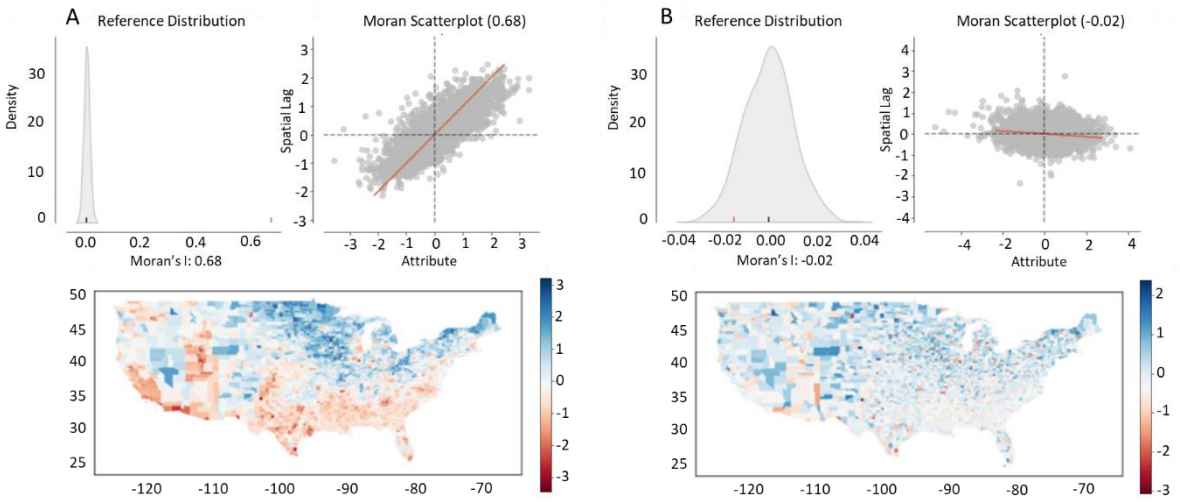


Fig. 11. Analysis of spatial autocorrelation of model residuals for A. Ordinary Least Squares linear regression and B. Autocovariate regression predicting 2012-2016 change in county-level voting patterns in US Presidential elections (McGovern et al. 2020). Bootstrapped Moran's I density plot (top left) and scatterplot (top right) generated by the splot package (Lumnitz et al. 2020) to visualize PySAL spatial analysis workflows (Rey and Anselin 2010). Figures in panel (A) show significant autocorrelation of model residuals, and figures in panel (B) show no significant autocorrelation of residuals. Including spatial autocorrelation as a covariate significantly decorrelated model residuals, and is associated with an increase in the standard error on model coefficients [Supplemental Appendix C, code adapted from (Wolf 2018)].

If spatial autocorrelation of residuals is present globally, as diagnosed by Moran's I, this needs to be clearly stated as a caveat to the results in the discussion, as it strongly indicates

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

that estimates of model skill are inflated because of unparameterized autocorrelation in the data and increases the likelihood of biased estimators (Fig. 12a) or generalized model misspecification (Fig. 12d). When either global or local multicollinearity is present, it is also very important to plot the residuals of the trained model predictions over space and time. Spatiotemporal clustering in residuals is one sign that spatiotemporally variable multicollinearity has created model instability that has impacted the generalizability of your predictions. This is particularly true if your residuals cluster in regions that had locally divergent collinearity (e.g., if your residuals, or errors, are spatially clustered where local correlation coefficients, CN, VIF, or VD are of different sign or magnitude than the global correlation coefficient, CN, VIF, VD) (Fig. 12b). Model residuals which are spatially concomitant with other unparameterized geographic covariates are indicative of omitted variable bias, which can be difficult to measure directly (Fig. 12c).

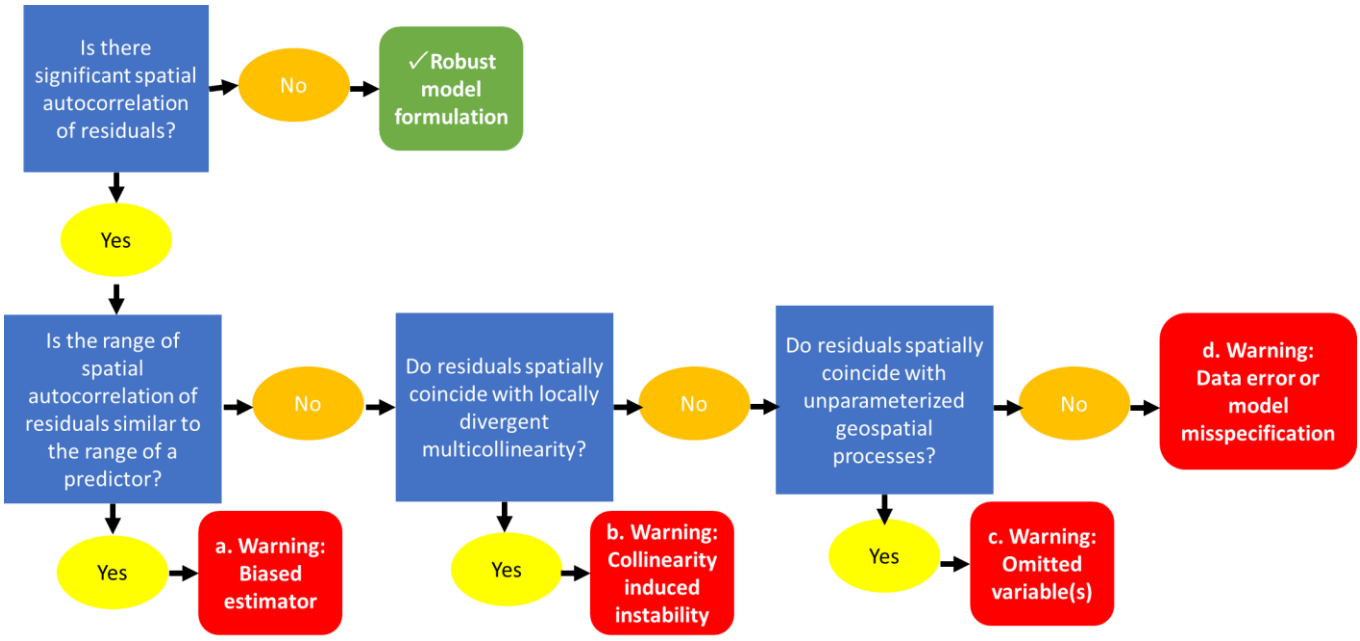


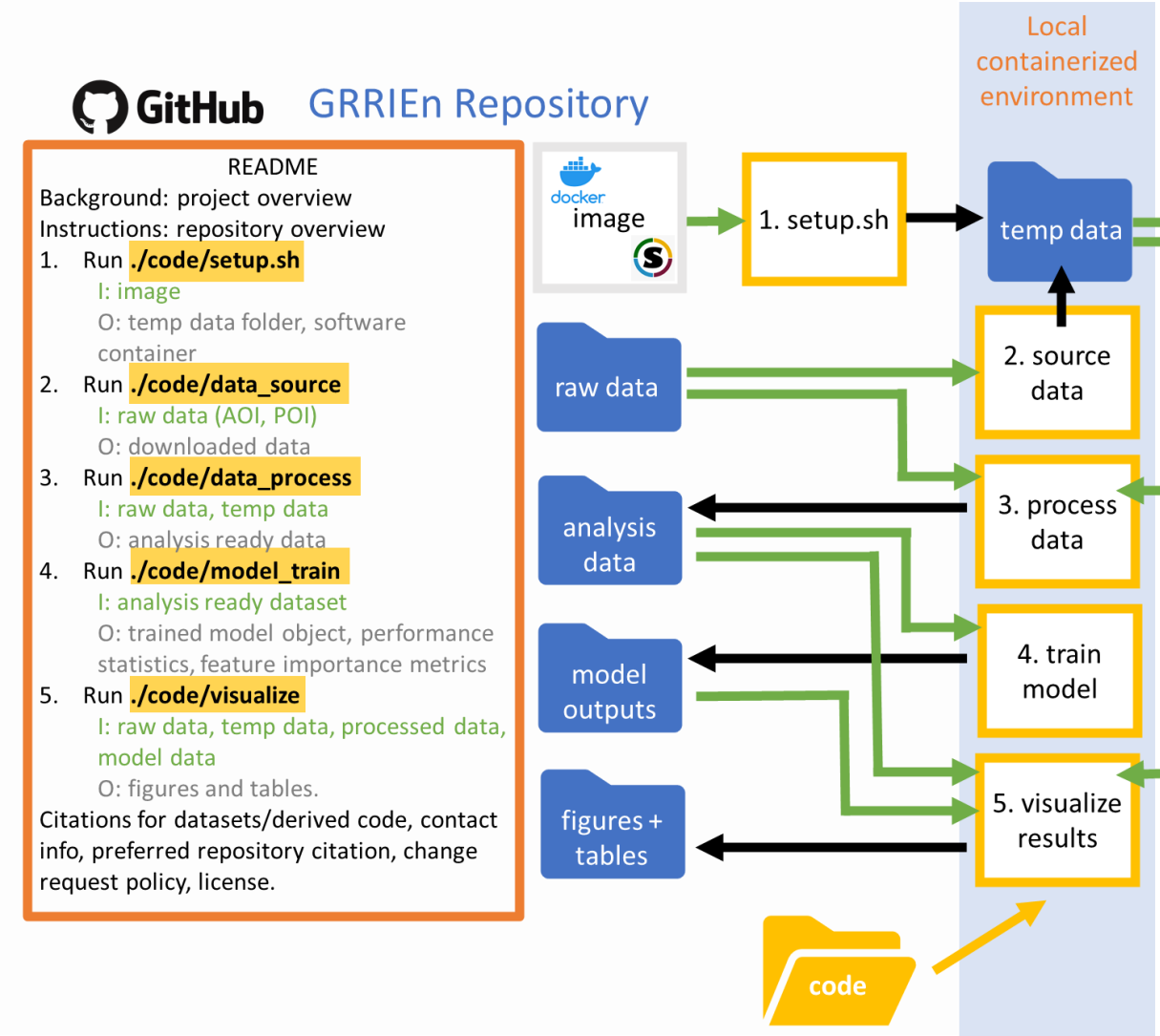
Fig. 12. Flowchart for interpretation of spatial autocorrelation of model residuals.

4. Reproducible

The term "reproducibility" is usually defined as obtaining the same results when others use the same datasets and methods of the original study. As outlined in the literature (National Academies of Sciences, 2019), reproducibility in research is often referred to as computational reproducibility: can another scientist understand your method sufficiently to

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

replicate data processing, model building, and validation in their computational environment? The reproducible research community utilizes an ever-expanding collection of computational tools to facilitate easy sharing of code and data. We summarize the current best practices relating to geospatial analysis as standard elements of a GRRIEn repository (Fig. 13), which facilitates easy replication of data acquisition, engineering, model training, and model evaluation, all of which can be published on a platform like GitHub alongside peer-reviewed publication of your results. More information on the software and hardware tools required to create a GRRIEn repository can be found in Supplementary Appendix A ("Required (Computational) Tools").



Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Fig. 13. Standard elements of a reproducible repository for end-to-and data engineering, modelling, and interpretation, utilizing global EOs from public geospatial data repositories as predictors. Inputs (I) and Outputs (O) of each code component are listed in README file, and are either contained in the GitHub repository or, in the case of large datasets, sourced from the internet.

The GRRIEn repository contains standard elements used commonly in open geospatial research to ensure maximum portability and effortless reproducibility and adaptability of supervised learning workflows utilizing global earth observations on public geodatabases. A GRRIEn repository should contain the following:

a. GRRIEn repository elements

- Raw data (folder): Contains your in-situ data, or labels, and standard delimiters of your AOI and POI (such as a start and end date, bounding box coordinates, or a shapefile) that can be used to collocate observations from global EOs.

- Analysis-ready data (folder): Contains the analysis-ready dataset merging your in-situ data and global proxy variables; most often a data frame, matrix, or array.

- Model outputs (folder): Contains trained model objects that can be called to make predictions, surrogate models used for model explanation (See Section 5: Interpreted), as well as any data objects describing model fit and stability.

- Figures and tables (folder): Contains graphical and tabular representations of results, including graphs, figures, maps, and tables in open-source file formats.

- Readme (html or Markdown file): A document providing detailed instructions on how to use elements of the repository for end-to-end analysis, the hardware and software required, the process for submitting merge requests, and license.

- Debian Linux-compatible software image (file): A container image is a static file that includes executable code that will initiate an instance in the local compute environment containing all version-controlled software, system libraries, and system tools to execute all code in a repository.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

- Version history (.gitconfig file) a detailed record of all edits and contributions made by individual team members to the repository over the life of the project.

- Code (folder): this contains all scripts used to interact with the other files and elements in the repository.

b. GRRIEn code elements

The code contained in the GRRIEn repository should be sufficient to replicate a software and dataset version controlled end-to-end analysis. The basic steps of working through the GRRIEn code can be seen in Fig. 13, and include programmatic workflows to set up required software in a local computing environment, searching and accessing raw data from quality-controlled open data repositories, processing input data into the data structure required by the software algorithm, training and validating algorithms, explaining algorithms, and visualizing results. For portability (i.e., can be run on any computer) and traceability (i.e., each step of the research process is documented), the GRRIEn repository contains the code required to access data from the cloud, instead of storing data in the repository itself. As such, it builds on the movement to make all research data publicly available on quality-controlled databases with stable digital object identifiers (DOIs).

- Setup.sh: bash script that initiates software image, installing all version-controlled software, libraries, and packages required by repository scripts, and creates a temporary data folder (temp data) in the user's local environment. By providing required software, the setup.sh script ensures the portability of the workflow. Reads from "Debian-Linux compatible software image;" exports to User's computer/compute platform.

- Data source: utilizes online geodatabase API to programmatically access data for AOI and POI given database catalog search parameters. The data sourcing code promotes the use of published datasets in version-controlled repositories external to the GitHub repository. Reads from: "raw data" folder, internet; exports to "temp data."

- Process data code: collocates raw data (predictand) and downloaded global EO data (predictor) in space and time, including coordinate reference system conversions, resampling functions, and zonal statistics functions; converts collocated data into an

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

analysis ready format (i.e. stacked raster, numpy array, pandas dataframe, R DataFrame); preprocesses analysis ready data for model training, including any feature reduction, variable transformations, and processing of missing data; calculates global and local multicollinearity; and global and local autocorrelation of all variables. Reads from "raw data" folder, "temp data" folder; exports to: "temp data" folder, "analysis data" folder.

- Model train and validate code: splits data into training and testing datasets; trains statistical model; derives statistics of model accuracy, stability, and summary statistics; calculates Moran's I, spatial variogram, and spatial plots of model residuals; and produces model explanations. Reads from "analysis data" folder; exports to "model data" folder.

- Figure and table generating code: code required to generate figures and tables. Reads from "processed data" folder, "model data" folder; exports to "figures and tables" folder.

Having these elements in your repository will enable end-to-end replication and portability of the research workflow, ensuring that other users can implement it without having to invest time compiling software, managing dependency chain issues, sourcing input data, or learning new computational techniques.

5. Interpreted

When trying to build models that are generalizable at landscape scales, it is not enough to rely purely on traditional metrics of model fit. One of the most awe-inspiring things about life on planet earth is its propensity towards uniqueness. It is very difficult to collect representative samples of landscape scale phenomena (Meyer and Pebesma 2022). Even in interpolation, patterns of interdependence between both parameters and observations in large-scale space/time systems are frequently both complex and dynamic in scale (Dormann et al. 2013). Because of global change, both natural and anthropogenic, forecasting and hindcasting in time will implicitly introduce new patterns in the scale dependence of and dependence between environmental variables (Refsgaard et al. 2014). Plus, in observational analysis we can never capture every factor that drives variability in our environmental process, either by proxy or direct measurement, so omitted variable bias is functionally

endemic. All this means that even after selecting the most important variables, training for model stability as well as model fit, ensuring that your data is not aliased, accounting for autocorrelation, and presenting your analytical pipeline in a format that can be vetted by your peers, your model may still produce erroneous interpolations and extrapolations in the wild.

To this end, the most scientifically relevant data emerging from your GRRIEn analysis pipeline is the trained model itself. The most critical tool you need to successfully implement GRRIEn analysis is your domain expertise as an earth scientist. In environmental analysis, you cannot reasonably conclude that your model will make the right predictions in different contexts unless you make sure those predictions are happening for the right reason: do your model weights and parameters reflect a physically plausible diagnosis of the environmental system? In other words, has your published model been *interpreted* within the context of your theoretical understanding of the system it purports to represent? It is the process of expert interpretation of the machine learning algorithm that leads to knowledge discovery in environmental data science.

a. Interpreted modelling step 1: form an interpretable hypothesis

Prior to modeling, use your domain expertise to write a series of hypothesis statements for every predictor variable used in model training, using very specific verbiage reflecting data and relationship type. This hypothesis should be motivated by examples from the experimental literature in your field. Things to consider in your hypothesis statement:

- *What kind of data are you using?* Are your variables continuous or categorical? If continuous, are the values bounded, and what are the data distributions? Has the data been log-transformed, normalized, or standardized prior to analysis? If categorical, are categories nominal or ordinal? Are there any categories that are arbitrary? Do you have class imbalances, or any categories that are not representatively sampled in the training data?

- *What types of relationships do you expect to find?* For continuous or ordinal data, is the relationship between the predictand and individual predictor linear or nonlinear? If nonlinear, is the relationship monotonic (predictand consistently either increases or decreases as predictor increases) or non-monotonic (predictand increases with increase in predictor for some range(s) of predictor value and decreases for other

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

range(s) of predictor value)? If non-monotonic, is the relationship global (e.g., polynomial relationship) or local (e.g., changepoints or peak over threshold response)? In multivariate datasets, are there interactions between predictors (e.g., maize yield response to air temperature is positive when precipitation is high, negative when precipitation is low)? If interactions are suspected, are these interactions isotropic (monotonic across both variables) or anisotropic (nonmonotonic across either variable)? When evaluating interactions between continuous and categorical data, what moment(s) of the distribution of the continuous variable are likely to modify, or be modified by, the categorical variable (e.g., does storm type change the mean, variance, skew, or kurtosis of precipitation rate)?

b. Interpreted modelling step 2: identify a model interpretation method.

Model explanation methods fall into two main categories: instance (local) explanation methods and model (global) explanation methods (Fig. 14).

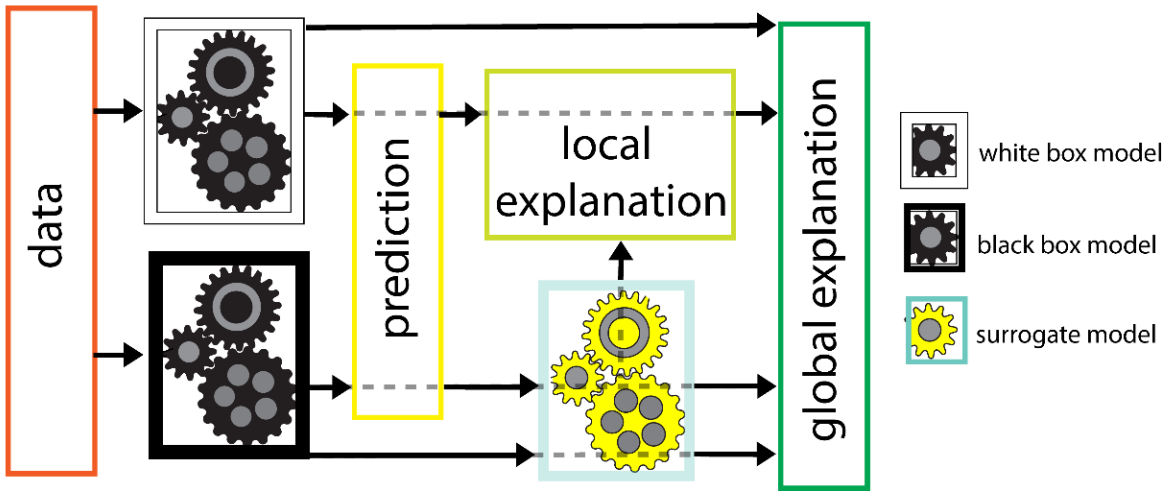


Fig. 14. The model explanation pipeline for black box and white box algorithms. For white box models, global explanations can be extracted from the model itself. For black box algorithms, global explanations must be inferred by way of a surrogate model, trained from both input data and predictions. Local explanations are derived from interpreting predictions across all possible values of predictors. Figure adapted from (Burkart and Huber 2021).

Instance (local) explanation methods allow the user to understand how the predictand responds to the predictor by producing local (to values of predictor) predictions for specific

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

ranges of input values (e.g., like one example [or sample] from the training dataset). Local explanation methods are useful for exploring and qualitatively characterizing non-linear (including non-monotonic and non-isotropic) relationships between input and output variables. Local feature importance metrics are often interpreted both quantitatively (for example, comparing the magnitude of variance in predictand associated with individual predictor variables to determine relative importance of a variable among multivariate predictors) and visually (for example, a graph of predictions along the range of possible values of a predictor). Methods for instance explanation include: locally-interpretable model-agnostic explanations (LIMEs), Shapely values, and local sensitivity analysis (Ryo et al. 2021).

Model (global) explanation methods quantify or summarize the importance of a predictor across all possible values of predictands. Generally, this is accomplished by providing an easy-to-understand function that can generalize how predictand will respond to all values of predictors. For the purpose of generating global explanations, we can divide supervised algorithms into two main classes: white box models and black box models (Fig. 14). With white box models, which include linear models (multivariate regression, logistic regression), decision trees, rule-based models, interactive models, and Bayesian networks (Burkart and Huber 2021), global feature importance can be deciphered from the model object itself. For example, the sign and magnitude of regression coefficients yield insight on the nature and strength of the relationship between predictor and predictand. Black box models, such as neural networks, map relationships between predictors and predictands through a series of weights applied to different transformations of data, and as such they produce model objects which are difficult for humans to quantitatively interpret. For black box models, we must generate *surrogate models* that approximate the trained model's function, often using coefficients associated with linear or nonlinear (i.e., polynomial, interactive terms) representations of input variables, to arrive at a human-interpretable global explanation. Methods for surrogate model imputation include the sum of Shapely values (Lundberg et al. 2020; Aas, Jullum, and Løland 2021), decision paths (Van Assche and Blockeel 2007; Sagi and Rokach 2020), and counterfactual explanations (Verma, Dickerson, and Hines 2020).

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Both local and global feature importance offer utility in helping scientists interpret black box machine learning algorithms for knowledge discovery. Local feature importance metrics can assist with providing a physical explanation for feature selection, comparing the relative importance of predictors, confirming predictand values associated with peak-over-threshold responses, and confirming qualitative theories of variability in systems. Global model explanations are useful if your goal is to learn a complex, unknown function describing your environmental system (such as to parameterize a process model). They are strongly encouraged anytime you will be using your machine learning method to extrapolate beyond the spatial/temporal domain of your training dataset, or interpolate to a different spatial/temporal resolution that may be associated with unique physical drivers.

c. Interpreted modelling step 3: use local and global explanation methods to confirm or reject original hypotheses.

In model interpretation, we use local and global feature importance metrics to evaluate our original hypothesis. Do local and global feature importance metrics confirm or reject your original data hypothesis, as it related to your data types and variable types? If not, does this represent a plausible new discovery of spatiotemporal variability in your system? If it does not, the model is not robust.

6. Conclusion: using GRRIEn for experimental design.

Most students are trained to write a research manuscript containing five core sections: introduction, methods, results, discussion, conclusion. This standard format has evolved alongside experimental science. Much like our GRRIEn repository structure (Section 4) ensures reproducibility of computational research, the format of a standard research manuscript is a well-worn, highly effective roadmap for reproducibility in experimental research. Each section corresponds to a stage of the analysis, and the guidelines for what must be included in these sections mirror details which must be attended to for a rigorous experimental analysis to occur. Fig. 15 translates this foundational manuscript structure for GRRIEn (Generalizable, Reproducible, Robust, and Interpreted Environmental) analysis.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

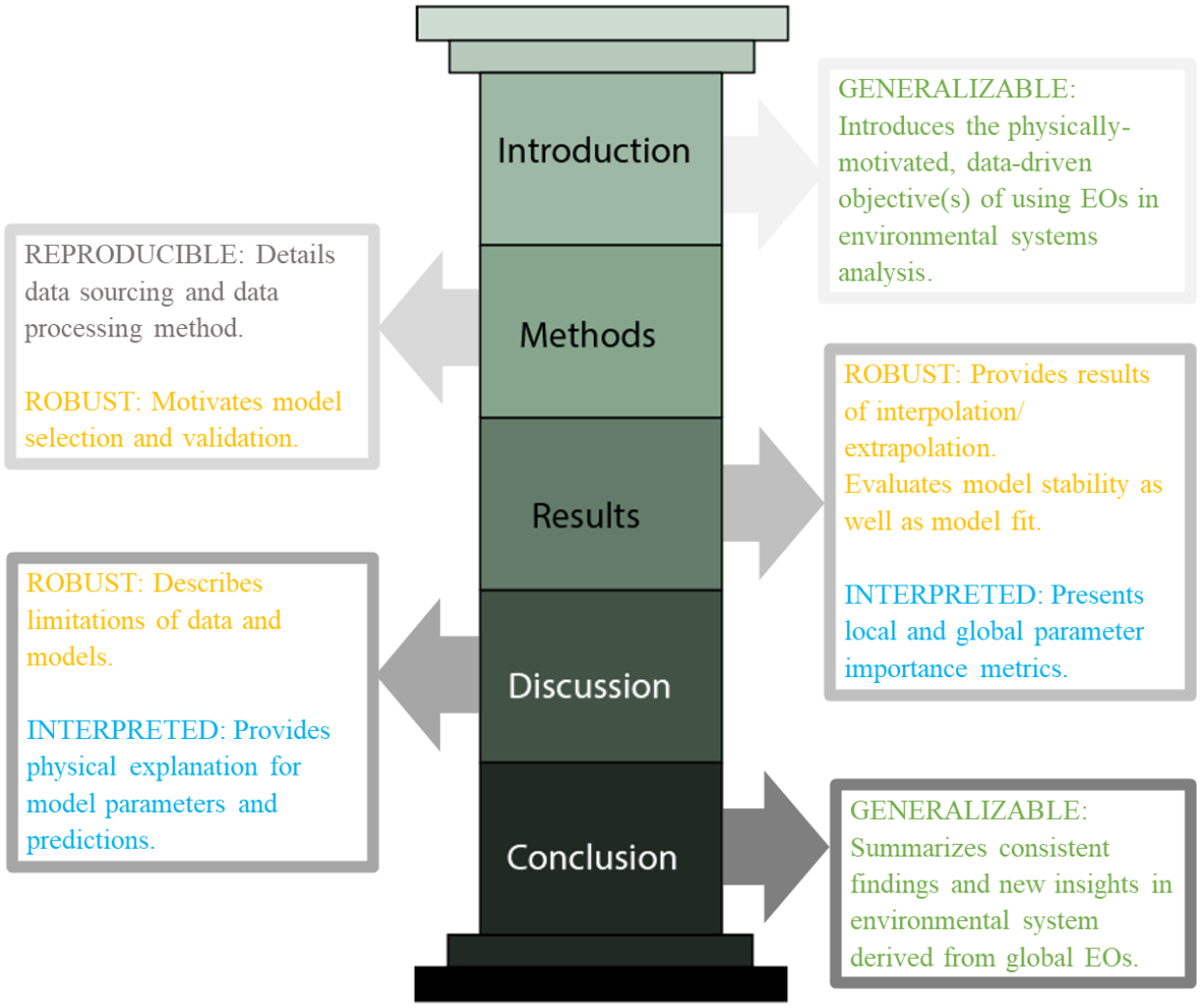


Fig. 15. Structuring the research manuscript for GRRIEn analysis.

Introduction: theoretically motivates the analysis, presents research objectives (interpolation, extrapolation, or diagnostic modeling), and decomposes this objective into a set of physically-motivated hypotheses relating the predictand (environmental process observations) to predictors (globally-available earth observations).

Methods: Introduces the predictand and predictors. Defines the area of interest and period of interest. Defines and theoretically motivates process and sampling frequencies. Demonstrates that training data reflects environmental process variability across the AOI and POI. Describes how predictand and predictors were co-associated into an analysis ready dataset, including any spatial/temporal resampling and coordinate reference system conversions. Describes how subsetting of testing and training data is representative of any

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

scale-dependent feature and observation dependence in the data. Motivates model selection, such as regularization, local calibration, or autocorrelation functions, using any feature and observation dependence in your data.

Results: Presents results of interpolation or extrapolation, quantifies model fit and stability, decomposes model error into bias and variance, provides spatial analysis of model residuals, and presents local/global model interpretation.

Discussion: Using original hypotheses as guides, contextualizes model interpretation in theoretical background discussed in introduction. Describes caveats in results of interpolation or extrapolation related to autocorrelation, multicollinearity, omitted variable bias, or nonsensical model diagnostics. Discusses limitations of current suite of global EOs in resolving research objectives that relate to sensor design, or data spatial or temporal resolution.

Conclusion: Summarizes consistent findings and knowledge discovery related to the environmental process.

Global earth observations provide an opportunity to analyze spatiotemporal variability in environmental systems at unprecedented levels of detail, but to unlock this information, earth scientists with must work with a suite of computational tools that are evolving alongside the data and require technical training. For earth science disciplines to maintain a legacy of rigor, access to these computational tools for research must not interfere with focused training in, or application of, skills in theoretical and applied sciences in a researcher's area of expertise. The goal of GRRIEn analysis is to anchor the role of the earth scientists embarking in unchartered territories of computational research in the traditions of rigor, consistency, transparency, and theoretical history that have long formed the foundation for scientific understanding by simplifying expectations for open spatiotemporal data science. The GRRIEn framework is not meant to be complete or static. The computer hardware, software, data, and algorithms that are used in environmental data science are all evolving at breakneck speed. Instead, by outlining universal components of computational experimental pipelines in the earth sciences, including dataset engineering, model training, and model interpretation, the GRRIEn framework seeks to set a baseline for community standards that formally integrates the methods of the scientific process within computational research. Earth scientists embarking on research in the digital age may find themselves in uncharted territory.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

The GRRIEn analysis framework is intended to serve as a packing list of critical computational tools and spatiotemporal statistics to enable traceable paths of true knowledge discovery on this journey.

Acknowledgments.

We want to acknowledge the anonymous reviewers of this manuscript, whose substantial contributions of time, expertise, and insight can be seen throughout the finished work. In addition, we acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for the Coupled Model Intercomparison Project (CMIP), and we thank the climate modeling groups (listed in Figure 5 of this paper) for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Data Availability Statement.

Figure 12 utilizes US Elections Data (McGovern et al. 2020), prepared by the Center for Spatial Data Science at the University of Chicago.

File generated with AMS Word template 2.0

REFERENCES

Adnan, M. S. G., M. S. Rahman, N. Ahmed, B. Ahmed, M. F. Rabbi, and R. M. Rahman, 2020: Improving Spatial Agreement in Machine Learning-Based Landslide Susceptibility Mapping. *Remote Sensing*, **12**, 3347.

Alin, A., 2010: Multicollinearity. *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 370–374.

Anselin, L., and X. Li, 2020: Tobler's law in a multivariate world. *Geogr. Anal.*, **52**, 494–510.

Audebert, N., B. Le Saux, and S. Lefevre, 2019: Deep Learning for Classification of Hyperspectral Data: A Comparative Review. *IEEE Geoscience and Remote Sensing Magazine*, **7**, 159–173.

Balsamo, G., and Coauthors, 2018: Satellite and In Situ Observations for Advancing Global Earth Surface Modelling: A Review. *Remote Sensing*, **10**, 2038.

Bárcena, M. J., P. Menéndez, M. B. Palacios, and F. Tusell, 2014: Alleviating the effect of collinearity in geographically weighted regression. *J. Geogr. Syst.*, **16**, 441–466.

Benesty, J., J. Chen, Y. Huang, and I. Cohen, 2009: Pearson Correlation Coefficient. *Noise Reduction in Speech Processing*, I. Cohen, Y. Huang, J. Chen, and J. Benesty, Eds., Springer Berlin Heidelberg, 1–4.

Berrar, D., 2019: Cross-Vlidation. *Encyclopedia of Bioinformatcis and Computational Biology*, Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C, Ed., Academic Press, 542–545.

Blei, D. M., and P. Smyth, 2017: Science and data science. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 8689–8692.

Bogaert, P., and D. Russo, 1999: Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resour. Res.*, **35**, 1275–1289.

Bond, C. E., A. D. Gibbs, Z. K. Shipton, S. Jones, and Others, 2007: What do you think this is?``Conceptual uncertainty''in geoscience interpretation. *GSA Today*, **17**, 4.

Brauner, N., and M. Shacham, 1998: Role of range and precision of the independent variable in regression of data. *AIChE J.*, **44**, 603–611.

Breiman, L., 1996a: Bagging predictors. *Mach. Learn.*, **24**, 123–140.

——, 1996b: Stacked regressions. *Mach. Learn.*, **24**, 49–64.

——, 2001: Random Forests. *Mach. Learn.*, **45**, 5–32.

Brenning, A., 2005: Spatial prediction models for landslide hazards: review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.*, **5**, 853–862.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Burkart, N., and M. F. Huber, 2021: A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.*, **70**, 245–317.

Campbell, C. A., W. L. Pelton, and K. F. Nielsen, 1969: Influence of solar radiation and soil moisture on growth and yield of Chinook wheat. *Can. J. Plant Sci.*, **49**, 685–699.

Carter, Herrera, and Steinschneider, 2021: Feature Engineering for Subseasonal-to-Seasonal Warm-Season Precipitation Forecasts in the Midwestern United States: Toward a Unifying Hypothesis of …. *J. Clim.*,.

Carter, E. K., J. Melkonian, S. J. Riha, and S. B. Shaw, 2016: Separating heat stress from moisture stress: analyzing yield response to high temperature in irrigated maize. *Environ. Res. Lett.*, **11**, 094012.

——, ——, S. Steinschneider, and S. J. Riha, 2018a: Rainfed maize yield response to management and climate covariability at large spatial scales. *Agric. For. Meteorol.*, **256–257**, 242–252.

——, S. J. Riha, J. Melkonian, and S. Steinschneider, 2018b: Yield response to climate, management, and genotype: a large-scale observational analysis to identify climate-adaptive crop management practices in high-input maize systems. *Environ. Res. Lett.*, **13**, 114006.

Challinor, a. J., J. Watson, D. B. Lobell, S. M. Howden, D. R. Smith, and N. Chhetri, 2014: A meta-analysis of crop yield under climate change and adaptation. *Nat. Clim. Chang.*, **27**, 1–5.

Chan, J. Y.-L., S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z.-W. Hong, and Y.-L. Chen, 2022: Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Sci. China Ser. A Math.*, **10**, 1283.

Chen, Y., Z. Lin, X. Zhao, G. Wang, and Y. Gu, 2014: Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **7**, 2094–2107.

Chen, Y., Y. Wang, Z. Dong, J. Su, Z. Han, D. Zhou, Y. Zhao, and Y. Bao, 2021: 2-D regional short-term wind speed forecast based on CNN-LSTM deep learning model. *Energy Convers. Manage.*, **244**, 114451.

Clark, J. S., and A. E. Gelfand, 2006: A future for models and data in environmental science. *Trends Ecol. Evol.*, **21**, 375–380.

Committee on Earth Observing Satellites, 2022: CEOS Database. *Instruments Table*,. http://database.eohandbook.com/index.aspx (Accessed May 23, 2023).

Cristiano, and Veldhuis, 2017: Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas–a review. *Hydrol. Earth Syst. Sci.*,.

Dormann, C. F., and Coauthors, 2013: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* , **36**, 27–46.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Dubin, R. A., 1998: Spatial Autocorrelation: A Primer. *J. Hous. Econ.*, **7**, 304–327.

F. Dormann, C., and Coauthors, 2007: Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* , **30**, 609–628.

Feng, X., D. S. Park, Y. Liang, R. Pandey, and M. Papeş, 2019: Collinearity in ecological niche modeling: Confusions and challenges. *Ecol. Evol.*, **9**, 10365–10376.

Ferraciolli, M. A., F. F. Bocca, and L. H. A. Rodrigues, 2019: Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. *Comput. Electron. Agric.*, **161**, 233–240.

Fortin, and Dale, 2009: Spatial autocorrelation. *The SAGE handbook of spatial analysis*,.

Freund, Y., and R. E. Schapire, 1999: A Short Introduction to Boosting. http://www.yorku.ca/gisweb/eats4400/boost.pdf (Accessed November 13, 2022).

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *aos*, **29**, 1189–1232.

Garrigues, S., D. Allard, F. Baret, and M. Weiss, 2006: Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sens. Environ.*, **103**, 81–96.

Getis, A., 2010: Spatial Autocorrelation. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, M.M. Fischer and A. Getis, Eds., Springer Berlin Heidelberg, 255–278.

Goldman, A. E., S. R. Emani, L. C. Pérez-Angel, J. A. Rodríguez-Ramos, J. C. Stegen, and P. Fox, 2021: Special Collection on Open Collaboration Across Geosciences. *Eos*, https://doi.org/10.1029/2021EO153180.

Graham, M. H., 2003: Confronting multicollinearity in ecological multiple regression. *Ecology*,.

Guo, J., Q. Xiong, J. Chen, E. Miao, C. Wu, Q. Zhu, Z. Yang, and J. Chen, 2022: Study of static thermal deformation modeling based on a hybrid CNN-LSTM model with spatiotemporal correlation. *Int. J. Adv. Manuf. Technol.*, **119**, 2601–2613.

Hembram, T. K., S. Saha, B. Pradhan, K. N. Abdul Maulud, and A. M. Alamri, 2021: Robustness analysis of machine learning classifiers in predicting spatial gully erosion susceptibility with altered training samples. *Geomatics, Natural Hazards and Risk*, **12**, 794–828.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049.

Hilborn, R., and M. Mangel, 2013: *The Ecological Detective*. Princeton University Press,.

Hoerl, A. E., and R. W. Kennard, 1970: Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55–67.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Jensen, and Neville, 2002: Linkage and autocorrelation cause feature selection bias in relational learning. *ICML*,.

Kalogirou, 2013: Testing geographically weighted multicollinearity diagnostics. *Paper presented at GISRUK 2013*,.

Kattenborn, T., F. Schiefer, J. Frey, H. Feilhauer, M. D. Mahecha, and C. F. Dormann, 2022: Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, **5**, 100018.

Kidd, C., A. Becker, G. J. Huffman, C. L. Muller, P. Joe, G. Skofronick-Jackson, and D. B. Kirschbaum, 2017: So, How Much of the Earth's Surface Is Covered by Rain Gauges? *Bull. Am. Meteorol. Soc.*, **98**, 69–78.

Kim, J. H., 2019: Multicollinearity and misleading statistical results. *Korean J. Anesthesiol.*, **72**, 558–569.

de Knegt, H. J., and Coauthors, 2010: Spatial autocorrelation and the scaling of species–environment relationships. *Ecology*, **91**, 2455–2465.

Lark, R. M., 2002: Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49–80.

Li, K., and N. S. N. Lam, 2018: Geographically weighted elastic net: A variable-selection and modeling method under the spatially nonstationary condition. *Annals of the American Association of*,.

Li, M., D. Niu, Z. Ji, X. Cui, and L. Sun, 2021: Forecast Research on Multidimensional Influencing Factors of Global Offshore Wind Power Investment Based on Random Forest and Elastic Net. *Sustain. Sci. Pract. Policy*, **13**, 12262.

Lu, B., P. Harris, M. Charlton, and C. Brunsdon, 2014: The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo Spat. Inf. Sci.*, **17**, 85–101.

Lumnitz, S., and Coauthors, 2020: splot - visual analytics for spatial statistics. *J. Open Source Softw.*, **5**, 1882.

Luvall, Lee, Stavros, and Glenn, 2017: Thriving on Our Changing Planet-A Decadal Strategy for Earth Observations from Space: Surface Biology and Geology Designated Observables. *United States Forest Service*,.

Ma, Y., and M. G. Genton, 2000: Highly robust estimation of the autocovariance function. *J. Time Ser. Anal.*, **21**, 663–684.

Mahadi, M., T. Ballal, M. Moinuddin, and U. M. Al-Saggaf, 2022: A Recursive Least-Squares with a Time-Varying Regularization Parameter. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, **12**, 2077.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

McGovern, T., S. Larson, B. Morris, and M. Hodges, 2020: County-level presidential election results for 2008, 2012, and 2016. https://doi.org/10.5281/zenodo.3975765.

McLeod, 1975: Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Appl. Stat.*,.

McMillen, D. P., 2003: Spatial Autocorrelation Or Model Misspecification? *Int. Reg. Sci. Rev.*, **26**, 208–217.

Meyer, H., and E. Pebesma, 2022: Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.*, **13**, 2208.

Mudereri, and Dube, 2019: A comparative analysis of PlanetScope and Sentinel-2 space-borne sensors in mapping Striga weed using Guided Regularised Random Forest classification …. *Remote Sens. Spat …*,.

Murakami, D., N. Tsutsumida, T. Yoshida, T. Nakaya, and B. Lu, 2021: Scalable GWR: A Linear-Time Algorithm for Large-Scale Geographically Weighted Regression with Polynomial Kernels. *Ann. Assoc. Am. Geogr.*, **111**, 459–480.

Murugan, P., and S. Durairaj, 2017: Regularization and Optimization strategies in Deep Convolutional Neural Network. *arXiv [cs.CV]*,.

National Academies of Sciences, Engineering, and Medicine, 2018: *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press,.

Nativi, S., P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, and O. Ochiai, 2015: Big Data challenges in building the Global Earth Observation System of Systems. *Environmental Modelling & Software*, **68**, 1–26.

Neville, J., and D. Jensen, 2005: Leveraging relational autocorrelation with latent group models. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, ieeexplore.ieee.org, 8 pp.-.

Neville, J., O. Simsek, and D. Jensen, 2004: Autocorrelation and relational learning: Challenges and opportunities.

Nickerson, R. S., 1998: Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.*, **2**, 175–220.

Oliver, M. A., and R. Webster, 1986: Semi-variograms for modelling the spatial pattern of landform and soil properties. *Earth Surf. Processes Landforms*, **11**, 491–504.

Ramezan, C. A., T. A. Warner, and A. E. Maxwell, 2019: Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, **11**, https://doi.org/10.3390/rs11020185.

Refsgaard, J. C., and Coauthors, 2014: A framework for testing the ability of models to project climate change and its impacts. *Clim. Change*, **122**, 271–282.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Rey, S. J., and L. Anselin, 2010: PySAL: A Python Library of Spatial Analytical Methods. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, M.M. Fischer and A. Getis, Eds., Springer Berlin Heidelberg, 175–193.

Risser, M. D., C. J. Paciorek, T. A. O'Brien, M. F. Wehner, and W. D. Collins, 2019: Detected Changes in Precipitation Extremes at Their Native Scales Derived from In Situ Measurements. *J. Clim.*, **32**, 8087–8109.

Runge, J., and Coauthors, 2019: Inferring causation from time series in Earth system sciences. *Nat. Commun.*, **10**, 2553.

Ryo, M., B. Angelov, S. Mammola, J. M. Kass, B. M. Benito, and F. Hartig, 2021: Explainable artificial intelligence enhances the ecological interpretability of black‐box species distribution models. *Ecography* , **44**, 199‐205.

Saha, S., R. Sarkar, J. Roy, T. K. Hembram, S. Acharya, G. Thapa, and D. Drukpa, 2021: Measuring landslide vulnerability status of Chukha, Bhutan using deep learning algorithms. *Sci. Rep.*, **11**, 16374.

——, A. Saha, T. K. Hembram, K. Mandal, R. Sarkar, and D. Bhardwaj, 2022: Prediction of spatial landslide susceptibility applying the novel ensembles of CNN, GLM and random forest in the Indian Himalayan region. *Stoch. Environ. Res. Risk Assess.*, https://doi.org/10.1007/s00477-022-02212-3.

Salazar, J. J., L. Garland, J. Ochoa, and M. J. Pyrcz, 2022: Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *J. Pet. Sci. Eng.*, **209**, 109885.

Segurado, P., M. B. Araujo, and W. E. Kunin, 2006: Consequences of spatial autocorrelation for niche-based models. *J. Appl. Ecol.*, **43**, 433–444.

Sergeev, A. P., A. G. Buevich, E. M. Baglaeva, and A. V. Shichkin, 2019: Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. *Catena*, **174**, 425–435.

Shannon, C. E., 1949: Communication in the Presence of Noise. *Proceedings of the IRE*, **37**, 10–21.

Shi, Y., W. Gong, Q. Duan, J. Charles, C. Xiao, and H. Wang, 2019: How parameter specification of an Earth system model of intermediate complexity influences its climate simulations. *Progress in Earth and Planetary Science*, **6**, 46.

Smith, G., 2018: Step away from stepwise. *Journal of Big Data*, **5**, https://doi.org/10.1186/s40537-018-0143-6.

Smith, P. F., S. Ganesh, and P. Liu, 2013: A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J. Neurosci. Methods*, **220**, 85–91.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Srinivasan, R., D. M. Tucker, and M. Murias, 1998: Estimating the spatial Nyquist of the human EEG. *Behav. Res. Methods Instrum. Comput.*, **30**, 8–19.

Srisa-An, C., 2021: Guideline of Collinearity - Avoidable Regression Models on Time-series Analysis. *2021 2nd International Conference on Big Data Analytics and Practices (IBDAP)*, ieeexplore.ieee.org, 28–32.

Stephens, G., and Coauthors, 2020: The Miniaturization Revolution in Earth Observing. *Bull. Am. Meteorol. Soc.*, **101**, 373–378.

Subba Rao, T., and G. Terdik, 2017: On the frequency variogram and on frequency domain methods for the analysis of spatio-temporal data. *J. Time Ser. Anal.*, **38**, 308–325.

Tamura, R., K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, and T. Matsui, 2017: BEST SUBSET SELECTION FOR ELIMINATING MULTICOLLINEARITY. 日本オペレーションズ・リサーチ学会論文誌, **60**, 321–336.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bull. Am. Meteorol. Soc.*, **93**, 485–498.

Thornton, P. E., S. W. Running, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.*, **190**, 214–251.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, **58**, 267–288.

Tredinnick, L., and C. Laybats, 2018: Big data archives. *Business Information Review*, **35**, 142–144.

US News and World Report, 2022: Best Earth Science Programs. *US News and World Report*,. https://www.usnews.com/best-graduate-schools/top-science-schools/earth-sciences-rankings/21775470034_variation (Accessed July 23, 2022).

Versloot, C., 2020: *What are L1, L2 and Elastic Net Regularization in neural networks? – MachineCurve*. https://github.com/christianversloot/machine-learning-articles/blob/main/what-are-l1-l2-and-elastic-net-regularization-in-neural-networks.md.

Wang, F., and Coauthors, 2021: Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility. *Alex. Eng. J.*, **60**, 5813–5829.

Wen, T., X. Niu, M. Gonzales, G. Zheng, Z. Li, and S. L. Brantley, 2018: Big Groundwater Data Sets Reveal Possible Rare Contamination Amid Otherwise Improved Water Quality for Some Analytes in a Region of Marcellus Shale Development. *Environ. Sci. Technol.*, **52**, 7149–7159.

Wheeler, D. C., 2009: Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: The Geographically Weighted Lasso. *Environ. Plan. A*, **41**, 722–742.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1

Wolf, L., 2018: Spatial Autocorrelation Functions. *Yet Another Geographer*,. https://www.ljwolf.org/post/spatial_acf/ (Accessed November 21, 2022).

Xu, J., Z. Ma, S. Yan, and J. Peng, 2022: Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. *J. Hydrol.* , **605**, 127353.

Young, M. T., M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D. Kaufman, 2016: Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression. *Environ. Sci. Technol.*, **50**, 3686–3694.

Zang, H., L. Liu, L. Sun, L. Cheng, Z. Wei, and G. Sun, 2020: Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. *Renewable Energy*, **160**, 26–41.

Zou, H., and T. Hastie, 2005: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **67**, 301–320.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-22-0065.1