2016

POLARIS: A 30-meter probabilistic soil series map of the contiguous United States

Nathaniel W. Chaney
Princeton University, nchaney@princeton.edu

Eric F. Wood
Princeton University

Alexander B. McBratney
The University of Sydney

Jonathan W. Hempel
National Soil Survey Center

Travis W. Nauman
Southwest Biological Science Center

See next page for additional authors

Geoderma 274 (2016) 54–67

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

POLARIS: A 30-meter probabilistic soil series map of the contiguous United States

Nathaniel W. Chaney [a,*], Eric F. Wood [b], Alexander B. McBratney [c], Jonathan W. Hempel [d], Travis W. Nauman [e], Colby W. Brungard [f], Nathan P. Odgers [c]

[a] Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA
[b] Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA
[c] Department of Environmental Sciences, Faculty of Agriculture and Environment, The University of Sydney, Sydney, Australia
[d] National Soil Survey Center, NRCS, Lincoln, NE, USA
[e] U.S. Geological Survey, Southwest Biological Science Center, Moab, UT, USA
[f] Department of Plants, Soils, and Climate, Utah State University, Logan, UT, USA

ARTICLE INFO

Article history:
Received 12 November 2015
Received in revised form 19 March 2016
Accepted 24 March 2016
Available online xxxx

Keywords:
Digital soil mapping
Environmental modeling
High performance computing

ABSTRACT

A new complete map of soil series probabilities has been produced for the contiguous United States at a 30 m spatial resolution. This innovative database, named POLARIS, is constructed using available high-resolution geospatial environmental data and a state-of-the-art machine learning algorithm (DSMART-HPC) to remap the Soil Survey Geographic (SSURGO) database. This 9 billion grid cell database is possible using available high performance computing resources. POLARIS provides a spatially continuous, internally consistent, quantitative prediction of soil series. It offers potential solutions to the primary weaknesses in SSURGO: 1) unmapped areas are gap-filled using survey data from the surrounding regions, 2) the artificial discontinuities at political boundaries are removed, and 3) the use of high resolution environmental covariate data leads to a spatial disaggregation of the coarse polygons. The geospatial environmental covariates that have the largest role in assembling POLARIS over the contiguous United States (CONUS) are fine-scale (30 m) elevation data and coarse-scale (~2 km) estimates of the geographic distribution of uranium, thorium, and potassium. A preliminary validation of POLARIS using the NRCS National Soil Information System (NASIS) database shows variable performance over CONUS. In general, the best performance is obtained at grid cells where DSMART-HPC is most able to reduce the chance of misclassification. The important role of environmental covariates in limiting prediction uncertainty suggests including additional covariates is pivotal to improving POLARIS' accuracy. This database has the potential to improve the modeling of biogeochemical, water, and energy cycles in environmental models; enhance availability of data for precision agriculture; and assist hydrologic monitoring and forecasting to ensure food and water security.

1. Introduction

Soil has an essential role in the function and structure of terrestrial ecosystems. It stores and supplies nutrients and water for plants, animals, and other living organisms; it provides a medium for plant growth and anchors human structures; it controls drainage, lateral flow, and storage of water; and it acts as a natural filtration system to regulate water quality (Grayson et al., 1997; Rodriguez-Iturbe and Porporato, 2004; Brady and Weil, 2008; Manzoni and Porporato, 2009; Crow et al., 2012; Lichstein et al., 2014). To inform decision makers and stakeholders for construction projects, highway building, natural resources planning, and crop management, soil surveyors map the complex spatial patterns in soils (Nauman and Thompson, 2014). Over the contiguous United States (CONUS), these surveys are primarily performed by the Natural Resource Conservation Service National Cooperative Soil Survey (NRCS-NCSS) and catalogued in the Soil Survey Geographic (SSURGO) database. This conventional soil map is a national vector and tabular database that provides detailed information on soil taxonomic classes and their related characteristic vertical profiles (Soil Survey Staff, 2014).

SSURGO's primary purpose is local and regional land use planning and it is an excellent resource for such decisions. However, there are challenges that limit SSURGO's use for other purposes including: variable quality and spatial detail between soil surveys, artificial discontinuities at political boundaries, and incomplete spatial coverage (Zhu et al., 2001; Gatzke et al., 2011; Thompson et al., 2012; Subburayalu and Slater, 2013; Du et al., 2014; Nauman and Thompson, 2014). These challenges must be addressed to fully use SSURGO in contemporary applications such as climate and hydrologic models that require high quality

* Corresponding author at: Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ 08544, United States.
E-mail address: nchaney@princeton.edu (N.W. Chaney).

http://dx.doi.org/10.1016/j.geoderma.2016.03.025
0016-7061/

and spatially complete soil information over continental extents (Wood et al., 2011; Bierkens et al., 2014; Chaney et al., 2014; Hengl et al., 2014). One promising path forward is through digital soil mapping (DSM). DSM capitalizes on the relationship between soil spatial patterns and the physical environment (e.g., lithology, relief, and land cover) to re-interpret existing legacy soil data (i.e., conventional soil maps) using existing high-resolution environmental data (McBratney et al., 2003).

DSM provides potential solutions to the existing challenges in SSURGO. First, DSM can be used to harmonize soil surveys and remove artificial discontinuities at political boundaries (Dobos et al., 2010; *Wei* et al., 2010). These algorithms (e.g., ensemble of decision trees) can be used to mine conventional soil maps and existing environmental data to reconstruct the rule sets used by the original surveyors. These reconstructed rule sets from multiple surveys can then be combined to find consensus among the surveys and thus minimize surveyor bias. Second, DSM can be used to gap-fill regions that lack data in existing legacy soil databases (Frazier et al., 2009; Meirik et al., 2010). The relationships developed between the legacy soil data and the environmental covariates where surveys exist are used to provide predictions for the areas that have not been surveyed. Third, DSM can take advantage of the rich metadata of conventional soil maps to spatially disaggregate the polygon map units in legacy soil databases. These algorithms provide higher resolution soil products that are less prone to misinterpretation and more suitable for contemporary applications of soil databases such as environmental models (Bui and Moran, 2001; Hansen et al., 2009; Yang et al., 2011; Thompson et al., 2012; Nauman and Thompson, 2014; Odgers et al., 2014; Subburayalu et al., 2014).

Although DSM can address the primary weaknesses of the SSURGO database, until now, computational constraints have limited the application of these algorithms to regional studies. This is no longer necessary with available high performance computing (HPC) resources. HPC is commonly used in numerous fields including computational fluid dynamics, astrophysics, and numerical weather forecasting by distributing computation across many computational cores to minimize computer wall-clock time (Michalakes and Vachharajani, 2008; Ferziger and Peric, 2012; Balaji, 2013). The conceptual design behind DSM algorithms makes them amenable to implementation on existing HPC to take advantage of thousands to tens of thousands of computation cores (Padarian et al., 2015). HPC makes it possible to run DSM over continental extents at very high spatial resolutions in a matter of hours.

This study capitalizes on a century of legacy soil data and readily available high-resolution environmental covariates to illustrate the potential benefits of using petascale HPC in digital soil mapping. The DSMART-HPC algorithm, an extension of the DSMART algorithm (Odgers et al., 2014), is applied using a moving window (termed a target) to spatially disaggregate, harmonize, and gap-fill the SSURGO database over CONUS. At each target in the moving window, DSMART-HPC uses available environmental covariates, legacy soil data, and a random forest model to reconstruct – and ultimately improve upon – the conceptual rule-based schemes that led to the original surveys. The resulting state-of-the-art product after applying DSMART-HPC over CONUS is the probabilistic remapping of SSURGO (POLARIS) dataset. POLARIS covers CONUS at a 30 m spatial resolution (~9 billion grid cells). At each grid cell, it provides the 50 most probable soil series (termed components) and their associated uncertainties (i.e., probabilities of occurrence).

2. Data

2.1. Legacy soil data: SSURGO

The Soil Survey Soil Geographic (SSURGO) database is a compilation of a century's worth of soil survey in the United States. It covers a large extent of the contiguous United States (CONUS) (Soil Survey Staff, 2014). SSURGO is a polygon format vector map; each polygon is assigned a map unit label. A relational database is used to connect each map unit to information on the observed soil and landscape characteristics of the survey area (e.g., soil texture). Polygons commonly share map unit labels; however, there are never two map units per polygon. Each map unit consists of a set of components (generally up to four) that are commonly shared among map units. These components can be either soil series with corresponding properties (e.g., vertical profile of soil pH) or other characteristic landscape features such as urban areas, water bodies, or rock outcrops. Each map unit summarizes the spatial properties of its components by providing their percentage areal coverage of the map unit and characteristic landscape features.

2.2. Soil covariates

The physical processes that drive the development of complex spatial patterns in soils also contribute to appreciable correlations between soils and environmental covariates. This is the basis of traditional soil-landscape models and a core element of digital soil mapping (Jenny, 1941; Hudson, 1992). McBratney et al. (2003) generalizes these concepts into the scorpan model where it is assumed that a soil class is a function (generally non-linear) of seven factors including soil properties (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and geographic position (n). This section offers an overview of the chosen datasets and environmental covariates used in DSMART-HPC; this information is further summarized in Table 1.

2.2.1. Relief

The USGS National Elevation Dataset (NED) provides elevation data over CONUS at a 30 m spatial resolution (Gesch et al., 2009). From this dataset the following terrain attributes are derived at a 30 m spatial resolution: accumulation area, local slope gradient, topographic wetness index, multi-resolution valley bottom flatness index (MRVBF; Gallant and Dowling, 2003), multi-resolution ridge top flatness index (MRRTF), total curvature, planiform curvature, profile curvature, slope aspect, topographic ruggedness index, and topographic position index (Hengl and Reuter, 2008).

2.2.2. Parent material and age

The U.S. Geological Survey (USGS) gamma aeroradiometric product is used as a proxy for parent material. This data was collected by measuring the gamma-rays emitted from radioactive isotopes in rocks and soils by instruments in low-flying aircraft. This information was then gridded to provide estimates of the mean surface concentration of uranium, thorium, and potassium over CONUS at a 2 km spatial resolution (Duval et al., 2005). Although the spatial resolution is much coarser than

Table 1
Summary of the environmental covariates used in the implementation of DSMART-HPC over CONUS.

Environmental covariate	Dataset	Variable	Spatial resolution
Relief	NED DEM	Elevation Slope Accumulation area Topographic index MRVBF MRRTF Topographic ruggedness index (TRI) Topographic position index (TPI) Curvature Planiform curvature Profile curvature Aspect	30 m
Parent material	USGS aeroradiometric	Uranium Thorium Potassium	2000 m
Organisms	NLCD	Land cover	30 m

the intended goal (i.e., 30 m), the important relationship between parent material and the gamma aeroradiometric variables (e.g., Odgers et al., 2014) makes it ill-advised to disregard these data solely due to spatial discrepancies.

2.2.3. Organisms

The 2006 National Land Cover Database (NLCD) is used to represent the spatial distribution of organisms (vegetation). The NLCD is a 20-class land cover classification spanning CONUS at a 30 m spatial resolution and originates from a decision-tree classification of Landsat data (Fry et al., 2011). The land cover classes are open water, perennial ice/snow, developed (open space), developed (low intensity), developed (medium intensity), developed (high intensity), barren land (rock/sand/clay), deciduous forest, evergreen needleleaf forest, mixed forest, dwarf scrub, shrub/scrub, grasslands/herbaceous, sedge/herbaceous, lichens, moss, pasture/hay, cultivated crops, woody wetlands, and emergent herbaceous wetlands.

2.2.4. Geographic position and climate

The disaggregation algorithm does not explicitly account for geographic position and climate as input variables. However, a moving window approach ensures the implementation over the continent implicitly accounts for them. In this study, the impact of sub-window heterogeneity of these covariates is assumed negligible when compared to other scorpan covariates.

2.2.5. Soil

This study does not use another soil product as a covariate in the scorpan model. The legacy soil database (SSURGO) is used as observations with which to train the DSMART-HPC algorithm.

2.3. NASIS

POLARIS is validated using the NRCS National Soil Information System (NASIS) database. NASIS contains soil observations made over the years by soil surveyors. Each point in NASIS is assigned a soil series—multiple points share soil series; this makes it suitable to validate POLARIS. It should be noted that the majority of NASIS observations are field soil transect observations used for developing SSURGO map unit concepts. As such, NASIS is not a completely independent validation of POLARIS. However, since no other comparable spatial database of soil observations exists for CONUS, NASIS is chosen as the best option for validation. NASIS sites with latitude and longitude values within CONUS, including areas not currently mapped by SSURGO, are used. Validation sites with component names that are not contained anywhere in SSURGO are discarded. The curated NASIS database used in this study consists of 294,746 point observations. Validation is performed by comparing each NASIS' site component name to the predicted 50 most probable components in POLARIS. The best case is when the rank match is one (highest probability) and the worst case is when the rank match is 50 (or missing altogether). This leads to 294,746 rank match values – one per site over CONUS.

3. Methods

3.1. Digital soil mapping: DSMART-HPC

This section provides an overview of the DSMART-HPC algorithm, which parallelizes the DSMART (Disaggregation and Harmonization of Soil Map Units Through Resampled Classification Trees) algorithm (Odgers et al., 2014) to take advantage of HPC resources (Fig. 1). For implementation on available HPC resources, CONUS is divided into 12,474 non-overlapping targets (boxes). The effective size of each target is approximately 30 km by 30 km. A minimum 60 km buffer is added on each side of each target to minimize artificial spatial discontinuities between the predictions of each target. The DSMART algorithm is applied to each target and its corresponding buffer independently. The 12,474 domains are distributed onto 12,474 computation nodes on the Blue Waters supercomputer (Bode et al., 2013). The effective time to run the entire algorithm (wall-clock time) for CONUS is approximately the time it takes to run DSMART on one of the targets.

The DSMART algorithm applied on each domain (target and buffer) follows the method presented in Odgers et al. (2014). The legacy soil data (SSURGO) and environmental covariates (Table 1) are randomly sampled for each domain. The number of sampling points per domain is set to 50,000. The percentage areal coverage of each map unit in the domain is used as a weight to assign the number of samples taken from each map unit. The minimum number of samples required per map unit is 100. This rarely leads to exceeding the allowed maximum number of sampling points (50,000). At each sampling point, the corresponding soil covariates and SSURGO component names are retrieved. Given that there are generally multiple components per map unit, choosing the component for each point can be uncertain. Following Odgers et al. (2014), each sampling point is assigned a component name by weighted random allocation. The weight of each component is the normalized proportion of occurrence of the component in the point's corresponding SSURGO map unit.

Having assembled the training dataset (soil covariates and component names) a random forest is trained using the scikit-learn package (Pedregosa et al., 2011). Random forests are an ensemble method in which a number of decision trees T are built using bootstrap samples of the original training data (Breiman et al., 1984). Fig. 1 shows a schematic of how to obtain the component predictions at a given site using a trained random forest. First, for a given grid cell, the environmental covariates are assembled into a vector $\boldsymbol{v}$. At each decision tree t, it is initially assumed that there is an equal probability of obtaining any component c used in the training dataset. The vector of covariates $\boldsymbol{v}$ is used to condition the probabilities until reaching the corresponding leaf on each decision tree. The final result per decision tree is a vector of conditional probabilities — one value per component. These probabilities are averaged to obtain the averaged probability per component $p_{\boldsymbol{T}} = \frac{1}{T}\sum_{t}^{T} p_t(c|\boldsymbol{v})$ across all decision trees. The vector of conditional probabilities is used to rank each component to provide the prediction and their corresponding probabilities (uncertainties).

After running DSMART on a given domain, maps of the 50 most probable components—these components can differ per grid cell—and their associated probabilities are created for the corresponding target area at a 30 m spatial resolution. Finally, maps of each target are assembled to create the final data product.

3.2. Feature importance

To compute the importance of each soil covariate, the relative feature (i.e., soil covariate) importance metric as defined in the scikit-learn package is used (Pedregosa et al., 2011). This metric is defined as the normalized average over all decision trees of the sample-weighted change in Gini impurity at each node that belongs to feature (i.e., covariate) i. Given a histogram per node, the Gini impurity informs how often a randomly chosen item (i.e., component) at a node will be misclassified. It is defined as the sum probability of obtaining each distinct item in the histogram times the probability of not obtaining the given item ($Gini = 1 - \sum_{i}^{n} p_i^2$).

3.3. Prediction uncertainty

DSMART-HPC estimates the probability of obtaining a given component at each 30 m grid cell over CONUS. At each grid cell, these probabilities are ranked to provide component predictions. However, if the

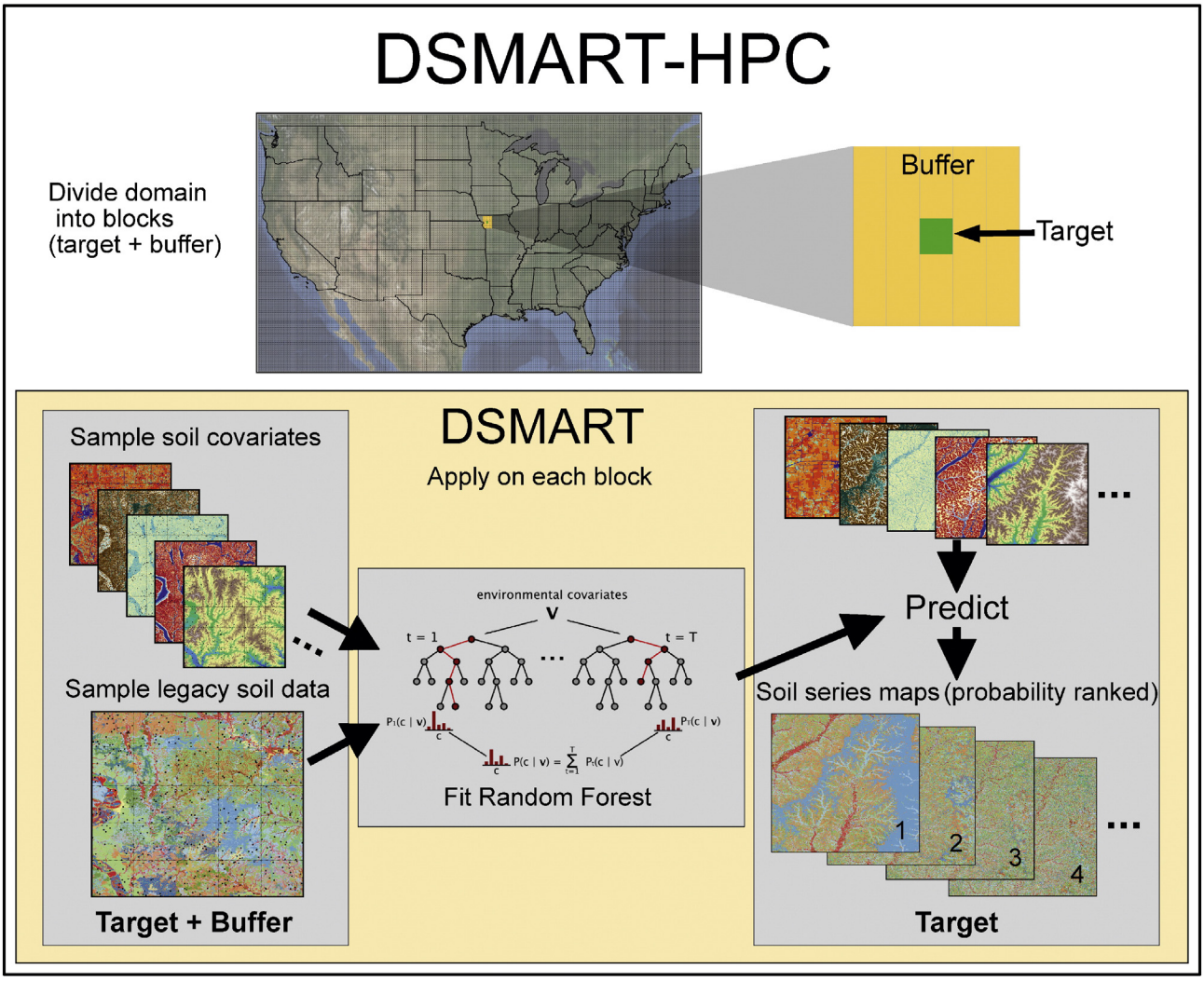

Fig. 1. Schematic of the DSMART-HPC algorithm. CONUS is split into a grid of targets which are distributed among different computational nodes. The DSMART algorithm is run on each target and a surrounding buffer (target + buffer = domain) to avoid artificial discontinuities. The steps in DSMART are: 1) sample the environmental covariates and the legacy soil data over the domain; 2) train a random forest; 3) estimate the components and their associated probabilities.

difference in probabilities between the top predictions is small, there is a considerable chance of misclassification. To assess the prediction uncertainty at each grid cell, the confusion index (Burrough et al., 1997) is calculated from the first and second most probable components ($CI = 100 - (P_1 - P_2)$). The CI values range from 0 to 100. The Gini impurity—as defined in Section 3.2—is also used to provide a measure of the chance of misclassification at a given grid cell; it accounts for the probabilities of the top 50 predictions per grid cell. The Gini impurity values range from 0 to 1. Higher confusion index and Gini impurity values indicate a higher chance of misclassification.

4. Results

4.1. Comparison between SSURGO and POLARIS

Visual comparisons between SSURGO and POLARIS are shown over CONUS in Fig. 2 and over 6 distinct regions of comparable size (~216 km by ~216 km) in Figs. 3 and 4. These regions are shown as they are able to illustrate the strengths and weaknesses of DSMART-HPC in gap-filling, harmonizing, and spatially disaggregating SSURGO.

Northern Mississippi (Fig. 3 — Top) — The artificial county boundaries in SSURGO all but disappear in POLARIS. Over the Grenada and Calhoun counties (central region), the legacy soil data is replaced by the soil information in the surrounding counties. The limited number of NASIS sites in these two counties limits the validation of the harmonization results. POLARIS' ability to conserve the boundary between the Mississippi river floodplain and the rest of the area is an example where the trained random forest can reproduce large-scale geological and topographic features. The confusion index over the entire region indicates that the best performance is obtained when predicting water bodies and soils on ridges. There is no apparent spatial organization to the rank match values except for a tendency for there to be lower rank match values in areas of low confusion index values.

Northern Nevada (Fig. 3 — Middle) — DSMART-HPC aims to spatially disaggregate SSURGO's coarse polygons in this region by distinguishing the different components in each map unit. Visual inspection of the resulting POLARIS map shows an increase in soil spatial heterogeneity when compared to SSURGO. However, the map of rank match values suggests that the disaggregated product, in general, does not agree with the NASIS observations, limiting the value of the spatial disaggregation. There are also entire areas that remain largely untouched after disaggregation. Further inspection shows that the components for the areas that remain largely untouched in POLARIS are defined as Water and Playa in SSURGO. Visual analysis of the NLCD database illustrates that these areas are clearly defined water bodies and barren land respectively. DSMART-HPC is able to capture this strong relationship

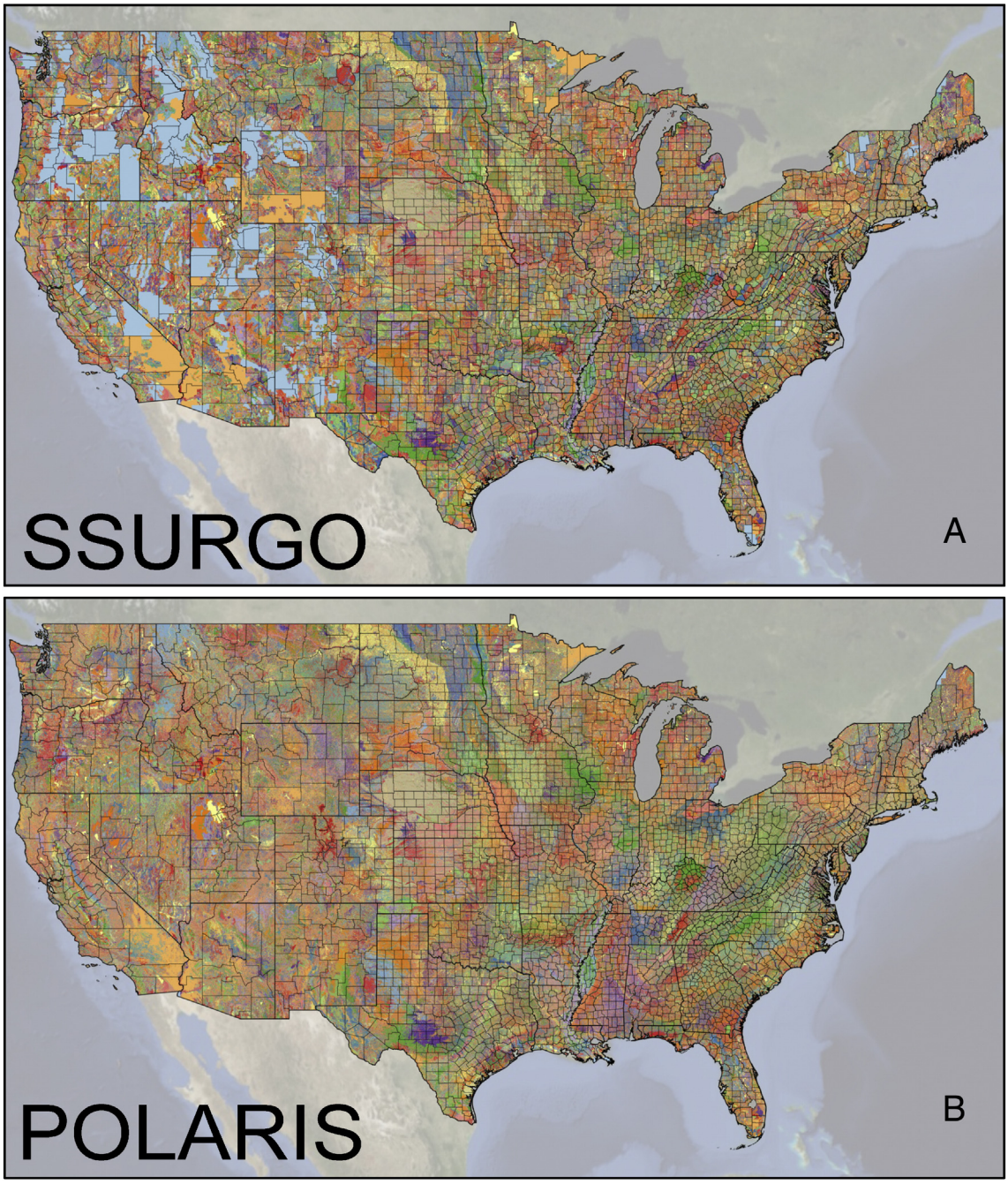

Fig. 2. Comparison of SSURGO's most dominant component (A) to POLARIS' most probable component (B). For visualization over CONUS, the 30 m POLARIS data product is upscaled to a 1 km spatial resolution using nearest neighbor interpolation. Light blue areas indicate regions where SSURGO lacks data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between NLCD and SSURGO over these areas and thus provide reliable predictions. This explains the lack of spatial heterogeneity over these areas in POLARIS.

Northern New Hampshire (Fig. 3 — Bottom) — SSURGO does not have estimates over parts of northern New Hampshire and western Maine. More specifically, it lacks data over the White Mountains. DSMART-HPC is able to gap-fill this region using information from adjacent areas. It fills in the missing regions using component information from smaller areas towards the north and northeast. Further inspection shows that the components that fill in most of the missing area are defined as Rock Outcrop and Saddleback. This result is encouraging since one would expect rock outcrops in the missing region; furthermore, the Saddleback series is also physically consistent as these soils are commonly found in mountainous regions in Maine, New Hampshire, and New York. The validation results over the unmapped areas in SSURGO are inconclusive since the NASIS observations are clumped together and do not uniformly sample the unmapped areas. For the entire region, one potential concern is the loss of certain components in favor of more predominant ones; DSMART-HPC spatially disaggregates the most frequent components while disregarding the less frequent ones altogether.

Western Washington (Fig. 4 — Top) — SSURGO does not have estimates for parts of the Skagit, Snohomish, King, and Pierce counties. To gap-fill the missing areas, DSMART-HPC appears to use the mountainous region to the east. Over the missing areas, the most common prediction is rock outcrop. These areas are also gap-filled with soils from the

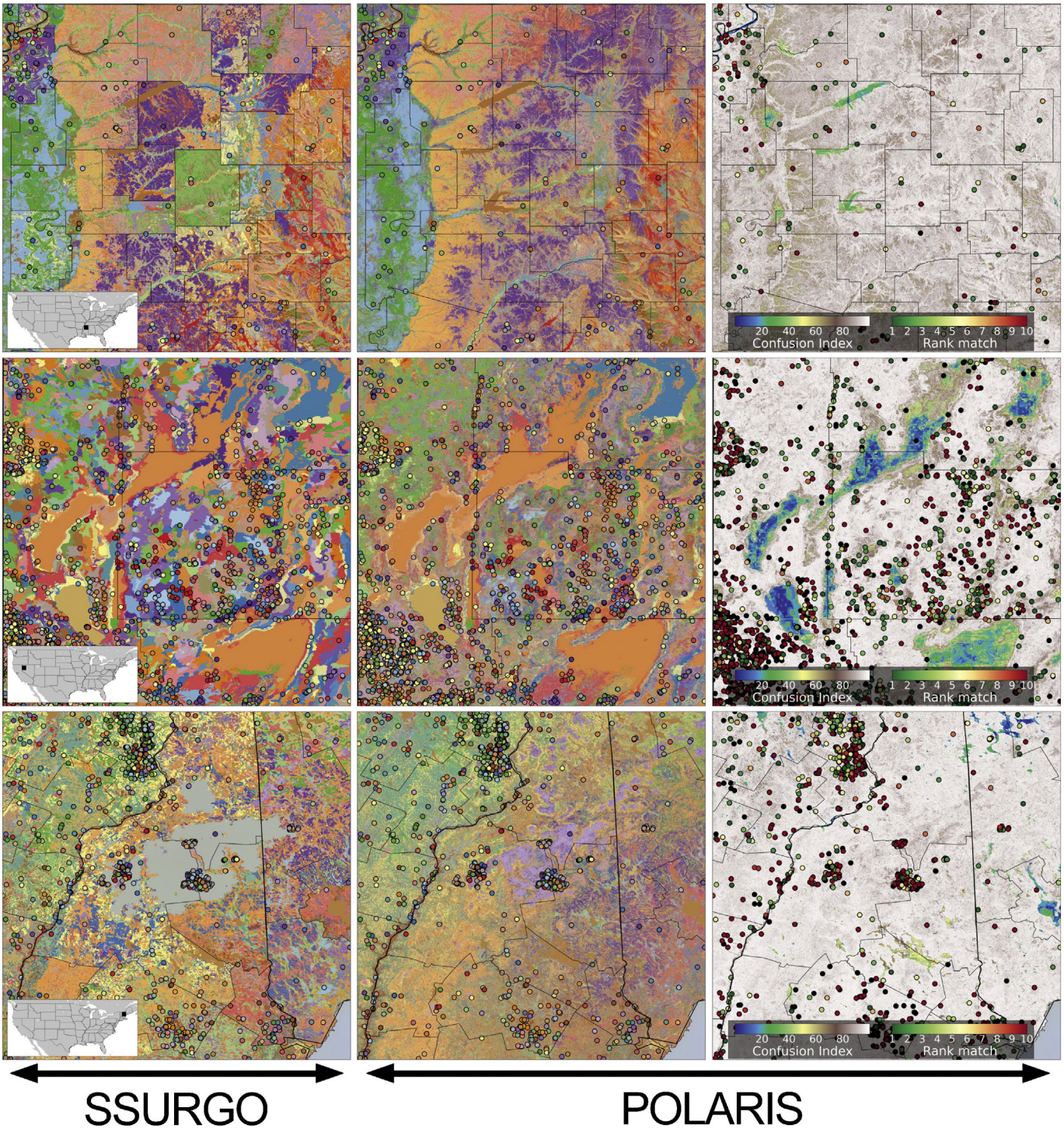

Fig. 3. Comparison between the most probable component of SSURGO (left) and POLARIS (middle) over northern Mississippi (top), northern Nevada (middle), and northern New Hampshire (bottom). Each color represents a different component. The confusion index (right) illustrates prediction uncertainty. Point symbols represent NASIS validation sites. Rank color indicates the rank at which POLARIS components match the corresponding NASIS site component. A site that does not have a POLARIS match is set to black. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Nimue and Playco soil series, both characteristic soils of the Cascades. This result suggests that although the position of the soils might be locally incorrect, the gap-filling appears to place components in appropriate areas. The loss of spatial complexity in the eastern section after applying DSMART-HPC is due to the predominant components being chosen over the minor components. This can be explained by the legacy soil data sampling scheme of the algorithm (see Section 3.1) and a lack of appropriate environmental covariates to properly differentiate the location of the minor components from the major components.

Central California (Fig. 4 — Middle) — DSMART-HPC uses the soil covariates that represent topography to transfer the available legacy soil surveys to the missing area (parts of Calaveras and Tuolumne counties) (see Section 4.5). The algorithm's ability to maintain the spatial structure of the western foothills is encouraging; however, validation results show

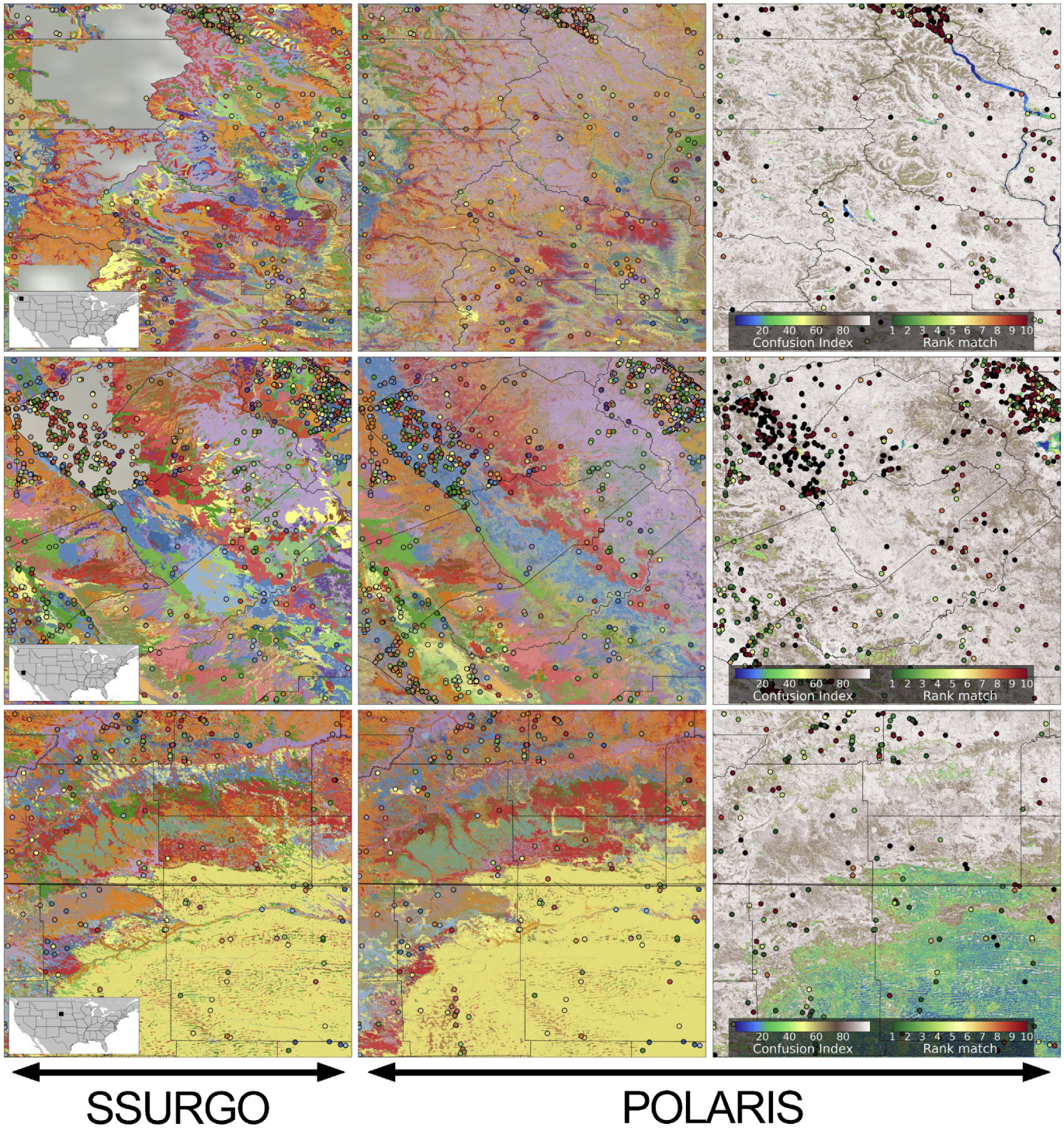

Fig. 4. Comparison between the most probable component of SSURGO (left) and POLARIS (middle) over western Washington (top), central California (middle), and northwestern Nebraska and southeastern South Dakota (bottom). Each color represents a different component. The confusion index (right) illustrates prediction uncertainty. Point symbols represent NASIS validation sites. Rank color indicates the rank at which POLARIS components match the corresponding NASIS site component. A site that does not have a POLARIS match is set to black. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that the predicted soil series in these gaps tend to not match NASIS. Instead of connecting similar soils in the north and south, the observations illustrate that the components characteristic in the Stanislaus National Forest extend further west into the foothills. The predicted dominant components only become a reasonable estimate towards the west.

Northwestern Nebraska and Southwestern South Dakota (Fig. 4 — Bottom) — In this region, DSMART-HPC does not add much information but instead mainly reproduces SSURGO. The close agreement between the two can be attributed to the gamma aeroradiometric variables that accurately depict the Sandhills region in northwestern Nebraska and southwestern South Dakota. The low confusion index values over the Sandhills help explain the strong agreement between SSURGO and POLARIS over this region. This example also shows how deficiencies in the covariates can lead to unrealistic artifacts in the POLARIS product—missing data in the gamma aeroradiometric variables cause the rectangular pattern in the northeast quadrant.

4.2. Prediction uncertainty

The quality of POLARIS' predictions is quantified via probabilities. For each target in the moving window, DSMART-HPC initially assumes that all components are equally probable; the random forest uses the environmental covariates to condition these probabilities. The predicted components in each grid cell are ranked according to these probabilities (rank one is assigned to the component with the highest probability). Probability maps for the first- (P_1) and second-most probable (P_2) components per grid cell are shown in Fig. 5. The probabilities are upscaled by box averaging to a 30 arcsec (~1 km) spatial resolution to allow for visualization over CONUS. Both the confusion index and the Gini impurity are then calculated to assess prediction uncertainty of the components in POLARIS for each grid cell.

Over the majority of CONUS, the probabilities of the rank one (most probable) component are between 10% and 40%. In regions where the probabilities do not exceed 20%, the differences between P_1 and P_2 can be very small. This leads to high confusion index values and low confidence in the predictions in many regions. The results are similar when accounting for the top 50 predictions per grid cell using the Gini impurity. Over CONUS, the highest confidence predictions occur over wide rivers, lakes, deserts, salt pans, and areas with regionally dominant components. Deterministic predictions over ridges are, in general, more reliable than those in valleys and riparian zones.

The high confusion index and Gini impurity values suggest that the chosen environmental covariates are not sufficient to make precise component predictions—many components share the same environmental covariate values. Another possible explanation for such low probabilities could be due to the inability of the decision trees in each local random forest to fully grow due to memory constraints on each computational node. This will limit the algorithm's ability to condition the probabilities and should be addressed in future updates of the POLARIS dataset. Future work should also analyze the taxonomic distance between the most probable components per grid cell. It is possible that components have different names but are taxonomically similar and occur in similar environmental conditions. This would help explain the high confusion index and Gini impurity values and provide guidance towards the need to account for taxonomic similarity in the predictions.

4.3. Preliminary validation

POLARIS is validated using 294,746 sites from the NRCS National Soil Information System (NASIS) database (see Section 2.3 for further details). As shown in Fig. 6, for many of the NASIS sites, a match is not found until after the tenth highest ranked soil class. Model performance is highly variable across CONUS with a limited set of spatial clusters in performance. The best performance appears to be achieved in Nebraska, Kansas, and northern Ohio, while POLARIS struggles over the Appalachians, the Mountain West, and most of the northern Midwest.

Empirical cumulative distribution functions (ecdf) of the rank matches for all sites in a given region are summarized in Fig. 7. The lower and upper bounds indicate the change in ecdfs when searching for the rank match over a three by three window surrounding each validation point and indicate the impact of fine-scale spatial noise. For all sites where SSURGO also has data, approximately 17% of the validation sites match at rank one, 55% have a match when including the first ten ranks, and around 68% have a match when including the first 50 ranks. When including all the POLARIS predictions in the surrounding three by three window, validation results can vary dramatically suggesting a high degree of spatial noise in the predictions. A similar comparison using SSURGO shows that 48% of the validation sites match at rank one and 61% match when including all components in a map unit. Although SSURGO does generally outperform POLARIS deterministically, there are many NASIS sites at which POLARIS does outperform SSURGO.

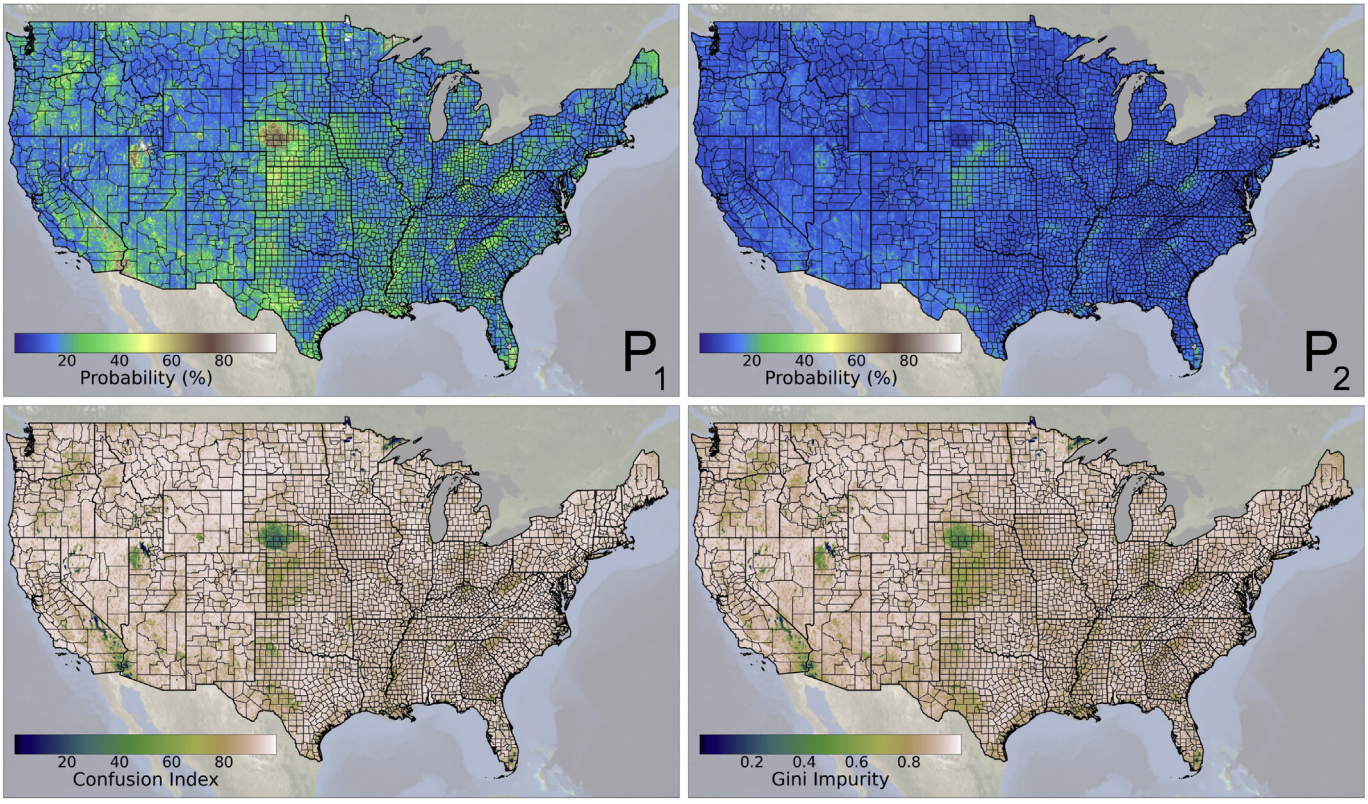

Fig. 5. Comparison between the probability for the first (P_1) and second (P_2) most probable components in POLARIS. The confusion index (bottom left) formalizes this comparison by assessing how close the probabilities P_1 and P_2 are at each grid cell. The Gini impurity index (bottom right) uses the probabilities of the 50 component predictions at each grid cell to provide a measure of the chance of misclassification. Results have been upscaled to 30 arcsec (~1 km) for display purposes.

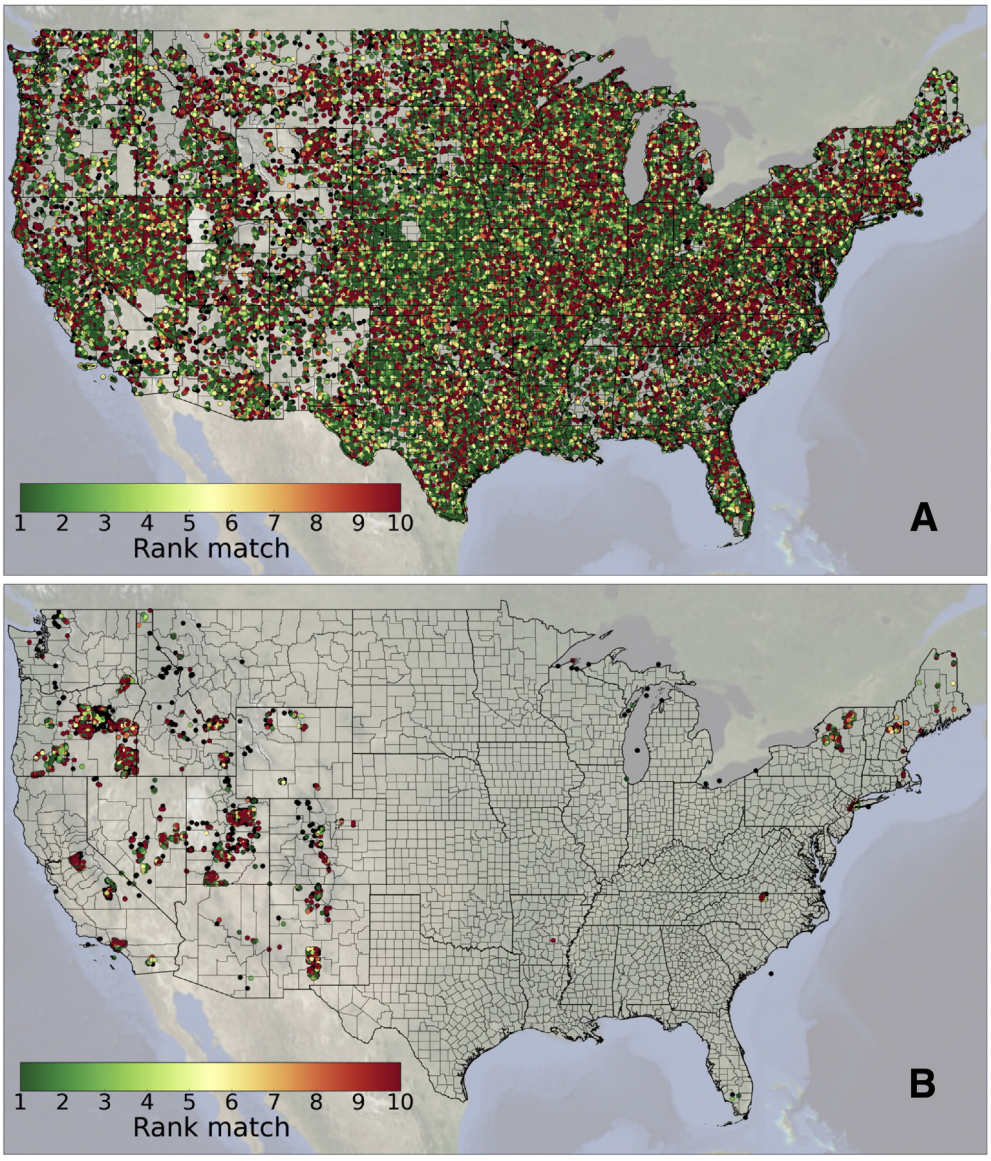

Fig. 6. Validation of the POLARIS database using the NASIS point database. The points shown are the ranks at which POLARIS predictions match the corresponding NASIS site component name. A site that does not have a match is set to black. The top panel (A) shows comparisons at sites where SSURGO has data while the bottom panel (B) shows comparisons at sites where SSURGO does not have data.

The results in Fig. 7 provide insight into POLARIS' strengths and weaknesses; although it provides less reliable deterministic predictions (most probable component) than SSURGO, when considering all ranks, it matches the validation observations more frequently. It is important to note that having a match for 78% of the NASIS sites at rank 50 does not mean that the predictions are random. Each local random forest is built on hundreds to thousands of components; being able to restrict this uncertainty down to only 50 components shows DSMART-HPC's ability to constrain the original uniform distribution.

Overall, SSURGO outperforms POLARIS over all regions when the goal is to obtain the most probable component. However, given the probable short taxonomic distance between many soil series—different soil series may occur in similar environmental conditions due to taxonomic similarities—the results are not conclusive. More exhaustive validation approaches should consider the taxonomic distance between soil series (Minasny and McBratney, 2007).

For the NASIS sites where SSURGO does not have data, approximately 5% of the validation sites match at rank one, 22% have a match when including the first ten ranks, and around 40% have a match when including all components in a map unit. These results are highly variable per region with the best performance in the Northeast and the worst performance in the Pacific West. The performance in the Northeast can be explained by the relatively small areal coverage of the unsurveyed areas in SSURGO (see Fig. 2). DSMART-HPC gap-fills these areas with survey data that is close in distance and has a similar physical environment and thus more likely to be appropriate for the missing regions. Over the Pacific West, Southwest, and Midwest, the unsurveyed areas have a larger areal coverage and can have large differences in their physical environment when compared to their adjacent surveyed areas. This explains the poor performance when gap-filling these areas.

4.4. Relationship between the prediction uncertainty and validation results

Visual inspection of Fig. 5 and Fig. 6 show a distinct relationship between the prediction uncertainty metrics (confusion index and Gini impurity) and the ranks at which the POLARIS predictions match the NASIS in-situ observations. Fig. 8 summarizes this comparison across the 6 regions used in Section 4.3. The Gini impurity values at the co-located grid cells of the NASIS validation sites are compared to the rank match

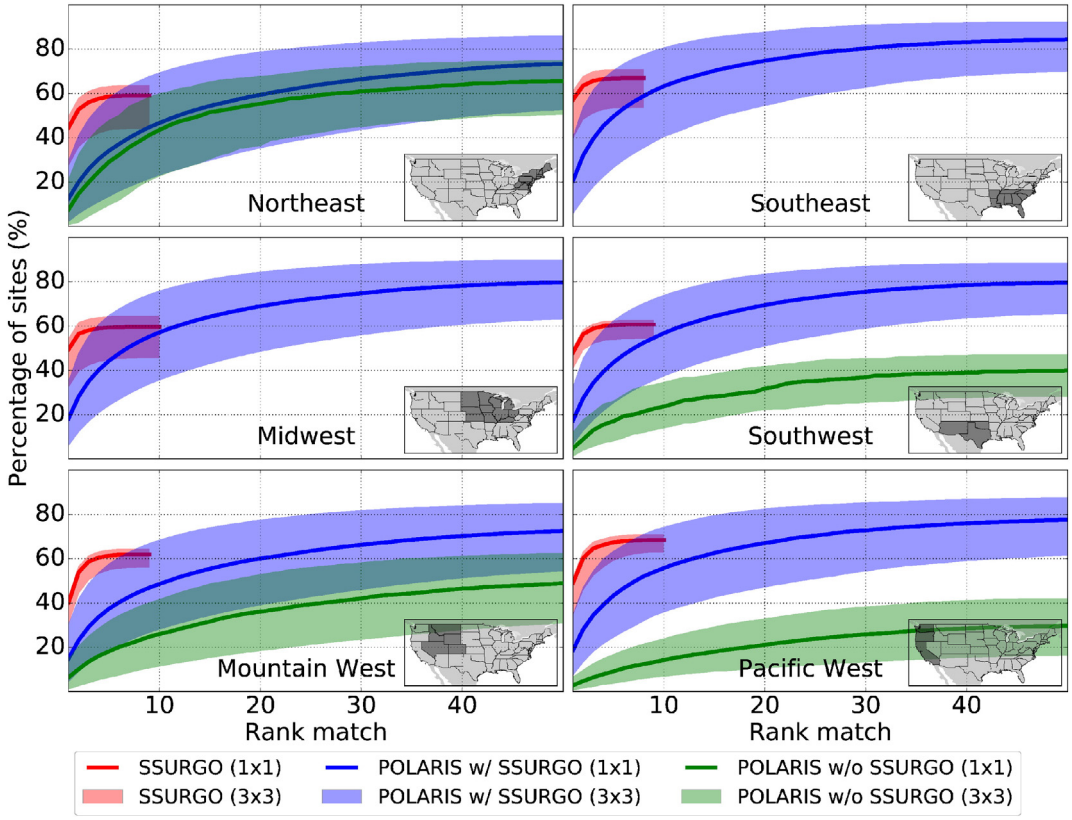

Fig. 7. Summary of the POLARIS validation using the NASIS point database. Rank matches of all validation sites in a given region are used to create an empirical cumulative distribution function. The results are shown for SSURGO and POLARIS where SSURGO has an estimate, and for POLARIS at sites where SSURGO does not have an estimate. The lower and upper bounds show the worst and best scenarios when evaluating the 3 × 3 grid cell window surrounding each NASIS site.

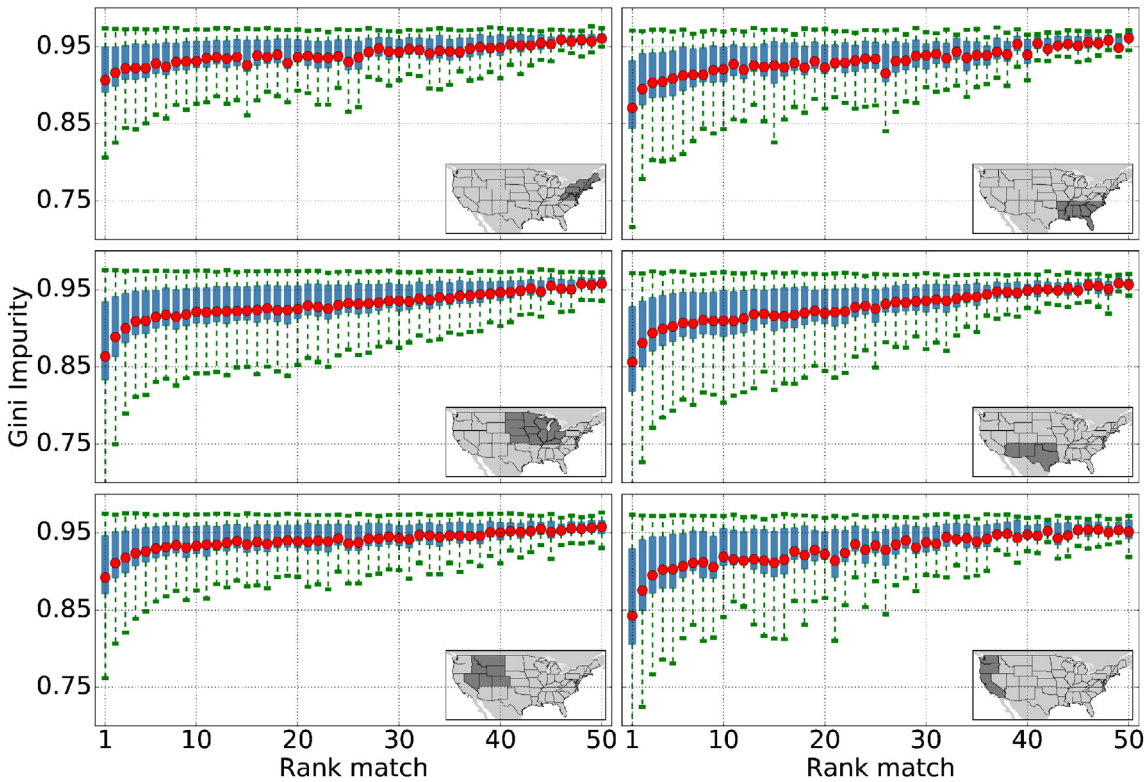

Fig. 8. Relationship between the rank matches of all validation sites in a given region and POLARIS' prediction uncertainty (Gini impurity). The Gini impurity values for all sites that belong to a given region are binned according to their rank match. Boxplots are used to display the distribution of Gini impurity values at each rank match. The mean Gini impurity value per rank match is shown as a red dot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

values; the Gini impurity metric is chosen since it provides a measure of the likelihood of misclassification when drawing from the 50 predicted series at each grid cell (see Section 4.2 for more details). Gini impurity values are binned according to the grid cell's rank match and are shown via boxplots. The results are conclusive; on average, as the rank match increases the prediction uncertainty increases. In other words, the less confidence DSMART-HPC has in a prediction, the more likely the prediction is erroneous. These results agree with the validations results in Section 4.3. The regions that have the highest Gini impurity values, in general, have the worst performance when compared to NASIS and vice-versa. These results provide strong evidence that if the probabilities of the predictions can be further constrained DSMART-HPC will provide more reliable deterministic predictions. This will most likely be accomplished through the inclusion of additional environmental covariates that more fully explain and represent the observed soil spatial patterns.

4.5. Covariate importance

Beyond component predictions, each target's random forest also estimates the role of the individual environmental covariates in the POLARIS prediction. High relative covariate importance values indicate that a given covariate plays a large role in constraining the initial uniform distribution in a random forest. The feature importance values are spatially interpolated to create spatial maps (Fig. 9) and summarized over CONUS (Fig. 10).

The covariate importance results are conclusive; elevation is the single most important covariate in POLARIS. It has a prominent role over the Mountain West, Southwest, and Pacific West and other regions with appreciable topographic relief. Elevation can also play a pivotal role even when the topographic relief is not as pronounced (e.g., Mississippi flood plain). Even though the gamma aeroradiometric variables (potassium, thorium, and uranium) have a relatively coarse spatial resolution (~2 km), their ability to represent the parent material makes them key covariates. The role of the potassium variable is especially notable over the Sandhills region in Nebraska. This most likely explains the low uncertainty in the predictions over this region (see Fig. 4). The MRVBF and MRRTF covariates play important roles in areas of low topographic relief. Land cover also has a defining role, although its impact is spatially variable. It has a high impact over

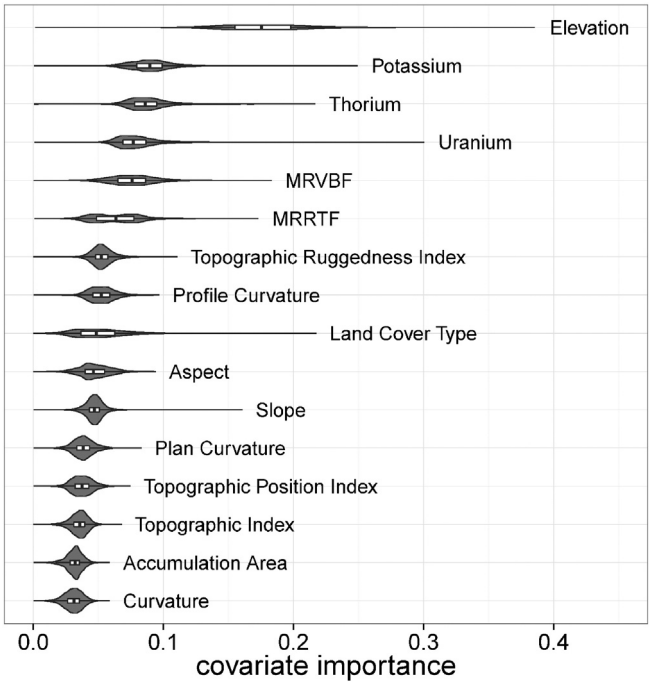

Fig. 10. Violin plots of the relative importance of each environmental covariate used in DSMART-HPC. The results summarize the maps in Fig. 9.

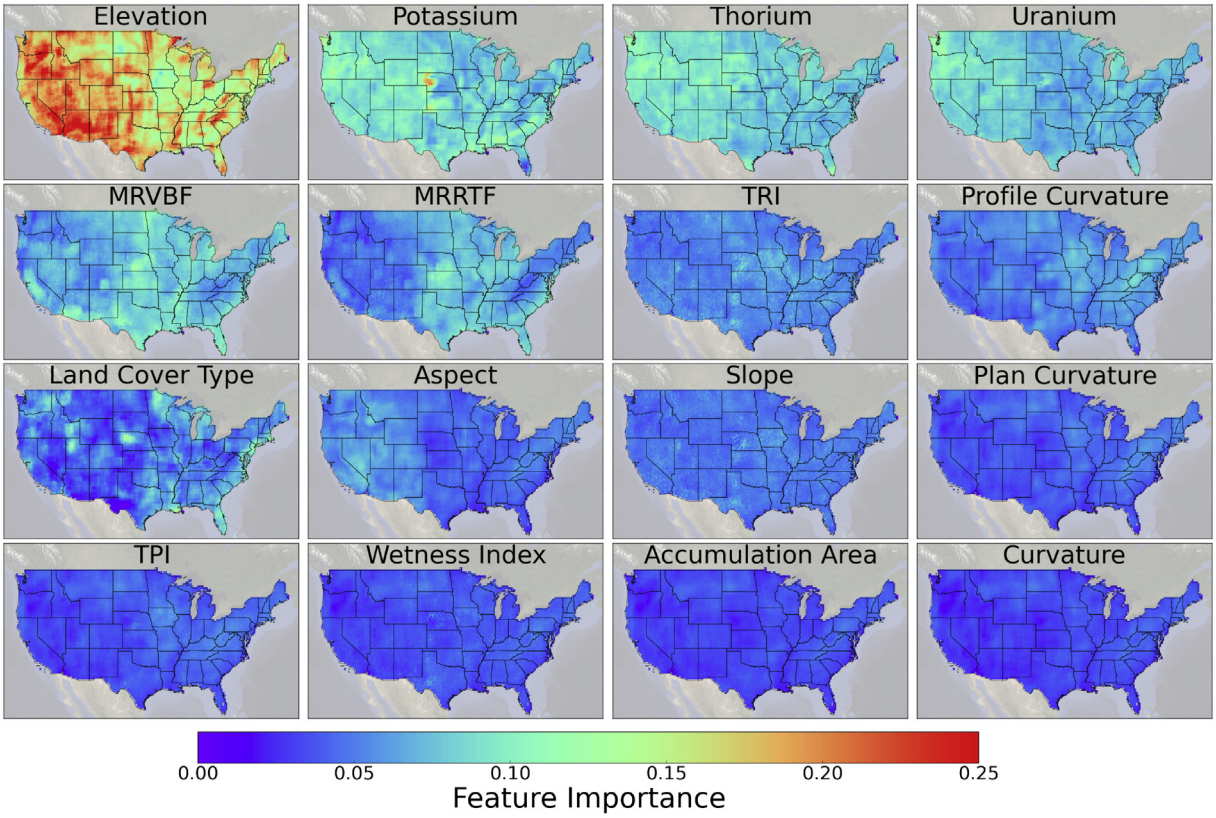

Fig. 9. Relative feature importance of each environmental covariate used in DSMART-HPC.

urban areas (e.g., San Francisco Bay Area and Northeast corridor), wetlands (e.g., Florida), rivers, lakes (e.g., Minnesota), and salt pans (e.g., Bonneville Salt Plains). This result suggests that if the miscellaneous components in SSURGO (e.g., urban areas and water bodies) are discarded, the relative importance of land cover will most likely become negligible over CONUS. Most of the remaining variables play a relatively minimal role in the overall picture.

5. Discussion

5.1. High performance computing in digital soil mapping

High performance computing is commonly used across multiple scientific fields (e.g., climate modeling). In the earth sciences, HPC is becoming essential to harness the data from the increasing number of satellite remote sensing missions that observe the earth system at high temporal and spatial resolutions (Plaza and Chang, 2007; Lee et al., 2011). The goal to produce state-of-the-art continental soil maps using these data necessitates integrating HPC in DSM—our study over CONUS demonstrates that this is feasible. The total computation required for the development of POLARIS was around 450,000 core-hours with a wall-clock time of 5 h on the Blue Waters supercomputer (Bode et al., 2013). This is negligible computer time at current HPC facilities that can handle 10 million (~1100 years) core-hour tasks. For example, existing HPC computing resources and available environmental covariates allow for the implementation of DSMART-HPC over the globe land surface at a 30 m spatial resolution or at a 10 m spatial resolution over CONUS. The DSM community should embrace these powerful resources. Such adoption will accelerate the DSM model development cycle and help meet the growing data needs of the scientific community, farmers, land managers, and policy makers for spatially complete, harmonized soil information.

5.2. Addressing the challenges in SSURGO

POLARIS provides potential solutions to the primary challenges in SSURGO. First, the artificial discontinuities due to county and state boundaries all but disappear. DSMART-HPC uses the environmental covariates and the legacy soil data to determine the most probable components in a given region while disregarding the political boundaries—an implicit covariate used to assemble the original soil surveys. Second, DSMART-HPC gap-fills the missing areas with information from adjacent areas. Having learned the relationships between the legacy soil data and the environmental covariates, the algorithm is generally able to predict physically consistent components in the missing regions. For example, missing areas in the White Mountains in New Hampshire are gap-filled with rock outcrops and the Saddleback soil series; both these components are commonly found in mountainous regions in Maine, New York, and New Hampshire. Third, DSMART-HPC spatially disaggregates the polygons by using the environmental data and different polygons to distinguish the different components in each map unit.

5.3. Future validation efforts

Although DSMART-HPC addresses the primary challenges in SSURGO, the preliminary validation results in Section 4.3 suggest that further work is necessary before POLARIS can be used as a deterministic soil map. As shown in Section 4.4, to improve model performance, constraining the probabilities of the predictions should be a priority in the development of future versions of POLARIS. However, it currently remains unclear what steps are necessary to accomplish this goal. This section discusses future validation efforts that will further elucidate the strengths and weaknesses of POLARIS and thus provide insight into necessary model improvements.

5.3.1. Comparison to previous DSM studies over CONUS

Over the past 15 years there have been multiple regional studies over CONUS that have sought to address the challenges in SSURGO using DSM (Zhu et al., 2001; Wei et al., 2010; Subburayalu and Slater, 2013; Nauman and Thompson, 2014; Nauman et al., 2014; Subburayalu et al., 2014). Although these studies cover relatively small spatial extents, they are a valuable resource to evaluate how POLARIS compares to other digital soil series products derived from SSURGO. A preliminary comparison suggests that these regional data products outperform POLARIS. These differences are most likely explained by the algorithm, the number of soil series predicted, and the environmental covariates used. Future work should more formally compare POLARIS with these data products to understand the differences and assess how DSMART-HPC can be improved and what environmental covariates should be added to improve future versions of POLARIS.

5.3.2. Taxonomic distance

Another challenge in the prediction of soil series using DSMART-HPC and the validation of POLARIS is accounting for taxonomic distance. Even though there are tens of thousands of distinct soil series within SSURGO, there is most likely a short taxonomic distance between many of them. As a result, these soil series may occur in similar environmental conditions, making it a challenge to adequately constrain the prediction probabilities using DSMART-HPC; this would help explain the high confusion index and Gini impurity values in POLARIS (see Section 4.2 for more details). Taxonomic distance is also a challenge when validating POLARIS since relying on a simple soil series name match (Section 4.3) will not account for the similarities between the soil series and thus not be a suitable comparison when the end goal is to produce soil property maps. Future validation efforts of POLARIS and future improvements of the DSMART-HPC algorithm should account for the taxonomic distance between soil series.

5.4. Additional environmental covariates

Since reducing prediction uncertainty has shown to be strongly related to improving the database's accuracy (see section 4.4), additional environmental covariates should be used in future versions of POLARIS. First, given the important role of the gamma aeroradiometric product, adding other parent material information (e.g., lithology), as an input environmental covariate, should be a priority. Second, the role of climate in soil spatial properties is mainly disregarded in this study; this is not necessary given the large availability of high-resolution meteorological datasets over CONUS. The next iteration should include climate datasets such as the National Land Data Assimilation System (NLDAS; Mitchell et al., 2004) and the Parameter-elevation Relationships on Independent Slopes Model (PRISM; Daly et al., 2008). Third, instead of only using the classic terrain attributes it would be helpful to use DEM derived variables that approximate the landform elements observed by the surveyors. Finally, databases that divide CONUS into areas that share similar soils, climate, and land use activities (e.g., Major Land Resource Areas) could be used as environmental covariates to help define more concrete boundaries to avoid placing soils in physically unrealistic regions. This division into natural landscape units could also be used to replace the square target areas used in DSMART-HPC.

5.5. Improving spatial disaggregation

Another driver of prediction uncertainty is the weighted random allocation scheme in DSMART-HPC (see section 3.1) used to spatially disaggregate the map units in SSURGO. Moving forward, the wealth of information in the SSURGO database and original soil survey manuscripts represent options to better train the target random forests. Most components in each map unit have various descriptions of the environmental context in which the surveyor observed the component; including slope shape and hillslope position, among others (Thompson et al.,

2010; Nauman and Thompson, 2014). Assuming that the raster rules can be developed to relate these descriptors to the available data over CONUS, this technique could replace the weighted random allocation scheme to spatially disaggregate the information prior to training the random forests. It is currently uncertain if this will significantly alter the predictions, however, it should help constrain the probabilities by reducing the randomness introduced by the current component assignment scheme.

6. Conclusion

This work is a breakthrough in digital soil mapping (DSM); it uses a state-of-the-art DSM model to reinterpret a complex legacy soil database over CONUS at a 30 m spatial resolution. The results demonstrate the potential of integrating petascale HPC in DSM. Nonetheless, POLARIS should not be seen as a replacement to SSURGO; indeed the current version seldom outperforms SSURGO when validated with the NASIS dataset. Its primary objective is to provide the scientific community with a probabilistic field-scale soil dataset over CONUS that is both harmonized and spatially complete. POLARIS provides a spatially continuous, internally consistent, quantitative prediction of soil series. This database has the potential to improve the modeling of biogeochemical, water, and energy cycles over land in numerical weather forecasting and climate models; assist drought and flood monitoring and forecasting to ensure food and water security; and enhance availability of data for precision agriculture. It is also meant as an exploratory dataset that when analyzed will provide insight into environments in which the drivers of spatial heterogeneity of soils are not well represented (e.g., riparian zones) and help guide future environmental covariate selection.

Data accessibility

The POLARIS database can be accessed at http://stream.princeton.edu/POLARIS

Acknowledgements

This study was supported by NSF grant 1144217 (Petascale Design and Management of Satellite Assets to Advance Space Based Earth Science). This work would not have been possible without the provision of data by the National Cooperative Soil Survey (NCSS) and the Blue Waters supercomputer. Alex McBratney acknowledges the support of the Australian Research Council through its Discovery Program. A special thanks to the many contributors that have helped understand the strengths and weaknesses of the POLARIS database and provide insight to future research including Skye Wills (USDA-NRCS), Tom Hengl (ISRIC), Budiman Minasny (University of Sydney), Dylan Beaudette (USDA-NRCS), James Thompson (West Virginia University), Stephen Roecker (USDA-NRCS), Sharon Waltman (USDA-NRCS), Tom D'Avello (USDA-NRCS), and David Hoover (USDA-NRCS), among many others. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Balaji, V., 2013. Scientific computing in the age of complexity. XRDS 19 (3).
Bierkens, M.F.P., et al., 2014. Hyper-resolution global hydrological modeling: what's next. Hydrol. Process. 29 (2), 310–320.
Bode, B., Butler, M., Dunning, T., Gropp, W., Hoefler, T., Hwu, W., Kramer, W., 2013. In: Vetter, J. (Ed.), The Blue Waters Super-System for Super-Science, in Contemporary HPC Architectures. Chapman and Hall/CRC.
Brady, N.C., Weil, R.R., 2008. The Nature and Properties of Soils. Fourteenth. Prentice Hall/Pearson Education.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.
Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103 (1–2), 79–94.
Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77 (2–4), 115–135.
Chaney, N.W., Roundy, J.K., Herrera Estrada, J.E., Wood, E.F., 2014. High-resolution modeling of the spatial heterogeneity of soil moisture: applications in network design. Water Resour. Res. 51 (1), 619–638. http://dx.doi.org/10.1002/2013WR014964.
Crow, W.T., Berg, A.A., Cosh, M.H., Loew, A., Mohanty, B.P., Panciera, R., de Rosnay, P., Ryu, D., Walker, J.P., 2012. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. Rev. Geophys. 50 (RG2002).
Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28 (16). http://dx.doi.org/10.1002/joc.1688.
Dobos, R., Bialko, T., Micheli, E., Kobza, J., 2010. In: Boettinger, J.L. (Ed.), Legacy Soil Data Harmonization and Database Development, in Digital Soil Mapping, Progress in Soil Science 2. Springer.
Du, F., Zhu, A.X., Band, L., Liu, J., 2014. Soil property variation mapping through data mining of soil category maps. Hydrol. Process. 29, 2491–2503.
Duval, J.S., Carson, J.M., Holman, P.B., Darnley, A.G., 2005. Terrestrial Radioactivity and Gamma-Ray Exposure in the United States and Canada: U.S. Geological Survey Open-File Report 2005-1413. Available online only http://pubs.usgs.gov/of/2005/1413/.
Ferziger, J.H., Peric, M., 2012. Computational Methods for Fluid Dynamics. Springer, Berlin.
Frazier, B.E., Rodgers, T.M., Briggs, C.A., Rupp, R.A., 2009. Remote area soil proxy modeling technique. Soil Surv. Horizons 50, 62–67.
Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., Wickham, J., 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States. PE&RS 77 (9), 858–864.
Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12), 1347–1360.
Gatzke, S.E., Beaudette, D.E., Ficklin, D.L., Luo, Y., O'Geen, A.T., Zhang, M., 2011. Aggregation strategies for SSURGO data: effects on SWAT soil input and hydrologic outputs. Soil Water Manag. Conserv. 75 (5).
Gesch, D., Evans, G., Mauck, J., Hutchinson, J., Carswell Jr., W.J., 2009. The National Map-Elevation: U.S. Geological Survey fact sheet, (2009-3053).
Grayson, R.B., Western, A.W., Chiew Francis, H.S., Blöschl, G., 1997. Prefered states in spatial soil moisture patterns: local and nonlocal controls. Water Resour. Res. 33 (12), 2897–2908.
Hansen, M.K., Brown, D.J., Dennison, P.E., Graves, S.A., Bricklemyer, R.S., 2009. Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data. Geoderma 150 (1–2), 72–84.
Hengl, T., Reuter, H.I. (Eds.), 2008. Geomorphometry: concepts, software, applicationsDev. Soil Sci. 33, 772.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvenlink, G.B.M., 2014. SoilGrids1km-Global soil information based on automated mapping. PLoS One 9 (8), e105992.
Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56 (3), 836–841.
Jenny, H., 1941. Factors of Soil Formation, A System of Quantitative Pedology. McGraw-Hill, New York.
Lee, C.A., Gasster, S.D., Plaza, A., Chang, C.-I., Huang, B., 2011. Recent developments in high performance computing for remote sensing: a review. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4 (3).
Lichstein, J.W., Golaz, N., Malyshev, S., Shevliakova, E., Zhang, T., Sheffield, J., Birdsey, R.A., Sarmiento, J.L., Pacala, S.W., 2014. Confronting terrestrial biosphere models with forest inventory data. Ecol. Appl. 24 (4).
Manzoni, S., Porporato, A., 2009. Soil carbon and nitrogen mineralization: theory and models across scales. Soil Biol. Biochem. 41, 1355–1379.
McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Meirik, E., Frazier, B., Brown, D., Roberts, P., Rupp, R., 2010. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), ASTER-based vegetation map to improve soil modeling in remote areas, in Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer.
Michalakes, J., Vachharajani, M., 2008. GPU acceleration of numerical weather prediction. Parallel Process. Lett. 18 (4).
Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital soil mapping of soil classes. Geoderma 142 (3–4), 285–293.
Mitchell, K.E., Lohmann, D., Houser, P.R., Wood, E.F., Schaake, J.C., 2004. The multi-institution North American Land Data Assimilation System (NLDAS): utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. J. Geophys. Res. 109.
Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399.
Nauman, T.W., Thompson, J.A., Rasmussen, C., 2014. Semi-automated disaggregation of a conventional soil map using knowledge driven data mining and Random Forests in the Sonoran Desert, USA. PE&RS 80 (4), 353–366.
Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91–100.
Padarian, J., Minasny, B., McBratney, A.B., 2015. Using Google's cloud-based platform for digital soil mapping. Comput. Geosci. 83, 80–88.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Plaza, A., Chang, C.-I., 2007. High Performace Computing in Remote Sensing. Taylor & Francis, Boca Raton, Florida.

Rodriguez-Iturbe, I., Porporato, A., 2004. Ecohydrology of Water-Controlled Ecosystems: Soil Moisture and Plant Dynamics. Cambridge Univ. Press, New York.
Soil Survey Staff, 2014. Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States. United States Department of Agriculture, Natural Resources Conservation Service. Available online at http://datagateway.nrcs.usda.gov.
Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio county soil map. Soil Sci. Soc. Am. J. 77 (4), 1254–1268.
Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio county soil survey map using possibilistic decision trees. Geoderma 213, 334–335.
Thompson, J.A., Prescott, T., Moore, A.C., Bell, J., Kautz, D., Hempel, J., Waltman, S.W., Perry, C.H., 2010. Regional Apprach to Soil Property Mapping Using Legacy Data and Spatial Disaggregation Techniques, in 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia.
Thompson J. A., T. W. Nauman, N. P. Odgers, N. Libohova, J. Hempel (2012), Harmonization of Legacy Soil Maps in North America: Status, Trends, and Implications for Digital Soil Mapping Efforts, in The 5th Global Workshop on Digital Soil Mapping, Digital Soil Assessments and Beyond, edited by A. B. McBratney, B. Minasny, B. Malone, Sydney, Australia.
Wei, S.A., McBratney, A., Hempel, J., Minasny, B., Malone, B., D'Avello, T., Burras, L., Thompson, J.A., 2010. Digital Harmonisation of Adjacent Soil Survey Areas — 4 Iowa Counties, in Proceedings of the 19th World Congress of Soil Science, Soils Solutions for a Changing World. Brisbane, Australia.
Wood, E.F., et al., 2011. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resour. Res. 47 (5).
Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. 75 (3), 1044–1053.
Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65 (5), 1463–1472.