Landscape and Urban Planning 190 (2019) 103599

Contents lists available at ScienceDirect

Landscape and Urban Planning

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

Research Paper

Soil swelling potential across Colorado: A digital soil mapping assessment

Emma Stell[a], Mario Guevara[b], Rodrigo Vargas[a,b,*]

[a] Department of Geography, University of Delaware, Newark, DE 19716, United States
[b] Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, United States

ARTICLE INFO

Keywords:
Machine learning
Shrink-swell
Expansive clays
Linear extensibility

ABSTRACT

Swelling soils contain high amounts of expansive clay minerals which swell upon wetting and shrink upon drying. Urban growth across swelling soils causes billions of dollars in infrastructure damage annually throughout the United States. Here, continuous spatial information of soil swelling properties is provided across Colorado, a state experiencing extensive urban growth across swelling soils. The objectives were: 1) model the spatial variability and associated uncertainty of soil swelling-related properties across Colorado; 2) generate a continuous statewide map of soil swelling potential and associated uncertainty at 1x1km resolution; and 3) identify urban areas prone to experience swelling conditions. A digital soil mapping (DSM) framework for extracting pedologically-relevant information from legacy maps was used to train machine learning models to analyze multiple sources of information to represent the soil forming environment (e.g., soils, climate, organisms, topography). Best predictions were based on regression trees and explain over 80% of soil swelling spatial variability across the state (10-fold cross validation strategy). Over 20% of urbanized areas were identified as being prone to experience swelling conditions. Special considerations must be undertaken to couple urban development with soil functionality in order to prevent future damages and economic losses.

1. Introduction

Swelling soils, those containing specific types of expansive clay particles which swell upon wetting and shrink upon drying, cause more than $2.3 billion worth of structural damage throughout the nation each year (Jones & Holtz, 1973). Consequently, swelling soils threaten economic prosperity and infrastructure maintenance. The state of Colorado is one of the most threatened regions in the United States because it is located within an arid/semi-arid region (Kottek, Grieser, Beck, Rudolf, & Rubel, 2006) which experiences large variability in soil moisture (e.g., contrasting drying and wetting cycles), increasing the probability of regional swelling (Jones & Jefferson, 2012). Therefore, it is critical to provide continuous information to provide insights for land planning and urban development across areas dominated by swelling soils.

Swelling occurs in soils that can absorb large amounts of water when wetted, but also become hard, shrinking and cracking when dry. This phenomenon is primarily related to the structure and presence of 2:1 clays (e.g., phyllosilicate minerals, such as smectite and vermiculite), in which water can get into interlayer spaces within clay minerals, causing swelling and shrinking at the molecular level (Vaught, Brye, & Miller, 2006). The presence of cracks upon the shrinking of soil allows water to enter more easily upon re-wetting, thus exacerbating the amount of water that can be absorbed. These types of soils are sensitive to changes in soil moisture content, which is influenced by precipitation variability, local site physical changes (e.g., construction), and/or removal of vegetation cover (Jones & Jefferson, 2012). Structural engineers, geotechnical engineers, and other home builders include procedures in case swelling soils are located within a construction site, but the presence of these type of soils is not always obvious (Houston, Dye, Zapata, Walsh, & Houston, 2011).

Shrinking and swelling behavior of soils also impacts infiltration of water by creating and sealing cracks in the soil that act as preferential flow paths (Stewart, Najm, Rupp, & Selker, 2016). Successive wetting and drying cycles on swelling soil causes crack locations to alternate, leading to more uniform wetting with depth as time goes on and a reduction of preferential flow paths (Wells et al., 2003). Within swelling soils, ponding and infiltration after the first rainstorm leads to the formation of cracks as soil dries. These cracks lead to increased ponding time and infiltration during the next rainstorm, causing cracks to deepen and widen, creating a positive feedback loop (Römkens & Prasad, 2006). Eventually, cracks in heavily drained soil may not close, even after prolonged wetting. Furthermore, soil swelling is subject to seasonal variation of soil moisture (Beven & Germann, 1982) and

* Corresponding author.
E-mail addresses: emstell@udel.edu (E. Stell), mguevara@udel.edu (M. Guevara), rvargas@udel.edu (R. Vargas).

https://doi.org/10.1016/j.landurbplan.2019.103599
Received 8 October 2018; Received in revised form 3 June 2019; Accepted 14 June 2019
Available online 02 July 2019

consequently influences local hydrology. Lack of accounting for swelling systems can cause estimation errors in water/solute flux, water balance, drainage, and aquifer recharge (Kirby, Bernardi, Ringrose-Voase, Young, & Rose, 2003; Smiles, 2000). Finally, swelling soils can also affect biogeochemical processes as cracks in the soil can contribute to reductions in soil organic carbon through effects on (decreasing) plant productivity and increasing decomposition of organic matter (Yoo, Amundson, Heimsath, & Dietrich, 2006).

Information of soil swelling potential not only provides insights for proper construction techniques to minimize future potential structural problems caused by soil swelling, but also provides information about ecohydrological patterns. The advent of a spatially consistent map of soil swelling can be used to provide a better understanding of soil-related hydrological properties and water infiltration processes. Additionally, the soil swelling information generated here combines soil data from the physical, organic and chemical soil phases, so it could be used to better inform the development of soil spatial indicators of land degradation, soil fertility, or soil quality.

The United States Geological Survey (USGS) generated in 1989 a map of swelling clays throughout the conterminous United States (CONUS) with a scale of 1:7,500,000. However, this product has low spatial detail for localized use and detailed analyses and is based on the type of bedrock present beneath the soil (Olive, Chleborad, Frahme, Schlocker, Schneider, & Schuster, 1989). At the state level, there is pedological and detailed information of soil swelling, but it has spatial discontinuities (see methods section) due to the combination of different soil sampling methods and soil survey strategies. This information is provided by the gSSURGO (gridded Soil Survey Geographic) geodatabase (Soil Survey Staff, 2016). Much of the information within the gSSURGO database appears in a grid-like pattern, indicating potential discrepancies in and around those areas (see Methods section). Grid-like patterns of values indicate potential errors, such as those that arise from field estimations, measurement error, and protocol/mapping distinctions between counties. Multiple spatial inconsistencies seem to run along county lines. This important and extremely valuable database was developed based on the official Soil Survey Geographic (SSURGO) database, which was developed by National Cooperative Soil Survey (NCSS) for use in regional, statewide, and/or national resource planning and soils analysis (Soil Survey Staff, 2016). This study proposes that there is need to provide alternative products within regions of interest (e.g., Colorado) to inform about the potential of swelling of soils.

Here, information is provided about swelling soils through a digital soil mapping (DSM) approach. DSM is a reference framework to model and predict soil properties and functions in areas where no information is available. This is achieved by coupling gridded environmental information with observational soil property data throughout statistical models (e.g., geostatistics, machine learning). For DSM, the SCORPAN model is used, which states that a soil attribute (or class) can be predicted as a function of the soil forming environment (McBratney, Mendonça Santos, & Minasny, 2003).

$$Sf^{\sim}(s, c, o, r, p, a, n) + \varepsilon \quad (1)$$

where S = swelling (or other attribute); s = other known soil attributes; c = climate; o = organisms; r = relief; p = parent material; a = age; n = space; ε = uncertainty. In short, soil-forming environmental attributes are together able to predict soil properties by using statistical models (f) such as hypothesis-driven models (i.e., linear methods) and machine learning (ML) data-driven models (i.e., random forest, RF). While linear methods are useful to explain linear relationships, ML is used to uncover patterns in large data sets using computer-based data-driven decisions (such as RF) (Heung et al., 2016). Thus, linear models and ML algorithms can be used on DSM to produce spatially continuous information (i.e., continuous maps) of soil swelling properties.

1.1. Research objectives

The objectives of this study are: 1) model the spatial variability and associated uncertainty of soil swelling-related properties across Colorado; 2) generate a continuous statewide map of soil swelling potential and associated uncertainty at 1x1km resolution; and 3) identify urban areas prone to experiencing swelling conditions. The following interrelated research questions are explored: 1) which are the best environmental prediction factors for soil swelling properties?; 2) how much spatial variability can be explained by using ML to predict the spatial variability of these properties across Colorado?; and 3) what is the percent of urban areas across Colorado located across soils with high swelling potential? The novelty of this study is that different modeling approaches are tested to predict swelling potential in soils across Colorado to provide value-added information with high spatial resolution (i.e., 1 × 1 km grids).

2. Methodology

A DSM approach for disaggregating soil swelling polygon information was applied as a way to predict soil swelling properties in places where no information is available or in places with spatial discontinuities due to the spatial aggregation of multiple polygon maps. Gridded environmental data sources such as remote sensing imagery, digital terrain analysis (i.e., geomorphometry), legacy and thematic maps (i.e., land use, soil type), and gridded climatology products (i.e., seasonal precipitation and temperature) were used as prediction factors for soil swelling properties: saturated hydraulic conductivity (ksat_r), cation exchange capacity (cec7_r), liquid limit (ll_r), plasticity index (pi_r), and total clay amount (claytotal_r). First, the main relationships of soil swelling properties and environmental conditions were quantified and then the soil swelling potential across the state of Colorado was predicted using 1 × 1 km grids. Fig. 1 summarizes the DSM framework used.

2.1. Study area

Colorado, though known as one of the “Rocky Mountain” states, has a large variety of landscape types, making general environmental generalizations difficult. The eastern part of the state is mostly composed of the High Plains, which become rolling hills as they move west to approach the mountains. The mountains are found in the middle of the state, with plateaus characterizing the western region (Doesken, Pielke, Roger, & Bliss, 2003).

Colorado has a large heterogeneity of soils, primarily Alfisols, Aridisols, Mollisols, and Entisols. Alfisols are found primarily in the western half of the state, while Entisols are primarily present in the southeastern corner and along the western border of the state. Mollisols and Aridisols are spread throughout Colorado (United States Soil Conservation Service, 1976). Colorado is known for having high amounts of clay minerals in the soils. In the Front Range, for example, smectite is mainly found in plains areas, vermiculite is usually found in the forested areas, and a mix of clay minerals is commonly found above the tree line (Birkeland, Shroba, Burns, Price, & Tonkin, 2003). These mineral deposits are due to the presence of expanding claystone bedrock near the surface in the region, whose effects are often exacerbated by steep dips within the bedrock layers (Chabrillat, Goetz, Krosley, & Olsen, 2002).

Colorado’s location in an arid/semi-arid region causes an exacerbation of swelling activities, as each precipitation event has the ability to dramatically alter the amount of soil moisture residing within the soil. As soil swelling can only occur if there is a change in soil moisture, arid regions that experience large changes in soil moisture with the advent of rainfall should be more heavily affected (Jones & Jefferson, 2012).

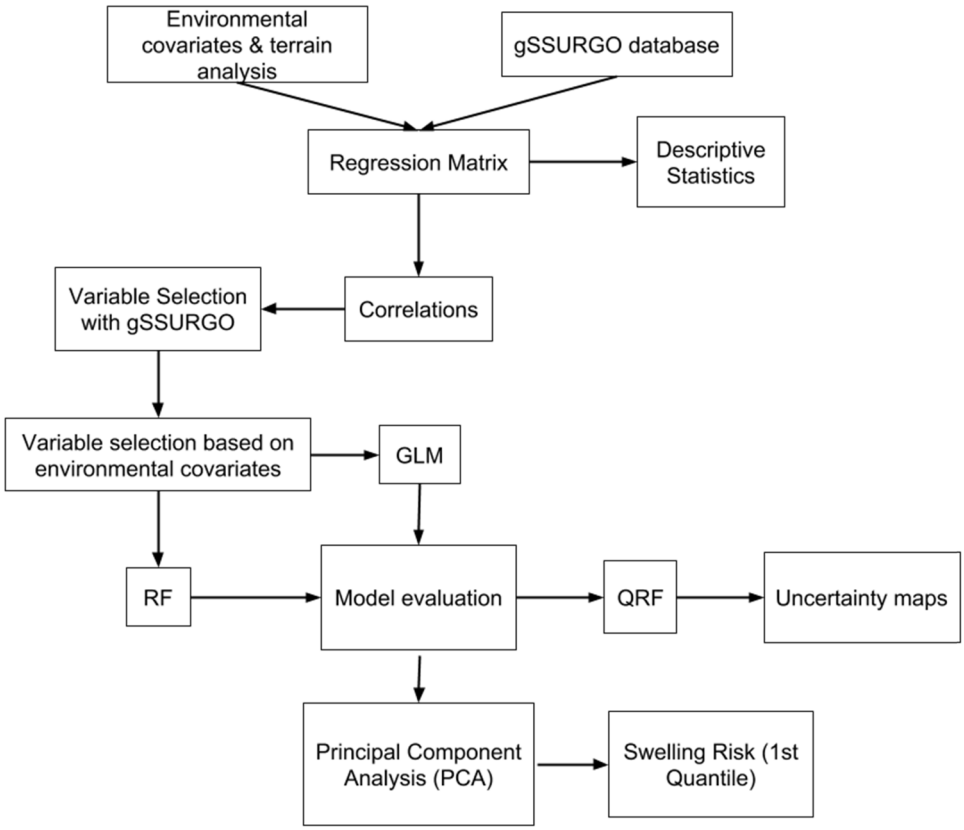

Fig. 1. Workflow designed for this study.

2.2. Digital soil mapping training data

A statewide gSSURGO geodatabase of Colorado from the US-Natural Resource Conservation Service-Geospatial Data Gateway (https://datagateway.nrcs.usda.gov/) (Soil Survey Staff, 2016) was used. Tabular data, based on soil properties stored in the National Soil Information System (NASIS), were merged with the vector-based soil spatial information from the gSSURGO geodatabase into statewide extents. This geo dataset contains soil type polygon maps of over 150 different soil attributes. From this extensive list of soil attributes, properties thought to relate to shrink-swell conditions were manually selected, including saturated hydraulic conductivity (ksat_r), cation exchange capacity (cec7_r), liquid limit (ll_r), plasticity index (pi_r), and total clay amount (claytotal_r) (see supplementary material S1). These variables are associated with the shrink/swell capacity of soil properties and were selected after literature review (Jones & Jefferson, 2012; Nayak & Christenson, 1971; Pruška & Šedivý, 2015; Thomas, Baker, & Zelazny, 2000). The variable lep_r was also selected, as it is identified in the National Soil Survey Handbook as "the linear expression of the volume difference of natural soil fabric at 1/3-bar or 1/10-bar water content and oven dryness" (USDA, 2018). As it is a common way of quantifying shrink/swell activity, the NRCS has assigned associated Shrink/Swell classes to values of linear extensibility (Table 1). For this reason, lep_r was assumed to be a direct estimate of soil swelling (USDA, 2018). See Supplementary Material 1 for definitions of all utilized gSSURGO variables.

Using conventional GIS procedures [i.e., Create Random Points and Extract Values to Points tool in ArcGIS], the known pedological information from the chosen gSSURGO attributes was extracted to 10,000 random spatial points throughout Colorado state boundaries (Fig. 2b). Extracting pedological information to 10,000 points simulated a random distribution of locations that were assumed to capture the variability of the gSSURGO map (around 26 points per squared km) across Colorado. These random points were consistent to replicate our results without major differences in predictions after multiple model realizations. These points were selected randomly to maximize the possibility of avoiding a systematic pattern of the spatial inconsistencies (i.e., across counties) in the gSSURGO map. The spatial inconsistencies present in the gSSURGO map are mostly comprised of zero values, so of the original points, those falling within zero values were eliminated in order to avoid similar inconsistent patterns in our final maps. Then, a regression matrix was generated containing georeferenced data of swelling-related variables (e.g., lep_r, ll_r, pi_r, cec7_r, and claytotal_r) and an extensive set of environmental prediction factors (see Section 2.3).

Table 1
NRCS Shrink/Swell Classes.

Shrink-Swell Class	LEP[1]	COLE[2]
Low	< 3.0	< 0.03
Moderate	3.1–5.9	0.03–0.06
High	6.0–8.9	0.06–0.09
Very high	> 8.9	> 0.09

[1] LEP = Linear Extensibility.
[2] COLE = Coefficient of linear extensibility.

2.3. Prediction factors for soil swelling

Prediction factors were represented by environmental information from worldgrids.org (last accessed Jan. 2018), an initiative of ISRIC Soil Information (Reuter & Hengl, 2012). A total of 118 1 km × 1 km environmental predictors were downloaded and masked to the Colorado state boundaries in order to quantitatively represent the soil-forming environment (i.e., climate, topography, living organisms, parent material). The *worldgrids* predictors come from three sources: remote

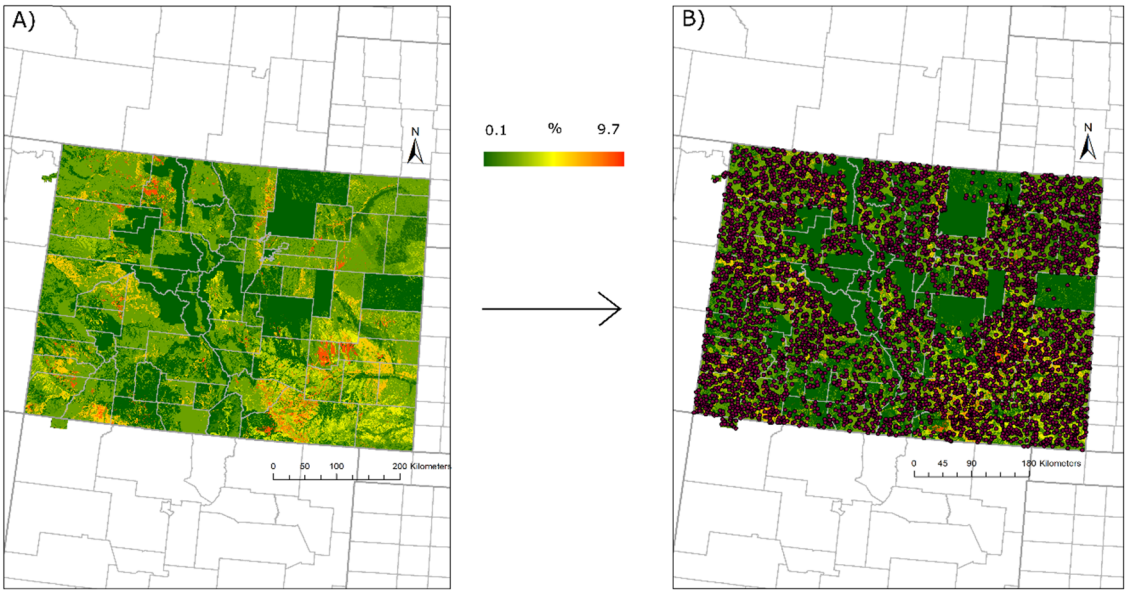

Fig. 2. (A) Original information linked and displayed from the gSSURGO database. Grid-like patterns of values indicate potential errors, such as those that arise from field estimations, measurement error, and protocol/mapping distinctions between counties. (B) gSSURGO database information extracted to 10,000 random points throughout the state of Colorado. The color scale indicates swelling percentage of soils, from 0.1% (low) to 9.7% (very high). Potential errors from distinctions between counties seemed to be composed of mostly zero values, so points with zero values were removed for analysis.

sensing, climate surfaces, and digital terrain analysis (see complete list and description of layers in Appendix C).

Additional environmental predictors included terrain parameters (e.g., terrain slope, aspect, channel network base level, topographic wetness index, length-slope factor, analytical hillshading, elevation, cross-sectional curvature, longitudinal curvature, convergence index, closed depressions, flow accumulation, vertical distance to channel network, valley depth, and relative slope position) calculated by running a basic digital terrain analysis on a USGS-Colorado DEM (1 × 1 km pixels) in SAGA GIS (System for Automated Geo-scientific-Analysis Geographical Information System) (Conrad et al., 2015). These parameters are indicators of the spatial variability of soil swelling properties, since they control the overall distribution of water across the landscape, the potential incoming solar radiation, and the overland flow from rain or irrigation water. Organic matter content was not included as a potential modifier of soil swelling because it represents the irreversible component of soil swelling, whereas shrink/swell through LEP is reversible (Seybold & Libohova, 2017).

2.4. Prediction of linear extensibility and model evaluation

The informational discontinuities in the gSSURGO variables were corrected using a modeling strategy based on environmental correlation methods following the SCORPAN reference framework for digital mapping.

First, the regression matrix (i.e., training data and prediction factors for those locations of training data) was used to calculate descriptive statistics. Second, a correlation analysis was performed to identify and explore "key" gSSURGO variables associated with lep_r. Once each of these variables were selected, they became the focus of a model using GLM, RF, or QRF, with eight *worldgrids* predictor variables chosen based on a subsequent correlation analysis. Then, two different models, a generalized linear model (GLM) and random forest (RF) model, were used to make predictions based on the best correlated environmental predictors. A GLM is a linear model used to test a response variable to predict the response under a wide range of independent variables (Lane, 2002). RF uses a combination of tree predictors, in which attributes at each node are randomly split to find the best predictions for that group of attributes (Breiman, 2001).

Each RF model was built using the eight best correlated environmental predictors (including topographic and *worldgrids* covariates) for each soil swelling variable. While the best correlated environmental predictors were similar across the selected gSSURGO variables, they were not exactly the same (see Section 3). A Quantile Regression Forest (QRF) model was used to map the RF model uncertainty for each soil swelling variable. Like RF, QRF is based on a combination of tree predictors with randomly split nodes. However, QRF considers the spread of values of the response variable at each node, instead of just the mean as RF does. This means QRF allows estimates of the distribution of the full response of training data to the prediction factors. In this study, QRF was used to provide a measure of model uncertainty and reliability. It has been shown that these models are robust for use in soil mapping applications with multiple sampling designs (Vaysse & Lagacherie, 2017). Finally, model performance was assessed using 10-fold cross validation to calculate the r^2 (i.e., explained variance) and the root mean squared error (RMSE). All analyses were performed using R software (R Core Team, 2016).

2.5. Integration of soil swelling predicted properties into a soil swelling map

First, the mean of all soil swelling predicted variables was scaled in a zero centered basis. Then a Principal Component Analysis (PCA) was used in order to integrate the 1 × 1 km predicted variables that dictate soil swelling properties across Colorado. According to McBratney et al. (2003), PCA is used when there are a large number of correlated predictor variables in order to arrange the original inputs linearly. Here, the PCA was used to reduce the soil swelling related variables to a single soil swelling potential map which combines information of all the variables included in the analysis (i.e., first principal component [PC1]). Inputs to the PCA were the outputs of five RF models of lep_r and its best correlated factors; those that are also included in the gSSURGO database (e.g., pi_r, ll_r, cec7_r, and claytotal_r). The PCA reduced variable redundancy in a potentially highly correlated multivariate space (e.g., correlated variables explaining soil swelling conditions). The resulting PC1 was characterized by quartiles in order to translate the scale from the PC1 unitless values to a comparative scale in order to distinguish between low and high soil swelling potential. Thus, swelling potential was able to be expressed as a linear

combination/integration of multiple variables in order to "rank" swelling areas.

2.6. Urban swelling risk

Remote sensing technologies allow scientists to monitor changes of human activity, and remote sensing of night lights has been used for decades to inventory the global distribution of human activity (Kruse & Elvidge, 2011). The predicted soil swelling map was overlapped with a night lights map across Colorado (worldgrids.org system; LN1DMS representing the first principal component of the long-term lights at night images; Fig. 7), to identify areas where urban development is prone to experience soil swelling conditions. The original lights at night data were gathered from 1992 to 2013 as cloud-free composites of DMSP-OLS (Defense Meteorological Satellite Program – Operational Linescan System) smooth resolution data collected by the Earth Observation Group of the National Oceanic and Atmospheric Administration.

3. Results

3.1. Descriptive statistics

The mean value of lep_r was relatively low in relation to the maximum, indicating that high amounts of swelling were probably mostly concentrated in a few select areas (Table 2). Ksat_r had the largest range of values, along with the largest standard deviation of the soil swelling properties. Lep_r and pi_r had the lowest range of values, along with the lowest standard deviations. Descriptive statistical covariate information can be found in Table 2. Ksat_r was expected to be a variable with potential influence on soil swelling, but it was found to not be closely correlated with lep_r.

3.2. Prediction factors for soil swelling

First, gSSURGO generated variables were tested against lep_r to find the best correlations and attributes related to swelling. The four variables with the highest correlations were claytotal_r (0.795), pi_r (0.787), ll_r (0.651), and cec7_r (0.587). Since these highly correlated gSSURGO variables come from the same source, they have similar spatial discontinuities as found in lep_r. Thus, each of these variables was evaluated through the SCORPAN model to achieve continuous predictions.

We tested the relationship between *worldgrids.org* and topographic variables with lep_r and all soil swelling related variables (i.e., claytotal_r, pi_r, ll_r, cec7_r) to find the best predictor variables and reduce model complexity. This was needed because the soil-forming environment was represented by variables from *worldgrids.org* and topographic variables in order to represent prediction factors at the available 1x1 km spatial resolution (Table 3). The environmental and topographic variables tested showed significant but low correlations ($r \leq 0.3$). For *worldgrids* variables, those associated with temperature were frequently among the most significant variables, such as "Mean value of the 8-day MODIS day-time LST time series data" for "-Apr/May," "-Jun/Jul," "-Aug/Sep," and "-Oct-Nov" (tx3mod3a, tx4mod3a, tx5mod3a, tx6mod3a); "Mean value the 8-day MODIS day-time LST time series data" (tdmmod3a); and "Standard deviation of the 8-day MODIS day-time LST time series data" (tdsmod3a). Other frequently correlated variables included soil-based variables, such as "Percent coverage Podzols" (gpzhws3a), "Percent coverage Leptosols" (glphws3a), and "Geological ages based on the surface geology" (geaisg3a). Appendix C provides a complete list of *worldgrids.org* covariates, acronyms, and descriptions.

Table 2
Descriptive Statistics of each gSSURGO variable. Number of observations (*n*), quartiles (Qu), median, mean and the standard deviation (Sdev) of the covariates. Lep_r = linear extensibility (%); ksat_r = saturated conductivity (um s^{-1}); ll_r = liquid limit (%); claytotal_r = % total clay; cec7_r = cation exchange capacity (meq 100 g^{-1}); pi_r = plasticity index (%).

Covariate	n	Min	1stQu	Median	Mean	3rdQu	Max	SDev
lep_r	5609	0.10	1.50	1.50	2.01	1.60	9.70	1.45
ksat_r	5609	0.01	9.00	9.17	28.30	28.22	423.07	51.91
ll_r	5238	0.00	26.00	27.50	28.53	34.00	63.00	10.24
claytotal_r	5587	0.00	14.00	20.00	19.72	23.50	52.50	9.44
cec7_r	5488	0.00	10.00	14.50	14.75	18.80	175.00	7.46
pi_r	5238	0.00	7.50	7.50	9.48	13.30	36.00	6.41

3.3. Model evaluation and explained variance

Using the best correlated prediction factors (Table 3), predictions from a generalized linear model (GLM) and a random forest (RF) model were compared. The RF model was able to explain at least 88% of variance for each soil swelling-related variable. The GLM showed lower levels of explained variance (between 25.0 and 39.6%), as it does not account for non-linear relationships. RF performed better than the GLM model and was able to explain higher amounts of spatial variability with lower RMSE values (Table 4). Lep_r showed the highest amount of explained variance at 92.1%, while ll_r showed the lowest amount of explained variance, at 88%.

Fig. 3 shows a visual representation of attribute predictions using the RF model. Generally, higher predicted amounts of each attribute were found in the southeastern and southwestern quadrants of the state, and sporadically along the western border. However, there were high amounts of ll_r and cec7_r throughout the state.

Fig. 4 shows selected areas for comparison of the gSSURGO lep_r polygon map and the RF lep_r prediction map. While gSSURGO has a higher resolution, RF reduced spatial inconsistencies present, especially those that follow strict county lines. By initially removing the zeros that generate spatial inconsistencies, the RF prediction was able to overcome (i.e., gap fill) these inconsistencies.

Supplementary Fig. 1 shows a rasterized differences map between the original gSSURGO lep_r map and the lep_r RF prediction. Small areas with higher differences are spread throughout the landscape, but with larger differences aggregated in the northwestern and southeastern portions of the study area. The spatial artifacts in the original gSSURGO map are highlighted by the differences map, but the RF prediction eliminated these inconsistencies.

3.4. Spatial patterns of model prediction uncertainty

Statewide maps of the model prediction uncertainty (i.e., the variance of all the trees in the random forest ensemble) were generated for each attribute per pixel (Fig. 5). Ll_r and claytotal_r showed the highest amount of variability, followed by lep_r and pi_r. Cec7_r showed the lowest amount of variability. Overall, higher amounts of model prediction uncertainty for the different attributes were mostly concentrated within the southeastern quadrant of the state. Additional high values were present along the western border of Colorado, especially near the north and southwestern corners.

Root mean square error (RMSE) was calculated for each best correlated attribute for all models. For the QRF model, RMSE had a wide range of values, from 1.32% for lep_r (below the first quantile, Table 2) to 9.82% for ll_r (below the first quantile, Table 2). Claytotal_r had the second highest RMSE of 8.72% (below the first quantile, Table 2), while cec7_r has a slightly lower RMSE of 8.35 meq 100 g^{-1} (below the first quantile, Table 2), and pi_r has the second lowest RMSE of 5.71% (below the first quantile, Table 2).

The RMSE values for the GLM and RF models were all below their first quantiles (Table 2). RMSE was overall lower for the RF models than for the GLMs. Lep_r had the lowest RMSE of 0.41%, while ll_r had the highest, one of 3.54%.

Table 3
Best correlated gSSURGO variables and their best correlations with *worldgrids.org* and topographic variables. The best correlated gSSURGO variables are found and shown in Table 3. These are used here to find the best environmental correlations. Numbers in parentheses indicate correlation values ranging from −1 to 1 between each variable and the best correlated predictors (i.e., Best Correlated Variable (Correlation)). Explained variance (r^2) represents the median explained model variance for each variable. RMSE represents the root mean square error associated with the QRF model for each variable.

Variable	Best Correlated Variable (Correlation)	Explained Variance (r^2)	RMSE
lep_r	tx6mod3a (0.311); tx3mod3a (0.297); tdmmod3a (0.296); tx4mod3a (0.293); tx5mod3a (0.279); tx2mod3a (0.276); tdhmod3a (0.271); geaisg3a (−0.255)	0.210	1.321
claytotal_r	tx4mod3a (0.292); tx5mod3a (0.288); tx6mod3a (0.285); tx3mod3a (0.274); gpzhws3a (−0.273); geaisg3a (−0.273); glphws3a (−0.263); tnsmod3a (0.254)	0.219	8.718
pi_r	geaisg3a (−0.295); tx6mod3a (0.265); tx4mod3a (0.245); glphws3a (−0.243); gpzhws3a (−0.241); tx3mod3a (0.240); tx5mod3a (0.237); tdmmod3a (0.231)	0.243	5.709
ll_r	geaisg3a (−0.224); tx4mod3a (0.194); tnsmod3a (0.194); tx5mod3a (0.182); tdsmod3a (0.173); gflhws3a (0.172); gpzhws3a (−0.168); tx6mod3a (0.167)	0.176	9.823
cec7_r	VerticalDistanceToChannelNetwork (−0.219); g12igb3a (0.207); g11igb3a (0.197); geaisg3a (−0.185); tx6mod3a (0.180); RelativeSlopePosition (−0.178); l01igb3a (−0.177); DEM (−0.176)	0.123	8.354

Table 4
Evaluated machine learning methods with r^2 and RMSE values. A RF (Random Forest) model was compared to a GLM (generalized linear model).

Variable	Method	Var explained	RMSE
lep_r	GLM	25.1%	1.26
	RF	92.1%	0.408
pi_r	GLM	38.1%	5.04
	RF	90.6%	1.97
claytotal_r	GLM	34.8%	7.63
	RF	88.6%	3.19
ll_r	GLM	39.6%	7.95
	RF	88.0%	3.54
cec7_r	GLM	28.1%	6.32
	RF	91.5%	2.18

3.5. Principal component analysis

This integrated map of swelling potential follows the expected areas of high lep_r values as predicted by the RF models, especially in the southeast quadrant and patterns throughout the midwestern part of the state (Fig. 6). It was found that the first PC (i.e., PC1) of all the highly correlated gSSURGO variables was able to explain nearly 80% of variability. This first PCA was classified in quantiles to characterize the soil swelling potential. Thus, the values of the predicted gSSURGO variables (e.g., lep_r, pi_r, ll_r, cec7_r, and claytotal_r) were classified according to overall swelling potential, comparatively ranking present soils' swelling capability from "low" to "very high." Overall trends of both areas of increased swelling and of little/to no swelling seem to follow those seen in the lep_r RF model predictions. Areas of high lep_r included much of the southeastern quadrant as well as in the southwest,

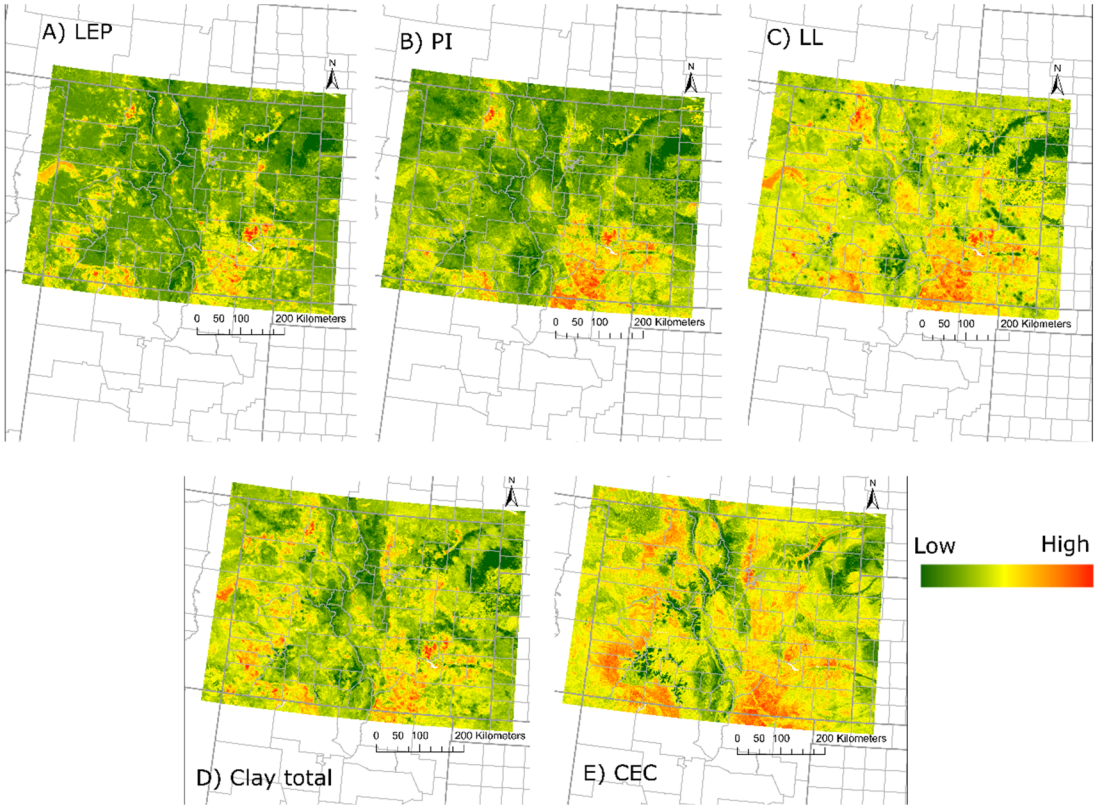

Fig. 3. Predictions of soil swelling related variables. Prediction maps calculated with the RF method for each soil swelling-related variable: (A) = prediction for lep_r; (B) = prediction for pi_r; (C) = prediction for ll_r; (D) = prediction for claytotal_r; (E) = prediction for cec_r. (A)-(D) are measured by %, while (E) is measured by meq $100\,g^{-1}$.

Fig. 4. Zoomed Comparisons between gSSURGO and Lep_r RF prediction. (A) gSSURGO lep_r polygon map, zoomed in at three areas that show spatial inconsistencies. (B) Lep_r RF prediction map, selected areas for comparison in the same three areas. Spatial inconsistencies are reduced by using RF.

and in distinct areas along the western border.

3.6. Swelling risk across urban areas

The lights at night map across Colorado (*worldgrids.org*; LN1DMS first principal component of the long-term lights at night images; Fig. 7a), was used as a filter for the PCA model in order to identify areas where urban development is prone to experience soil swelling conditions (Fig. 7b). Results show that over 20% of the state of Colorado urbanized areas (i.e., where night lights indicate urban human activity) are prone to experience swelling conditions. The majority of these swelling-prone urbanized areas indicate very high swelling potential, indicated by dark green (Fig. 7b).

4. Discussion

A DSM framework was developed to generate continuous information about swelling soils to overcome spatial inconsistencies (e.g., unrealistic spatial patterns) on current available information. The spatial variability of soil swelling properties was modeled, and continuous predictions were provided across Colorado in a 1x1km grid, including a spatially continuous measure of model uncertainty. High amounts of lep_r are most likely concentrated across a few select areas suggesting that more detailed surveys and analyses are needed within these "hotspots".

The results regarding the best correlated attributes support trends identified in previous research. For example, clay total and PI are

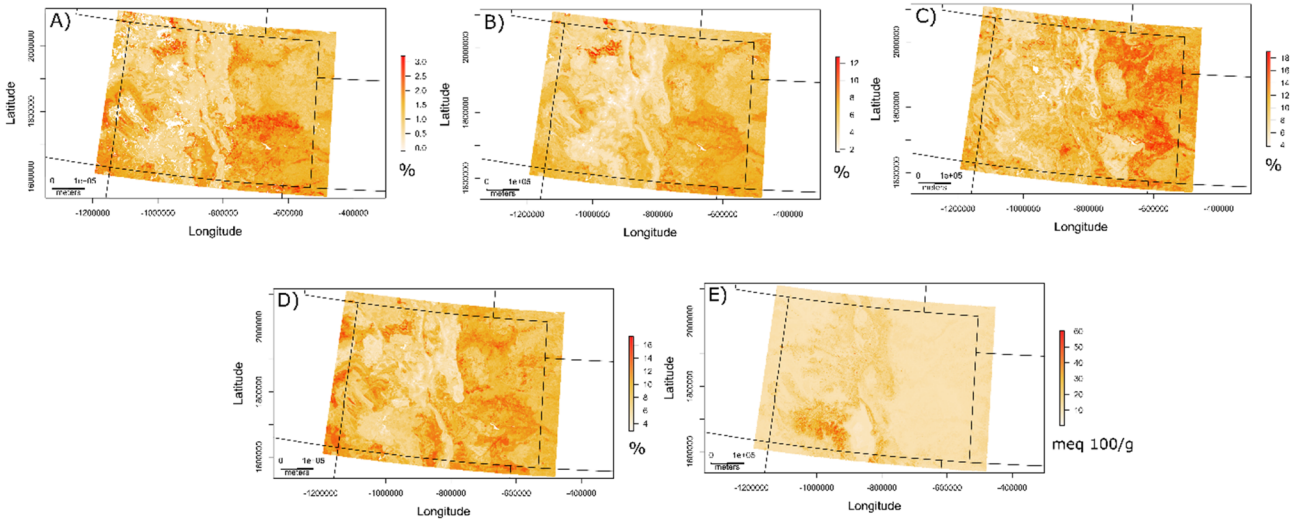

Fig. 5. Uncertainty maps. Standard deviation (per pixel) of the full conditional distribution of each swelling variable derived from the quantile random forest (QRF) method. These maps are surrogates of model uncertainty. (A) = standard deviation of lep_r; (B) = standard deviation pi_r; (C) = standard deviation ll_r; (D) = standard deviation claytotal_r; (E) = standard deviation cec_r. (A)-(D) are measured by %, while (E) is measured by meq 100 g^{-1}.

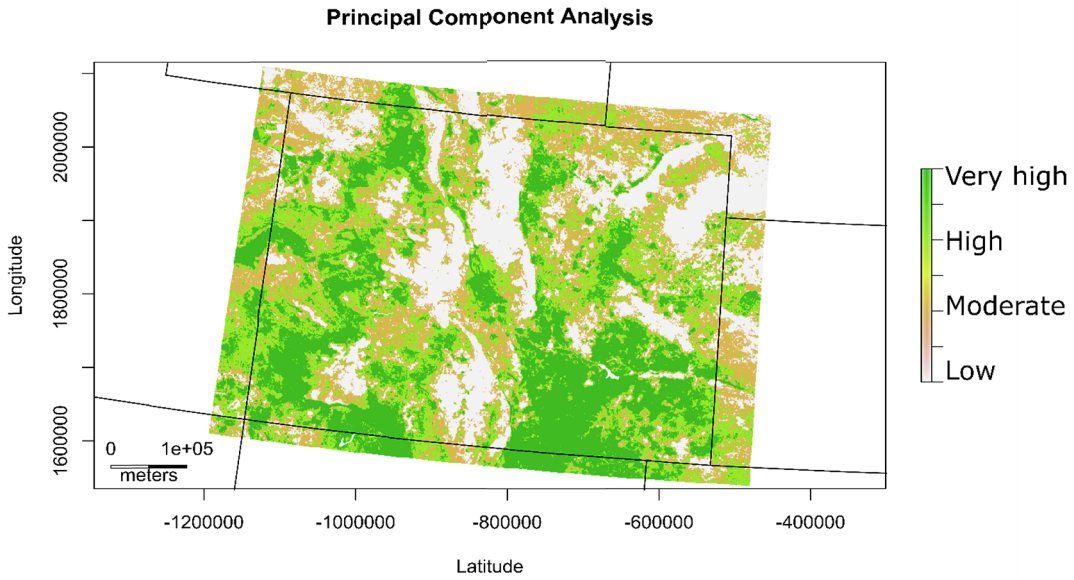

Fig. 6. Integrated PCA map of swelling potential in the first quartile using swelling attributes. This map linearly integrates attributes to show an overall representation of the soil swelling risk across Colorado. Over 80% of swelling potential was explained in the 1st principal component (see Supplementary Figs. 2 & 3).

known to be highly correlated with LEP, and the highest correlated gSSURGO attributes were found to be claytotal_r (r^2 = 0.795) and pi_r (r^2 = 0.787). Previous work has reported high correlation between swelling and PI, usually in conjunction with other identified highly correlated attributes (Nayak & Christenson, 1971; Jayasekera & Mohajerani, 2003; Kariuki & van der Meer, 2004). Furthermore, CEC has been reported to have a strong correlation with swelling percent (Christidis, 1998; Kariuki, Woldai, & Meer, 2004; Thomas et al., 2000; Yilmaz, 2006). However, it is important to distinguish between clay and carbonate clay contributors to CEC (carbonate clays do not have as much of an influence), which affects CEC contribution to LEP (Seybold & Libohova, 2017). The results showed that cec7_r was the fourth most significant attribute, but with a lower correlation of 0.587. Overall, the results follow general trends of high correlations between soil swelling and PI, clay content, LL, and CEC as identified in previous research.

QRF models of each gSSURGO variable all show high concentrations of large values in the southeastern portion of Colorado. Furthermore, very low values are concentrated over the Rocky Mountains, which are expected due to the large amounts of impervious rock characteristic of mountain ranges. These high concentration areas of low and high values are reflected in the final map derived from the PCA. High amounts of uncertainty in predicted attributes are also present in the southeastern quadrant, such as for claytotal_r and ll_r for the QRF model. This may indicate that further research may be needed to reduce the uncertainty of these attributes with soil swelling. One important implication of the present study is the demonstration that pedologically relevant information contained in soil polygon maps can be extracted and coupled with other sources of information to continuously predict swelling conditions using statistical models (i.e., Random Forest).

The presence of swelling soils can cause large economic losses dealing with infrastructure damage, so spatial information of soil swelling conditions provides insights to inform urban development strategies. Previous studies related to swelling soils within Colorado have focused on the urban Front Range, instead of a more comprehensive view of the entire state (Chabrillat et al., 2002; Noe, 1997). The average cost of building a single-family home in Colorado is \$1152/m^2 (Siniavskaia, 2014). There are 10,000 square meters in one hectare, meaning each hectare has a maximum potential economic loss of over \$11.5 million if swelling is not accounted for. Private home and land owners, as well as state contractors can use this map in order to determine if further site testing for swelling soils would be advisable.

Beyond the direct implication in urban growth, soil swelling also has important ecological/hydrological implications, regulating the amount of water in the soils and its capacity to infiltrate it to deeper layers or its availability for plant growth. Thus, soil swelling also affects land productivity and hydrological process such as aquifer recharge. Not taking swelling into account can cause errors in measurements of aquifer recharge, water storage, water/solute flux, drainage, and other hydrological variables (Kirby et al., 2003; Smiles, 2000). Large amounts of error in these variables can cause large uncertainties of available human water supplies.

This study was able to provide a complementary addition to gSSURGO lep_r maps by generating a continuous map that adjusts unrealistic spatial artifacts present in the original gSSURGO database. That said, this study has a coarser spatial resolution of 1 km, but it could be improved if higher resolution environmental covariates are available for the region of interest. As the original gSSURGO map is a model in itself, the results here cannot be better than the original model. Ground-truth data and information within soil swelling "hotspots" and across gSSURGO missing information will provide the means for a full validation and improvement of our approach.

Future work could include consideration of the clay mineralogy (as in Seybold & Libohova, 2017), as certain clay minerals swell more than others – e.g., smectite and illite, with 2:1 clay crystalline structures, swell more than kaolinite, a clay mineral with a 1:1 structure (Goetz, Chabrillat, & Lu, 2001; Yitagesu, van der Meer, & Werff van der, 2009; Kariuki et al., 2004; Taboada, 2004). In addition, spatially detailed trends in regional soil moisture should be taken into consideration, as changes in soil moisture are the absolute determinant in actual amount of swelling that takes place. The research framework can be implemented at a state level throughout the United States as a complement to the valuable gSSURGO lep_r state maps, and theoretically on a larger, country-wide scale as well.

5. Conclusion

This study provides a pedologically-driven DSM framework to represent the state-level information of soil swelling and provide insights about potential links between swelling and various environmental attributes. The framework can be used to gap-fill and generate continuous maps of swelling soils at the state level. Through this framework, the swelling potential across Colorado was able to be characterized,

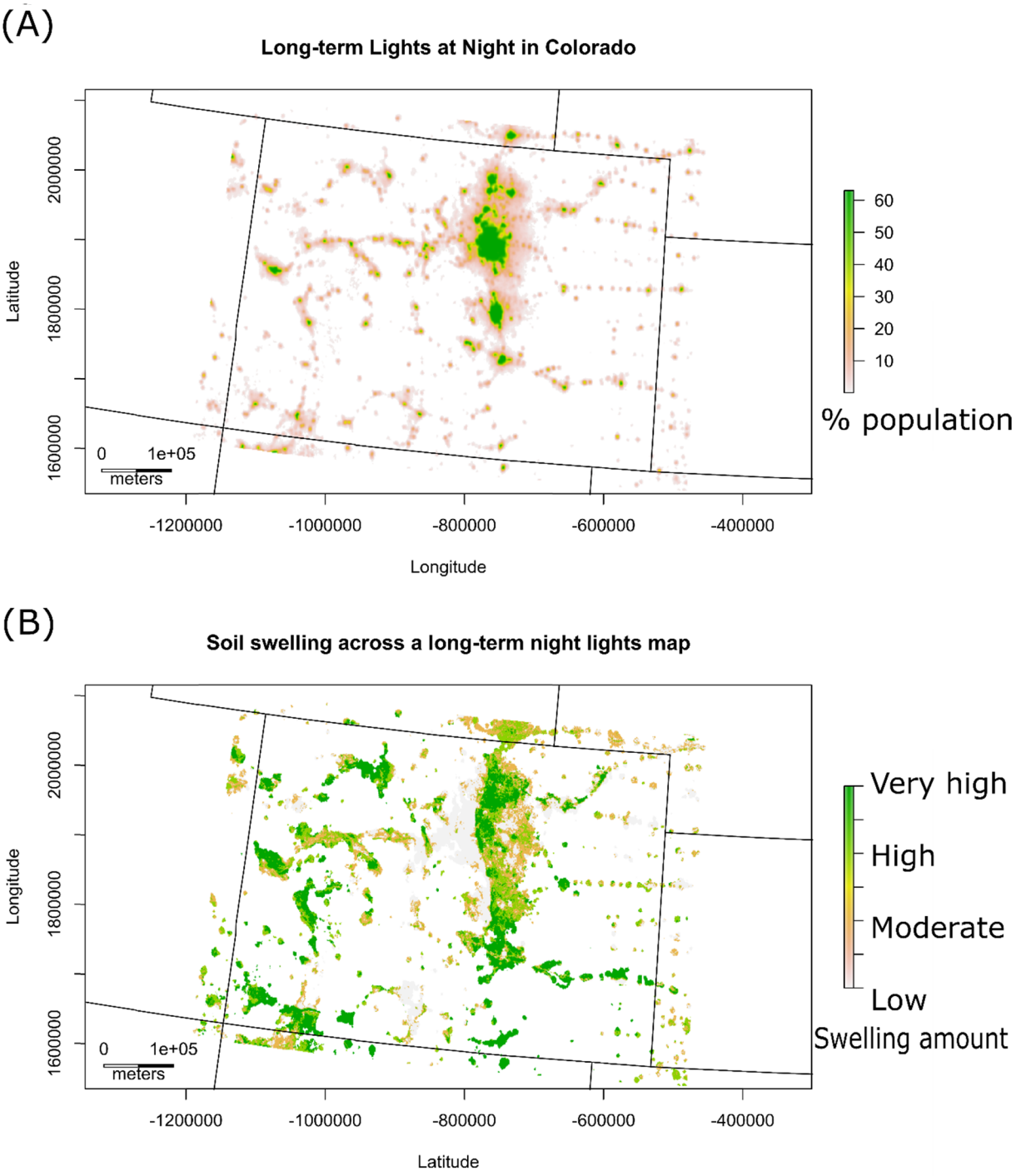

Fig. 7. Soil swelling across a long-term night lights map based on satellite imagery (ln1dms3a, from worldgrids.org). (A) Lights at night (ln1dms3a) over Colorado. This map is a surrogate of human activity and urban areas (population density); (B) ln1dms3a was overlaid with the swelling map and essentially used as a filter for the PCA model. The colors show the quantiles of the first principal component of the soil swelling integration. In dark green, the fourth quantile dominated areas are more vulnerable to swelling across areas dominated by human settlements and production activities representing nearly 20% of the total state land area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

removing unrealistic spatial artifacts of nationwide available products. A random forest machine learning approach was able to predict over 80% of spatial variability in swelling attributes. The areas with highest swelling risks were spread throughout the southeast, southwest, and western portions, with a large spatial match with urban development (> 20%). This information is important for both ecohydrological implications and to inform urban development. Prevention through planning, when applied strategically, can prevent ecological damages and economic losses by identifying suitable conditions or risks for urban development.

Acknowledgements

MG acknowledges partial support from a Conacyt fellowship from Mexico, and RV acknowledges support from the United States NSF (1724843).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.landurbplan.2019.103599.

References

Beven, K., & Germann, P. (1982). Macropores and water flow in soils. *Water Resources Research, 18*(5), 1311–1325. https://doi.org/10.1029/WR018i005p01311.

Birkeland, P. W., Shroba, R. R., Burns, S. F., Price, A. B., & Tonkin, P. J. (2003). Integrating soils and geomorphology in mountains—An example from the Front Range of Colorado. *Geomorphology, 55*(1), 329–344. https://doi.org/10.1016/S0169-555X(03)00148-X.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chabrillat, S., Goetz, A. F. H., Krosley, L., & Olsen, H. W. (2002). Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution. *Remote Sensing of Environment, 82*(2), 431–445. https://doi.org/10.1016/S0034-4257(02)00060-3.

Christidis, G. E. (1998). Physical and chemical properties of some bentonite deposits of Kimolos Island Greece. *Applied Clay Science, 13*(2), 79–98. https://doi.org/10.1016/S0169-1317(98)00023-4.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., ... Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development, 8*(7), 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.

Doesken, Nolan J., Pielke, Sr., Roger, A., & Bliss, Odilia A. P. (2003). *Climate of Colorado.* Colorado Climate Center, Atmospheric Science Department, Colorado State University Retrieved from http://ccc.atmos.colostate.edu/pdfs/climateofcoloradoNo.60.pdf.

United States Soil Conservation Service. (1976). General Soil Map Colorado. Map. 1:1500000. Retrieved from https://esdac.jrc.ec.europa.eu/images/Eudasm/US/us_18.jpg.

Goetz, A. F. H., Chabrillat, S., & Lu, Z. (2001). Field reflectance spectrometry for detection of swelling clays at construction sites. *Field Analytical Chemistry & Technology, 5*(3), 143–155. https://doi.org/10.1002/fact.1015.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma, 265*, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.

Houston, S. L., Dye, H. B., Zapata, C. E., Walsh, K. D., & Houston, W. N. (2011). Study of expansive soils and residential foundations on expansive soils in Arizona. *Journal of Performance of Constructed Facilities, 25*(1), 31–44. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000077.

Jayasekera, S., & Mohajerani, A. (2003). Some relationships between shrink-swell index, liquid limit, plasticity index, activity and free swell index. Retrieved from *Australian Geomechanics: News Journal of the Australian Geomechanics Society. 38*(2), 53–58. https://www.researchgate.net/publication/282820933_Some_relationships_between_shrink-swell_index_liquid_limit_plasticity_index_activity_and_free_swell_index.

Jones, D. E., & Holtz, W. G. (1973). Expansive soils – The hidden disaster. *Journal of Materials in Civil Engineering, 43*(8), 49–51 https://trid.trb.org/view/133235.

Jones, L. D., & Jefferson, Ian F. (2012). Expansive soils. In J. Burland, T. Chapman, & H. Skinner (Eds.). *ICE manual of geotechnical engineering: Volume 1, geotechnical engineering principles, problematic soils and site investigation* (pp. 413–441). ICE Publishing Retrieved from https://www.researchgate.net/publication/267764450_Expansive_soils.

Kariuki, P. C., Woldai, T., & Meer, F. V. D. (2004). Effectiveness of spectroscopy in identification of swelling indicator clay minerals. *International Journal of Remote Sensing, 25*(2), 455–469. https://doi.org/10.1080/0143116031000084314.

Kariuki, P. C., & van der Meer, F. (2004). A unified swelling potential index for expansive soils. *Engineering Geology, 72*(1), 1–8. https://doi.org/10.1016/S0013-7952(03)00159-5.

Kirby, J. M., Bernardi, A. L., Ringrose-Voase, A. J., Young, R., & Rose, H. (2003). Field swelling, shrinking, and water content change in a heavy clay soil. *Australian Journal of Soil Research, 41*(5), 963. https://doi.org/10.1071/SR02055.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift, 259–263*. https://doi.org/10.1127/0941-2948/2006/0130.

Kruse, F. A., & Elvidge, C. D. (2011). *Identifying and mapping night lights using imaging spectrometry.* IEEE1–6 https://doi.org/10.1109/AERO.2011.5747396.

Lane, P. W. (2002). Generalized linear models in soil science. *European Journal of Soil Science, 53*(2), 241–251. https://doi.org/10.1046/j.1365-2389.2002.00440.x.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Nayak, N. V., & Christenson, R. W. (1971). Swelling characteristics of compacted, expansive soils. *Clays and Clay Minerals, 19*(4), 251–261. https://doi.org/10.1346/CCMN.1971.0190406.

Noe, D. C. (1997). Heaving-bedrock hazards, mitigation, and land-use policy; Front Range Piedmont, Colorado. *Environmental Geosciences, 4*(2), 48–57 Retrieved from https://store.coloradogeologicalsurvey.org/product/heaving-bedrock-hazards-mitigation-land-use-policy-front-range-piedmont-colorado/.

Olive, W. W., Chleborad, A. F., Frahme, C. W., Schlocker, J., Schneider, R. R., & Schuster, R. L. (1989). Swelling clays map of the conterminous United States (USGS Numbered Series No. 1940). Retrieved from http://pubs.er.usgs.gov/publication/i1940.

Pruška, J., & Šedivý, M. (2015). Prediction of soil swelling parameters. *Procedia Earth and Planetary Science, 15*, 219–224. https://doi.org/10.1016/j.proeps.2015.08.052.

R Core Team (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing http://www.R-project.org/.

Reuter, H.I. & Hengl, T. (2012). Global Soil Information Facilities-Component Worldgrids.org. EGU General Assembly Conference Abstracts. Retrieved 9 September, 2018 from https://www.researchgate.net/publication/233540147_Global_Soil_Information_Facilities-Component_Worldgrids_org.

Römkens, M. J. M., & Prasad, S. N. (2006). Rain Infiltration into swelling/shrinking/cracking soils. *Agricultural Water Management, 86*(1), 196–205. https://doi.org/10.1016/j.agwat.2006.07.012.

Seybold, C. A., & Libohova, Z. (2017). Soil survey: Pedotransfer function of linear extensibility percent for soils of the United States. *Soil Science, 182*(1), 1. https://doi.org/10.1097/SS.0000000000000191.

Siniavskaia, Natalia. (2014). "Regional Differences in New Homes Started in 2013," Special Studies, National Association of Home Builders. Retrieved from https://www.nahb.org/en/research/housing-economics/special-studies/regional-differences-in-new-homes-started-in-2013.aspx.

Smiles, D. E. (2000). Hydrology of swelling soils: A review. *Australian Journal of Soil Research, 38*(3), 501. https://doi.org/10.1071/SR99098.

Soil Survey Staff (2016). *Gridded Soil Survey Geographic (gSSURGO) Database for Colorado.* United States Department of Agriculture, Natural Resources Conservation Service Retrieved from https://gdg.sc.egov.usda.gov/. (FY2016 official release).

Stewart, R. D., Najm, M. R. A., Rupp, D. E., & Selker, J. S. (2016). Modeling multidomain hydraulic properties of shrink-swell soils. *Water Resources Research, 52*(10), 7911–7930. https://doi.org/10.1002/2016WR019336.

Taboada, M.A. (2004). Soil shrinkage characteristics in swelling soils (INIS-XA-989). Skidmore, E.L. (Ed.). International Atomic Energy Agency (IAEA). Retrieved from http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/38/100/38100131.pdf.

Thomas, P. J., Baker, J. C., & Zelazny, L. W. (2000). An expansive soil index for predicting shrink-swell potential. *Soil Science Society of America Journal, 64*(1), 268–274. https://doi.org/10.2136/sssaj2000.641268x.

U.S. Department of Agriculture, Natural Resources Conservation Service. National soil survey handbook, title 430-VI. Retrieved September 7, 2018 from http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054242.

Vaught, R., Brye, K. R., & Miller, D. M. (2006). Relationships among coefficient of linear extensibility and clay fractions in expansive, stoney soils. *Soil Science Society of America Journal, 70*(6), 1983–1990. https://doi.org/10.2136/sssaj2006.0054.

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma, 291*, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017.

Wells, R. R., DiCarlo, D. A., Steenhuis, T. S., Parlange, J.-Y., Römkens, M. J. M., & Prasad, S. N. (2003). Infiltration and surface geometry features of a swelling soil following successive simulated rainstorms. *Soil Science Society of America Journal, 67*(5), 1344. https://doi.org/10.2136/sssaj2003.1344.

Yilmaz, I. (2006). Indirect estimation of the swelling percent and a new classification of soils depending on liquid limit and cation exchange capacity. *Engineering Geology, 85*(3), 295–301. https://doi.org/10.1016/j.enggeo.2006.02.005.

Yitagesu, F. A., van der Meer, F., & Werff van der, H. (2009). Prediction of volumetric shrinkage in expansive soils (role of remote sensing). *Advances in Geoscience and Remote Sensing.* https://doi.org/10.5772/8328.

Yoo, K., Amundson, R., Heimsath, A. M., & Dietrich, W. E. (2006). Spatial patterns of soil organic carbon on hillslopes: integrating geomorphic processes and the biological C cycle. *Geoderma, 130*(1), 47–65. https://doi.org/10.1016/j.geoderma.2005.01.