Building an Open Science Framework to Model Soil Organic Carbon

Edward Flathers* and Paul E. Gessler

Abstract
As funding agencies embrace open science principles that encourage sharing data and computer code developed to produce research outputs, we must respond with new modes of publication. Furthermore, as we address the expanding reproducibility crisis in the sciences, we must work to release research materials in ways that enable reproducibility—publishing data, computer code, and research products in addition to the traditional journal article. Toward addressing these needs, we present an example framework to model and map soil organic carbon (SOC) in the cereal grains production region of the northwestern United States. Primarily associated with soil organic matter, SOC relates to many soil properties that influence resiliency and soil health for agriculture. It is also critical for understanding soil–atmospheric C flux, a significant part of the overall C budget of the Earth. The scorpan technique for modeling soil properties uses seven categories of environmental input data to make predictions: known soil attributes, climatic values, organisms present, relief, parent material, age, and spatial location. We gather data representing these categories from various public sources. The map is produced using a random forest statistical model with scorpan inputs to predict SOC content on a 30-m spatial grid. All modeling components including input data, metadata, computer code, and output products are made freely available under an explicit open-source license. In this way, reproducibility is supported, the methods and code released are available to be reused by other researchers, and the research products are open to critical review and improvement.

Core Ideas
• We created a spatial model of SOC, which is important for agriculture and climate change.
• Open science principles and practices lead to better science and reproducibility.
• It is important to make readers aware of the hidden limitations of data and models.

As funding bodies for research (e.g., USDA, National Science Foundation, and National Institutes of Health, among others) embrace free and open publication of research data, many science disciplines are developing a new culture of data sharing. For example, a recent study using magnetic resonance imaging data has uncovered a flaw in a common analytical technique. The authors note: “through the introduction of international data-sharing initiatives in the neuroimaging field, it has become possible to evaluate the statistical methods using real data” (Eklund et al., 2016). Data sharing has enabled scientists to check methods and improve methods in ways that have not previously been possible, allowing rigorous science to advance more quickly.

In addition to data sharing, researchers are developing software systems to help enable a more complete sharing culture. For example, the Center for Open Science developed the Open Science Framework (https://osf.io) to organize components of research projects and enable collaboration and sharing of project materials including data, code, and text both during and after projects. The adoption of open science practices, including sharing research data and software, has the potential to improve the progress of science in the same way as publication of methods and results in journal articles, as scientists learn and take inspiration from the works of their peers. However, open science practices can also improve science in other ways. Gezelter (2015) argues that “as numerical experiments become more complex and the datasets become larger, calculations that are reproducible in principle are no longer reproducible in practice without access to the code, data, and the metadata that describes how the data is organized.”

A 2016 Nature survey showed that 52% of researcher respondents (no information was given about their fields of study) agreed that there is a “significant crisis of reproducibility” across the sciences (Baker, 2016). The reproducibility crisis is caused by factors common to all science: “Problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the...
results” (Open Science Collaboration, 2015). Openly sharing science data, code, and products provides an avenue for reproducing results in any discipline (McNutt, 2014).

One obstacle to the embracing of open data sharing is the danger of being scooped by other researchers who use shared data to publish papers. This type of scooping has allegedly occurred in genomics, when Massachusetts Institute of Technology researchers published a paper using data made publicly available by the Woods Hole Marine Biological Laboratory in a way contrary to the restrictions imposed on the data by the authors (Marshall, 2002). One approach to mitigating this danger is a publication embargo, which grants original researchers time to publish using their data before they become public (Cragin et al., 2010). However, the length of the embargo should not extend beyond the useful life of data and code. Ideally, the embargo provides researchers a head start on publication while also allowing access to others while products are still relevant.

Reproducibility is required for rigorous science and demonstrates the fundamental stability of the methods applied in a study. Without the ability to reproduce an experiment, scientists have no way to judge the veracity of the results. Asendorpf et al. (2013) define reproducibility as providing a set of outputs that researchers must produce to enable reproduction of their studies: raw data, metadata, and the actual code used to perform the analyses.

Using this definition of reproducibility, this paper documents a dataset and associated analytical products that meet Asendorpf’s criteria. The USDA-funded Regional Approaches to Climate Change for Pacific Northwest Agriculture (REACCH-PNA) is a project focused on the potential impacts of climate change on cereal grain production in the northwestern United States. One product for the study area is a derived map of soil organic carbon (SOC) as a base from which C dynamics can be mapped and monitored. Soil organic C is primarily associated with soil organic matter and relates to many soil properties that influence resiliency and soil health. It is also critical for understanding soil-atmospheric C flux, which is a significant part of the overall C budget of the Earth (Raich and Schlesinger, 1992). Although there are ongoing efforts to produce global maps of SOC (FAO, 2017), the scope of this model is limited to a smaller geographic area.

The soil C map is produced by applying a scorpan technique to create a random forest statistical model to predict SOC content for a spatial grid of 30-m resolution. The scorpan model is a more recent implementation of Jenny’s (1941) quantitative pedology work into a framework for predicting soil types and properties (Florinsky, 2012). Calibration and evaluation of the model is performed using point-based SOC observations. The explanatory variables are gridded geospatial data describing soil, climate, organisms, topography (relief), parent material, age, and spatial position (McBratney et al., 2003). Because soil respiration creates a flux of soil C that is partly dependent on an erosion–deposition cycle (Doetterl et al., 2016), topography-derived hydrological and geomorphological layers are also included in the model (Gessler et al., 2000). All inputs to the model are collected from data that various agencies and researchers have made freely available online.

“Big data” is an overloaded term in research—it is used to describe data that are large in volume, variety, velocity, value, or complexity (Kaisler et al., 2013). Although the explanatory variables involved in the scorpan model are large (~180 GB), the volume of data is not the greatest challenge in collating the inputs. The more significant big data challenge is the variety of data: a collection of geospatial data produced by a diversity of organizations at different times and different spatial and temporal scales, using varied units of measurement, to name just a few of the differences. The extract, transform, load (ETL) process is designed to ease the integration of the data (Vassiliadis, 2009). Despite this, some artifacts of the disparate origins of the data remain. The “ecological fallacy” describes a misinterpretation of statistical data in which a characteristic of aggregate data is simply applied to groups within the aggregate (Selvin, 1958; Piantadosi et al., 1988). Because some of the gridded input data for the model are collected at larger spatial scales (4- to 30-m cells), we commit the ecological fallacy when we assign the attributes of a 4-km cell to the 30-m cells contained within it. Despite this, we proceed with the analysis as a real-world compromise; it is often the case that data do not fit elegantly together for analysis, and it is important to be explicit about potential sources of error.

The model output product layers can be used as inputs to other spatial analysis projects (Moore et al., 1993; Gessler et al., 1995, 2000). Instruments located around the REACCH-PNA study area are monitoring CO₂ flux and combining stored soil C data with CO₂ flux data for a better understanding of how soil-atmospheric CO₂ flux relates to stored soil C, and how fluxes change over time. Additionally, other members of the REACCH-PNA team are working on projects such as CropSyst, a crop simulation model that outputs, among other attributes, soil organic matter (Stöckle et al., 2003). Soil C maps have the potential to provide inputs or serve as a basis for comparing inputs and outputs of CropSyst and other models. These maps also provide for combination with climate change scenarios and known agroecosystem domains that suggest shifting of cropping systems as a result of climate change. This effort helps develop the building blocks for such analyses across the region.

The aim of this paper is to create and demonstrate a repeatable, reusable framework for applying a scorpan model for mapping SOC over the REACCH study area. This is an initial step toward developing accurate, spatially explicit soil C maps to support analyses understanding that the initial map is likely inaccurate. Explicit publication of data and methods provides a framework to refine the modeling and improve the outputs. The focus is to demonstrate the concepts of open science and a reusable and modifiable framework that can be improved on or applied in other spatial and temporal contexts and scales. All modeling components, including input data, metadata, computer code, and output products, are made freely available under an explicit open-source license. In this way, Asendorpf’s criteria are explicitly met, the methods and code released are available for reuse, and research products are plainly open to critical review and improvement.

Data Development Process
Software and Algorithms
The methods listed in this section are all implemented in the statistical programming language R (https://www.r-project.org/) and the more general-purpose programming language Python (https://www.python.org/), combined with the Esri ArcGIS Python application program interface (API) (https://
Collected Data Products

The input products downloaded for use in the model are listed below. The general workflow for each data input is:

1. Download data from the provider.
2. Preprocess the data:
   a. Project, if necessary, into Albers Equal Area projection.
   b. Clip data to remove any cells outside the REACCH bounding box.
   c. Write the resulting raster to disk.

Two of the input data products involve processing beyond this general workflow. Using the digital elevation model, we derive layers for slope and topographic wetness index. The point-location soil samples require more extensive processing, described below. Once this workflow is complete for each input, the covariate raster data are combined into a comma-separated value (CSV) file for each county that intersects the REACCH bounding box. The CSV files for each county are joined using identifiers for each observation, and then values of interest (SOC percentage weight, soil bulk density, and rock content) are extracted for each sample. Samples lacking any of the values of interest are discarded. Soil organic C content (g m\(^{-2}\)) is computed following the method described by Bliss et al. (1995) for each soil layer and summed to find a total SOC value for each sample location. Sample locations outside the REACCH bounding box are removed, and all remaining samples are added to a point-based shapefile. The process is repeated for each county, building on the shapefile until all counties are processed.

Soil: USDA-NRCS gSSURGO

The USDA Natural Resources Conservation Service (NRCS) produces the Gridded Soil Survey Geographic (gSSURGO) Database, a 10-m spatial resolution gridded dataset with associated tabular data describing soil series and associated characteristics across the United States. The gSSURGO dataset is based on the SSURGO polygon vector dataset, a product of rasterizing the SSURGO polygon vector dataset, a product of rasterizing the [SSURGO] vector polygons in an Albers Equal Area projection, which is to say, the vector polygons of SSURGO have been divided into 10-m grid cells and all the values of each polygon transferred to each grid cell (USDA NRCS, 2016). Because gSSURGO is based on county soil survey data, the temporal scale varies depending on the update frequency of the counties. The associated tabular data include a measure of SOC (g m\(^{-2}\)) and are used primarily as a basis for comparison of the model output of the scarpan process.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, and Washington) and are processed by mosaiccing the state grids together and then extracting the grid cells from within the area of the REACCH bounding box, and finally the SOC value from the associated table is joined to the grid.
Climate: GRIDMET

Given the importance of precipitation on SOC dynamics, a precipitation and temperature dataset has been included from Abatzoglou’s Gridded Surface Meteorological Data (GRIDMET) (Abatzoglou, 2013). Mean annual temperature and mean annual precipitation have been shown to be significant predictors of SOC variability (Morrow, 2014). These data are raster data with 4-km spatial resolution describing average precipitation and minimum and maximum temperature from 1979 to 2010. Using these data with a 4-km spatial resolution involves commission of the ecological fallacy; however, the importance of these climate variables on C dynamics in combination with the difficulty of obtaining higher resolution data make these data the best currently available for the purpose.

The data are downloaded from the Northwest Knowledge Network’s (https://www.northwestknowledge.net) Thematic Real-time Environmental Distributed Data Services (THREDDS) server (http://www.unidata.ucar.edu/software/thredds/current/tds/), which allows the user to specify spatial and temporal bounds, as well as aggregation criteria. Therefore, the downloaded data already represent the correct variables and spatial extent and require no preprocessing other than reprojection.

Organisms: USDA-NASS NCDL

The USDA National Agricultural Statistics Service (NASS) produces the National Cropland Data Layer (NCDL) yearly as a grid with 30-m spatial resolution aligned to the 30-m National Elevation Dataset (NED) grid. The NCDL algorithmically classifies individual grid cells into agricultural cover types using satellite imagery, supervised classification techniques, and ground truthing. Since the area of interest of this study area is primarily agricultural land, the NCDL crop categories provide detailed information about land cover within the region.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, and Washington) and are processed by mosaicking the state grids together, then extracting the grid cells from within the area of the REACCH bounding box. Sinks are filled (Reuter et al., 2009) and a slope grid is generated, followed by flow direction, flow accumulation, and TWI.

Relief: USGS NED

The USGS NED is a gridded digital elevation model for the United States with 30-m spatial resolution. The varying topography of the study area influences erosional and depositional patterns of SOC across the landscape. Several products are derived from the NED: slope, a depression-filled digital elevation model (O’Callaghan and Mark, 1984), flow accumulation and flow direction (Jenson and Domingue, 1988), and a topographic wetness index (TWI) (Quinn et al., 1991; Moore et al., 1993; Gessler et al., 1995). Flow accumulation and direction are intermediate layers in this model; only elevation, slope, and TWI are used as model inputs.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, and Washington) and are processed by mosaicking the state grids together, then extracting the grid cells from within the area of the REACCH bounding box. Sinks are filled (Reuter et al., 2009) and a slope grid is generated, followed by flow direction, flow accumulation, and TWI.

Parent Material and Age: USGS Aeroradiometric Grids

The USGS aeroradiometric grids are 2-km-spatial-resolution maps of potassium, thorium, and uranium concentration in the top 30 cm of the Earth’s surface (Duval et al., 2005). The aeroradiometric data are thought to be a convenient proxy for parent materials (Gessler et al., 1995; Bierwirth et al., 1996). The data were collected between 1973 and 1981 by various contractors at various spatial resolutions and were combined into a single dataset for the conterminous United States by the USGS with some data loss (Hill et al., 2009). Because of the heterogeneous composition of these data, their variable spatial and temporal
Table 2. Land cover classification mapping. Old classes are mapped to new classes to reduce the total number of classes. Pixel count is included as an indication of relative frequency of old classes.

<table>
<thead>
<tr>
<th>New class</th>
<th>Old class</th>
<th>Pixel count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>86</td>
<td>Sunflower</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>Mint</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>6,785</td>
<td>Other crops</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>2,297</td>
<td>Clover and wildflowers</td>
<td></td>
</tr>
<tr>
<td>224</td>
<td>71</td>
<td>Vetch</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>773,482</td>
<td>Sugarbeets</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>1,667</td>
<td>Miscellaneous vegetables and fruits</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0</td>
<td>Watermelons</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>Cucumbers</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>0</td>
<td>Tomatoes</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0</td>
<td>Caneberrys</td>
<td></td>
</tr>
<tr>
<td>206</td>
<td>7,040</td>
<td>Carrots</td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>0</td>
<td>Asparagus</td>
<td></td>
</tr>
<tr>
<td>208</td>
<td>0</td>
<td>Garlic</td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>0</td>
<td>Cantaloupe</td>
<td></td>
</tr>
<tr>
<td>216</td>
<td>541</td>
<td>Peppers</td>
<td></td>
</tr>
<tr>
<td>219</td>
<td>477</td>
<td>Greens</td>
<td></td>
</tr>
<tr>
<td>221</td>
<td>0</td>
<td>Strawberries</td>
<td></td>
</tr>
<tr>
<td>227</td>
<td>2,054</td>
<td>Lettuce</td>
<td></td>
</tr>
<tr>
<td>242</td>
<td>0</td>
<td>Blueberries</td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>0</td>
<td>Cabbage</td>
<td></td>
</tr>
<tr>
<td>246</td>
<td>2,120</td>
<td>Radishes</td>
<td></td>
</tr>
<tr>
<td>247</td>
<td>3,937</td>
<td>Turnips</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>5,265</td>
<td>Peaches</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>710</td>
<td>Christmas trees</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>0</td>
<td>Other tree crops</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>0</td>
<td>Walnuts</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>311</td>
<td>Pears</td>
<td></td>
</tr>
<tr>
<td>218</td>
<td>767</td>
<td>Nectarines</td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>681</td>
<td>Plums</td>
<td></td>
</tr>
<tr>
<td>223</td>
<td>0</td>
<td>Apricots</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>121</td>
<td>2,561,272</td>
<td>Developed and open space</td>
</tr>
<tr>
<td>122</td>
<td>1,021,456</td>
<td>Developed and low intensity</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>394,022</td>
<td>Developed and medium intensity</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>44,875</td>
<td>Developed and high intensity</td>
<td></td>
</tr>
<tr>
<td>222</td>
<td>0</td>
<td>Squash</td>
<td></td>
</tr>
<tr>
<td>229</td>
<td>160</td>
<td>Pumpkins</td>
<td></td>
</tr>
<tr>
<td>249</td>
<td>0</td>
<td>Gourds</td>
<td></td>
</tr>
</tbody>
</table>

resolutions, and their age relative to the rest of the input products, they may serve as a weak proxy for present-day parent material but are nonetheless included as the best nationwide aeroradiometric product that currently exists. As with the climate data, the mismatch in spatial resolution between these data and other gridded inputs represents commission of the ecological fallacy.

The data are downloaded as a spatial grid for each element (potassium, thorium, and uranium) for the continental United States and are processed by extracting the grid cells from within the area of the REACCH bounding box.

Reproducibility

To enable reproducibility, input data, metadata, computer code, and output products will be packaged together and made freely available for download. The input data for this model are publicly available datasets; however, as new versions of datasets are released, the specific data used in this project may be eliminated, altered, or replaced. Therefore, input data are included in the packaging in their raw form, in addition to processed products of those input data.

Computer code that was used to download and/or perform ETL and preprocessing steps to prepare the input data for the modeling process is also included. The ETL programs developed are able to collect data from the necessary repositories, process the data, and generate metadata that describes the data-processing steps used to prepare data for model ingestion. For example, the NED data are used to derive other products such as slope and TWI, which then go on to inputs to the modeling process. These intermediate, processed data are also included in the project package to demonstrate that the ETL code verifiably produces the intermediate data, and that the intermediate data are then used as inputs to the modeling process.

The R code that implements the modeling process itself is also included. The random forest process depends on a pseudorandom number generator to execute. Owing to different implementations of pseudorandom number generators, and of floating point operations and representation across different computer architectures, model outputs may not be exactly identical to the output produced by this project. However, the architecture and software used to produce these outputs are documented in the project metadata files, and a random number seed is fixed within the model code to maximize the replicability of the exact model outputs.

The publishable results of this project will be the product layers, the ETL and model programs used to create the product layers, appropriate metadata documentation, and this paper describing the development and implementation of the software and product layers. All products are freely available in an online repository and licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0) (https://creativecommons.org/Licenses/by-sa/4.0/legalcode). In short, this license expresses that others who wish to use any part of this project are free to do so in any context they wish, so long as they agree to provide attribution to the authors in any publication they make and agree to make their derived products likewise accessible.

Analysis

McSweeney et al. (1994) provide a foundation for modeling soil characteristics using GIS and soil horizon characteristics. We combine this approach with the scorpan model, formalized by McBratney et al. (2003), which describes a set of covariates used as inputs to digital soil models. These inputs are soil, climate, organisms, topography, parent material, age, and spatial position. Odgers et al. (2014) applied the scorpan model in combination with an algorithm called DSMART (Disaggregation and Harmonization of Soil Map Units through Resampled Classification Trees). Chaney et al. (2016) extended the DSMART algorithm to work in supercomputing environments (DSMART-HPC) and applied the scorpan method to develop a soil series map of the contiguous United States. As an example analysis of the environmental data collected, and following the general methods of these papers, we develop a model using scorpan inputs to develop a map of SOC. Because soil C is a continuous variable, a classification algorithm is not an appropriate
predictive tool, and a random forest regression algorithm is substituted. The *scorpan* approach explicitly supports the use of modeling continuous attributes of soil (McBratney et al., 2003).

With the exception of the point-based measurements of SOC used to train the model, all of the explanatory variables take the form of geospatial gridded datasets.

**Model Selection, Performance, and Diagnostics**

A model selection process was followed to choose up to seven input variables (this limit was chosen as the maximum due to memory constraints). Each of mean annual temperature, mean annual precipitation, NCDL classification, elevation, slope, TWI, potassium, thorium, and uranium were introduced into the model, replacing the least important terms when the limit was reached. By this method, a random forest model was specified to predict the natural log of SOC for the full available soil depth using elevation, thorium (Th), uranium (U), NCDL classification, slope, mean annual temperature, and mean annual precipitation:

\[
\ln(SOC) \sim \text{elevation} + \text{Th} + \text{U} + \text{NCDL} + \text{slope} + \text{temperature} + \text{precipitation}
\]

Table 3 shows these variables in order of importance measured as the increase in mean squared error (MSE) when their values are randomly permuted during model training. The size of the random forest was fixed at 512 trees, per Oshiro et al. (2012) showing that increasing the number of trees far beyond 128 is likely to show little benefit for the increase in processing time due to the low number of covariates in the model leading to asymptotically small performance gains of additional trees; as processor speed has improved since 2012, this threshold was multiplied. The random forest approach produces a pseudo-\(R^2\) value that is equal to 1 – (MSE/variance), and it indicates the model's benefit over a null model of using the grand mean of the independent variable as the prediction. This model produces a pseudo-\(R^2\) of 20.49%, which is acceptable for the proof-of-concept purpose of this paper. The model mean squared residuals value of 0.25 is low, although the value is reported in log-transformed units and may represent overfitting of the model. Figure 2 is a map of predicted SOC for the REACCH study area. Figure 3 is a map of the same area using SSURGO SOC values for comparison. In both maps, a general west-to-east increase in SOC within the REACCH study area is shown. An east-to-west precipitation gradient over the area is a primary SOC driver (Morrow, 2014), reflecting the importance of precipitation in the model. The maps are particularly different in the southeastern portion of the study area within Idaho, where very few training sample locations were found. Figure 4 is a map of the variance of the random forest estimator for each point on the grid. The relatively high variance in that same area of Idaho indicates less stability in the random forest estimator in that area.

**Discussion**

**Covariates**

The pool of covariates were chosen because of their historical inclusion in similar modeling exercises (Jenny, 1941; McBratney et al., 2003; Morrow, 2014), and an exhaustive effort to identify additional or different covariates was not undertaken. The primary goal of this paper is to develop and describe a repeatable framework as a first step toward a more accurate SOC model. It would be informative to repeat the modeling process with a different land cover classification product such as the National Land Cover Database. Additional covariates could also be added, including the Multiresolution Valley Bottom Flatness Index (MRVBF) (Gallant and Dowling, 2003), geological data that could help characterize the parent material at a spatial resolution that is better or meets the 30-m resolution of the other covariate layers, and other covariates as needed. The publication of the input data and processing code required to execute the model make it feasible for the community to further evolve the model with new and different covariates or different modeling approaches. The intent of the framework approach is, at least initially, to produce a foundation for development and comparison of approaches to SOC modeling (and potentially other soil- or agriculture-centric climate analyses).

**The Ecological Fallacy**

The 4-km grid cell footprint of the climatic variables and the 2-km grid of the USGS aeroradiometric data appear prominently in areas throughout the modeled output product. This is a consequence of the assumption that the values of these large grid cells are representative of the values of the 30-m cells that are overlaid from the other gridded products. This assumption is an example of the ecological fallacy, in which a characteristic of aggregate data is simply applied to groups within the aggregate (Selvin, 1958; Piantadosi et al., 1988).

Two approaches to eliminating this issue from the result are to aggregate the other input layers from 30 m up to 4 km, thus creating a complete set of inputs that are comparable in spatial scale, or to use alternate data layers for climate and aeroradiometrics (possibly even derived from the current layers using a statistically valid downscaling technique) to provide influence from climate and parent material in the model. The latter approach is beyond the scope of this paper, and the former approach was not chosen to highlight the existence of this common issue in spatial modeling and to illustrate the effects of a third choice—using the data as available and explicitly noting the incompatibility and the existence of statistical flaws in the approach. Since much of the spatial data that we use are gathered from external providers like the USGS, there are limitations to the compatibility of our various layers. In some cases, we may find that our models work best with data that are incompatible in some ways, and it is important to advertise these incompatibilities to readers and potential users of our products.
Spatial Extent and Scale

The spatial extent of this modeling effort was driven by the boundaries of the area of interest to the REACCH-PNA project. Given that the boundaries were established according to cereal production capacities and practices in the region, there is reduced variability within some or all of the variables of the model that suggests the suitability of this specific model and its inputs may be significantly different when applied to different geographic areas. Nonetheless, the general applicability of the scorpion model and the relatively nonspecific implementation framework presented here could be readily adapted to other spatial extents.

The 30-m spatial resolution of the input and output products was chosen as a result of the availability of input products at that resolution, driven primarily by the USGS elevation grid, which has been adopted by the creators of other gridded datasets to allow for convenient overlay. Depending on the spatial scale of
physical processes involved in the model and the intended application of any output products of the modeling process, the 30-m grid may or may not be an appropriate resolution. The USGS also makes available a 10-m gridded elevation layer, which may be more suited to certain types of analysis; however, it may be difficult or impossible to assemble necessary covariate layers at that resolution. Again, the code framework described here is readily adapted to process data of various spatial resolutions.

Temporal Considerations

One challenge of building these types of models is the availability of data products that are collected at temporal scales similar to their natural variability. Some of the covariate layers, such as elevation, are not likely to change dramatically over short periods of time. Other layers, such as land cover classification, may change significantly from year to year, particularly in agricultural areas where cropping systems drive the rotation of different crops into fields over time, and where crop selection may be market driven. The NCSS soil samples have been taken over a range of years that is not necessarily expressed in the data tables, and although some areas may experience relatively stable soil conditions, this is not necessarily the case in agricultural areas under various management regimes. The mean annual temperature and precipitation layers are the two most important layers to the model and are also of great interest, considering changes expected under various likely regimes of climate change. It is unclear how much the variability of some input products over time influences the model output, but assessing the stability of the modeling process over time could be informative.

Summary

The reproducibility crisis is spreading in the sciences, and in light of its inherent complications, particularly with respect to climate change research, it is important that researchers embrace open science principles. Science is a process of building on existing work, and by publishing all components of research, including input, processing code, and output, we make explicit the foundation on which new science can be built.

To that end, this paper describes a framework of input data, processing code, and outputs designed around the concept of modeling SOC for the cereal grain producing region of the northwestern United States. The framework can be improved iteratively with updated versions of its existing covariates and with new covariates as they become available, with alternate data processing tools, and with improved statistical models. Run periodically for covariates representing different time periods, the framework could be used to model C dynamics over time. The framework can furthermore be altered to focus on different geographic areas or scales and to model other environmental variables, related to soil or otherwise. The flexibility and reusability of the framework makes it a potential foundation for more extensive modeling efforts, but it also makes explicit the processes that have gone into producing its results.

As we address reproducibility and the rapid pace of modern science, we can help ourselves by embracing open science by:

1. Publishing input data, ensuring that older versions of the data remain available and uniquely identifiable to preserve replicability
2. Publishing computer code, ideally in languages, APIs, and tools that are freely available, and ideally complete enough to replicate published results with relative ease
3. Publishing processed data that allows others to cross-check processing code, processed outputs, and result data
4. Publishing a paper that traditionally describes research motives, methods, and results
5. Applying an explicit license to all published products to ensure that downstream users are aware of their rights and responsibilities when using data or code

Fig. 4. Variance map of the soil organic carbon (SOC) estimator. We can see that the variance is higher in the southeastern area of the study area, where the model predicted higher SOC than the Soil Survey Geographic (SSURGO) model reports.

Variances (SOC)
High: 6488
Low: 191

Journal of Environmental Quality 733
References


Eklund, A., T.E. Nichold, and H. Knutsson. 2016. Cluster failure: Why fMRI data and code are all accessible using this DOI: 10.7923/G4XP72ZB. The computer code is also available on GitHub at https://github.com/flathers/soilCarbonFramework


Jenson, S.K., T.E. Nichold, and H. Knutsson. 2016. Cluster failure: Why fMRI data and code are all accessible using this DOI: 10.7923/G4XP72ZB. The computer code is also available on GitHub at https://github.com/flathers/soilCarbonFramework

Acknowledgments

This material is based on work that is supported by the National Institute of Food and Agriculture, USDA, under Award no. 2011-68002-30191.