[American Journal of Science, Vol. 319, December, 2019, P. 819-845, DOI 10.2475/10.2019.01]

American Journal of Science

DECEMBER 2019

RECURSIVE PARTITIONING IMPROVES PALEOSOL PROXIES FOR RAINFALL

WILLIAM E. LUKENS*,**,***,†, GARY E. STINCHCOMB§, LEE C. NORDT***, DAVID J. KAHLE§§, STEVEN G. DRIESE***, and JACK D. TUBBS§§

ABSTRACT. The bulk elemental composition of soil subsurface (B) horizons is influenced by environmental, biological, geological, and climatic factors. Because fossil soils (paleosols) are common in the geologic record, quantitative models that link climate to paleosol geochemistry are highly desirable in the paleoclimate community. Error associated with these models is typically reported as the root mean square error (RMSE) of a regression analysis and reflects the variability imparted by non-climatic influences on soil formation and the uncertainty associated with model fitting. However, for prediction purposes, the RMSE is well known to underestimate model uncertainty. In this work we re-evaluate a widely used transfer function for mean annual precipitation (MAP) based on the chemical index of alteration minus potassium (CIA-K) using data science best practices on two continental-scale soil data sets. Data set inter-comparisons and cross-validation of exponential regression models indicate that the root mean square prediction error (RMSPE) between CIA-K and MAP for soils representative of climates across the continental United States is around 299 mm, significantly higher than the currently accepted 182 mm RMSE. Further, CIA-K is unable to predict perhumid (>2000 mm MAP) climate zones. We show that transitioning from a simple regression framework to one of recursive partitioning via random forests can significantly increase prediction accuracy while automating variable selection. We introduce two new, widely applicable random forest models for MAP (RF-MAP) that use 10 elemental oxides as input variables and were calibrated on the Baylor University Soil Informatics (BU-SI) data set. RF-MAP version 1.0 (RF-MAP$_{1.0}$) was generated using the entire BU-SI data set (n = 685) and can predict MAP values up to 6865 mm with a RMSPE of 395 mm. RF-MAP version 2.0 (RF-MAP$_{2.0}$) was generated using a modification of the BU-SI data set (n = 642) and can predict MAP values up to ~1600 mm with a RMSPE of 209 mm. Pruned regression trees provide insight into the mechanisms driving the random forest models and demonstrate the first empirical confirmation of the sensitivity of soil elemental responses to global climate zones. The RF-MAP$_{1.0}$ and RF-MAP$_{2.0}$ models predict MAP values comparable to independent proxy estimates for a range of deep-time paleosols. We advocate for application of RF-MAP$_{1.0}$ in settings where no *a priori* information on paleoclimate is available, and encourage the use of either RF-MAP$_{1.0}$ or RF-MAP$_{2.0}$ if users have independent constraints that paleo-MAP was below 1600 mm.

Keywords: paleosol, paleoclimate proxy, rainfall, machine learning, weathering, soil

* Department of Geology and Environmental Science, James Madison University, Harrisonburg, Virginia 22807

** School of Geosciences, University of Louisiana at Lafayette, Lafayette, Louisiana 70504

*** Terrestrial Paleoclimatology Research Group, Department of Geosciences, Baylor University, Waco, Texas 76798

§ Watershed Studies Institute & Department of Earth and Environmental Sciences, Murray State University, Murray, Kentucky 42071

§§ Department of Statistical Science, Baylor University, Waco, Texas 76798

† Corresponding author: lukenswe@jmu.edu

INTRODUCTION

Soils form in open communication with local climate state, and therefore paleosols (fossil soils) act as *in situ* archives of past climates. In recent decades, a growing number of studies have focused on modeling relationships between soils and climate to develop predictive paleoclimate models that can be applied on paleosols (for example, Stiles and others, 2001; Sheldon and others, 2002; Retallack, 2005; Nordt and others, 2006; Sheldon, 2006; Cleveland and others, 2008; Retallack and Huang, 2010; Nordt and Driese, 2010; Óskarsson and others, 2012; Gallagher and Sheldon, 2013; Hyland and others, 2015; Nordt and others, 2015; Stinchcomb and others, 2016). Many of these studies have found robust relationships between the bulk elemental composition of subsurface (B) illuvial horizons and climate by regressing weathering indices (elemental ratios) against mean annual precipitation (MAP) or temperature (MAT). The resulting pedotransfer functions are now routinely applied throughout the geologic record (for example, Prochnow and others, 2006; Hamer and others, 2007; Retallack, 2008a; Adams and others, 2011; Secord and others, 2012; Hyland and Sheldon, 2013; Myers and others, 2014; Beverly and others, 2015; Liivamägi and others, 2015; Nordt and others, 2015; Driese and Ashley, 2016; Sheldon and others, 2016; Lukens and others, 2017a, 2017b).

The error associated with these pedotransfer functions is derived from three sources: (1) analytical uncertainty in measuring predictor variables (that is, elemental composition), (2) natural variability in the relationship between soil properties and climate, which is derived from other (non-climatic) influences on pedogenesis, and (3) uncertainty resulting from statistical methods used in proxy development, including modeling techniques and the type of data included in training data sets. For soil bulk elemental composition, modern analytical methods have minimized the influence of (1) above, with recent analyses accurate to <0.1 weight percent for most elements. Minimizing the error of statistical analyses is therefore essential to understanding the uncertainty in soil-climate relationships, and by extension, directly impacts the fidelity of paleoclimate reconstructions from paleosols.

Differences in calibration data sets and modeling techniques among studies have generated numerous functions that vary greatly in regression metrics (table 1; see table 2 for statistical terms). In general, adding unbiased data to an existing data set should decrease error and increase the proportion of explained variance of a model (for example, Retallack, 1994 vs. Retallack, 2005 in table 1). However, as we will demonstrate, adding data that was not represented in a small data set will increase the overall variance in the relationship between a predictor and response variable, resulting in higher error and lower explained variance for a regression model. When this effect is observed, the resulting model is likely more robust and more widely applicable, as it incorporates real-world variance and reduces bias from limited sampling.

Modeling methodology—which includes data pre-processing, strategy (for example, linear regression, spline fitting, regression trees, *et cetera*), and validation steps—has varied between paleoclimate proxies and thereby inhibits true comparison between models and adds confusion to model selection and application. Currently no standard method exists for development and testing of pedotransfer functions for climate, leaving researchers to search for correlations that may be specific to their data sets. These efforts have led to the selection of predictor variables and regression functions that reflect the composition of any one data set, rather than more universal processes. In statistical terms, modeling methods that focus exclusively on the dataset at hand result in overfit models and provide overconfident estimates for future scenarios. For example, the CALMAG proxy for Vertisols was calibrated on a climosequence where parent material composition and MAT were held constant (Nordt and Driese, 2010). This theoretically restricts the application of CALMAG to paleosols

TABLE 1

A selection of existing pedotransfer functions for climate

Model	Predictor variable	Response variable	Training set size	Regression type	Regression r^2	RMSE	Reference[§]
Mean annual precipitation							
Morphology	Depth to Bk*	MAP	26	linear	0.75	170	1
Morphology	Depth to Bk	MAP	106	linear	0.81	220	2, 3
Morphology	Depth to Bk	MAP	317	quadratic	0.78	330	3
Morphology	Depth to Bk	MAP	807	quadratic	0.52	147	4
Geochemistry	CALMAG	MAP	14	linear	0.90	108	5
Geochemistry	CIA-K	MAP	126	exponential	0.72	182	6
Geochemistry	CIA-K	MAP	479	exponential	0.27	564	This study
Mean annual temperature							
Geochemistry	CIW	MAT	36	linear	0.81	0.5[†]	7
Geochemistry	NaK	MAT	126	linear	0.37	4.4	6
Geochemistry	PWI	MAT	158	logarithmic	0.57	2.1	8

Note: Model error is root mean square error (RMSE) of regression functions, given in mm for mean annual precipitation and °C for mean annual temperature.

*Depth in profile to the occurrence of nodular carbonates (Bk horizon).

[†] Value reported by Retallack, 2018.

[§] References: 1) Arkley (1963), 2) Jenny (1941), 3) Retallack (1994), 4) Retallack (2005), 5) Nordt and Driese (2010), 6) Sheldon and others (2002), 7) Óskarsson and others (2012), 8) Gallagher and Sheldon (2013).

forming on similar parent materials and under similar temperature regimes, the latter of which is typically poorly constrained in paleosol successions. However, it is possible that the CALMAG pedotransfer function tracks similar processes that govern other MAP proxies (for example, CIA-K, depth to Bk; table 1). If this is true, the regression line drawn through the CALMAG training data set would not capture the full variability in the soil-climate relationship because of the limitations in parent material composition and MAT. Finally, paleosol proxy development studies commonly report root mean square error (RMSE) from regression as a measure of predictive capacity. The RMSE describes the average variability of training data about a regression function (table 2) and is well known to be an underestimate of error when models are applied on external data for prediction purposes (for example, Shmueli, 2010). Incorporation of out-of-sample model assessment strategies such as cross-validation is a necessary step in paleosol proxy development that few researchers have yet to incorporate, though it is commonplace in most scientific disciplines (for example, Birks and Birks, 2006; Mitchell and others, 2013; Sharma and others, 2014).

Multivariate techniques offer a more practical solution for modeling climate using soil geochemistry, as many soil-forming processes are decidedly multifactorial, nonlinear, and complex (for example, Chadwick and Chorover, 2001; Slessarev and others, 2016; Rasmussen and others, 2018; Lukens and others, 2018). Another reason multivariate and dimension-reduction techniques are useful is because many of the analyzed elements in soils covary. For example, the base oxides (CaO, MgO, K_2O, Na_2O) tend to accumulate in arid environments and leach from soils in humid environments. Residual enrichment of refractory metals (TiO_2, ZrO_2) and Al- and Fe-oxides (Al_2O_3, Fe_2O_3) occurs in highly weathered soils in warm-wet climates. Modeling these compounds in isolation ignores a substantial amount of well-known interdependence and statistically manifests in nuanced ways, in some cases underutilizing data and in others over-drawing conclusions from them. In systems where predictors (for example, elemental oxides noted above) covary, multivariate models can optimize the predictive capacity of those variables, thereby accessing information

TABLE 2

Definitions of statistical terms

Term	Abbreviation	Explanation
Training data set	N.A.	Data set used to calibrate a model; the set that contains all internal data for a given model.
Testing data set	N.A.	External data set used to calculate a model's prediction error; excludes training data.
Cross-validation	N.A.	The process of calculating prediction error by applying a calibrated model on a testing data set.
Correlation coefficient	r	The strength of a linear relationship between two variables.
Coefficient of determination	r^2, R^2	The proportion of the variance in the dependent variable explained by the independent variable. Lowercase is used for linear correlations with two variables (for example, linear regression lines), whereas uppercase is used for nonlinear functions and correlations with more than two variables.
Mean square error	MSE	Model error measured using internal data; the average of the squared differences between the predictor (x) and response variable (y)
Root mean square error	RMSE	Model error measured using internal data; the square root of MSE.
Root mean square prediction error	RMSPE	Model error measured using cross-validation on testing data; calculated as the RMSE between observed (y) and model-fitted values ($\hat{y}$) using external data. See text for equation.
Sum of squares	SS	The sum of the squared differences of each observation from the overall mean value, calculated across all observations. See text for equation.

Note: NA = not applicable.

typically overlooked in simple linear regression. This approach was taken by two recent studies using the Baylor University Soil Informatics (BU-SI) data set (Nordt and Driese, 2013). Stinchcomb and others (2016) developed the paleosol-paleoclimate model version 1.0 ($PPM_{1.0}$) using a partial least squares regression (PLSR) and thin-plate spline to model 11 elemental oxides for simultaneous prediction of MAP and MAT. Lukens and others (2018) used principal components analysis (PCA) to study correlations between soil pH, elemental oxide groupings, and soil-forming factors to elucidate the response of elemental constituents to climate via soil pH. These results demonstrate that climatic information is stored in a wide array of elemental constituents, and that using large, diverse data sets of soils can result in robust predictive models.

Recursive partitioning is a machine learning technique that commonly outperforms other multivariate approaches and therefore may improve climate estimations for soils and paleosols beyond the capability of current models (Prasad and others, 2006; Cutler and others, 2007; Oliveira and others, 2012; Rodriguez-Galiano and others, 2012). In contrast to PLSR, recursive partitioning does not create linear combinations of predictors and does not weight response variables. Instead, recursive partitioning iteratively divides a data set based on a single predictor (for example, elemental oxide) that maximally reduces the variance in the response (for example, MAP). By repeatedly and exhaustively subdividing a soil-climate data set, recursive partitioning can access subsets of the data that may have unique relationships between predictors and a response that are not apparent across the entire range of observations. We hypothesize that regression via recursive partitioning will reduce the uncertainty in the soil-climate relationship and will form a more flexible and robust model than univariate and PLSR methods.

In this paper, we review and test the role that data set size and composition have on soil-climate models. We then re-examine standard regression approaches to developing paleoclimate models, and for simplicity, we focus our efforts on the widely-applied relationship between CIA-K and MAP (Sheldon and others, 2002) that has been used in paleopedology for over 15 years (for example, Prochnow and others, 2006; Hamer and others, 2007; Retallack, 2008a; Gutierrez and Sheldon, 2012; Atchley and others, 2013; Hyland and Sheldon, 2013; Liivamägi and others, 2015; Sheldon and others, 2016; Lukens and others, 2017a; Driese and others, 2018; Liutkus-Pierce and others, 2019; Lukens and others, 2019). We then introduce strategies from the statistical machine learning community, recursive partitioning and random forests (RF), as a new method for paleosol proxy development. Finally, the random forest models are tested using three deep-time paleosol applications to compare model results to the CIA-K pedotransfer function and the $PPM_{1.0}$.

BACKGROUND

Continental Soil Data Sets

The performance of predictive models is directly linked to training data set composition and modeling technique. Sheldon and others (2002) utilized the Marbut (1935) data set for development of MAP and MAT proxies using elemental weathering indices. Since then, the Marbut (1935) data set has become a standard training set for paleosol proxy development (Gulbranson and others, 2011; Gallagher and Sheldon, 2013; Passchier and others, 2013; Nordt and others, 2015). A more recent soil and climate data set (BU-SI) was developed to replace the Marbut (1935) data set. It is therefore useful to provide a brief overview of the motivations behind the compilation of the Marbut (1935) and BU-SI data sets.

In the early 20th century, Curtis Marbut developed a system of soil classification that relied heavily on earlier systems pioneered by Russian soil scientists (for example, Sibirtsev, 1895, 1966; Glinka, 1914). Their philosophy focused on environmental (for example, climate, biota, topography) rather than geological (bedrock composition) influences on pedogenesis (Marbut, 1921, 1928; Baldwin and others, 1938; Paton and Humphreys, 2007; Soil Science Division Staff, 2017). A key underpinning of Marbut's perspective on soil development was that all soils, given the right conditions, will become mature (that is, well-drained and leached) and that only mature soils are truly mappable across large geographic areas ("zonal" soils). For example, Marbut largely overlooked Vertisols as an important soil product because he viewed them as immature soils due to their association with marl bedrock (Marbut, 1928). Overall, Marbut's perspectives likely biased the collection of laboratory samples for his national data set of soils toward those that met his criteria for soil maturity, and therefore excluded many soils that are now understood to be common in the geologic record, including alluvial soils and soils with variable drainage conditions (Sheldon and Tabor, 2009; Atchley and others, 2013).

Sheldon and others (2002) revisited the Marbut (1935) geochemical and environmental factor data and combined it with other data (vegetation, topography, parent material, and duration of soil formation estimates) to develop the CIA-K pedotransfer function for MAP, which has been widely used in the literature and has lead the way for subsequent paleoclimate proxy development. Of importance to this study is the observation that 20 of the 126 soils compiled in the Marbut (1935) data set have pedogenic carbonate based on reported carbonate measurements and high CaO contents. Although this is not noted in Sheldon and others (2002), the presence of carbonate suggests that the CIA-K climofunction is not strictly a feldspar-weathering

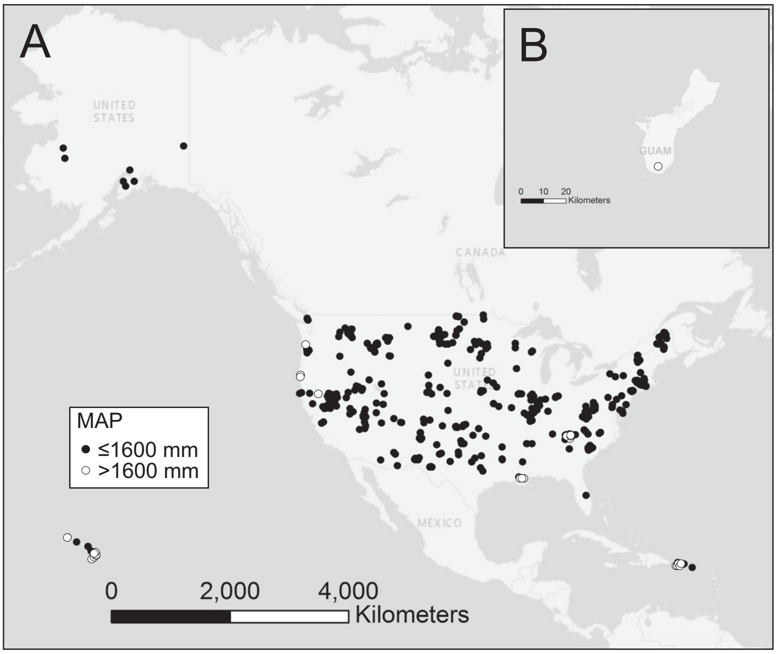

Fig. 1. Map of modern soils included in the Baylor University Soil Informatics (BU-SI) data set.

index, as it was originally designed to be (Nesbitt and Young, 1982; Harnois, 1988; Maynard, 1992). Most soils from the Marbut (1935) data set are from the Mollisol, Inceptisol, Ultisol and Alfisol orders, with a few Aridisols and Spodosols and one Vertisol. The data set does not include Oxisols, Gelisols, Andisols, Entisols, and Histosols. MAP values in the Marbut (1935) data set range from 179 to 1564 mm, with a median of 1035.

In an effort to enhance the study of deep-time Critical Zones, the Baylor University Soil Informatics (BU-SI) data set was introduced (Nordt and Driese, 2013). This data set consists of physical, chemical, biological, and mineralogical data from modern soils, pooled from the National Soil Information System (NASIS) database, managed by the Natural Resources Conservation Service (NRCS). In addition to these soil data, climate, land cover, topography, parent material, and duration of soil formation estimates were aggregated from various sources (See table 1 in Stinchcomb and others, 2016). The BU-SI data set does not include samples from the Marbut (1935) data set.

The BU-SI was modified whereby uppermost B horizons were selected from all soils in the data set, resulting in 685 records (fig. 1; Stinchcomb and others, 2016). Although the complete BU-SI data set has been reported elsewhere (Nordt and Driese, 2013) and contains nearly 6000 records of most soil horizon types, we focus on the 685 uppermost B horizons and simply refer to it hereafter as the BU-SI data set. These B horizons were used to build the $PPM_{1.0}$ model for MAP and MAT (Stinchcomb and others, 2016). Unlike the Marbut (1935) data set, soils with variable states of maturity and drainage ("azonal" soils) were not removed. This is an important distinction because deep-time paleosols that are commonly found in the sedimentary record meet these criteria (for example, Vertisols and Inceptisols that weathered mixed alluvium). The uppermost B horizon was chosen because (1) A horizons are rarely recognized in the rock record and (2) the geochemistry of B horizons more closely reflects the pedogenic state than underlying subsoil horizons that may have more parent material contributions. Gelisols and Histosols are not included in the uppermost B horizon subset of the BU-SI data set. Most B horizons consist of some type of Bt or Bw horizon (including Btk, Btss, *et cetera*). MAP values in the BU-SI data set range from 130 to 6866 mm, with a median of 1006. Further information regarding these samples can be found in Stinchcomb and others (2016) and Lukens and others (2018).

Modeling the Soil-Climate Relationship

The CIA-K Proxy for MAP.—The CIA-K weathering index was originally developed by Nesbitt and Young (1982) as a modification of the chemical index of alteration (CIA):

$$\mathrm{CIA} = 100 \times [\mathrm{Al_2O_3}/(\mathrm{Al_2O_3} + \mathrm{CaO} + \mathrm{Na_2O} + \mathrm{K_2O})], \tag{1}$$

where oxides are in molar percent (that is, weight percent divided by molecular mass). Potassium was omitted from the index by Harnois (1988) and dubbed the chemical index of weathering (CIW); however, Maynard (1992) argued that CIW should be termed CIA-K to avoid genetic interpretations of weathering state. Thus:

$$\mathrm{CIA\text{-}K} = 100 \times [\mathrm{Al_2O_3}/(\mathrm{Al_2O_3} + \mathrm{CaO} + \mathrm{Na_2O})]. \tag{2}$$

The CIA-K index was conceived as a feldspar weathering proxy by Nesbitt and Young (1982), with values trending toward 100 with increasing hydrolysis and leaching of base oxides, and toward 0 with increasing carbonate and/or apatite content. Initially, only silicate-bound CaO (termed CaO*) was used in these indices to remove secondary (soil-formed) carbonate from primary mineral compositions. However, standard characterization of soils includes measurement of total elemental composition on the < 2 mm grain size fraction, and therefore includes carbonate-bound CaO in soil data sets (Soil Survey Laboratory Staff, 1992). The CIA-K index is therefore useful for tracking soil-climate relationships due to the inverse relationship between calcium carbonate content of soil B horizons and mean annual rainfall amount—dryland soils tend to be alkaline and contain higher concentrations of calcite, whereas wetter environments tend to have acidic soils (Sheldon and others, 2002; Lukens and others, 2018). The inclusion of $\mathrm{Na_2O}$ in the weathering index should add to the correlation between CIA-K and MAP because Na leaching is dependent on climate (Dere and others, 2013), and Na-salts accumulate in desert soils (Amit and others, 1993).

Sheldon and others (2002) later developed regression functions that relate CIA-K to MAP using the Marbut (1935) data set. The maximum value for CIA-K is 100, which corresponds to a soil buffered by Al-oxyhydroxides (pH $\sim$ 4–5) with kaolinite clay and few base cations in the exchange complex. Such soils are common at MAP values beginning at around 1600 mm, which is approximately the maximum value in the Marbut (1935) data set. This effectively emplaces an asymptote at CIA-K = 100; therefore, an exponential function is preferred over a linear regression (Sheldon and others, 2002). In developing the pedotransfer functions for climate, Sheldon and others (2002) reported that numerous possible MAP predictors were sought in regression analysis, including different weathering ratios, ratios between horizons, depth functions, and horizon types. This undoubtedly required much time, and current and future researchers would benefit from an automated approach to such efforts.

Partial least squares regression and thin-plate spline for MAP and MAT.—The $\mathrm{PPM_{1.0}}$ was developed by applying a combined PLSR and a thin-plate spline modeling approach on the 685 uppermost B horizons from the BU-SI data set (Stinchcomb and others, 2016). The logic for choosing this modeling approach was based on several reasons and this is discussed in the original work. In general, however, the approach for building the $PPM_{1.0}$ was based on the well-known observation that chemical weathering is the addition, loss, transformation and translocation of elements within a soil that involves several linear and non-linear processes operating simultaneously at different rates. Thus, the relation between weathering and climate is complex and should be modeled in a similar manner.

The PLSR was first performed to reduce the dimensions of the 11 oxides measured for each B horizon sample. This dimension reduction was likened to the approach of developing oxide weathering ratios (for example, CIA-K), which also reduce the dimensions of the factors by creating a single metric, and is one way of attempting automated variable selection. The PLSR resulted in four regressors, R_1-R_4, that correlated MAP and MAT with the oxides. R_1, the strongest regressor, was inferred to track base loss, desilication and residual enrichment of elements with increasing MAT and MAP. R_2 was thought to relate MAT to temperature-dependent dissolution of Na- and K-bearing minerals. R_3 was thought to relate increasing MAP to decalcification and the retention of Si in less humid environments. R_4 was thought to relate MAP to Mg-retention in mafic-rich parent materials.

Climate (MAT and MAP) was modeled as a joint response on R_1 through R_4 using a semiparametric thin-plate spline:

$$y_{(MAP,\ MAT)} = f(R_1,...\mathrm{R}_4) + \varepsilon_i,\ i = 1,...685, \tag{3}$$

where $y_{(MAP,\ MAT)}$ are the combined climate responses MAP and MAT; $f(R_1,...R_4)$ is the non-parametric smoothing function of the model that relies on the R_1-R_4 as the smoothing variables; and ε_i is the independent, zero-mean random errors from the training set, i. Simulations showed that the $PPM_{1.0}$ has a root mean square prediction error (RMSPE) of 512 mm for MAP and 3.98 °C for MAT. Unlike CIA-K, this model does not address issues of diagenesis (K_2O metasomatism or illitization). Given the large prediction error, the developers suggested using the model for major climate transitions marked by large changes in MAP and MAT (Stinchcomb and others, 2016). One advantage of using a multivariate thin-plate spline approach, as in the $PPM_{1.0}$, is that more than one set of predictor values could theoretically yield the same MAP and MAT. This notion is realistic, as soils can weather differently under the same climate.

Recursive Partitioning

Recursive partitioning is a nonparametric modeling technique that has been widely used across the sciences for decades (Breiman, [1984] 2017; Friedl and Brodley, 1997; Rawls and Pachepsky, 2002; Pachepsky and others, 2006). The models are constructed by selecting an individual variable that splits the observations into daughter groups such that the variance in the response variable (y) in each daughter group is less than that of the parent group (fig. 2A). The process is then separately repeated on each daughter group, drawing from the same list of possible predictors that were attempted on the previous split. Each daughter group is termed a "node" and is defined by some criterion, which, in the case of soil chemistry, may be a threshold concentration of a given elemental oxide. Recursive partitioning continues until a stopping criterion is met, typically a minimum number of observations in each terminal node. In the end, the model structure takes the form of an inverted tree, wherein branches emanate from nodes and the final predictions exist as the "leaves." Because the model is typically grown to be overly complex, a cross-validation procedure is necessary to "prune" and simplify the tree, which seeks a balance between optimizing predictive power and preventing overfitting.

Random forest (RF) is a machine learning ensemble algorithm based on the concepts of recursive partitioning and bootstrap aggregation, or bagging (Breiman,

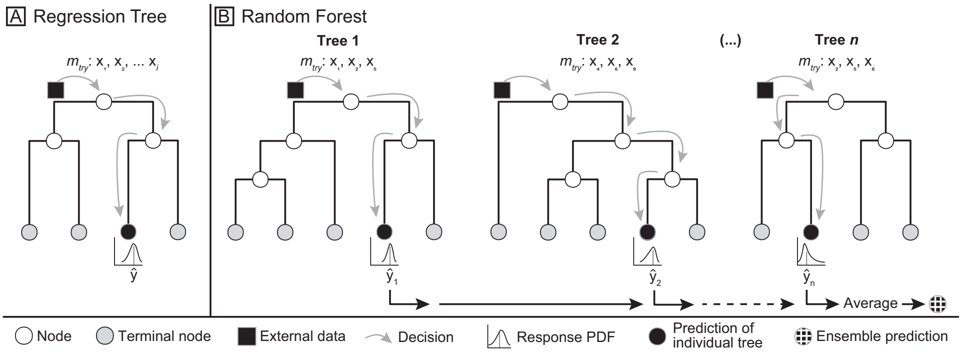

Fig. 2. Schematic diagrams of recursive partitioning models. For each tree model, external data are fed through a series of nodes to reach a prediction. (A) Regression tree; note that the response ($\hat{y}$) is the mean of the data distribution in a terminal node. (B) Random forests consist of an ensemble of trees, each constructed with a bootstrapped sample and a subset of predictors (m_{try}). The ensemble prediction is the average of responses from all trees. Modified after Fig. 4.1 in Criminisi and others (2011). PDF = probability density function.

1996, 2001). As in bagging, an RF model is constructed by growing a large number of regression trees that are each generated on bootstrapped samples of a training set; however, RF is unique in that each tree only draws on a subset of available predictor variables, typically set as the square root of the total available predictors (Friedman and others, 2001). This technique is particularly useful for relationships that are driven by primarily one strong predictor (for example, the CaO-MAP correlation in soils, Lukens and others, 2018) because the RF model omits any specific predictor in roughly 70 percent of the trees that comprise the RF (Friedman and others, 2001). This greatly improves RF performance over any single regression tree and reduces bias in the overall model. Finally, the RF prediction is the average response of all fitted trees in the forest, thereby reducing variance in the model (Hastie and others, 2009).

For soil and paleosol proxy development, recursive partitioning offers numerous benefits over simple regression. The first advantage is that the training data set is subdivided into groups that behave independently. This decreases the need for stratifying data sets prior to analysis—for example, by soil type (Nordt and Driese, 2010; Gallagher and Sheldon, 2013)—because the splitting algorithm will empirically divide the data set to improve homogeneity in the response variable across nodes. This is particularly useful for soils, as they form as a function of many interdependent state factors (Jenny [1941] 1994) and independent pedogenic pathways can potentially create similar weathering products (Holliday, 2004). Recursive partitioning therefore provides a straightforward solution to the endless possibilities of subdividing data sets prior to modeling.

The second advantage of tree-based models is that of variable selection and data transformation. Elemental oxides are typically transformed into molecular weathering ratios to improve regression statistics, but the oxides that are included in such ratios are left to the developer to decide. For example, Al_2O_3 is typically used as a recalcitrant oxide that normalizes mobile oxide concentrations. However, other oxides that behave like Al_2O_3 may be just as useful in simple regressions (for example, Fe_2O_3 and TiO_2, Lukens and others, 2018). Because regression trees can draw on all input variables independently when creating each node, more flexibility exists in predictor variable selection for different areas of the data set. The result is that variables that have little significance across the entire data set can become effective predictors for subsets of the data. For example, the CIA-K proxy for MAP (Sheldon and others, 2002) uses CaO and

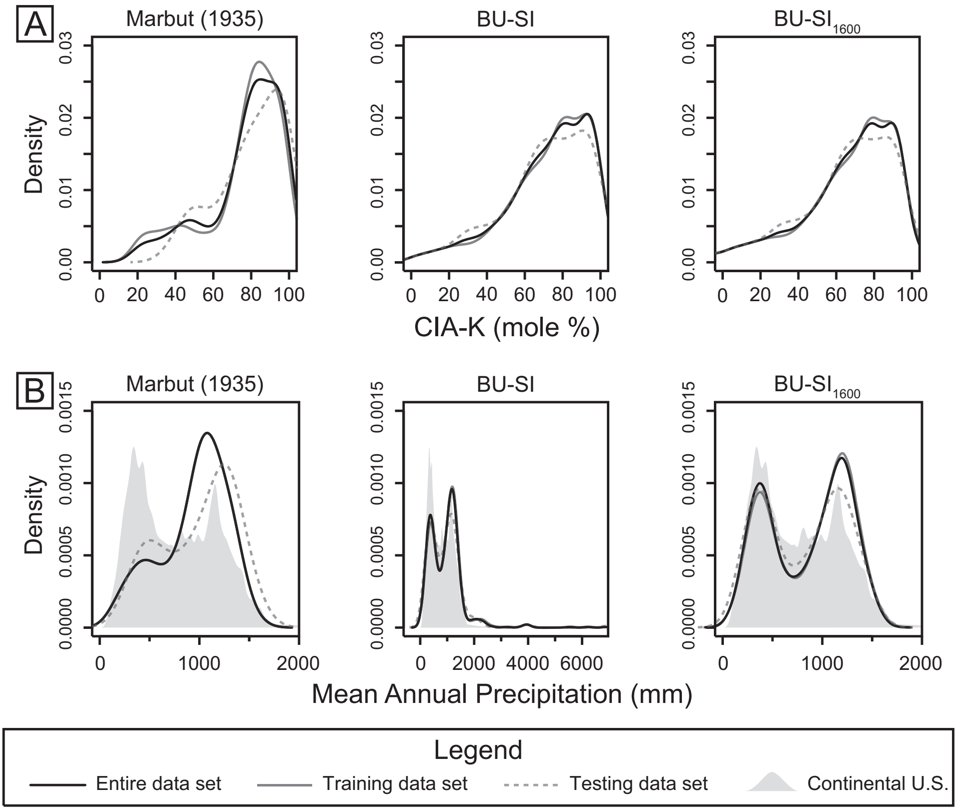

Fig. 3. Kernel density functions for (A) CIA-K values and (B) mean annual precipitation, across each of the three models, with training and testing data subsets indicated. Data for gray shaded area is the distribution of MAP for the continental USA from the 1981 to 2010 observation period (Daly and others, 1994). Note: for MAP distributions, training data (solid gray lines) overlap with entire data sets (black lines).

Na_2O as predictors, whereas the CALMAG proxy (Nordt and Driese, 2010) substitutes MgO for Na_2O. It is possible that Na_2O and MgO concentrations have predictive capacity for MAP but in different subsets of the data that correspond to unique soil-forming environments. The recursive partitioning approach works to detect these subsets of data by drawing on all possible predictors when creating splits in each subsequent node for every tree. Inclusion of either Na_2O or MgO in regressions of unstratified data sets could overlook their power in specific scenarios.

MATERIALS AND METHODS

Data Sets

In order to directly compare the BU-SI and Marbut (1935) data sets, a third data set (BU-SI_{1600}) was generated by selecting all observations from the BU-SI data with MAP less than 1600 mm (fig. 3). We used reported elemental oxide concentrations from the BU-SI and Marbut (1935) data sets, including Al_2O_3, CaO, Fe_2O_3 (total Fe), K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2, and TiO_2. We exclude ZrO_2 because it is not commonly measured on paleosols. These ten oxides were originally analyzed on the < 2 mm grain size fraction (Marbut, 1935; Stinchcomb and others, 2016, and references therein). MAP values in the BU-SI_{1600} data set range from 130 to 1583 mm, with a median of 957. Only one value (1583 mm) is above the highest value (1564) in the Marbut (1935) data set. We note that the median MAP values are similar for the Marbut (1935), BU-SI, and BU-SI$_{1600}$ data sets.

A review of the BU-SI data set published as supplementary data from Stinchcomb and others (2016) (http://earth.geology.yale.edu/%7eajs/SupplementaryData/2019/Lukens) shows four pedons with incorrect MAP values likely due to prior transcription errors. These MAP values were identified during ongoing model development efforts and corrected using the most recently available PRISM data (table S2). The MAP values were changed as follows: Pedon 02N0081, MAP was reduced from 3349 mm to 293 mm; Pedon 04N0501, MAP was reduced from 3349 mm to 458 mm; Pedon 04N0498, MAP was reduced from 3349 mm to 455 mm; and Pedon 02N0081, MAP was reduced from 4225 mm to 3206 mm.

Modeling Methods

All statistical analyses were performed in JMP version 14 (*JMP*, 2019) and RStudio (R Core Team, 2019). Each data set was divided into a training and testing set using a random 70/30 split for exponential regressions and regression trees; however, entire data sets were used for random forest model development.

Exponential regression.—Exponential regression analyses were run in JMP using CIA-K (mole %) as a predictor and MAP as a response. The regression function follows the form:

$$\mathrm{MAP} = \mathrm{a}e^{\mathrm{b(CIA\text{-}K)}} \tag{4}$$

where a is the scaling factor (y intercept), *b* is the growth rate, and *e* is Euler's number (2.718. . .).

Regression trees.—Individual regression trees offer insight into the decisions driving random forest models, which are otherwise "grey box" models. We therefore carefully grew and then pruned back regression trees for each of the data sets to understand the relationships between elemental inputs and MAP responses. Regression trees were generated using the rpart package (Therneau and others, 2018a), which implements most of the original recursive partitioning functions published by Breiman and others (1984). Input variables for each model included 10 commonly measured major and minor oxides (in wt. %): Fe_2O_3, Al_2O_3, SiO_2, TiO_2, CaO, MgO, K_2O, Na_2O, MnO, and P_2O_5. We omit ZrO_2 from analyses because it is not commonly measured on paleosols and showed little predictive capacity for MAP in early versions of regression tree models.

Splitting was performed using the analysis of variance (ANOVA) method (Therneau and others, 2018b). The splitting procedure first calculates the total sum of squares (SS_T) for the starting group (the node), which is the squared differences of each observation from the overall mean, summed across all observations:

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5}$$

where y is the MAP value for each sample i, $\bar{y}$ is the average MAP value in the group, and n is the number of observations in the group. The splitting criterion is then selected to maximize between-group variance ($SS_L + SS_R$, where L is the left group and R is the right group) (Therneau and others, 2018b).

The tradeoff between tree size and error is tuned using the complexity parameter (*cp*). As a regression tree is grown, splits are generated only if the model R^2 value increases by a factor equal to the *cp* value, which essentially serves as a pre-pruning step prior to cross-validation. For each model, two stopping criteria were used: 1) a split would be attempted if a minimum of 20 observations were present in a node, and 2) the minimum leaf (terminal node) size was one third of this minimum value (20/3 ≈ 7). Thus, even if 20 observations were available for node splitting, if either of the resultant leaves would have less than seven observations the split would be abandoned. The *rpart* package developers note that very small nodes are typically pruned away in cross-validation, so these default parameters save computational time (Therneau and others, 2018b, p. 24). For any tree, the reported values at each node and leaf are the means of the group. Cross-validation is necessary in order to estimate the prediction error associated with the regression tree models (Hastie and others, 2009).

Regression trees were pruned using a *k*-fold cross-validation procedure ($k = 10$). The training data set was split into 10 equal subdivisions and a tree was fit on each fold. The *cp* value associated with the tree that had the lowest root mean square prediction error (RMSPE) was then selected as the *cp* for a final, pruned regression tree grown using the entire training data set. For each tree, relative cross-validated error was analyzed as a function of *cp* value, with the expectation that error decreases with increasing complexity. In the event that high *cp* values resulted in higher cross-validated error—a sign of overfitting—the *cp* value was manually reduced to minimize error. Pruned and vetted regression trees are presented using carefully chosen *cp* values. Relative error is calculated as 1-R^2, whereas cross-validated error is calculated after 10-fold cross-validation and is a measure similar to the predicted residual error sum of squares (PRESS) statistic (Thernaeu and others, 2018b).

Random forests.—RF models were built in RStudio using the randomForest package (Liaw and Wiener, 2002). Each regression tree for the forest was generated through the following procedure. First, a bootstrapped sample was drawn from the training set by random sampling (with replacement) of roughly 2/3 of the observations. The 10 elemental oxides used for growing regression trees (See Regression Trees) were used as potential predictors. The number of predictor variables used for any individual tree is defined by the variable $\mathrm{m}_{\mathrm{try}}$, and by convention is set to one third of the available predictors (10/3 ≈ 3). Each tree was therefore grown with three randomly selected predictors using a random sample of the training data set, and no pruning procedure was applied. All samples not included in the bootstrapped sample ("out-of-bag" samples, or OOB) were then sent down the tree to calculate the OOB error estimate, which is an RMSPE for the random forest models. This procedure was carried out a total of 500 times for each of the three training sets (Marbut, BU-SI, and BU-SI_{1600}) to ensure that a sufficient number of trees were constructed to minimize model error. The randomForest package calculates predictor variable importance by sending OOB samples through a model constructed on a version of the data set wherein each variable is individually permuted. The difference in MSE before and after variable permutation is used as a metric to rank the predictive importance of each variable. When the RF models are applied on external data, response values are given as the average from all trees in the forest for any input value.

Prediction error.—A general cross-validation procedure was performed to estimate prediction error when an RMSPE was not part of a modeling algorithm (that is, excluding random forest). The exponential regression and regression tree models were calibrated using a training set (70% of observations) and then applied on the validation set (30% of observations). The RMSPEs were calculated according to the following formula:

$$\sqrt{\sum_{i}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \qquad (6)$$

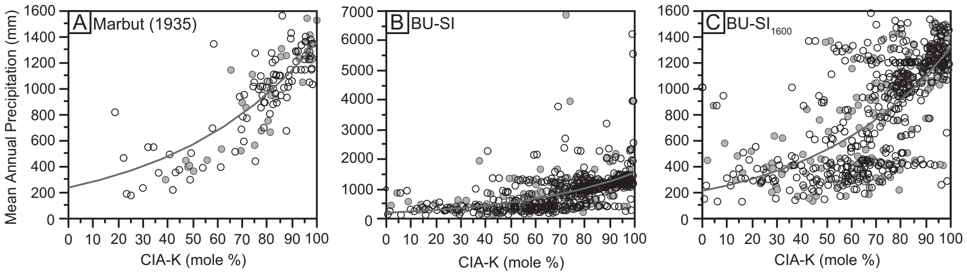

Fig. 4. Exponential regression models with CIA-K as the predictor and mean annual precipitation (MAP) as the response for the (A) Marbut (1935) data set, (B) BU-SI data set and (C) BU-SI$_{1600}$ data set. Open symbols = training data; gray symbols = testing data, red lines = exponential fit.

where $\hat{y}$ is the climate value predicted by the model, y is the known climate value, i is a given observation, and n is the number of observations in the validation set. The RMSPE values are taken as the best estimations of error when each model is applied on unknown data rather than RMSE values. Prediction errors associated with the random forest models are the one exception to this approach, as OOB error estimates provide robust prediction error values (Breiman, 2001). Splitting of a training and testing set is unnecessary for random forest models. Each random forest model was therefore generated on entire data sets (not just training sets) and reported RMSPEs are the OOB errors calculated via the random forest procedure.

RESULTS

Data Set Distributions

The online supplement contains the data sets analyzed in this study. Distributions for CIA-K and MAP values are presented for each data set in figure 3. MAP values measured from across the contiguous United States are shown for comparison (Daly and others, 1994). The random splitting of training and testing sets does not create systematic biases in any case. The Marbut (1935) data set shows two major differences compared to either the BU-SI or BU-SI$_{1600}$ data sets: (1) CIA-K values are more heavily skewed toward high values, and (2) arid regions are underrepresented. In contrast, the BU-SI and BU-SI$_{1600}$ data sets correspond more closely with the population of MAP values across the conterminous United States and tend to have a broader distribution of CIA-K values.

Exponential Regression Results

Exponential regression analyses are shown in figure 4, and summary statistics for each function are reported in table 3. The scaling factor and growth rate are very similar for each function, with overlapping values within 1σ confidence intervals (table 3). The minimum estimate for each function (CIA-K = 0) is equal to the scaling factor (a in table 3), as the exponential term in the regression equation reduces to one (eq 4).

Data set size is inversely proportional to the total variance explained by exponential models (r^2) and directly proportional to model error (RMSE) and prediction error (RMSPE). Prediction errors are similar or greater than model errors except for the Marbut (1935) data set, for which RMSPE is actually lower than RMSE. This effect is likely due to low sample size and biased sample coverage. Regression standard errors reported for the CIA-K index from previous efforts to model MAP are ~182 mm (Sheldon and Tabor, 2009), similar to results presented here. The BU-SI and BU-SI$_{1600}$

TABLE 3

Exponential regression model summaries

Data set	Training set (70%)	Testing Set (30%)	Scaling factor (a)	Growth factor (b)	Model r^2	RMSE* (mm)	RMSPE† (mm)
Marbut§	126	NA	221	0.020	0.72	182	NA
Marbut (1935)	87	38	239 ± 34	0.017 ± 0.002	0.68	191	178
BU-SI	479	206	194 ± 26	0.021 ± 0.002	0.27	564	620
BU-SI_{1600}	450	192	212 ± 16	0.018 ± 0.001	0.48	293	299

Note: For each exponential regression, the chemical index of alteration minus potassium (CIA-K) weathering index was used to predict mean annual precipitation (MAP). The BU-SI_{1600} data set includes all samples from the BU-SI data set in areas below 1600 mm MAP. All regression coefficients and models have a probability (p value) < 0.01. Scaling factor (a) and growth factor (b) are shown with 1σ errors.

*RMSE = Root mean square error.

†RMSPE = Root mean square prediction error.

§Exponential regression reported by Sheldon and others (2002).

data sets are more variable than the Marbut (1935) data set, with many samples occurring at high CIA-K values in arid to semi-arid climates (MAP <500 mm). There are also a greater number of samples with moderate CIA-K values that occur in subhumid to humid climates (MAP >1000).

Regression Tree Results

Growing and pruning regression trees.—The un-pruned regression tree constructed using the Marbut (1935) training data set shows a rapid increase in apparent R^2 (calculated before cross-validation) for the first 1-2 splits, followed by a "plateau" marking little increase in explanatory power with increasing tree size (fig. 5A). Cross-validation resulted in lower R^2 values that mirror the general pattern of apparent R^2 across tree size. Relative error (1-R^2) calculated after cross-validation indicates that error is minimized for a tree with two splits. Most of the variance in the relationship between geochemistry and MAP is explained by the first split (fig. 5A). The cp value of 0.034 was chosen for the final, pruned regression tree for the Marbut data set because cross-validated error and model R^2 are invariant after two splits (fig. 5A).

The un-pruned regression tree grown on the BU-SI training data set shows a steady increase in apparent R^2 for up to three splits (fig. 5B). Unlike the regression tree built on data from Marbut (1935), the variance between geochemistry and MAP for the BU-SI data set is more evenly distributed across tree splits. Cross-validated R^2 is consistently lower than apparent R^2. The steady decrease in cross-validated error with upwards of three splits justifies the use of a *cp* value of 0.049 in the pruned regression tree for the BU-SI data set.

The un-pruned regression tree grown on the BU-SI_{1600} training data set has a sharp increase in apparent R^2 after the first split with little change after a tree size of two splits (fig. 5C). Cross-validation resulted in maximized relative R^2 values and minimized error at a tree size of two splits. Accordingly, we used a *cp* value of 0.032 to prune the final regression tree.

Regression trees for each data set are presented in figure 6. A key is shown to aid the user in navigating the flow-chart style of decisions down each tree (also see fig. 2). Splitting criteria are labelled at the top of each node, and numbers in the circles overlain on each node identify the mean value of all observations in the node. The number of samples in each node are identified under the mean value in gray font. Each regression tree begins at the top-center of each diagram, where the first node is labelled with the first splitting criterion. Samples that meet said criterion proceed toward the left, whereas those that fail proceed to the right, and so on down to the

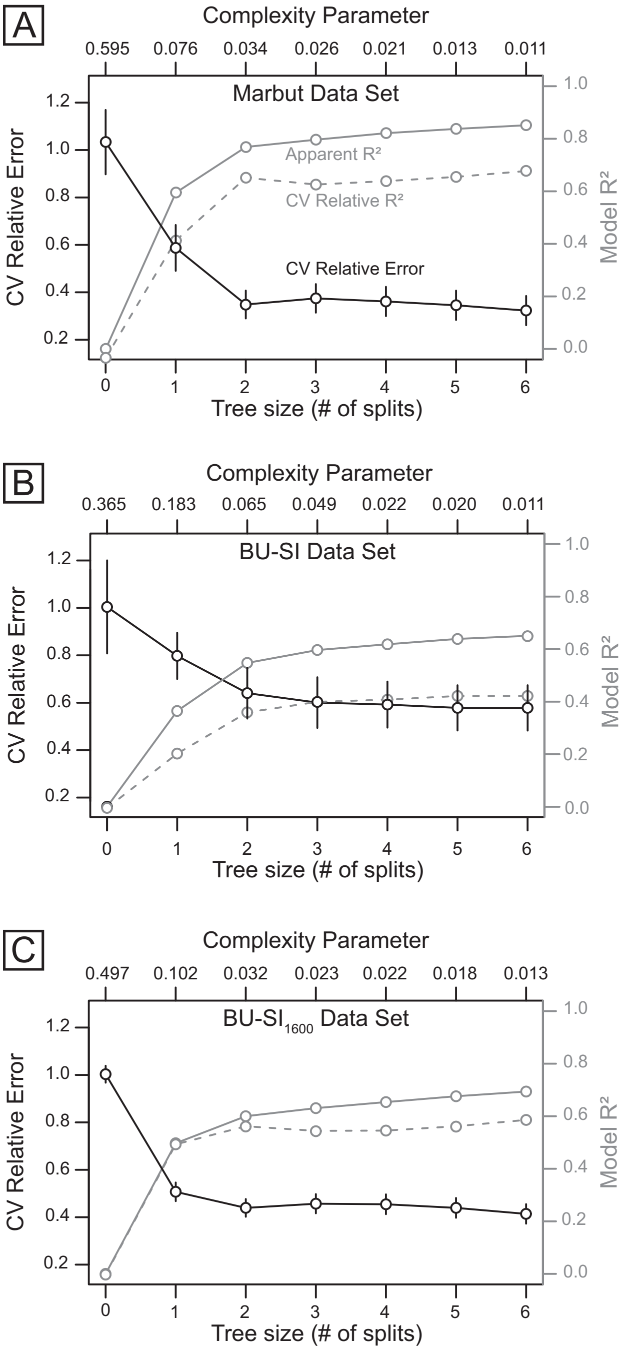

Fig. 5. Regression tree diagnostic plots. Black lines are the cross-validated (CV) relative error. Solid gray line is the apparent R^2 and dashed gray line is the cross-validated relative R^2. Complexity parameter (cp) is a measure of tree size and is used to prune each regression tree to prevent overfitting. "Plateaus" indicate minimization of explained variance. The first cp value associated with minimized CV relative error was chosen for tree pruning to prevent overfitting.

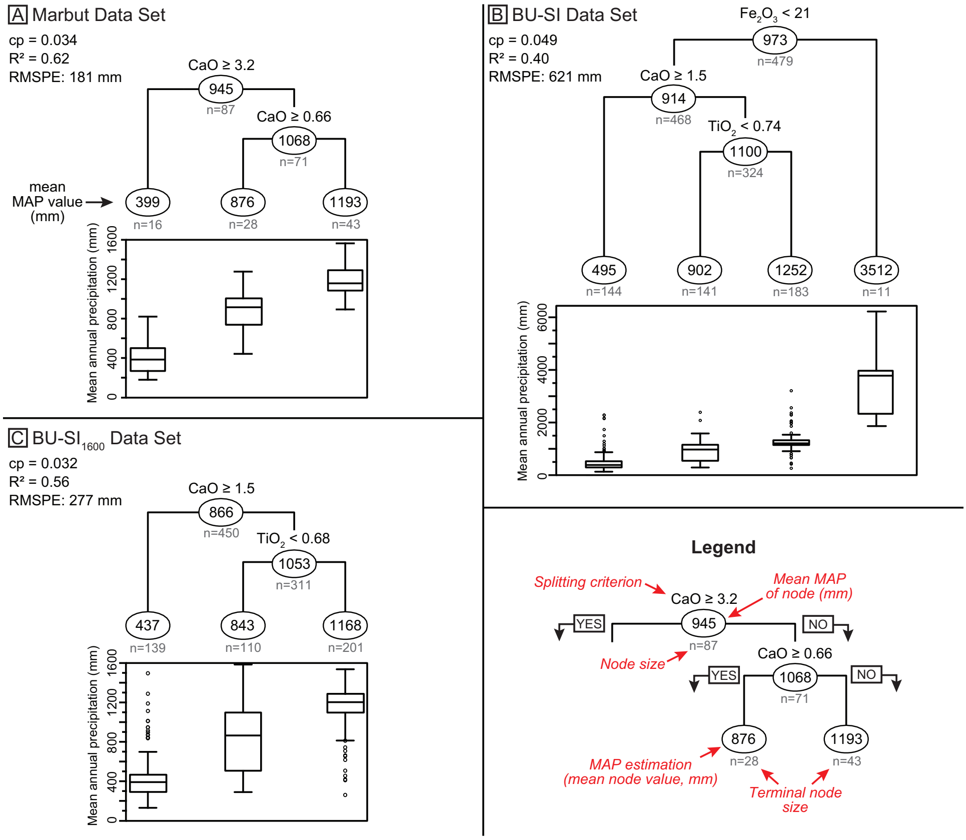

Fig. 6. Regression trees for MAP using 10 elemental oxides as potential predictors. All models are pruned to prevent overfitting (see text and fig. 5 for details). The legend explains how decisions are guided through the nodes to reach MAP predictions. The distribution of values in each terminal node is shown as boxplots below each tree.

terminal nodes. The final decision is given as the mean value of all samples in each terminal node. Box plots are shown for each regression tree to show the distribution of values in each terminal node. Each tree is plotted so that MAP values increase toward the right in each hierarchical partition.

The Marbut (1935) regression tree used only CaO as a MAP predictor and resulted in three terminal nodes corresponding to generally semi-arid, sub-humid, and humid climate zones. The BU-SI$_{1600}$ regression tree is similar but substitutes TiO_2 for CaO in differentiating sub-humid and humid climates. The BU-SI regression tree first splits soils from areas of high MAP using Fe_2O_3 and uses CaO and TiO_2 to predict MAP for drier climates. The BU-SI regression tree separated MAP responses into semi-arid, sub-humid, humid, and perhumid climate zones at the terminal nodes (fig. 6B).

Random Forest Results

Variable importance plots for the RF models trained on each data set are shown in figure 7. For each plot, the x-axis is the percent increase in mean square prediction error calculated by replacing permuted values over the data set for each variable. As expected from the other modeling techniques, CaO is the most powerful predictor of MAP. However, all other oxides hold some amount of predictive capacity for MAP, similar to findings of other multivariate techniques (Stinchcomb and others, 2016).

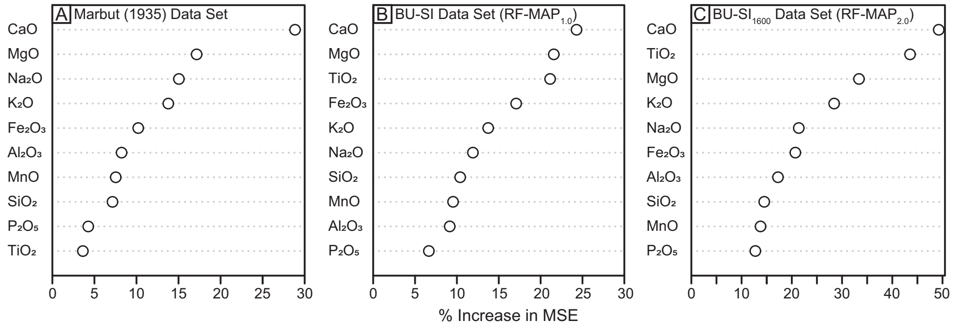

Fig. 7. Variable importance rankings for RF models calibrated on each data set. The x-axis is a measure of the increase in mean square error (MSE) between the model and a version of the model constructed on permuted values of each oxide. For each model, CaO is the most important predictor of MAP, but all oxides contribute to error reduction in the models.

The top four predictors for MAP in the Marbut (1935) data set are the four mobile base oxides (CaO, MgO, Na_2O, and K_2O). MgO and Na_2O have similar predictive capacity. All other oxides gradually decrease in importance. In contrast, input variables show a different ranking of importance in BU-SI and BU-SI$_{1600}$ RF models. After CaO, TiO_2 and MgO are the most powerful predictors for MAP in both models, whereas TiO_2 is the least valuable predictor in the Marbut RF model. Na_2O is relatively less important in predicting climate for both BU-SI-based models than for the Marbut model. Fe_2O_3 has more predictive capacity in the BU-SI model than the BU-SI$_{1600}$, likely due to the inclusion of samples from soils weathering under wetter climates. In all models, SiO_2, P_2O_5, Al_2O_3, and MnO are relatively less important in predicting MAP than the other oxides.

Model Comparison

A summary of modeling results for each data set and technique is presented in table 4. Exponential regression and regression tree results are generally similar, though exponential regression results in less variance explained and higher errors for the BU-SI and BU-SI$_{1600}$ data sets. RF provides the highest R^2 and lowest prediction errors for all data sets.

TABLE 4

Comparison of models for mean annual precipitation

Data set	Exponential regression		Regression tree		Random forest	
	r^2	RMSPE* (mm)	R^2	RMSPE (mm)	R^2	RMSPE (mm)
Marbut (1935)	0.68	178	0.62	181	0.76	154
BU-SI	0.26	619	0.40	621	0.58	395[†]
BU-SI1600	0.27	620	0.56	277	0.71	209[§]

Note: For each exponential regression, the chemical index of alteration minus potassium (CIA-K) weathering index was used to predict mean annual precipitation (MAP). Random forest models were run on entire data sets, not just training data. RMSPE for random forest are reported as out-of-bag error estimates; see text for details.

*RMSPE = Root mean square prediction error.

[†] Also referred to as RF-MAP$_{1.0}$.

[§] Also referred to as RF-MAP$_{2.0}$.

DISCUSSION

The study of paleosols is fraught with challenges, primarily due to limited knowledge about soil-forming factors in deep-time records. Paleosol proxies for environment and climate therefore must be widely applicable and developed with few *a priori* assumptions. The BU-SI data set was compiled specifically to include a large, diverse set of soils that have formed under a wide range of state-factor combinations, and to have detailed attributes available for sample selection and stratification (Nordt and Driese, 2013; Stinchcomb and others, 2016). Differences between the composition of the Marbut (1935) and BU-SI data sets are readily apparent upon comparison and are the direct result of the motivations behind each data compilation. We found that the Marbut (1935) data set of North American soils is biased toward soils forming in sub-humid and humid climates (800–1500 mm MAP), where soils tend to have high CIA-K values ($>$ 70) (fig. 3). This is likely the product of Marbut's stated focus on analyzing soils he determined to be "mature," but may also be due to the relative remoteness and paucity of soil science institutions in the arid western United States in the early 20th century. The BU-SI and BU-SI_{1600} data sets more closely reflect the bimodal distribution of MAP across the conterminous United States and have broader distributions of CIA-K values. We therefore propose that the Marbut (1935) data set be abandoned for future paleoclimate model development, and researchers should instead utilize the BU-SI and BU-SI_{1600} data sets.

Comparison of models trained on each of these data sets demonstrates that the inherent biases in the Marbut (1935) data leads to prediction errors that are unrepresentative of natural variability in soil-climate relationships. We reproduce earlier results of Sheldon and others (2002) in our exponential regression analyses of MAP versus CIA-K. The intercept and slope of regression functions are very similar between training data sets and indicates that the regression models are tracking similar phenomena (table 3). However, the lower r^2 and higher error (RMSE and RMSPE) for the BU-SI_{1600} data set relative to the Marbut (1935) data set cannot be attributed to a difference in MAP range, and instead is an effect of data set composition—namely, more variable soil types being present in the BU-SI_{1600}. This effect signals that biases exist in the coverage of soil and climate space in the Marbut (1935) data set, and RMSE and RMSPE values for models built on such data will be artificially low. Thus, for exponential pedotransfer functions, the prediction error of 299 mm is a more realistic uncertainty between CIA-K and MAP, rather than the 182 mm previously reported from the model trained on the Marbut (1935) data (Sheldon and Tabor, 2009).

The upper threshold of prediction for all exponential models is around 1600 mm. This is true for the model trained on the BU-SI data set, even though soils are included from climates with $>$6000 mm MAP. This observation supports the theoretical upper limit of CIA-K sensitivity to MAP of ~1600 mm proposed by Sheldon and others (2002). This limit has now been replicated on three independent data sets—the Marbut (1935) data, the BU-SI data, and the Vertisol climosequence of Nordt and Driese (2010). These results provide strong evidence that traditional approaches to modeling MAP using pedotransfer functions based on weathering indices are unable to predict perhumid ($>$2000 mm MAP) climate zones. Furthermore, the exponential regression models do not capture the sample space consisting of soils with relatively low MAP and high CIA-K values (fig. 4), suggesting that the exponential form is not an ideal fit for the relationship.

Recursive partitioning is a potentially more robust approach to modeling climate from soil bulk elemental composition, as demonstrated by our regression tree and RF results. Individual regression trees offer a simplified perspective on the drivers behind prediction in RFs and are therefore useful to study for insight into RF models. The regression trees trained on the Marbut (1935) and BU-SI_{1600} data sets are very similar—both rely on CaO to predict MAP in arid soils and only three climate zones are differentiable (fig. 6). The BU-SI regression tree is more complex, likely due to the larger number of soils and expanded MAP regime. Fe_2O_3 is important for differentiating humid soils from the rest of the data set, which mirrors results reported by Stinchcomb and others (2016). The importance of TiO_2 in predicting MAP for the BU-SI and BU-SI_{1600} data sets and not the Marbut (1935) data set suggests that limited coverage of soil types in the latter data set did not capture the relationship between high MAP and residual enrichment of refractory metals. We note that these results do not demonstrate that CaO is the *only* elemental oxide sensitive to MAP in the Marbut (1935) data set; rather, the recursive partitioning procedure found CaO to be the most important individual oxide for minimizing variance in MAP between nodes at all levels.

The RF results add deeper insight into the relationship between the soil elemental composition and MAP. As expected, CaO was the most important MAP predictor for each data set. The weatherable base oxides MgO, Na_2O, and K_2O were the most important predictors after CaO for the Marbut (1935) soils, confirming the findings of Sheldon and others (2002) that CIA-K, the bases to Al_2O_3 ratio, and the CaO to Al_2O_3 ratio were sensitive to MAP. The RF models trained on the BU-SI and BU-SI_{1600} data sets have a different order of predictor importance than the RF model trained on the Marbut (1935) data set, and TiO_2 and Fe_2O_3 were found to be important predictors for MAP. The finding that all 10 bulk elemental oxides hold some level of predictive capacity for MAP across all three data sets suggests that simple weathering indices only capture subsets of the manifold geochemical transformations in soils that are driven by climatic processes. This demonstrates the need for multivariate and non-parametric approaches for maximizing the relationship between soil bulk elemental composition and climate responses (for example, Stinchcomb and others, 2016).

Other paleosol proxies that relate soil elemental oxides to climate have used a variety of pretreatment and data transformation techniques, including the structuring of oxides into weathering indices or elemental ratios, and by weighting predictors using coefficients (for example, Sheldon and others, 2002; Retallack, 2008b; Nordt and Driese, 2010; Gallagher and Sheldon, 2013). Recursive partitioning is unique as a modeling tool because predictors are drawn individually and iteratively tested on subsets of the training data set. Thus, direct comparison between regression tree or RF models to simple regression models is not tenable. For example, the ordering of MgO above Na_2O in the RF models does not imply that CALMAG is more sensitive to MAP than CIA-K because both indices involve transformation of data into weathering indices that are traditionally applied across the full range of observations.

The mean values predicted for each terminal node of regression trees, with their associated error, offer insight into empirical climate zones that are predictable by soil bulk elemental oxide composition. In studies of paleoclimate, paleopedologists commonly frame reconstructed MAP values in terms of climate zones, including arid (< 250 mm), semi-arid (250–500 mm), sub-humid (500–1000 mm), humid (1000–2000 mm), and perhumid (> 2000 mm) (after Bull, 1991). However, there is currently no empirical justification for linking these climate zones to pedogenic responses, and elemental oxides in soil B horizons may not record measurable differences that correspond to these delineations. Depending on the size of the training data set, the *rpart* algorithms detect less or more climate zones than those proposed by Bull (1991) (fig. 6). For the regression trees grown on the Marbut (1935) and BU-SI_{1600} data sets, only three responses remained after pruning and correspond to semi-arid, sub-humid, and dry humid zones. One additional zone (perhumid) is a possible response in the BU-SI regression tree. These analyses are possibly the first empirical confirmation of the sensitivity of soil elemental responses to Bull's (1991) MAP zones.

The most widely applicable model resulting from our analyses is the RF model calibrated on the BU-SI data set, named here as RF-MAP$_{1.0}$. The prediction error for RF-MAP$_{1.0}$ is 395 mm, which is markedly less than the modelled RMSPE of 512 mm for PPM$_{1.0}$ (Stinchcomb and others, 2016). For settings in which MAP was likely less than 1600 mm, we recommend that researchers use the RF model developed on the BU-SI$_{1600}$ data set, named here as the RF-MAP$_{2.0}$ model. We differentiate version numbers for these models based on the differences in training data set composition that result in unique model structures, which, in turn, can produce different MAP estimates. The RF-MAP$_{2.0}$ model has a prediction error of 209 mm, lower than the CIA-K exponential function (299 mm) and substantially lower than RF-MAP$_{1.0}$ and the PPM$_{1.0}$ models. However, justification for assuming MAP values to be less than 1600 mm would need to be given for use of RF-MAP$_{2.0}$ on paleosols. This may include the presence of pedogenic carbonate or soluble salts in paleosol profiles, or an independent MAP estimate from the same locality (for example, from paleobotanical proxies; Peppe and others, 2011).

APPLICATION

Caveats for Model Applications

Code for implementing the RF-MAP$_{1.0}$ and RF-MAP$_{2.0}$ models is available in two sources: 1) a folder with all relevant files is in the online supplement (http://earth.geology.yale.edu/%7eajs/SupplementaryData/2019/Lukens); and 2) through a public GitHub repository (https://github.com/dkahle/rf-map). When users re-create the RF-MAP models in R or RStudio, the random number generator seed must be set to 42 (set.seed(42)), otherwise a unique random forest model will be generated and predicted values will not be reproducible across model versions. We strongly dissuade researchers from using figure 6 as a flow-chart for predicting MAP values. Splitting criteria are rounded in figure 6 but have several more decimal places in the regression tree model, which could result in inaccurate predictions. Further, the random forest approach is much more accurate than any individual regression tree and offers a continuous response, rather than a limited number of solutions (terminal nodes).

RF-MAP$_{1.0}$ and RF-MAP$_{2.0}$ are applicable on any soil or paleosol similar to those in the training data sets (table S2, http://earth.geology.yale.edu/%7eajs/SupplementaryData/2019/Lukens). These models should only be applied on the uppermost B horizon of paleosols and the cross-validated prediction errors of 395 mm and 209 mm, respectively, should be incorporated with all predictions. We no longer support the application of the CIA-K transfer function for MAP, as our data set inter-comparison demonstrates that functions trained on the Marbut (1935) data set are not appropriate for most applications; however, any future studies that utilize the CIA-K exponential function should use a prediction error of 299 mm. Paleosols exhibiting diagenetic alteration should be excluded from model application, especially in cases where post-burial additions or losses of CaO, MgO, Fe_2O_3, or TiO_2 may have occurred. Researchers should also follow laboratory techniques similar to those performed on the training data set, which include isolating the fine-earth fraction (<2 mm) and not pre-treating samples with acids to remove carbonates before analysis (Soil Survey Laboratory Staff, 1992).

Case Studies

We evaluate the performance of the RF-MAP models using three approaches to deep-time paleosol application. In the first case study, RF-MAP$_{1.0}$, RF-MAP$_{2.0}$, PPM$_{1.0}$, and the CIA-K exponential transfer function were used to predict MAP for a selection of paleosols with external climate control (table 5). The Ngira Paleosol is a highly weathered early Miocene Vertisol that formed on basaltic alluvium and colluvium in

Table 5

Application of RF-map on paleosols and model intercomparison

Formation and Locality	Age	Paleosol Type	Horizon	MAP Estimate (mm)				
	(Ma)			CIA-K	PPM1.0	RF-MAP1.0	RF-MAP2.0	Other
				± 299*	± 512	± 395	± 209	
Karungu Fm., Ngira, Kenya†	>17.8	Oxisolic Vertisol	Bss1	1309	1769	2105	993	1394–2618
Ogallala Fm., TX, USA§	6.6	Vertisol	Bss1	1091	1462	752	579	992
Redonda Fm., NM, USA#	~200	Aridisol	Bk1	458	357	387	376	282–438

Note: MAP estimates are based on the uppermost B horizon for each case study.

*RMSPE value from cross-validation on the BU-SI1600 data set.

†From Driese and others (2016), using Ngira unit 1; external estimates are based on nearby contemporaneous paleoflora (Michel and others, 2014).

§From Lukens and others (2017a), using the Non-Calcic Vertisol pedofacies; external estimates are based on the proportion of hyspodont taxa in a vertebrate faunal assemblage at the locality (Fraser and Theodor, 2013).

#From Cleveland and others (2008), using the Red18 paleosol; external estimates are based on depth to carbonate (Retallack, 2005; Cleveland and others, 2008).

equatorial East Africa (Driese and others, 2016). Authigenic kaolinite, highly decomposed silicate minerals, and large root traces indicate that the paleosol formed under an intense climatic regime with likely forested vegetation. Results from the RF-$MAP_{1.0}$ and $PPM_{1.0}$ models are similar to MAP estimates from closed canopy, seasonal forests reconstructed for nearby strata on Rusinga Island (table 5; Michel and others, 2014). The RF-$MAP_{2.0}$ and CIA-K estimates of sub-humid (993 and 1309 mm, respectively) climate are too dry to account for the extensive weathering state and paleogeographic position of this paleosol.

The RF-$MAP_{1.0}$ and $PPM_{1.0}$ models do not show agreement in MAP estimates for a late Miocene (6.6 Ma) Vertisol from the Coffee Ranch fossil locality in the Texas Panhandle (table 5; Lukens and others, 2017a). However, RF-$MAP_{1.0}$ and the CIA-K transfer function are within error or very close to a MAP estimate from a transfer function based on the proportion of hypsodont taxa in the Coffee Ranch fossil assemblage (Fraser and Theodor, 2013). Modern Vertisols forming in central Texas that are analogous to the Coffee Ranch Vertisols have a domain boundary at 1150 mm MAP, above which MnO_2 shows net accumulation relative to parent materials (Stiles and others, 2003). Mass-balance analysis of the paleosols at the site revealed no MnO_2 gains, which indicates the $PPM_{1.0}$ results are high but within error of this upper boundary for possible MAP (Lukens and others, 2017a). In this case, the RF-$MAP_{2.0}$ model estimates lower MAP values than are reasonable. This result underscores the need for considering RMSPEs as realistic uncertainty envelopes, and further indicates that the $PPM_{1.0}$ should primarily be used to study major climate changes and that multiple models should be cross-compared for deep-time paleoclimate proxy applications. Lastly, an Aridisol from the Late Triassic Chinle Formation in northwest New Mexico was previously documented to have semi-arid (282–438 mm) MAP based on two depth-to-carbonate transfer functions (Retallack, 2005; Cleveland and others, 2008). Results are consistent across all models (table 5).

In our second case study, we apply the two RF-MAP models to previously documented paleosols from the late Pennsylvanian Paganzo Group of southern Gondwana (modern NW Argentina) that were interpreted to have formed in moist to wet forested environments (fig. 8; Gulbranson and others, 2015). There is good agreement between the RF-$MAP_{1.0}$, CIA-K, and RF-$MAP_{2.0}$ for MAP values between 1300 to 1600 mm for most of the paleosols. However, RF-$MAP_{1.0}$ estimates values of 1670 to 1749 mm for four Vertisols that CIA-K predicts to be between 1311 to 1461 (fig. 8A). When mean annual temperature (MAT) estimates of 13 to 15°C reported by Gulbranson and others (2015) are considered, these rainfall values are within the

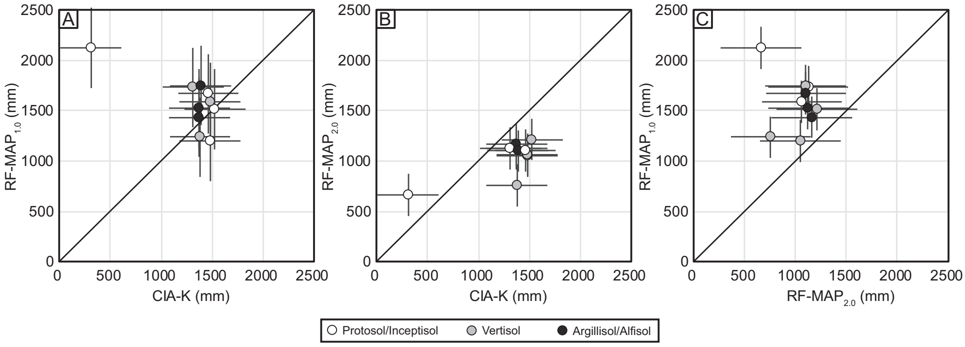

Fig. 8. Comparison of mean annual precipitation (MAP) estimates across from RF-MAP$_{1.0}$, RF-MAP$_{2.0}$, and the CIA-K exponential function of Sheldon and others (2002). (A) One paleosol with high CaO and Fe_2O_3 yields low MAP values predicted by CIA-K but high MAP values predicted by RF-MAP$_{1.0}$. Several samples are estimated to be above 1600 mm by RF-MAP$_{1.0}$. (B) Results from RF-MAP$_{2.0}$ and CIA-K are generally in agreement; however, RF-MAP$_{2.0}$ predicts values lower than CIA-K for most samples. (C) The behavior of RF-MAP$_{2.0}$ is similar to that of CIA-K, and underscores the need to use RF-MAP$_{1.0}$ in settings where MAP may be higher than 1600 mm. RMSPE for the CIA-K function is shown as 299 mm based on cross-validation of the BU-SI$_{1600}$ data set.

range of modern temperate forests and, for the more humid values estimated by RF-MAP$_{1.0}$, are within error of temperate rainforests (Peppe and others, 2011). The RF-MAP$_{1.0}$ model therefore will add great value in predicting wetter humidity provinces using the methods of Gulbranson and others (2011), which currently rely on MAP estimates from CIA-K. One paleosol shows extreme disagreement between RF-MAP$_{1.0}$ and both CIA-K and RF-MAP$_{2.0}$ (fig. 8). This sample is high in CaO (4.59 wt. %), which results in very low CIA-K and RF-MAP$_{2.0}$ predictions, but also is high in Fe_2O_3 (69.67 wt. %), which results in very high estimates from RF-MAP$_{1.0}$. It is therefore possible that any sample that contains discrete nodules of carbonate and/or Fe oxides or hydroxides will have highly disparate results from RF-MAP$_{1.0}$ and either CIA-K or RF-MAP$_{2.0}$. In these examples, the behavior of RF-MAP$_{2.0}$ is similar to that of CIA-K, and underscores the need to use RF-MAP$_{1.0}$ in settings where MAP may be higher than 1600 mm.

Our third application focuses on Late Triassic paleosols from Petrified Forest National Park, Arizona (Nordt and others, 2015). Nordt and others (2015) summarized extensive research efforts focused on understanding the evidence for and driving mechanisms of a protracted monsoonal climate system of western equatorial Pangea. The middle Norian climate shift at *ca.* 214.7 Ma initiated gradual regional aridification, as evidenced by the appearance of calcic paleosols. In this setting, paleo-MAP is known to have been variable, and a large number of paleosol samples are available for model application. We advocate for only applying RF-MAP$_{1.0}$ in this case, rather than selectively isolating samples to run in either RF-MAP$_{1.0}$ or RF-MAP$_{2.0}$.

In general, RF-MAP$_{1.0}$, PPM$_{1.0}$, and CIA-K show the same trend across the middle Norian climate shift, with sub-humid to humid MAP transitioning to more highly variable climate states marked by pronounced arid intervals after 214.7 Ma (fig. 9). Prior to the middle Norian climate shift, the CIA-K and PPM$_{1.0}$ models predict consistently humid to wet sub-humid climates. However, RF-MAP$_{1.0}$ predicts drier MAP values, consistently in the sub-humid range, but within error of the dry end of the humid climates. After the middle Norian climate shift, arid intervals are similarly estimated across each model. A key difference in model behavior emerges for paleosols between *ca.* 215 to 207 Ma: for wet phases, the PPM$_{1.0}$ predicts humid MAP values, whereas the CIA-K and RF-MAP estimates are near the sub-humid to humid boundary

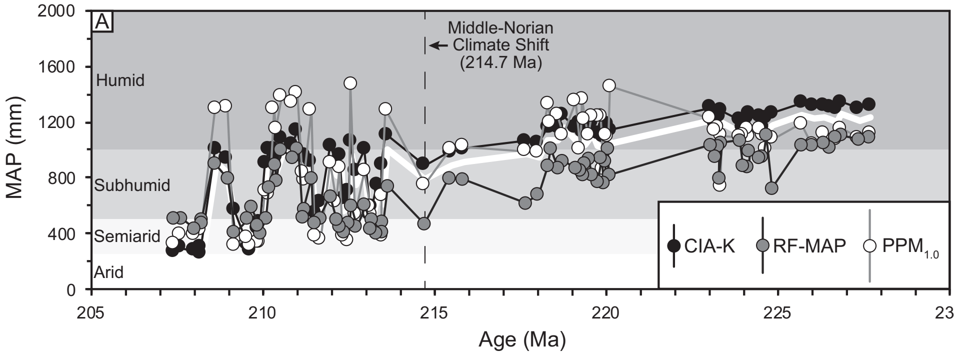

Fig. 9. Late Triassic paleosols from Petrified Forest National Park. (A) Mean annual precipitation (MAP) estimates from the RF-MAP$_{1.0}$, PPM$_{1.0}$, and the CIA-K exponential function of Sheldon and others (2002). RMSPE for the CIA-K function is shown as 299 mm based on cross-validation of the BU-SI$_{1600}$ data set. The bold white line is the MAP curve from Nordt and others (2015).

or as low as semi-arid values at times. The RF-MAP$_{1.0}$ model results suggest that semi-arid phases may have begun far earlier than previously documented, starting at around *ca.* 215 Ma, and that the middle Norian climate shift (214.7 Ma) was marked by a 42 percent decrease in rainfall from 795 ± 395 to 464 ± 395 (fig. 7). Further, the new MAP values estimated by the RF-MAP$_{1.0}$ model indicate that wet intervals before and after the middle Norian climate shift were similar in MAP value, and that the climate transition is characterized as the introduction of arid episodes rather than a secular drying trend. These notable wetting and drying cycles present after 214.7 Ma were identified across all models and occur on the order of 10^6 years.

conclusion

Predicting paleoclimate using the geochemistry of paleosols is a challenging endeavor and requires multivariate and non-linear modeling approaches and diverse data sets that reflect the complex nature of weathering. The Marbut (1935) national soil data set was compiled using philosophies of pedology that were novel at the time, but they do not apply to a wide range of soils. This collection of relatively limited soil types across climate space is not characteristic of the modern continental United States, Caribbean, or Pacific island soils. Nor is the Marbut (1935) data set representative of many paleosols observed in deep-time. Past uses of these data have resulted in pedotransfer functions with unrealistically low prediction errors for simple regression models and are therefore not ideal for application to paleosols. Embracing the more diverse assemblage of soils in the BU-SI data set allows for wider applicability of geochemical models for climate state, but the added variability in the soil-climate relationship for such data requires more complex modeling techniques to reduce prediction error. Application of recursive partitioning via random forest machine learning boosts the predictive capacity and minimizes error of MAP models based on B horizon elemental composition, and has the added benefit of automated variable selection. We have shown progress in modeling MAP using random forest, and advocate for using the RF-MAP$_{1.0}$ model in settings where either paleoclimate constraints are either unavailable, or if a large number of samples are to be run in the model. In settings where MAP values are known to have been less than 1600 mm, the RF-MAP$_{1.0}$ and RF-MAP$_{2.0}$ models should be applied to paleosols rather than the CIA-K proxy. We recommend that future attempts to develop paleosol proxies and pedotransfer functions should use the BU-SI data set rather than the biased Marbut (1935) data set. Proxy developers should also incorporate cross-validation and multivariate approaches in order to meet modern standards of best practices in statistical science.

ACKNOWLEDGMENTS

This research was supported by the Baylor University College of Arts and Sciences and the Department of Geosciences. We thank Greg Retallack, Tim White, and the editors for their constructive reviews of this manuscript.

REFERENCES CITED

Adams, J. S., Kraus, M. J., and Wing, S. L., 2011, Evaluating the use of weathering indices for determining mean annual precipitation in the ancient stratigraphic record: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 309, n. 3–4, p. 358–366, https://doi.org/10.1016/j.palaeo.2011.07.004

Amit, R., Gerson, R., and Yaalon, D. H., 1993, Stages and rate of the gravel shattering process by salts in desert Reg soils: Geoderma, v. 57, n. 3, p. 295–324, https://doi.org/10.1016/0016-7061(93)90011-9

Arkley, R. J., 1963, Calculation of carbonate and water movement in soil from climatic data: Soil Science, v. 96, n. 4, p. 239–248, https://doi.org/10.1097/00010694-196310000-00003

Atchley, S. C., Nordt, L. C., Dworkin, S. I., Ramezani, J., Parker, W. G., Ash, S. R., and Bowring, S. A., 2013, A linkage among Pangean tectonism, cyclic alluviation, climate change, and biologic turnover in the Late Triassic: The record from the Chinle Formation, southwestern United States: Journal of Sedimentary Research, v. 83, p. 1147–1161, https://doi.org/10.2110/jsr.2013.89

Baldwin, M., Kellogg, C. E., Thorp, J., and Hambridge, G., 1938, Soil classification, *in* Soils and Men: Yearbook of Agriculture 1938, USD: Washington, D. C., U.S. Government Printing Office, p. 979–1001.

Beverly, E. J., Driese, S. G., Peppe, D. J., Arellano, L. N., Blegen, N., Faith, J. T., and Tryon, C. A., 2015, Reconstruction of a semi-arid late Pleistocene paleocatena from the Lake Victoria region, Kenya: Quaternary Research, v. 84, n. 3, p. 368–381, https://doi.org/10.1016/j.yqres.2015.08.002

Birks, H. H., and Birks, H. J. B., 2006, Multi-proxy studies in palaeolimnology: Vegetation history and Archaeobotany, v. 15, n. 4, p. 235–251, https://doi.org/10.1007/s00334-006-0066-6

Breiman, L., 1996, Bagging predictors: Machine Learning, v. 24, n. 2, p. 123–140, https://doi.org/10.1007/BF00058655

——— 2001, Random forests: Machine Learning, v. 45, n. 1, p. 5–32, https://doi.org/10.1023/A:1010933404324

——— 2017, Classification and regression trees: New York, Routledge, 368 p., https://doi.org/10.1201/9781315139470

Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1984, Classification and regression trees: Boca Raton, Florida, Taylor and Francis Group, 358 p.

Bull, W. B., 1991, Geomorphic responses to climate change: Oxford University Press, 326 p.

Chadwick, O. A., and Chorover, J., 2001, The chemistry of pedogenic thresholds: Geoderma, v. 100, n. 3–4, p. 321–353, https://doi.org/10.1016/S0016-7061(01)00027-1

Cleveland, D. M., Nordt, L. C., and Atchley, S. C., 2008, Paleosols, trace fossils, and precipitation estimates of the uppermost Triassic strata in northern New Mexico: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 257, n. 4, p. 421–444, https://doi.org/10.1016/j.palaeo.2007.09.023

Criminisi, A., Shotton, J., and Konukoglu, E., 2011, Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning: Microsoft research technical report TR-2011-114.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J., 2007, Random forests for classification in ecology: Ecology, v. 88, n. 11, p. 2783–2792, https://doi.org/10.1890/07-0539.1

Daly, C., Neilson, R. P., and Phillips, D. L., 1994, A statistical-topographic model for mapping climatological precipitation over mountainous terrain: Journal of Applied Meteorology, v. 33, p. 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2

Dere, A. L., White, T. S., April, R. H., Reynolds, B., Miller, T. E., Knapp, E. P., McKay, L. D., and Brantley, S. L., 2013, Climate dependence of feldspar weathering in shale soils along a latitudinal gradient: Geochimica et Cosmochimica Acta, v. 122, p. 101–126, https://doi.org/10.1016/j.gca.2013.08.001

Driese, S. G., and Ashley, G. M., 2016, Paleoenvironmental reconstruction of a paleosol catena, the Zinj archeological level, Olduvai Gorge, Tanzania: Quaternary Research, v. 85, n. 1. p. 133–146, https://doi.org/10.1016/j.yqres.2015.10.007

Driese, S. G., Peppe, D. J., Beverly, E. J., DiPietro, L. M., Arellano, L. N., and Lehmann, T., 2016, Paleosols and paleoenvironments of the early Miocene deposits near Karungu, Lake Victoria, Kenya: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 443, p. 167–182, https://doi.org/10.1016/j.palaeo.2015.11.030

Driese, S. G., Medaris Jr., L. G., Kirsimäe, K., Somelar, P., and Stinchcomb, G. E., 2018, Oxisolic processes and geochemical constraints on duration of weathering for Neoproterozoic Baltic paleosol: Precambrian Research, v. 310, p. 165–178, https://doi.org/10.1016/j.precamres.2018.02.020

Fraser, D., and Theodor, J. M., 2013, Ungulate diets reveal patterns of grassland evolution in North America: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 369, p. 409–421, https://doi.org/10.1016/j.palaeo.2012.11.006

Friedl, M. A., and Brodley, C. E., 1997, Decision tree classification of land cover from remotely sensed data: Remote Sensing of Environment, v. 61, n. 3, p. 399–409, https://doi.org/10.1016/S0034-4257(97)00049-7

Friedman, J., Hastie, T., and Tibshirani, R., 2001, The elements of statistical learning: New York, New York, Springer, Springer Series in Statistics, v. 1, n. 10, https://doi.org/10.1007/978-0-387-21606-5_1

Gallagher, T. M., and Sheldon, N. D., 2013, A new paleothermometer for forest paleosols and its implications for Cenozoic climate: Geology, v. 41, n. 6, p. 647–650, https://doi.org/10.1130/G34074.1

Glinka, K. D., 1914, Die Typen der Bodenbildung, ihre Klassifikation und Geographische Verbreitung: Berlin, Germany, Borntraeger Brothers, 365 p.

Gulbranson, E. L., Montañez, I. P., and Tabor, N. J., 2011, A Proxy for Humidity and Floral Province from Paleosols: The Journal of Geology, v. 119, n. 6, p. 559–573, https://doi.org/10.1086/661975

Gulbranson, E. L., Montañez, I. P., Tabor, N. J., and Limarino, C. O., 2015, Late Pennsylvanian aridification on the southwestern margin of Gondwana (Paganzo Basin, NW Argentina): A regional expression of a global climate perturbation: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 417, p. 220–235, https://doi.org/10.1016/j.palaeo.2014.10.029

Gutierrez, K., and Sheldon, N. D., 2012, Paleoenvironmental reconstruction of Jurassic dinosaur habitats of the Vega Formation, Asturias, Spain: GSA Bulletin, v. 124, n. 3–4, p. 596–610, https://doi.org/10.1130/B30285.1

Hamer, J. M. M., Sheldon, N. D., and Nichols, G. J., 2007, Global Aridity during the Early Miocene? A Terrestrial Paleoclimate Record from the Ebro Basin, Spain: The Journal of Geology, v. 115, n. 5, p. 601–608, https://doi.org/10.1086/519780

Harnois, L., 1988, The CIW index: A new chemical index of weathering: Sedimentary Geology, v. 55, n. 3–4, p. 319–322, https://doi.org/10.1016/0037-0738(88)90137-6

Hastie, T., Tibshirani, R., and Friedman, J., 2009, The elements of statistical learning, Second edition: New York, Springer, 745 p., https://doi.org/10.1007/978-0-387-84858-7

Holliday, V. T., 2004, Soils in archaeological research: New York, Oxford University Press, 464 p., Wiley Online Library.

Hyland, E. G., and Sheldon, N. D., 2013, Coupled CO_2-climate response during the Early Eocene Climatic Optimum: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 369, p. 125–135, https://doi.org/10.1016/j.palaeo.2012.10.011

Hyland, E. G., Sheldon, N. D., Van der Voo, R., Badgley, C., and Abrajevitch, A., 2015, A new paleoprecipitation proxy based on soil magnetic properties: Implications for expanding paleoclimate reconstructions: GSA Bulletin, v. 127, n. 7–8, p. 975–981, https://doi.org/10.1130/B31207.1

Jenny, H., (1941) 1994, Factors of soil formation: A system of quantitative pedology: Courier Corporation, 281 p.

JMP, 2019, Cary, NC, SAS Institute Inc.

Liaw, A., and Wiener, M., 2002, Classification and regression by randomForest: R News, v. 2–3, p. 18–22.

Liivamägi, S., Somelar, P., Vircava, I., Mahaney, W. C., Kirs, J., and Kirsimäe, K., 2015, Petrology, mineralogy and geochemical climofunctions of the Neoproterozoic Baltic paleosol: Precambrian Research, v. 256, p. 170–188, https://doi.org/10.1016/j.precamres.2014.11.008

Liutkus-Pierce, C. M., Takashita-Bynum, K. K., Beane, L. A., Edwards, C. T., Burns, O. E., Mana, S., Hemming, S., Grossman, A., Wright, J. D., and Kirera, F. M., 2019, Reconstruction of the Early Miocene Critical Zone at Loperot, Southwestern Turkana, Kenya: Frontiers in Ecology and Evolution, v. 7, n. 44, https://doi.org/10.3389/fevo.2019.00044

Lukens, W. E., Driese, S. G., Peppe, D. J., and Loudermilk, M., 2017a, Sedimentology, stratigraphy, and paleoclimate at the late Miocene Coffee Ranch fossil site in the Texas Panhandle: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 485, p. 361–376, https://doi.org/10.1016/j.palaeo.2017.06.026

Lukens, W. E., Lehmann, T., Peppe, D. J., Fox, D. L., Driese, S. G., and McNulty, K. P., 2017b, The Early Miocene Critical Zone at Karungu, Western Kenya: An Equatorial, Open Habitat with Few Primate Remains: Frontiers in Earth Science, v. 5, p. 87, https://doi.org/10.3389/feart.2017.00087

Lukens, W. E., Nordt, L. C., Stinchcomb, G. E., Driese, S. G., and Tubbs, J. D., 2018, Reconstructing pH of Paleosols Using Geochemical Proxies: The Journal of Geology, v. 126, n. 4, p. 427–449, https://doi.org/10.1086/697693

Lukens, W. E., Fox, D. L., Snell, K. E., Wiest, L. A., Layzell, A. L., Uno, K. T., Polissar, P. J., Martin, R. A., Fox-Dobbs, K., and Peláez-Campomanes, P., 2019, Pliocene Paleoenvironments in the Meade Basin, Southwest Kansas, USA: Journal of Sedimentary Research, v. 89, n. 5, p. 416–439, https://doi.org/10.2110/jsr.2019.24

Marbut, C. F., 1921, The contribution of soil surveys to soil science: Society for the Promotion of Agricultural Science, Proceedings, v. 41, p. 116–142.

——— 1928, Soil, their genesis, and classification and development: A course of lectures given in the Graduate School of the United States Department of Agriculture, February to May, 1928, v. 1.

——— 1935, Atlas of American agriculture III, Soils of the United States: Washington, D.C., Government Printing Office.

Maynard, J. B., 1992, Chemistry of modern soils as a guide to interpreting Precambrian paleosols: The Journal of Geology, v. 100, n. 3, p. 279–289, https://doi.org/10.1086/629632

Michel, L. A., Peppe, D. J., Lutz, J. A., Driese, S. G., Dunsworth, H. M., Harcourt-Smith, W. E., Horner, W. H., Lehmann, T., Nightingale, S., and McNulty, K. P., 2014, Remnants of an ancient forest provide ecological context for Early Miocene fossil apes: Nature Communications, v. 5, p. 3236, https://doi.org/10.1038/ncomms4236

Mitchell, E. A., Payne, R. J., van der Knaap, W. O., Lamentowicz, Ł., Gąbka, M., and Lamentowicz, M., 2013, The performance of single-and multi-proxy transfer functions (testate amoebae, bryophytes, vascular plants) for reconstructing mire surface wetness and pH: Quaternary Research, v. 79, n. 1, p. 6–13, https://doi.org/10.1016/j.yqres.2012.08.004

Myers, T. S., Tabor, N. J., and Rosenau, N. A., 2014, Multiproxy approach reveals evidence of highly variable paleoprecipitation in the Upper Jurassic Morrison Formation (western United States): GSA Bulletin, v. 126, n. 7–8, p. 1105–1116, https://doi.org/10.1130/B30941.1

Nesbitt, H. W., and Young, G. M., 1982, Early Proterozoic climates and plate motions inferred from major element chemistry of lutites: Nature, v. 299, p. 715–717, https://doi.org/10.1038/299715a0

Nordt, L. C., and Driese, S. D., 2010, New weathering index improves paleorainfall estimates from Vertisols: Geology, v. 38, n. 5, p. 407–410, https://doi.org/10.1130/G30689.1

——— 2013, Application of the Critical Zone Concept to the Deep-Time Sedimentary Record: The Sedimentary Record, v. 11, n. 3, p. 4–9, https://doi.org/10.2110/sedred.2013.3

Nordt, L., Orosz, M., Driese, S., and Tubbs, J., 2006, Vertisol Carbonate Properties in Relation to Mean Annual Precipitation: Implications for Paleoprecipitation Estimates: The Journal of Geology, v. 114, n. 4, p. 501–510, https://doi.org/10.1086/504182

Nordt, L., Atchley, S., and Dworkin, S., 2015, Collapse of the Late Triassic megamonsoon in western equatorial Pangea, present-day American Southwest: GSA Bulletin, v. 127, n. 11–12, p. 1798–1815, https://doi.org/10.1130/B31186.1

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., and Pereira, J. M. C., 2012, Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest: Forest Ecology and Management, v. 275, p. 117–129, https://doi.org/10.1016/j.foreco.2012.03.003

Óskarsson, B. V., Riishuus, M. S., and Arnalds, Ó., 2012, Climate-dependent chemical weathering of volcanic soils in Iceland: Geoderma, v. 189–190, p. 635–651, https://doi.org/10.1016/j.geoderma.2012.05.030

Pachepsky, Y. A., Rawls, W. J., and Lin, H. S., 2006, Hydropedology and pedotransfer functions: Geoderma, v. 131, n. 3–4, p. 308–316, https://doi.org/10.1016/j.geoderma.2005.03.012

Passchier, S., Bohaty, S. M., Jiménez-Espejo, F., Pross, J., Röhl, U., Van De Flierdt, T., Escutia, C., and Brinkhuis, H., 2013, Early Eocene to middle Miocene cooling and aridification of East Antarctica: Geochemistry, Geophysics, Geosystems, v. 14, n. 5, 1399–1410, https://doi.org/10.1002/ggge.20106

Paton, T. R., and Humphreys, G. S., 2007, A critical evaluation of the zonalistic foundations of soil science in the United States, Part I: The beginning of soil classification: Geoderma, v. 139, n. 3–4, p. 257–267, https://doi.org/10.1016/j.geoderma.2007.01.020

Peppe, D. J., Royer, D. L., Cariglino, B., Oliver, S. Y., Newman, S., Leight, E., Enikolopov, G., Fernandez-Burgos, M., Herrera, F., Adams, J. M., Correa, E., Currano, E. D., Erickson, J. M., Hinojosa, L. F., Hoganson, J. W., Iglesias, A., Jaramillo, C. A., Johnson, K. R., Jordan, G. J., Kraft, N. J. B., Lovelock, E. C., Lusk, C. H., Niinemets, Ü., Peñuelas, J., Rapson, G., Wing, S. L., and Wright, I. J., 2011, Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications: New Phytologist, v. 190, n. 3, p. 724–739, https://doi.org/10.1111/j.1469-8137.2010.03615.x

Prasad, A. M., Iverson, L. R., and Liaw, A., 2006, Newer classification and regression tree techniques: Bagging and random forests for ecological prediction: Ecosystems, v. 9, n. 2, p. 181–199, https://doi.org/10.1007/s10021-005-0054-1

Prochnow, S. J., Nordt, L .C., Atchley, S. C., and Hudec, M. R., 2006, Multi-proxy paleosol evidence for middle and late Triassic climate trends in eastern Utah: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 232, n. 1, p. 53–72, https://doi.org/10.1016/j.palaeo.2005.08.011

R Core Team, 2019, A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing.

Rasmussen, C., Heckman, K., Wieder, W. R., Keiluweit, M., Lawrence, C. R., Berhe, A. A., Blankinship, J. C., Crow, S. E., Druhan, J. L., Pries, C. E. H., Marin-Spiotta, E., Plante, A. F., Schädel, C., Schimel, J. P., Sierra, C. A., Thompson, A., and Wagai, R., 2018, Beyond clay: Towards an improved set of variables for predicting soil organic matter content: Biogeochemistry, v. 137, n. 1, p. 297–306, https://doi.org/10.1007/s10533-018-0424-3

Rawls, W. J., and Pachepsky, Y. A., 2002, Soil consistence and structure as predictors of water retention: Soil Science Society of America Journal, v. 66, n. 4, p. 1115–1126, https://doi.org/10.2136/sssaj2002.1115

Retallack, G. J., 1994, The environmental factor approach to the interpretation of paleosols, *in* Amundson, R., Harden, J., and Singer, M., editors, Factors of Soil Formation: A Fiftieth Anniversary Retrospective: Soil Science Society of America, Special Publication, v. 33, p. 31–31.

——— 2005, Pedogenic carbonate proxies for amount and seasonality of precipitation in paleosols: Geology, v. 33, n. 4, p. 333–336, https://doi.org/10.1130/G21263.1

——— 2008a, Cool-Climate or Warm-Spike Lateritic Bauxites at High Latitudes?: The Journal of Geology, v. 116, n. 6, p. 558–570, https://doi.org/10.1086/592387

——— 2008b, Soils of the past: An introduction to paleopedology, 2nd edition: Oxford, England, John Wiley & Sons, 520 p., https://doi.org/10.1002/9780470698716

——— 2018, The oldest known paleosol profiles on Earth: 3.46 Ga Panorama Formation, Western Australia: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 489, p. 230–248, https://doi.org/10.1016/j.palaeo.2017.10.013

Retallack, G. J., and Huang, C., 2010, Depth to gypsic horizon as a proxy for paleoprecipitation in paleosols of sedimentary environments: Geology, v. 38, n. 5, p. 403–406, https://doi.org/10.1130/G30514.1

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P., 2012, An assessment of the effectiveness of a random forest classifier for land-cover classification: ISPRS Journal of Photogrammetry and Remote Sensing, v. 67, p. 93–104, https://doi.org/10.1016/j.isprsjprs.2011.11.002

Secord, R., Bloch, J. I., Chester, S. G. B., Boyer, D. M., Wood, A. R., Wing, S. L., Kraus, M. J., Mclnerney, F. A., and Krigbaum, J., 2012, Evolution of the Earliest Horses Driven by Climate Change in the Paleocene-

Eocene Thermal Maximum: Science, v. 335, n. 6071, p. 959–962, https://doi.org/10.1126/science.1213859

Sharma, A., Weindorf, D. C., Man, T., Aldabaa, A. A. A., and Chakraborty, S., 2014, Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH): Geoderma, v. 232–234, p. 141–147, https://doi.org/10.1016/j.geoderma.2014.05.005

Sheldon, N. D., 2006, Quaternary Glacial-Interglacial Climate Cycles in Hawaii: The Journal of Geology, v. 114, n. 3, p. 367–376, https://doi.org/10.1086/500993

Sheldon, N. D., and Tabor, N. J., 2009, Quantitative paleoenvironmental and paleoclimatic reconstruction using paleosols: Earth-Science Reviews, v. 95, n. 1–2, p. 1–52, https://doi.org/10.1016/j.earscirev.2009.03.004

Sheldon, N. D., Retallack, G. J., and Tanaka, S., 2002, Geochemical climofunctions from North American soils and application to paleosols across the Eocene-Oligocene boundary in Oregon: The Journal of Geology, v. 110, n. 6, p. 687–696, https://doi.org/10.1086/342865

Sheldon, N. D., Grimes, S. T., Hooker, J. J., Collinson, M. E., Bugler, M. J., Hren, M. T., Price, G. D., and Sutton, P. A., 2016, Coupling of marine and continental oxygen isotope records during the Eocene-Oligocene transition: GSA Bulletin, v. 128, n. 3–4, p. 502–510, https://doi.org/10.1130/B31315.1

Shmueli, G., 2010, To explain or to predict?: Statistical Science, v. 25, n. 3, p. 289–310, https://doi.org/10.1214/10-STS330

Sibirtsev, N. M., 1895, The basis of genetical soil classification: Mem. Inst. Agron. Novo-Alex, v. 9, n. 7.

——— 1966, Selected Works, Volume 1: Jerusalem, Soil Science, Israel Program for Scientific Translation.

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y., Schimel, J. P., and Chadwick, O. A., 2016, Water balance creates a threshold in soil pH at the global scale: Nature, v. 540, p. 567–569, https://doi.org/10.1038/nature20139

Soil Survey Laboratory Staff, 1992, Soil survey laboratory methods manual: Soil Survey Investigations Report, v. 42.

Soil Science Division Staff, 2017, Soil survey manual, *in* Ditzler, C., Scheffe, K., and Monger, H. C., editors, USDA Handbook 18: Washington, D. C., Government Printing Office, 639 p.

Stiles, C. A., Mora, C. I., and Driese, S. G., 2001, Pedogenic iron-manganese nodules in Vertisols: A new proxy for paleoprecipitation?: Geology, v. 29, n. 10, p. 943–946, https://doi.org/10.1130/0091-7613(2001)029<0943:PIMNIV>2.0.CO;2

Stiles, C. A., Mora, C. I., Driese, S. G., and Robinson, A. C., 2003, Distinguishing climate and time in the soil record: Mass-balance trends in Vertisols from the Texas coastal prairie: Geology, v. 31, n. 4, p. 331–334, https://doi.org/10.1130/0091-7613(2003)031<0331:DCATIT>2.0.CO;2

Stinchcomb, G. E., Nordt, L. C., Driese, S. G., Lukens, W. E., Williamson, F. C., and Tubbs, J. D., 2016, A data-driven spline model designed to predict paleoclimate using paleosol geochemistry: American Journal of Science, v. 316, n. 8, p. 746–777, https://doi.org/10.2475/08.2016.02

Therneau, T., Atkinson, B., and Ripley, B., 2018a, rpart: Recursive Partitioning and Regression Trees, R package, https://CRAN.R-project.org/package=rpart.

——— 2018b, An Introduction to Recursive Partitioning Using the RPART Routines.