Palaeogeography, Palaeoclimatology, Palaeoecology 518 (2019) 155–171

Contents lists available at ScienceDirect

Palaeogeography, Palaeoclimatology, Palaeoecology

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

Mammalian faunas, ecological indices, and machine-learning regression for the purpose of paleoenvironment reconstruction in the Miocene of South America

Jackson P. Spradley[a,b,*], Bryan J. Glazer[c], Richard F. Kay[b,d]

[a] University of North Carolina-Wilmington, Department of Biology and Marine Biology, Wilmington, NC 28403, United States of America
[b] Duke University, Department of Evolutionary Anthropology, Durham, NC 27708, United States of America
[c] Vanderbilt University, Department of Biomedical Informatics, Nashville, TN 37232, United States of America
[d] Duke University, Division of Earth and Ocean Sciences, Nicholas School of the Environment, 27710, United States of America

ARTICLE INFO

Keywords:
Gaussian process regression
Random forest
Santa Cruz
La Venta

ABSTRACT

Reconstructing paleoenvironments has long been considered a vital component for understanding community structure of extinct organisms, as well as patterns that guide evolutionary pathways of species and higher-level taxa. Given the relative geographic and phylogenetic isolation of the South American continent throughout much of the Cenozoic, the South American fossil record presents a unique perspective of mammalian community evolution in the context of changing climates and environments. Here we focus on one line of evidence for paleoenvironment reconstruction: ecological diversity, i.e. the number and types of ecological niches filled within a given fauna. We propose a novel approach by utilizing ecological indices as predictors in two regressive modeling techniques—Random Forest (RF) and Gaussian Process Regression (GPR)—which are applied to 85 extant Central and South American localities to produce paleoecological prediction models. Faunal richness is quantified via ratios of ecologies within the mammalian communities, i.e. ecological indices, which serve as predictor variables in our models. Six climate/habitat variables were then predicted using these ecological indices: mean annual temperature (MAT), mean annual precipitation (MAP), temperature seasonality, precipitation seasonality, canopy height, and net primary productivity (NPP). Predictive accuracy of RF and GPR is markedly higher when compared to previously published methods. MAT, MAP, and temperature seasonality have the lowest predictive error. We use these models to reconstruct paleoclimatic variables in two well-sampled Miocene faunas from South America: fossiliferous layers (FL) 1–7, Santa Cruz Formation (Early Miocene), Santa Cruz Province, Argentina; and the Monkey Beds unit, Villavieja Formation (Middle Miocene) Huila, Colombia. Results suggest general concordance with published estimations of precipitation and temperature, and add information with regards to the other climate/habitat variables included here. Ultimately, we believe that RF and GPR in conjunction with ecological indices have the potential to contribute to paleoenvironment reconstruction.

1. Introduction

Reconstructing paleoenvironments has long been considered a vital component for understanding the paleobiology of extinct organisms, as well as the evolutionary pathways of species and higher-level taxa. However, the ability for researchers to reconstruct past environments is necessarily limited by the available lines of evidence; including geochemical indicators of the environment (Wilson, 1974; Hayes et al., 1990; Kohn, 2010; Cerling et al., 2011), the adaptations of the organisms present in fossil assemblages (Andrews and Van Couvering, 1975; Vizcaíno et al., 2017), and the ecological richness of the assemblage (Fleming, 1973; Andrews et al., 1979; Kay and Madden, 1997; Reed, 1998; Kay et al., 2012a; Di Giacomo and Fariña, 2017). For the purposes of this study, we focus on this latter form of evidence, ecological richness. That is, the number and types of ecological niches (Hutchinson, 1957) filled within any given faunal community. This study proposes a novel approach for the reconstruction of climate parameters in the fossil record by utilizing two machine-learning algorithms—the random forest (RF) and Gaussian process regression (GPR)—which are applied to faunal diversity and richness, and climate data from several extant Central and South American localities to produce a paleoecology prediction model. In order to gauge the

* Corresponding author at: University of North Carolina-Wilmington, Department of Biology and Marine Biology, Wilmington, NC 28403, United States of America.
E-mail address: spradleyj@uncw.edu (J.P. Spradley).

https://doi.org/10.1016/j.palaeo.2019.01.014
Received 13 September 2018; Received in revised form 8 January 2019; Accepted 9 January 2019
Available online 15 January 2019
0031-0182/

predictive power of the generated models, the models were used to estimate climate parameters at several extant Australian sites based on their faunal assemblages. Australia serves as an ideal test for our data for two primary reasons: 1) due to the relative geographic isolation of Australia, the mammalian fauna there is phylogenetically distinct from that of South America, and thus similar ecologies and morphologies between the two faunas are almost certainly due to convergence rather than a signal of phylogenetic relationship; and 2) Australia is a large continent with a great diversity of climates and habitats, much like South America.

To the degree that we find the accuracy of our model for estimating modern climate parameters to be accurate between the two phylogenetically distinct faunas, we have gained confidence for its use for reconstructing paleoclimatic variables in two well-sampled Miocene faunas from South America, the Santa Cruz fauna of Argentina and the La Venta fauna of Colombia. The former is an unusually well sampled (from the perspective of species richness and fossil completeness) late Early Miocene fauna from the far southern tip of South America. The latter is the only Middle Miocene neotropical locality for which we have a relatively complete sample of well-preserved mammal taxa. For both formations, we are restricting our analysis to specific well-sampled stratigraphic levels that have previously been identified as approximating a contemporary fauna, fossiliferous levels 1–7 (FL 1–7) of the Santa Cruz formation (Kay et al., 2012b) and the Monkey Beds of the Honda Group in La Venta (Kay and Madden, 1997). We compare our results to those of Kay and Madden (1997) and Kay et al. (2012b).

2. Background

That mammal species occupying similar habitats have convergent adaptations is one of the foundational blocks of natural philosophy, including Cuvier (Gould, 1982) and Darwin (Darwin, 1859), and serves as the theoretical basis for the use of ecological richness as an avenue for understanding and reconstructing paleoecology. Referred to by Charles Elton in 1927 as ecological “vicars”, Darwin, Elton, and others envisioned species as filling roles (or niches) within an ecosystem, roles that would be present and filled in similar environments regardless of the taxonomic affiliation of the species that are actually present; just as the duties of a village vicar would be performed regardless of what individual actually assumed the role (see also: Hutcinson, 1965). While the metaphor can only go so far, when describing the niche structure of a community it nevertheless becomes useful to describe species not by name but rather by what they do (e.g. how large are they, what do they eat, what part of the environment they inhabit, etc.). This premise is validated by the strong correlation between aspects of phenotype and environment as documented by countless studies designed to test this relationship (e.g., Anthony and Kay, 1993; Boyer, 2008; Boyer et al., 2013; Crompton and Lumsden, 1970; Ross and Kirk, 2007). While this implies an adaptive link between form and function (Kay and Cartmill, 1977), it is not dismissive or at odds with the reservations of Gould and Lewontin (1979) who suggested that not all morphology is adaptive, since such research is extremely circumscribed in how it defines aspects of morphology that can be candidates for adaptive processes (e.g., Kay and Cartmill, 1977; Kay et al., 2004; Seiffert et al., 2009; Boyer et al., 2013).

The link between form and function allowed for studies such as the seminal work of Andrews et al. (1979) who sought to identify patterns in the distributions and frequencies of various niches within mammalian communities, particularly as they relate to the general structure of the environments they inhabit. While it was Harrison (1962) and Fleming (1973) who established that mammalian communities vary predictably with latitude, Andrews et al. (1979) demonstrated that even when latitude was held constant, mammalian community structure is significantly correlated with habitat type. And when both latitude and habitat type are similar, mammalian niche structure is “remarkably constant” (Andrews et al., 1979). Other studies have gone on to show that niche structure covaries predictably with precipitation (Kay et al., 1997). (For a recent example see Robinson et al., 2017). The former approach of Andrews et al. (1979), for reconstructing the habitat type reveals a potential flaw of the approach of Kay et al. (1997); namely, focusing on a specific climate variable fails to consider that a climate variable (such as precipitation) is related to community structure only insofar as it is also correlated with a certain habitat type (specified by Andrews et al. as the vegetational structure). In other words, by using community structure to reconstruct a single climate variable like rainfall, we are potentially losing track of the important matter of describing the habitat in which these animals lived. On the other hand, using quantitative variables eliminates the need to subjectively characterize a habitat into one category or another. To reconcile these two approaches, we include a number of different climatic variables (temperature, precipitation, and seasonality) as well as two quantitative variables directly associated with the botanic habitat (canopy height and net primary productivity). Combined, these variables can give a specific description of the habitat without relying on somewhat subjective qualitative descriptions of habitat.

In the context of paleoclimate reconstruction, it is thus proposed that the kinds and proportions of species within specific niches within a mammalian fauna as a whole can serve as a proxy for habitat structure or climate, as has been done in a number of previous studies (e.g., Andrews et al., 1979; Gingerich, 1989; Fariña, 1996; Vizcaíno et al., 2010; Kay et al., 2012a, 2012b; Robinson et al., 2017). Such studies are limited by a multitude of factors, specifically related to the difficulties inherent to sampling, both in living faunas and in the fossil record. In relation to living faunas, an incomplete sampling effort can lead to the exclusion of rare species from a faunal list. It can also be difficult to give an accurate estimation of the general climate or habitat, depending on the length of the observation. In order to overcome these potential pitfalls, long-term field studies can be used. Such stations collect data on the fauna present as well as track various climatic variables (e.g., precipitation, temperature, etc.) over decades, providing an approximation for the ecosystem that is robust to aberrant seasons or years.

Similar issues arise when defining a fossil fauna in both space and time. When recording the species in a modern mammalian community, one can control the size of the area sampled, and be confident that all species are contemporaneous, i.e., not time averaged. With a fossil fauna, however, assumptions must be made as to the fidelity of the fossil assemblage as compared to the prehistoric life assemblage. Taphonomic studies in mammals suggest a bias against recovery of small mammals in bone assemblages (Behrensmeyer and Dechant-Boaz, 1980; Damuth, 1982; Sept, 1994). As an illustrative example, Behrensmeyer and Dechant-Boaz (1980) found that only 60% and 21% of small (< 1 kg) herbivores and carnivores, respectively, that were recorded in a six-year study of a mammalian community were also recorded in the bone assemblage. That is compared to large mammals (> 15 kg), of which 95–100% were also recorded in the bone assemblage.

A second commonly mentioned obstacle to assessing fossil community structure, time averaging, may be the solution to the first. Time averaging, or the process by which distinct events appear to be synchronous in the geological record in part because the processes associated with burial are generally slower than population turnover (Walker and Bambach, 1971). Time averaging potentially creates over-completeness in an assemblage instead of under-completeness (Kowaleski, 1996). While the ways in which time-averaging can affect a fossil assemblage are varied and distinct (Kidwell and Flessa, 1996), in the context of this study the biggest concern is that a single fossil assemblage, by being a sampling of extinct species from an interval of time (versus a single time slice, as in a modern fauna) might encompass an interval of environmental change, thereby blunting the environmental signal in the data. However, such extreme cases are rare (Kidwell and Flessa, 1996) and are typically recognized by incongruences within a geological context (e.g., unconformities). When

temporal resolution is constrained, a limited degree of time-averaging allows for even rare species within a fossil assemblage to be recorded, thus providing a presumably complete inventory of the fossil community.

Finally, Croft (2013) warned that, in the context of paleoecological analyses, the area sampled while compiling faunal lists for extant communities must be considered in relation to the presumed area of the fossil locality, based on the reasoning that ecological richness within a community will have a positive correlation with the area sampled. While this is certainly true, as Croft (2013) himself notes, this is an inherently difficult task and the degree to which it affects the results of such analyses is difficult to gauge. In addition, estimates for the area sampled are frequently excluded from published faunal lists. This presents a major obstacle as to the incorporation of sampling area in any paleoecological analysis.

3. Methods

We focus here on several techniques for paleoecological reconstruction that take advantage of regression statistics for prediction—or more appropriately, retrodiction. Until recently, studies seeking to reconstruct past habitat and climate from the mammalian fossil record have used linear regression models and Principal Components analyses using variables drawn from those regression analyses (Andrews et al., 1979; Kay and Madden, 1997; Kay et al., 2012a, 2012b; Rowan et al., 2016). This study tests the applicability and accuracy of newer, more computationally complex regression techniques that have been used in the ecology literature, including regression tree analysis, random forests, and Gaussian process regression.

Beginning with a dataset of niche metrics of mammalian species from modern South American localities, we test each of the methods described below for their accuracy and predictive power, as applied to both a test sample from South America and one from Australia. Our objective is to shed light on the potential of more complex machine-learning algorithms for paleoecological reconstruction, as well as how they compare to previously published regression techniques.

3.1. Multivariate linear regression

Included in their methods, Kay and Madden (1997) and Kay et al. (2012a, 2012b) use a multivariate approach to linear regression. Working with a dataset similar to our own, they assign each species from the localities in their study a categorical description of its ecological niche. This includes assignment to categories of diet, body size, locomotion and substrate preference. Such broad categories allow for comparisons with extinct species, for which more precise ecological data is not easily deduced (Kay et al., 2012b). After compiling lists of species and their niche metrics, they create what they refer to as “ecological indices”, i.e. ratios of particular ecological niches that are used to represent the ecological diversity of a given faunal community. After performing a PCA of the ecological indices included in their study, they use the first principal component axis (PC1) as an independent variable in a simple bivariate correlation. A linear regression of this correlation provides a simple linear model for reconstruction (in the case of Kay et al. (2012b), a reconstruction of rainfall in the Santa Cruz Fossiliferous Levels 1–7). While we do not include all of the ecological indices of Kay and Madden (1997) or Kay et al. (2012b), we include this multivariate approach in our study for comparison, and generally refer to it as “linear regression”.

The primary differences between this paper and the studies of Kay et al. (2012a, 2012b) is as follows: 1) Kay et al. (2012b) uses 25 modern South American localities by which to create their regression models; 2) Regression techniques are limited to linear regression, so as to make the output easy to comprehend and illustrate; 3) Climate variables were limited to just temperature and rainfall, and no habitat variables were included; 4) Kay et al. excluded any localities from above 1000 m above sea level. Thus, we compare our results from GPR, RF, and RTA to these previous estimates and interpret those differences in the context of the paleoenvironment of the fossiliferous levels 1–7 of the Santa Cruz Formation.

3.2. Regression tree analysis

While standard regression techniques assume some specified relationship between the predictor variable(s) and the response variable (e.g. a straight line, quadratic, etc.), regression tree analysis (RTA) (Efron and Tibshirani, 1991; De'ath and Fabricius, 2000) builds a series of rules for partitioning the data into homogenous subsets based on values of the predictor variables. RTA has shown particular value in modeling the complex relationships commonly found in ecological analyses (Michaelson et al., 1994; De'ath and Fabricius, 2000; De'ath, 2002). This series is best illustrated as nodes, branches, and leaves—hence the “tree”. At each node, the best predictor variable is chosen from all other predictor variables to split the data into its largest homogenous groups, these groups are then further split—i.e. “recursive partitioning”—using added predictor variables until the data can no longer be split. Each leaf then is a cell that carries all data points that have the same partitioning course. The resultant tree can then be used to predict new values in an efficient, easy to understand (and illustrate) manner (see Fig. 1 for an illustrative example using hypothetical data).

The advantages that RTA carries over classical regression are immediately clear. First and foremost among these is the ability to incorporate non-linear relationships between a response and a predictor variable; an attribute that is particularly useful in biology wherein complicated, non-linear relationships are common (Moore et al., 1991; Kay and Madden, 1997). In addition, the simplistic output of the tree allows for the relative importance of the various predictors on the distribution of the data to be clearly shown, providing insight into the spatial influence of the predictors (Iverson and Prasad, 1998; De'ath and Fabricius, 2000).

However, RTA is not without its disadvantages, even when compared to conventional regression. First, any linear functions that happen to be included in the analysis is only approximated, and because of the nature of RTA these linear functions tend to be binned in a manner that oversimplifies the linear relationship, leading to a “coarseness” of the RTA output (Moisen, 2008). Those values that lie close to the boundary of two bins are more likely to be misidentified. Additionally, results from RTA are highly dependent on the input data, meaning they are unstable and small changes in the data can produce highly divergent trees. While this can be overcome with an appropriately large dataset, particularly one in which the variables are normally distributed, such a dataset is frequently hard to obtain in ecology. For a summary of these issues, as well as other potential issues related to classification, please refer to Moisen (2008), pp. 586–587.

3.3. Random forests

Introduced by Brieman (2001), random forests (RF) is a machine-learning approach based heavily on RTA. Borrowing heavily from the description provided by Liaw and Wiener (2002), RF analysis works in the following steps (this is further illustrated in Figs. 2 and 3): 1. The original data is bootstrapped into a specified number of re-sampled data sets. 2. Regression trees are produced for each bootstrapped sample (Fig. 3). Node partitioning (i.e. segregating the data into homogenous groups) is performed on a random sampled subset of the resampled predictor variables, and the best predictor is chosen among that subset. Note that this differs from the traditional regression tree, which has all predictor variables available when partitioning the data at each node. 3. New predictions are made through aggregating and averaging all the trees of the bootstrapped samples (Fig. 3). Data is predicted by averaging the predictions of all the trees of the bootstrapped samples. While the result from any given tree may differ dramatically from that of

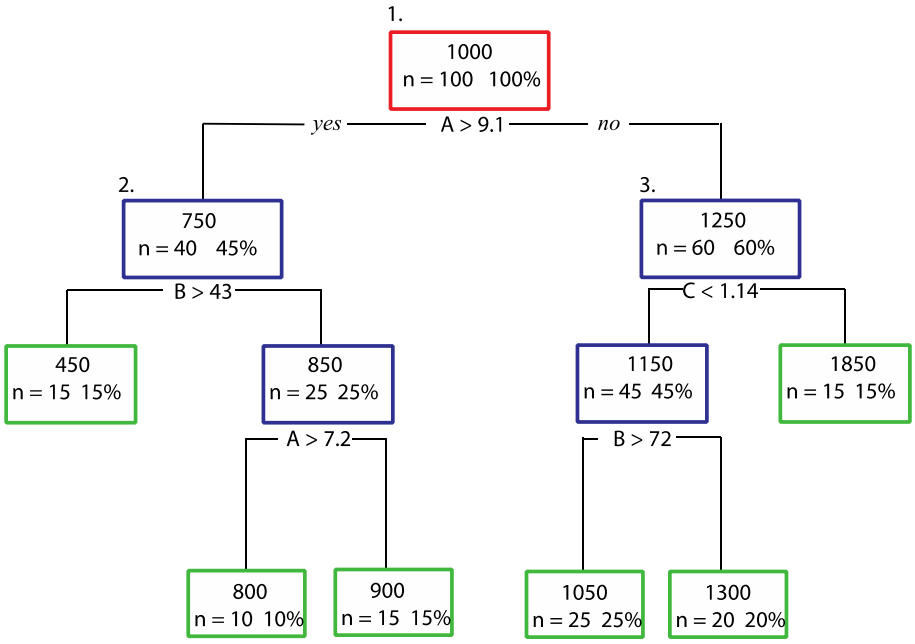

Fig. 1. Illustrative regression tree constructed using hypothetical data. In each box there are three numbers. The first number, n = ## indicates the total number of individual data points represented in the subset at that node. For instance, in the first node (colored red), there are 100 data points represented at this node. The number to the right of the N value indicates the percentage of the total dataset represented at that node. In the case of the first node, this would be the complete dataset, 100%. The number above these two is the average value of the response variable (the variable that we are trying to predict) of that subset. Below this box, we also see "A > 9.1". This represents the predictor value that best subdivides the data into homogenous groups. In this hypothetical data, any data point that has a value of "A" > 9.1 is now represented in node 2. As the algorithm proceeds through the nodes—colored in blue—eventually it comes upon subsets of data that effectively cannot be further subdivided into homogenous groups. These are the "leaves" of a regression tree, and are colored in green here. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

another tree, the high number of trees that are typically grown in RF— ≥ 1000, hence the "forest"—allows it to effectively "learn" the best fit of the data.

One of the complaints frequently leveled at RF is its apparent lack of interpretability, that it is a "black box" (Cutler et al., 2007). This criticism stems from the fact that RF lacks the illustrative efficiency of more traditional regression techniques. Instead of producing one tree, RF predicts "out-of-bag" observations by calculating the average of hundreds (or even thousands) of fully-grown, bootstrapped trees, meaning that the relationship between variables is impossible to demonstrate.

However, RF carries many advantages over traditional methods. Chief among these is that instead of having to select one or two predictor variables from a set of highly correlated predictors, RF can include all variables and even determine the relative importance of each variable (Cutler et al., 2007). Also, because a large number of trees are grown, the potential of the training data error to be radically different from the underlying model error for a given population is small, minimizing the potential of overfitting (Prasad et al., 2006). Additionally, while RF does not require an a priori model selection, there are parameters that must be chosen, namely the number of predictor variables randomly chosen per node (m_{try}), and the number of trees (n_{tree}). Though the chosen values of these parameters can influence the results (i.e. the predictive accuracy of the RF model), this effect has been shown to be relatively negligible (Brieman, 2001; Liaw and Wiener, 2002; Cutler et al., 2007).

While choosing the values for the pertinent parameters for random forest algorithms included in this study the recommendations of

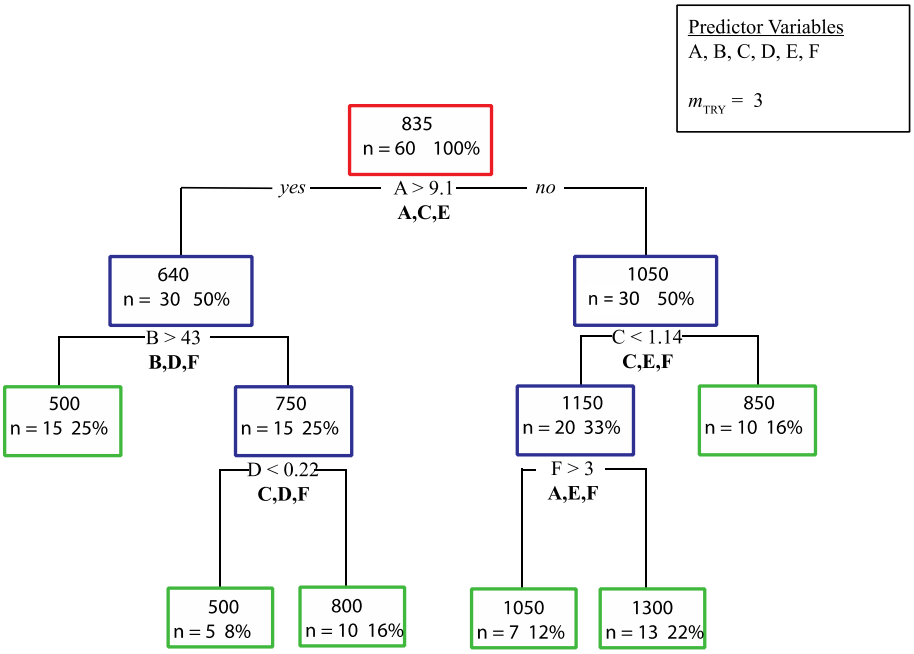

Fig. 2. Example of a single tree from a random forest analysis. In this example, at each node three (m_{try}) of the six predictor variables are randomly chosen and the variable that best splits the data into homogenous groups is chosen.

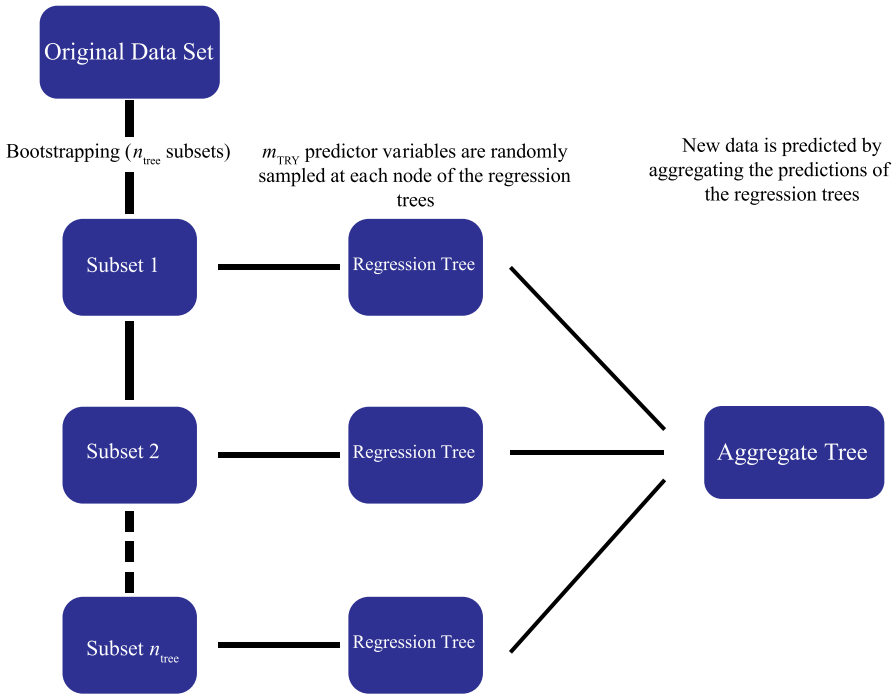

Fig. 3. Overview of the procession of the random forest algorithm from the original data set to predicting new data.

previous literature were followed when available (Liaw and Wiener, 2002; Stitson et al., 1999). For RF, m_{try} was set to 4, representative of $p/3$, where p is the number of predictor variables (Liaw and Wiener, 2002), and n_{tree} was set to 1000, the default for RF.

3.4. Gaussian process regression

Gaussian process regression (GPR) is a non-linear, non-parametric, Bayesian regression technique. Originally developed as "kriging" in the field of geostatistics, GPR has not seen wide application in ecology or paleoecology (Golding and Purse, 2016), but is nonetheless useful in these applications because of its expressiveness and flexibility.

GPR can be viewed as an extension of simple linear regression.

$$f(x) = x^{\intercal}w \tag{1}$$

$$y = f(x) + \varepsilon \tag{2}$$

Here, $x^{\intercal}w$ is a dot product between a transposed vector of dependent variables, $x^{\intercal}$, and their associated weights, w. The error term ε is modeled as a Gaussian distribution. Thus, the likelihood function (i.e., fit of model to the data given the weights) is given as:

$$p(y \mid X, w) = \prod_{i=1}^{n} p(y_i \mid x_i, w) \tag{3}$$

$$y_i \sim N(x_i^{\intercal}w, \sigma_{\varepsilon}) \tag{4}$$

where σ_{ε} is the variance of the error term.

Further, the distribution of the linear regression coefficients or weights (**w**) is given a multivariate Gaussian distribution:

$$w \sim N(0, \Sigma_p) \tag{5}$$

where Σ_p is the covariance matrix of the weights. This prior distribution over the weights also leads to an interpretation of Gaussian processes as specifying a distribution over functions. Each function represents a Gaussian distribution centered at the input point—in this case at zero because the distribution is z-transformed—with all of these functions having a joint Gaussian distribution. Due to the tractability of computing integrals of the Gaussian distribution, the posterior and predictive distributions have closed-form solutions, as shown in Rasmussen and Williams (2006).

This Bayesian linear regression model (Eq. (2)) is the basis for Gaussian process regression but is limited by the inflexibility of the linear model. To overcome this limitation, the predictors are projected into "feature space" using a non-linear function (denoted by $\phi(x)$). A simple example is the polynomial expansion, which projects a 1d vector x into a 3d vector:

$$\phi(x) = [x, x^2, x^3] \tag{6}$$

In the case of Gaussian processes, the projection functions are always used in the context of dot-products (known formally as inner-products). For example, the following expression, which is derived in Rasmussen and Williams (2006), is a dot product of a projection function $\phi(x)$ and a covariance matrix Σ_p:

$$\phi(x)^{\intercal}\Sigma_p\phi(x) \tag{7}$$

The use of projection functions in dot products allows the use of the "kernel trick" to replace the projection function. This common operation in machine learning models, stemming from Mercer's theorem (Mercer, 1909), allows projection functions in inner products to be represented by a covariance function. Thus, the complex projection function into feature space does not have to be explicitly calculated. Instead one can operate only on a function that represents the covariance between two points projected into the feature space. This allows significantly more complex, non-linear interactions between predictors to be represented in the model, using only a computationally simple covariance function.

The Gaussian radial basis function is a commonly used kernel (Chang et al., 2010) that is derived from the Gaussian distribution. It has a single parameter *sigma* that controls the smoothness of the resulting regression function. Additionally, the ANOVA kernel was tested. This kernel was selected based on results from Stitson et al. (1999) showing good performance on a similar multivariate regression problem. The ANOVA kernel has two parameters that can be tuned, *sigma* and *degree*. While harder to interpret than the single *sigma* parameter of the Gaussian radial basis function, these parameters roughly control the smoothness and magnitude of the regression function, respectively.

Of course, Gaussian processes and random forests are not the only non-linear regression models available. We did not explore modeling frameworks such as neural networks or hierarchical Bayesian models. Neural networks have shown impressive performance in a variety of tasks; however, they typically require large amounts of training data and are prone to "overfitting" unless special care is given to prevent this. Hierarchical Bayesian models are popular in ecology and the life sciences (Royle and Dorazio, 2008; Ogle, 2009; Cressie et al., 2009) as they allow the modeler to encode their a priori knowledge into the model structure and provide confidence intervals for predictions and model parameters. However, this flexible modeling approach requires more upfront effort to yield good predictive performance.

Finally, in order to compare the predictive performance of our different modeling approaches, we use mean absolute error (MAE) and root mean square error (RMSE), the two most commonly used error metrics in ecology when testing predictive performance (Willmott et al., 2009; Chai and Draxler, 2014). The calculation of MAE is straightforward; the absolute values of the errors are summed ("total error") and then divided by n (Willmott and Matsuura, 2005). RMSE on the other hand begins first with the sum of the individual squared errors, which is divided by n. This calculation results in the "mean squared error (MSE)"; RMSE is then calculated by taking the square root of MSE (Chai and Draxler, 2014). While RMSE has typically been used as the standard throughout ecological and climatological literature, MAE has been argued as the more effective descriptor of error (Willmott and Matsuura, 2005; Willmott et al., 2009). Conversely, Chai and Draxler (2014) argued for the continued use of RMSE. Rather than debate the merits of one over the other, we chose to present both metrics in this paper.

3.5. Data collection

Faunal lists for a total of 85 localities across Central and South America (Fig. 4), as well as 14 localities from across Australia (Fig. 5), were compiled from the literature (see supplemental documents for references) as well as the Information Center for the Environment (ICE) Biological Inventories of the World's Protected Areas (*http://www.ice.ucdavis.edu/bioinventory/bioinventory.html*). ICE is an online archive of the world's national parks, and features mammalian faunal lists for each park included in its archive. Because this study is concerned with community structure and depends on the listed community being more or less complete, only faunal lists that were recorded by ICE as being either "complete" or "essentially complete" were included in this study. While it is likely that a few species go unrecorded in these lists, this procedure provides a more reliable representation of any given fauna at a locality than does the use of distribution maps (Spradley et al., 2015).

For each modern locality, the presence or absence of 617 species of non-volant mammals was recorded. Basic ecological information for each species was compiled from the literature (see supplemental documents for references), including average body mass, primary diet, and preferred locomotion, and allows for a general characterization of a species' position in niche space. However, the precise characterization of an extant species' ecology is dependent on the degree to which that species has been studied in the wild and/or captivity, casting doubt on any precise characterization of ecology in poorly-studied species. In addition, the identification of any of these variables in the fossil record is inherently even more problematic and not as precise as their identification in living species. Thus, we follow Kay and Madden (1997) and Kay et al. (2012a, 2012b) for the categorization of ecological variables, which somewhat alleviates this potential complication by placing species in appropriately broad categories associated with the expected ecological resolution gathered from museum specimens. These categories are summarized in Table 1. In addition, we include as variables a predator/prey ratio (Croft, 2006), calculated as the number of faunivores divided by primary consumers with a body mass between 500 g–500 kg, as well as total species richness.

After assigning every species an appropriate category, a series of indices were used to describe the relative abundance of a specific ecology (or ecologies) within each mammalian community (Kay and Madden, 1997). These indices include the Arboreality Index (AI: the relative number of arboreal and scansorial species to the total number of species in a fauna), the Frugivore Index (FI), the Grazing Index (GI), and the Browsing Index (BI). These indices serve as our primary variables of interest in relation to climatic variables (see below). Faunal lists, along with ecological information for each species, are presented in the supplementary documents. Data for each locality is also included in the supplemental documents.

Spatial maps for six bioclimatic variables were downloaded from *WorldClim.org*. These spatial maps are produced from data collected at weather stations across the continent and represent averages of 20–50 years. Areas between weather stations are filled in with interpolated data. In addition, spatial maps for net primary productivity (NPP) and canopy height were downloaded from the Numerical Terradynamic Simulation Group (NTSG; http://www.ntsg.umt.edu/project/mod17) and the NASA Spatial Data Access Tool (SDAT: http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10023_1), respectively. Net primary productivity is a measure of carbon usage by plants, and is calculated by the amount of carbon dioxide consumed by plants during photosynthesis minus the amount of carbon dioxide released during respiration. It is presented here as grams of carbon per square meter per day. Data for canopy height is obtained by spaceborne LiDAR, and has been shown to be accurate on regional scales (Sawada et al., 2015). To account for potential error in point measurements, we calculated mean canopy height (meters) by taking the average value of surrounding points (these points being 1 km^2 squares). Except in the case of coastal localities, this typically resulted in the average canopy height of nine measurements from a 9 km^2 grid. These spatial maps were then uploaded into QGIS® (QGIS, 2017) and the values for each variable were recorded for each locality at their associated GPS coordinates using the "point sampling" plugin in QGIS. Bioclimatic variables for each locality are presented in a supplementary table.

We excluded the variable of site area from our analyses. As described in the "Background" section of this paper, there are inherent difficulties in trying to estimate the sampling area of a fossil locality, as well as obtaining sampling area from many field studies. Given that this difficulty affects all of our sampled localities, and that the area sampled is variable across all habitats, we assume that the result should be a broad error rather than specific skewing of the results.

All statistical analyses were performed in the open-source statistical software, R (version 3.4.4). The "rpart" package (Therneau et al., 2015) was utilized for RTA and the "randomForest" package (Liaw and Wiener, 2002) utilized for RF. The "gGPRlot2" and "forestFloor" (Welling et al., 2016) packages were used to visualize the results from these analyses. The R script for our analyses is included in the supplemental documents for this paper.

4. Study sites

While the above methods have been used previously in ecological studies (e.g., Okubo and Levin, 1980; Dunning et al., 1995; Monestiez et al., 2005; Rigby et al., 2010; Verrelst et al., 2012) and ecomorphological studies (Mendoza and Palmqvist, 2008; Cassini et al., 2011), they have yet to be applied to the fossil record for the purposes of environmental/climatic reconstruction. Heretofore, linear regression has been the favored method due to its simplicity of output, application, and interpretation (Andrews et al., 1979; Kay and Madden, 1997; Reed, 1998; Kay et al., 2012b; Rowan et al., 2016). Here, we test these other regression methods along with the more conventional linear regression in order to explore their potential in paleoenvironmental reconstruction.

Additionally, we test all included methods on two well-sampled fossil localities in South America, the La Venta and Santa Cruz FL 1–7 Miocene faunas, in order to compare the estimates of these methods to

Fig. 4. Map of South American illustrating the faunal localities used in this study.

one another and, if possible, to previously published reconstructions. Brief descriptions of each of these fossil localities are presented in the following two paragraphs.

4.1. La Venta

4.1.1. Paleogeography

The La Venta faunal list used in this study comes from the Monkey Beds of the Villavieja Formation (Honda Group) in the Magdalena river valley in Colombia. The Honda Group, at about 5°N latitude, attains a thickness of 1150 m (Guerrero, 1996) and is constrained by radiometric and paleomagnetic evidence to a period beginning at approximately 13.5 Ma and ending at approximately 11.8 Ma (Flynn et al., 1996; Guerrero, 1996; Madden et al., 1996).

The Monkey Beds comprise a well-exposed richly fossiliferous paleosol layer of the Villavieja Formation with a thickness of approximately 14.8 m (Guerrero, 1996; Kay and Madden, 1997). Magnetostratigraphic evidence constrains the time interval of the Monkey Beds to ~15,000 years from between 13 and 12.85 Ma (Flynn et al., 1996; Kay and Madden, 1997).

In the mid-Miocene, the La Venta region was situated within 5° of the equator and was on a peninsula in northern South America that was bordered to the west by the Pacific Ocean and to the north and east by the Caribbean Sea that extended southward into what is now the llanos of Colombia (Whitmore and Stewart, 1965; Duque-Caro, 1990; Hoorn et al., 1995; Kay and Madden, 1997). Fossil vertebrate evidence, namely from amphibians and reptiles, suggests low elevation (< 100 m) in the middle Miocene La Venta, which stands in contrast to the region today, which is largely dominated by the Colombian Andes—though the actual fossil locality lies in the Magdalena river valley (approximately 500 m in elevation) between the Central and Eastern Cordilleras (Kay and Madden, 1997).

Previous paleoecological analyses at La Venta (Kay and Madden, 1996; Kay and Madden, 1997) have suggested a forested environment within a complex system of meandering rivers, and a mean average rainfall of between 1500 and 2000 mm/year. Seasonality was determined to not be a significant component of the middle Miocene at La Venta (Guerrero, 1996; Kay and Madden, 1997).

4.1.2. Fauna

The mammalian fauna from La Venta, Colombia remains the only well-sampled Tertiary tropical paleofauna to be found in South America, and as such is of essential importance to the understanding of the evolution of South American tropical communities. Kay and Madden (1997) reconstructed the rainfall at La Venta to be between 1500 and 2000 mm/year, and posited that the habitat was marked by

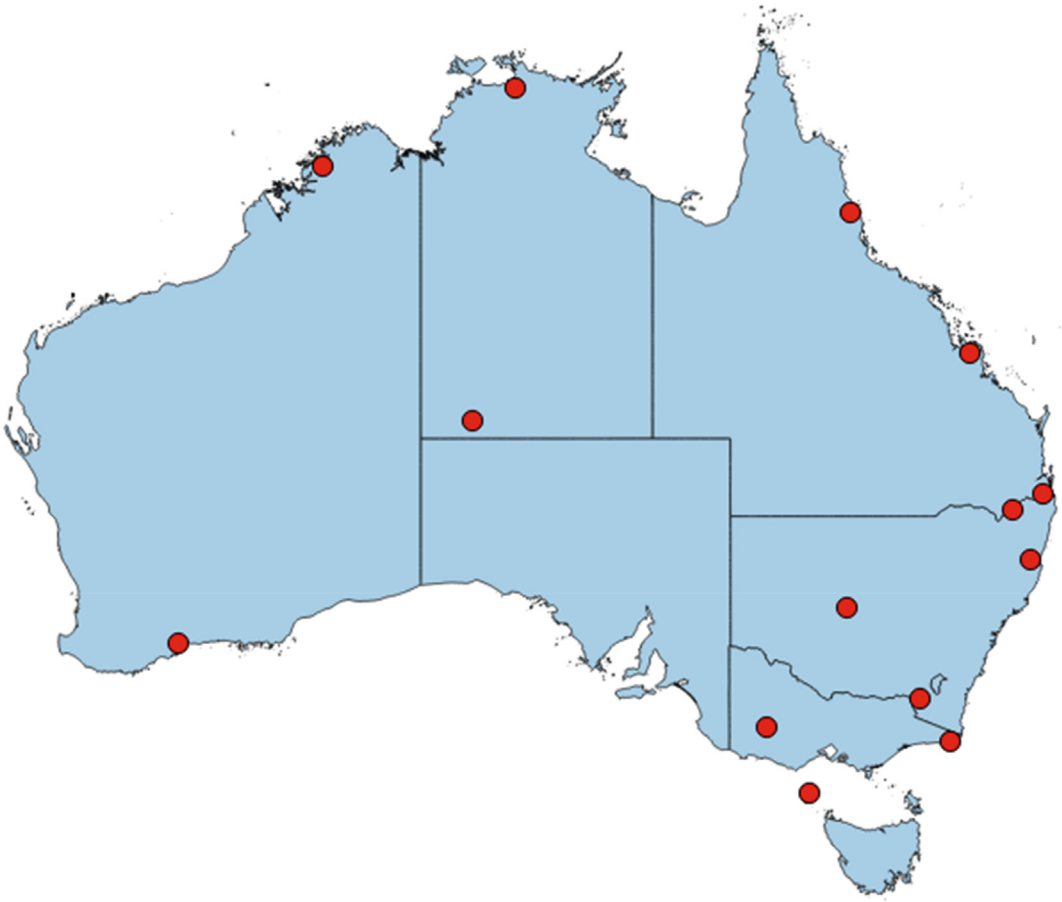

Fig. 5. Map of Australian faunal localities used in this study.

Table 1
Summary of niche categories.

Body mass category	Definition	Locomotor category	Definition
1	10–100 g	1	Large terrestrial (> 1 kg)
2	100 g–1 kg	2	Small terrestrial (< 1 kg)
3	1–10 kg	3	Arboreal (including gliding)
4	10–100 kg	4	Arboreal/terrestrial (scansorial)
5	100–500 kg	5	Semi-aquatic
6	> 500 kg	6	Fossorial (including semi-fossorial)

Dietary category	Definition
1	Vertebrate prey
2	Ants and termites
3	Insects (with some fruit)
4	Fruit (with some animals)
5	Small seeds of grasses (and other plants or insects)
6	Fruit with leaves
7	Leaves (browse)
8	Stems and leaves of grasses (graze)
9	Tree gums (with fruit/insects)

heterogeneity in both ground cover and the distribution of river systems. Croft (2001) came to a similar conclusion using a cenogram analysis. The conclusions of Kay and Madden (1997) were drawn from least-squares and polynomial regressions involving niche metrics obtained from 18 extant lowland mammalian communities of South America (all of which are included in this study, as well).

4.2. Santa Cruz

4.2.1. Paleogeography

The Santa Cruz Formation (SCF) in southern Argentina is an exposure of Early Miocene deposits stretching from the Atlantic coast of Patagonia in east to the Andes in the west, including exposures along the Santa Cruz River, presently between ~47° and 52°S latitude. Despite a significant amount of longitudinal movement of the South American continent throughout the Cenozoic, the SCF lies in approximately the same latitude today as it did in the early Miocene (Vizcaíno et al., 2012).

The fauna that we analyze here is found in the southern part of the SCF along the Atlantic coast south of Coy inlet in Santa Cruz province (Vizcaíno et al., 2012). Tauber (1994) identified fourteen fossiliferous levels (FL) 1–14 in this region, concluding that levels were successively younger from north to south. Tephrostratigraphic correlations and sedimentation rates reported by Perkins et al. (2012) established that the first seven levels (FL 1–7) are very close in age, spanning a period of about 100,000 years (~17.5–17.4 Ma).[1] Based on this information, Kay et al. (2012b) chose to treat FL 1–7 as a single paleofauna, as we do here.

[1] The fossiliferous levels (FL) 1–7 clearly have a lower temporal resolution than that from the Monkey beds. However, it is worth noting that we are interested in comparing these two sites only in the context of testing our models, rather than any detailed comparison between the faunal compositions of the two sites. Thus, this difference in temporal resolution can effectively be dismissed.

4.2.2. Fauna

The fossiliferous levels 1–7 of the SCF—considered by Kay et al. (2012a, 2012b) to be essentially a contemporaneous fauna—represent another of the most well sampled fossil mammalian localities in South America. Previous paleoecological analyses of the FL 1–7 of the SCF indicate a mosaic environment of open grasslands and patches of forests, with a high degree of temperature seasonality, rainfall of approximately 1000–1500 mm/year, and mean annual temperatures > 14 °C (Kay et al., 2012b). This reconstruction was based on a variety of different sources and techniques, including paleosol records indicating fluvial deposits (Matheos and Raigemborn, 2012), paleofloral records (Brea et al., 2012), the presence of both terrestrial and arboreal ecologies in the mammalian fauna (Degrange et al., 2012; Prevosti et al., 2012; Bargo et al., 2012; Candela et al., 2012; Vizcaíno et al., 2012a, 2012b; Abello et al., 2012; Cassini et al., 2012), the presence of an extant species of frog with known habitat tolerance (Fernicola and Albino, 2012), the presence of primates (Kay et al., 2012b), and community level analyses similar to what we present in this study (Kay et al., 2012b). Unlike La Venta, the SCF is of a much higher latitude, and presumably represents a cooler, drier climate with day/night seasonality. The SCF thus presents us with a chance to compare results between two localities that *should* have much different estimates for climate.

5. Results

5.1. Spatial autocorrelation

No statistically significant autocorrelation among the climatic variables across the South American localities was found using Mantel's r_M test statistic. This indicates general independence between our localities, but as a conservative measure any potential spatial autocorrelation between closely associated localities in our dataset was accounted for in the models by including latitude and longitude as predictor variables.

5.2. South American extant faunas

Results from the models obtained from the training data as applied to the test data are presented in Table 2. To quantify the error from these predictions, both root mean square error (RMSE) and mean absolute error (MAE) are presented. Generally, linear regression (LR) outperforms regression trees (both unpruned and pruned) but is surpassed in predictive performance by both random forests (RF) and Gaussian process regression (GPR). RF has significantly lower error terms for both mean annual precipitation (MAP) and precipitation seasonality. GPR has significantly lower error terms for both net primary productivity (NPP) and temperature seasonality. GPR and RF are roughly equivalent in their predictive performance in respect to both mean annual temperature (MAT) and canopy height, though GPR has slightly lower error in both cases. Figs. 6–8 illustrate predicted values vs. actual values derived from the test data for three of bioclimatic variables (only half of the variables are presented for the sake of brevity).

5.3. Australian faunas

Results from the models obtained from the South American training data as applied to the Australian data are presented in Table 3. In general, the error of predicted values for Australian localities is slightly greater than that from the values of South American localities (Table 2). For instance, the root mean square error for MAT using RF and GPR in South America is 3.3 °C and 3.7 °C, respectively, but is 5.5 °C and 5.0 °C in Australia. However, this is not true for MAP, for which the RTA, RF, and GPR all have lower error terms than what is seen in the South American test sample (Table 3, Fig. 9). For example, RF has a RMSE of 440 mm/year for the South American test sample, but only 246 mm/year in the Australian sample. For all climatic variables except MAP, a pruned RTA outperforms—that is, demonstrates the lowest mean error—all other modeling techniques, likely due to the small number of faunas sampled in the Australian dataset (discussed further below). Our results do suggest that the model derived from South America is broadly applicable to geographically and phylogenetically distinct fauna.

Table 2
Error metrics for climatic models.

Method	MAE	RMSE
	Mean annual temperature (°C)	
Baseline average	6.6	8.0
Linear regression	3.6	4.3
Regression tree (RTA)	4.2	5.8
Random forests (RF)	2.6	3.3
Gaussian process regression (GPR)	2.8	3.7
	Temperature seasonality (°C)	
Baseline average	16.0	18.4
Linear regression	8.5	11.1
Regression tree (RTA)	6.5	9.1
Random forests (RF)	6.7	8.6
Gaussian process regression (GPR)	5.3	6.8
	Mean annual precipitation (mm/year)	
Baseline average	802.5	944.3
Linear regression	410.5	512.3
Regression tree (RTA)	536.2	647.3
Random forests (RF)	536.2	647.3
Gaussian process regression (GPR)	377.7	439.1
	Precipitation seasonality (%)	
Baseline average	25.5	38.2
Linear regression	25.1	38.1
Regression tree (RTA)	22.3	30.9
Random forests (RF)	20.8	31.1
Gaussian process regression (GPR)	21.4	31.2
	Net primary productivity (gC/m^2/day)	
Baseline average	3722.2	5115.3
Linear regression	3194.5	4115.0
Regression tree (RTA)	3461.8	4494.5
Random forests (RF)	3461.8	4494.5
Gaussian process regression (GPR)	2686.4	3597.3
	Canopy height (m)	
Baseline average	8.0	9.3
Linear regression	6.1	7.6
Regression tree (RTA)	7.8	10.0
Random forests (RF)	7.8	10.0
Gaussian process regression (GPR)	5.4	6.9

Root mean square error (RMSE) and mean absolute error (MAE) indicate the average amount of error between the predicted values of bioclimatic variables and the actual recorded values.

5.4. Paleoclimatic reconstruction

Estimations of the six climatic variables for both the La Venta and Santa Cruz faunas are presented in Tables 4 and 5, respectively. Results from our models suggest general agreement with previously published paleoecological analyses of the Monkey Beds of La Venta (Kay and Madden, 1997) and Santa Cruz FL 1–7 (Kay et al., 2012b).

5.4.1. La Venta

With respect to the Monkey Beds, all models suggest rainfall between 1500 and 2000 mm/year, though the two machine-learning algorithm models suggest slightly higher predictions, around 2000 mm/year (Table 4). With the exception of the multivariate linear regression model which has a much lower estimates of ~21 °C, estimations of MAT are between 25 and 26 °C. There is disagreement between the models for the other bioclimatic variables. GPR and RTA both predict much lower values of temperature seasonality relative to linear regression and

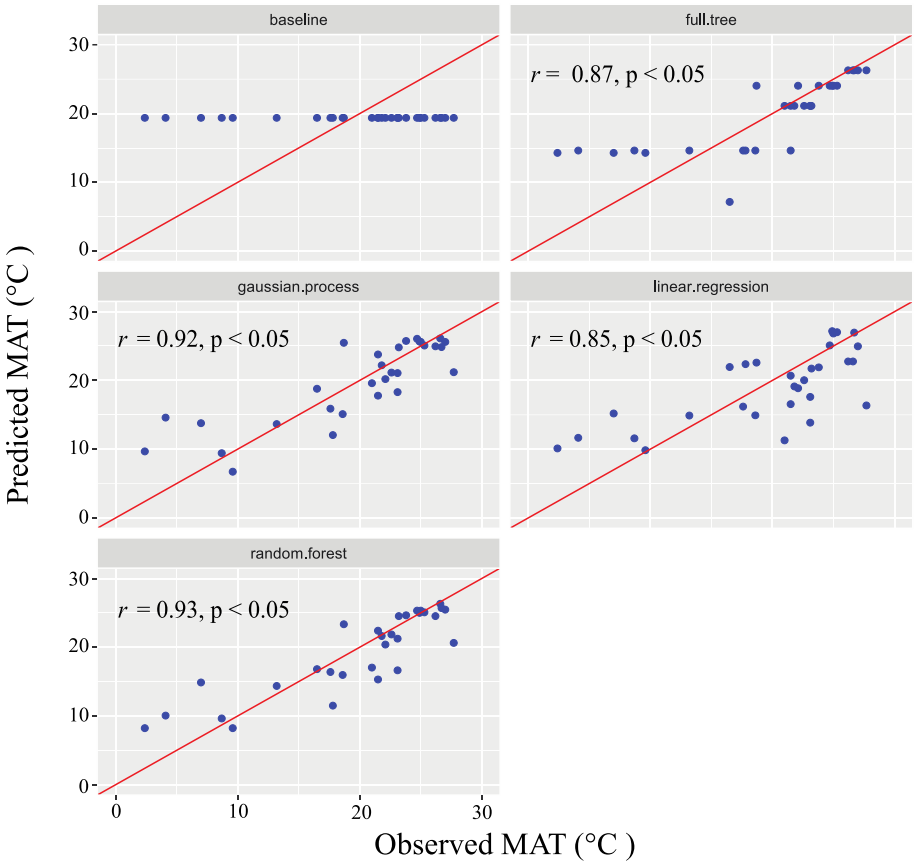

Fig. 6. Actual MAT (x-axis) and predicted MAT (y-axis) values for the test dataset. MAT values are presented as ˚C. The plots represent the five different regression techniques—and also a baseline mean (top left corner) that represents the average value for MAT across all localities—included in this study. The red line represents a hypothetical 1:1 relationship between predicted and actual values, and would suggest a perfect fit for the values predicted by the models and the observed values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

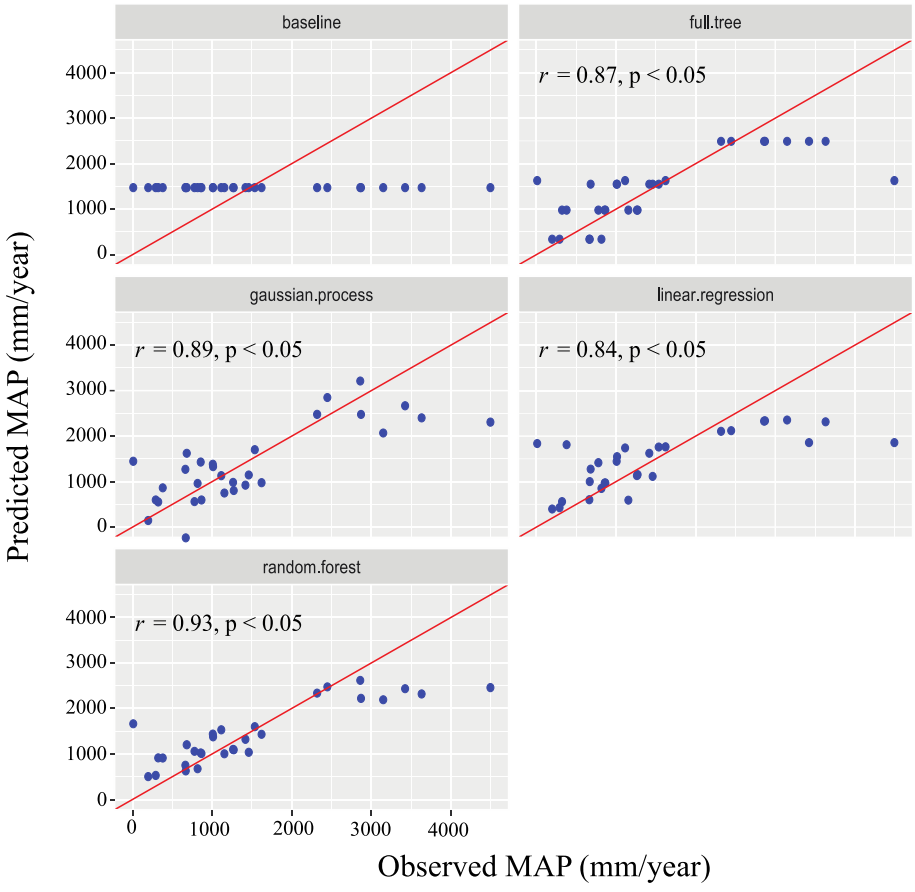

Fig. 7. Actual MAP (x-axis) and predicted MAP (y-axis) values for the test dataset. MAP values are presented as mm/year. The plots represent the five different regression techniques—and also a baseline mean (top left corner)—included in this study. The red line represents a hypothetical 1:1 relationship between predicted and actual values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

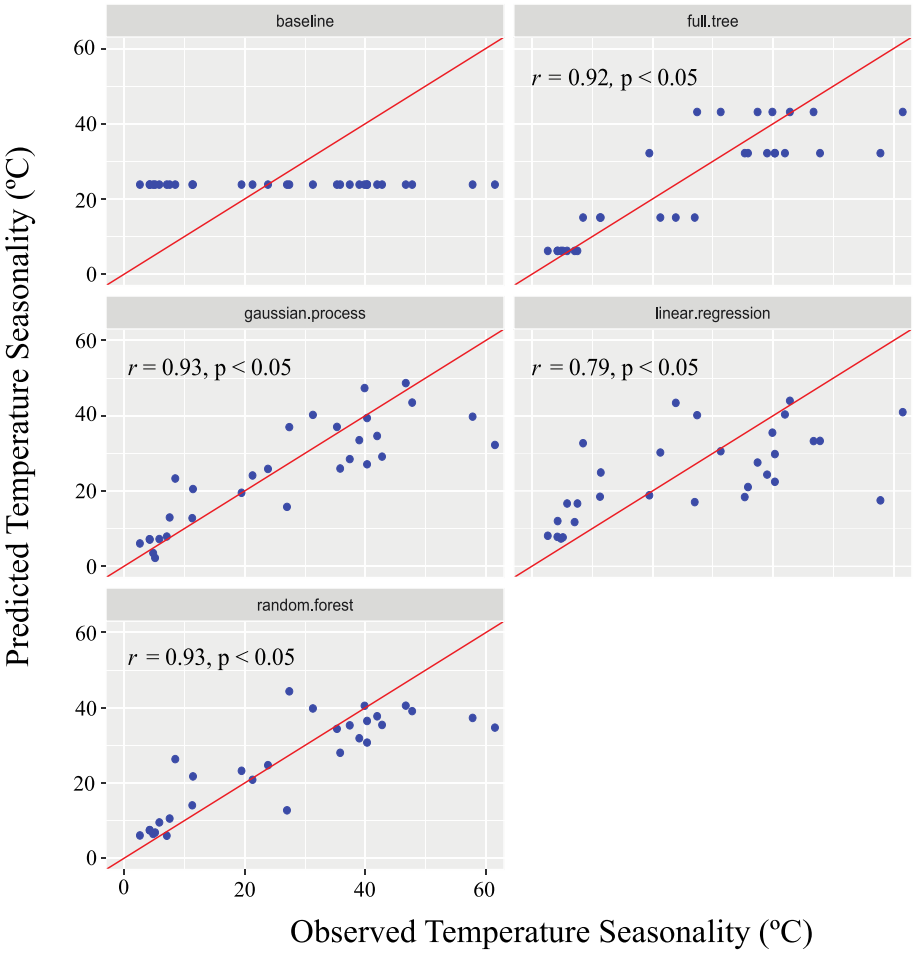

Fig. 8. Actual temperature seasonality (x-axis) and predicted temperature seasonality (y-axis) values for the test dataset. The plots represent the five different regression techniques—and also a baseline mean (top left corner)—included in this study. The red line represents a hypothetical 1:1 relationship between predicted and actual values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

RF. RTA estimates both the precipitation seasonality and NPP at the Monkey Beds as much higher than the other models. RTA also predicts that the canopy height is considerably lower (by at least 4 m) than the other models. On the other hand, GPR reconstructs the canopy height as significantly higher (~30 m) than the linear regression or RF (between 22 and 23 m).

5.4.2. Santa Cruz

For FL 1–7 of the Santa Cruz formation, all models suggest rainfall between 800 and 1350 mm/year, with the two machine-learning algorithm models reconstructing the MAP slightly higher (between 1250 and 1350 mm/year) than linear regression and regression tree analysis (Table 5). Most of the models suggest a MAT between 15 and 18 °C; the exception is RTA, which predicts the MAT closer to 23 °C. All models suggest substantial temperature seasonality. There is substantial disagreement between the models in regards to precipitation seasonality and NPP, which follows the significant amount of error we found with the testing of the models. Finally, all models suggest a canopy height of at least 16 m, with GPR having the highest estimate (25.8 m).

6. Discussion

The relationship between the community structure of a mammalian fauna and environment is undoubtedly complex, and it is impossible to take every potential confounding variable into consideration. Previous efforts to reconstruct paleoenvironments from mammalian data have acknowledged this complexity but have otherwise sought to establish linear relationships that can provide simple, easy to interpret estimations from the mammalian fossil record. Here, we have incorporated many more variables and considered more complex, non-linear relationships than simple least-squares regression. Given the importance of applying all available evidence to reconstruct paleoclimates and paleoenvironments, we seek to test the efficacy of sophisticated regression techniques on the estimation of climate variables in modern localities across South America solely from ecological information obtained from the local mammalian faunas. These methods allow researchers to include several uncertainty parameters (discussed in the introduction) and allow for more limited error ranges for any given, new data point, as the simplifying assumptions that are inherent to techniques like linear regression are not necessary in machine-learning algorithms.

6.1. Performance of regression techniques

Our results show a consistent improvement in accuracy for the two machine-learning algorithms, random forests (RF) and Gaussian process regression (GPR), over all other included methods. These two methods clearly benefit from the ability to apply data learning to the model, thus better accounting for the variation of the data. This is distinct from the linear regression, which only uses a best-fit line to all data. A posterior distribution allows GPR to more accurately account for non-linear relationships between predictor variables and the response variable, as well as provides useful confidence intervals (Williams, 1998; Cutler et al., 2007), while random bagging allows RF to test a large number of

Table 3
Error metrics for model testing of Australian localities.

Method	MAE	RMSE
	Mean annual temperature (°C)	
Baseline average	4.3	5.4
Linear regression	8.1	9.5
Regression tree (RTA)	4.9	6.1
Random forests (RF)	4.8	5.5
Gaussian process regression (GPR)	4.3	5.0
	Temperature seasonality (°C)	
Baseline average	10.3	12.6
Linear regression	18.0	21.8
Regression tree (RTA)	7.8	10.2
Random forests (RF)	5.7	9.6
Gaussian process regression (GPR)	5.9	9.7
	Mean annual precipitation (mm/year)	
Baseline average	443.7	525.3
Linear regression	641.8	765.7
Regression tree (RTA)	230.0	315.3
Random forests (RF)	183.2	246.0
Gaussian process regression (GPR)	207.6	269.2
	Precipitation seasonality (%)	
Baseline average	23.8	30.2
Linear regression	25.0	29.5
Regression tree (RTA)	11.8	15.2
Random forests (RF)	17.8	21.2
Gaussian process regression (GPR)	27.3	33.4
	Net primary productivity ($gC/m^2/day$)	
Baseline average	5249.9	5856.9
Linear regression	4783.8	5552.4
Regression tree (RTA)	5401.7	6891.3
Random forests (RF)	4400.9	5933.5
Gaussian process regression (GPR)	3533.1	5130.0
	Canopy height (m)	
Baseline average	7.3	9.2
Linear regression	6.9	8.7
Regression tree (RTA)	7.3	9.0
Random forests (RF)	5.9	8.3
Gaussian process regression (GPR)	7.2	8.4

potential tree-like models before deciding upon which produces the best fit.

There are, however, important a priori parameters in each technique before running the analyses, each of which will have significant effects on the results. Qualitatively, choosing good parameters was easier for random forests. Interpreting the parameters (number of trees, number of predictor variables, etc.) was straightforward, and there were a manageable number from which to choose. With regard to the number of response variables (m_{try}) randomly chosen at each node during RF modeling, setting m_{try} too high results in overfitting the data—discussed further in relation to regression trees below—while setting m_{try} too low also results in a significant decrease in performance. The current standard, used here, is to take the square root of the total number of predictor variables (Brieman, 2001: Liaw and Wiener, 2002). Further, a wide range of choices for the parameters resulted in models that gave reasonably accurate predictions. GPR had a larger number of parameters from which to choose, and some kernel functions made interpretation difficult. Many kernel functions resulted in poor models that we could not improve due to uninterpretable results.

Additionally, the percentage split of data randomly chosen for either the training or testing dataset when constructing the models can have a significant impact on the results. While the percentage split is arbitrarily chosen, intuitively one would be best served by choosing a majority of the data for the training set. The compromise lies in choosing enough of the data to have a meaningful effect on the trained model so that it essentially captures the variation of the entire dataset, but also having enough of the data set aside for testing so that interpretations from the results are meaningful. In general, previous studies have chosen a threshold of 70–80% for the training dataset (Cutler et al., 2007), and our choice of 70% is thus in keeping with previous studies. The second decision on the other hand is again chosen a priori on the basis of the structure of the data. Given that we are working with multidimensional regression, the recommended kernel function is the ANOVA function (Hoffman et al., 2008).

Interestingly, regression trees (included those that were "pruned" by cross-validation) do not perform as well as least-squares regression in the majority of the response variables included in this study. This is likely due to the fact that the output of the regression tree is binning the data too broadly, resulting in estimations that fit into a limited number of bins. Each bin has its own mean and its own confidence interval, so the result is that the estimations are themselves limited to just a handful of means equivalent to the number of bins. Compiling reliable confidence intervals for any given prediction (as we present with the two fossil localities) becomes complicated by the lack of confidence that any prediction was correctly placed in the right bin. The relatively few number of bins may be related to the relatively small number of localities in this study (< 100), which is significantly less than other similar studies (> 300) that have used regression tree analysis in relation to communities and environment (De'ath and Fabricius, 2000; De'ath, 2002). While a larger tree could potentially fix this problem, the standard protocol for determining tree size in previous studies is to choose the tree that has the smallest mean square error (De'ath, 2002), in a process known as "pruning". In our results, the pruned tree and the unpruned tree give the same results.

6.2. Interpretation of results of modern localities

6.2.1. Mean annual precipitation (MAP)

An interesting aspect of our results is that the localities with particularly high MAP (> 2500 mm/year) are consistently underestimated, regardless of which regression model is used. This suggests two things: 1) the various ecological indices used do not have a linear relationship with MAP, particularly at higher values, as has been suggested by previous studies (Kay et al., 1997; Spradley et al., 2015); and 2) the amount of data available for localities of such high MAP is limited, such that even the machine-learning based models only have a few data points with which to extrapolate (or interpolate). This is a difficulty acknowledged by previous studies of mammalian community structure (Andrews et al., 1979; Kay and Madden, 1997; Kay et al., 1997). In the context of the fossil record, it makes environments with such extreme amounts of rainfall difficult to identify, and it is likely only through indicator species (such as aquatic forms) that such an environment might be identified (Kay et al., 1997). It is also worth noting that in predicting MAP, the linear regression technique of Kay et al. (2012b) performed better even than GPR. This was the only climatic variable in which the linear regression outperformed either of the machine-learning approaches.

6.2.2. Mean annual temperature (MAT)

In general, MAT is estimated with high accuracy in modern South American localities, particularly by GPR and RF. In fact, in our results GPR and RF display a low amount of error for almost all MAT values for the entire test dataset. The only exceptions are two localities, the Ulla Ulla National Reserve in Bolivia and the Lauca National Park in Chile, which happen to correspond to the two localities with the highest elevation in our study, at ~5000 m and ~4500 m respectively. Of the localities found in this study, 89.4% are found at altitudes *below* 1500 m, so the inability of the models to accurately estimate the climate variables at high altitudes should come as no surprise, as high altitudes produce much different physiological demands on the organisms (Lenfant, 1973; Monge and Leon-Velarde, 1991). In general, our results suggest that mammalian community ecological indices are closely related to temperature. This bears close resemblance to plant communities, among which diversity and richness have also been shown to

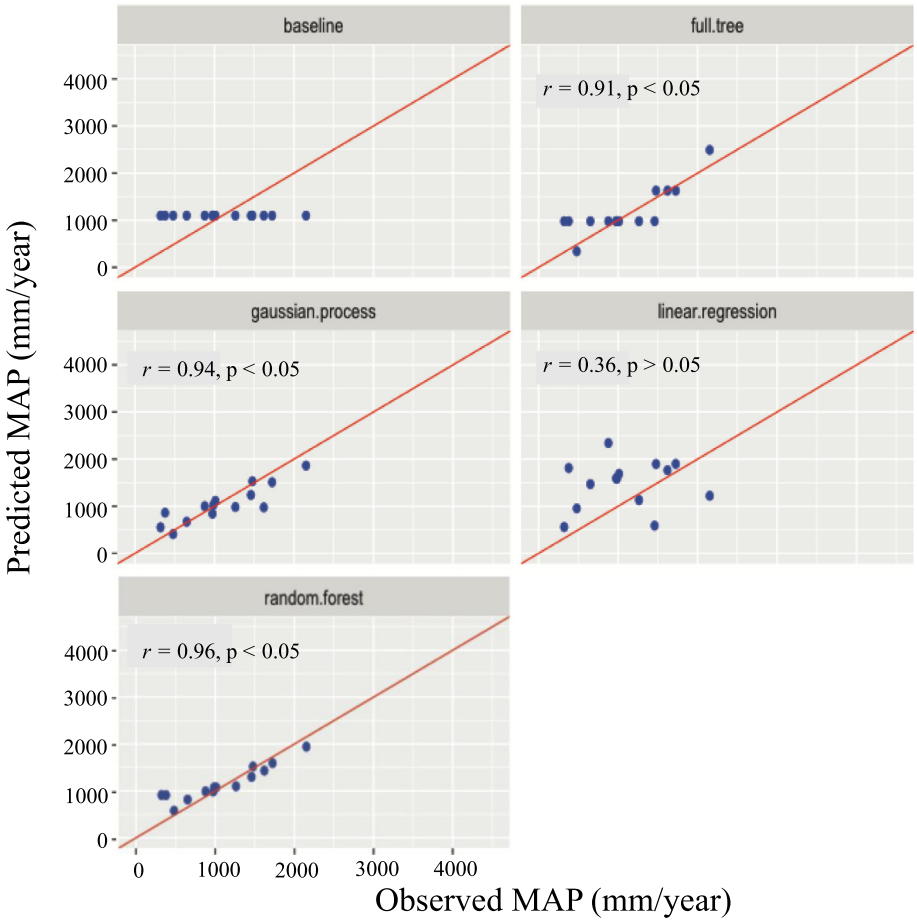

Fig. 9. Actual MAP (x-axis) and predicted MAP (y-axis) values for the Australian dataset. MAP values are presented as mm/year.

have a positive correlation with temperature (see Pausas and Austin, 2001 for review of studies of plant species richness) and is also in agreement with the energy-water hypothesis of species richness of Hawkins et al. (2003).

6.2.3. *Temperature seasonality*

Most of the regression models (with the exception of the pruned regression tree) predict temperature seasonality very well, GPR in particular. This result is consistent with similar studies which found that temperature seasonality was a good predictor of animal species richness (Danell et al., 1996; Andrews and O'Brien, 2000; Hawkins et al., 2003; Hurlbert and Haskell, 2003). Temperature seasonality has a particularly strong relationship with latitude relative to the other climatic variables, and thus may be a key driver of the latitudinal gradient of species richness. In a review of species richness and climate, Hawkins et al. (2003) noted that lower temperatures not only drive plant productivity down by limiting energy for photosynthesis, but they also create higher metabolic demands on the fauna. Both of these factors are likely limiting the ecological richness of the mammalian fauna in these localities. Kay et al. (2012b) also suggested that certain ecological niches in these highly seasonal environments might be filled by avian species that can then migrate during seasons of low plant productivity. This hypothesis is supported by Aizen and Ezcurra (1998), who noted that seed dispersal in temperate forests is accounted for by just a few bird species, many of them migratory. Therefore, a seasonal but productive environment may not carry a strong signal in the mammalian fauna.

Regardless, our results suggest that temperature seasonality can be accurately predicted using ecological indices, with the best performing predicting variables being the Frugivore Index, the Arboreality Index, and the Grazing Index.

6.2.4. *Precipitation seasonality*

Seasonality in precipitation displays a much more complex

Table 4
Estimations of climatic variables in the Miocene of La Venta.

Model	MAP	MAT	Temp. seasonality	Precip. seasonality	NPP	Canopy height
Linear regression	**1851** (424.7–3277.0)	**22.4** (14.0–30.9)	**17.0** (0–37.9)	**54.7** (0–115.5)	**10,411** (1236–19,587)	**23** (6.4–39.9)
Regression tree	**2344**	**26.1**	**5.9**	**67.9**	**7458**	**21**
Random forest	**2396** (907–4265.5)	**25.7** (21.6–27.8)	**10.3** (3.6–20.7)	**69.1** (28–142)	**9092** (2797–18,909)	**23** (10.7–34.3)
Gaussian processes	**2882** (2762.2–3002.6)	**27.3** (23.7–30.9)	**9.2** (0–24.2)	**63.9** (41.9–85.9)	**11,355** (11067–11,633)	**30** (17.6–43.0)

**Numbers in parentheses represent 95% confidence intervals for each prediction.

Table 5
Estimations of climatic variables in the Miocene of Santa Cruz.

Model	MAP	MAT	Temp. seasonality	Precip. seasonality	NPP	Canopy height
Linear regression	**987** (0–2416.1)	**15.0** (6.6–23.5)	**32.3** (11.4–53.2)	**57.2** (0–118.2)	**6264** (0–15,462)	**16** (0–33.0)
Regression tree	**840**	**23.3**	**44.0**	**24.7**	**2554**	**21**
Random forest	**1344** (222–3424)	**18.3** (4.5–27.4)	**32.1** (11–48)	**44.4** (12–97)	**6546** (608–16,923)	**19** (8–35)
Gaussian processes	**1264** (1122.2–1406.6)	**15.7** (11.5–19.9)	**33.2** (15.2–51.2)	**8.3** (0–33.7)	**8977** (8637–9317)	**26** (10.9–30.7)

relationship with mammalian faunal composition. This is illustrated by the fact that RF performs only slightly better than the baseline (the average precipitation seasonality value from across all 85 localities), and GPR even performs slightly worse. Unlike temperature seasonality, precipitation seasonality does not have a consistent relationship with any other climatic variable included in this study, suggesting that it might be more dependent on local geography than on temperature, or that short-term climatic events that occurred during the decades of data collection on which BioClim is based are artificially inflating the precipitation seasonality values of multiple localities. That this result stands in contrast with that from temperature seasonality may suggest that precipitation seasonality in and of itself is not as limiting to plant productivity (and thus faunal richness) as is temperature seasonality. While these causes for this poor relationship are purely speculative, the fact remains that based on our results from mammalian niche structure as we have characterized it, precipitation seasonality would be difficult to predict accurately.

6.2.5. Net primary productivity (NPP)

As it is the most direct measure of the energy available to herbivores and up through the food web, NPP should be expected to have a strong correlative relationship with many of our ecological indices. However, results from our models suggest that it is nevertheless a difficult variable to predict using those same indices. This is perhaps best illustrated by a simple bivariate plot of NPP and the Frugivore Index (Fig. 10), showing that the relationship between these two variables is extremely poor ($r = 0.20$). This is a surprising result considering that NPP is supposedly an indirect measure of the metabolism of plants as they absorb carbon dioxide during the production of sugars and starches during photosynthesis (specifically, NPP is the difference between CO_2 taken in during photosynthesis and CO_2 released during respiration and is measured in grams of carbon per square meter per day). Higher NPP values would suggest that plants are converting more and more carbon dioxide into energy; energy that would presumably be available to the mammalian fauna for consumption (Kay et al., 1997). We might expect, then, that primary consumers in the community should be particularly species rich in an environment with high NPP. That we do not find such a relationship may be indicative of three things: 1) NPP is a poor predictor of plant species richness, such that NPP is being driven by a relatively small number of plant species with high abundance (and thus ecological richness in the mammalian fauna is similarly low); 2) methods used to measure NPP (NASA's Moderate Resolution Imaging Spectroradiometer, or MODIS) are inexact for the level of resolution that is present in the other variables included in this study; or 3) mammalian species richness is not inherently linked to NPP. This last point is complicated by previous analyses of productivity and species richness (Mittelbach et al., 2001; Sandom et al., 2013), which showed that such relationships in mammals are highly variable, displaying both positive and negative correlations depending on the study region. This variability is also likely to affect richness of secondary consumers in mammals, which has been shown to have a strong relationship with the richness of prey species (Cassini and Vizcaíno, 2006; Sandom et al., 2013). Regardless of the cause, NPP, as measured remotely, would appear to be a difficult variable to predict accurately from our mammalian faunal indices.

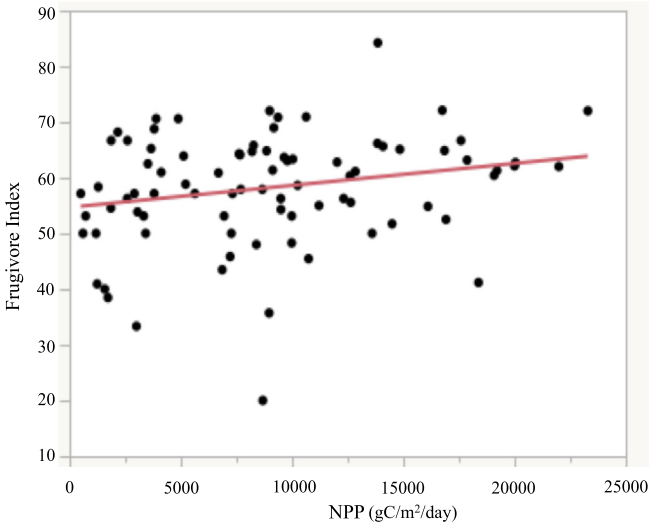

Fig. 10. NPP versus Frugivore Index for the South American dataset. Solid red line represents least-squares regression ($r = 0.70$, $p < 0.05$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.2.6. Canopy height

Results from our regression models suggest that both GPR and RF are capable of predicting canopy height to within less than five meters. This, unsurprisingly, is being driven by the arboreality index and the browsing index, two indices that directly (or at least indirectly) measure the richness of arboreal species in the community. This is in accordance with the findings of Louys et al. (2015), who found that tree cover (likely related to canopy height) had a positive correlation with similar ecological indices. That these variables may be able to be reconstructed in the fossil record suggests the potential to reconstruct fine scale differences in paleoenvironments, especially in combination with geochemical traces in the fossil record (Cerling et al., 2011). In turn, this would permit the study of such differences on the evolution of different arboreal clades.

6.3. Australian faunas

In order to compare our results from South America to another continent, we also include data on a number of localities representative of diverse environments from across Australia. One interesting result from this dataset is that the regression tree, and particularly the pruned regression tree performs well, especially regarding the mixed results from these same models in the South American dataset. This is likely because of the small number of localities from Australia and the binning effect that limited the accuracy of the regression trees in South America is actually manifested as a positive effect on the accuracy in Australia. We would suggest that including a greater number of localities from

Australia might decrease the accuracy of the regression tree and would be more comparable to the results from South America.

In general, however, our results do suggest that models derived from the South American data can be used to predict the climate variables in Australian localities. This in turn suggests that the patterns we see in South America are generalizable to other continents and are not just a regional phenomenon in the South American continent, and also increases out confidence that we can extrapolate our findings to the SA fossil record even though the phylogenetic composition of middle Cenozoic SA faunas is radically different from that of today.

6.4. Paleoenvironment reconstructions

Finally, we applied these models to two well-sampled Miocene faunas from the South American fossil record, one from the Santa Cruz Formation of southern Argentina and the other from the La Venta Monkey Beds of the Villavieja Formation, Colombia. In addition to being well sampled, the two faunas are different from one another in apparent ecology and taxonomic composition (see: Kay and Madden, 1997; Kay et al., 2012a, 2012b) making them an ideal test of our modeling techniques.

6.4.1. La Venta

Our own estimates of rainfall for La Venta exceed those of Kay and Madden (1997) and suggest rainfall in the range of 2000–2500 mm/year, with the lowest estimate suggesting rainfall in the range of 1800–2000 mm/year. In general, the reconstructed climatic variables suggest an environment that is much more likely to be characterized as a tropical rainforest, with high average temperatures (> 25 °C), high NPP, and a relatively tall canopy (> 23 m). These estimates most closely resemble those of the extant lowland Amazon sites in Brazil, Peru, or Venezuela. This reconstruction differs substantially from Kay and Madden (1997) and suggests an environment much more like the modern environments from a similar latitude. This is also in accordance with other modern localities that have a similarly high number of primates (Spradley et al., 2015), which are heavily forested with very few breaks in the canopy.

However, our estimations for temperature seasonality for La Venta are substantially higher than most modern tropical lowland localities in the Amazon (the same can be said of precipitation seasonality, though as mentioned above, the difficulty we had in accurately predicting precipitation seasonality with our models gives us pause in claiming anything definitive about this variable in the fossil record). This high seasonality may help to explain the conclusions of Kay and Madden (1997) and Croft (2001) that the La Venta fauna seems to have inhabited a significantly different environment compared to modern tropical environments at similar latitudes and geography and may still suggest a habitat unlike that of the modern Orinoco and Amazon Basins. This may be related to the uplift of the Eastern Cordillera of the Colombian Andes, which has been dated to approximately the same period as the La Venta fauna (Barke and Lamb, 2006) or even earlier (Horton et al., 2010), and has been suggested as a major driver of environmental change in the Miocene of this region (Hoorn et al., 1995). On the other hand, other studies have suggested that this uplift is considerably younger than the La Venta fauna (Guerrero, 1996).

6.4.2. Santa Cruz

Kay et al. (2012b) summarize the climate of the SCF as a markedly seasonal environment with rainfall > 1000 mm/year and a MAT of > 14 °C. Unsurprisingly, given the considerable overlap in our methods with those of Kay et al. (2012b), their estimates are generally in line our results derived from GPR, RF, and linear regression of PC1, providing further evidence for these paleoclimatic reconstructions. There is considerable difference, however, with the results from RTA, which has a significantly lower estimate for MAP and NPP. Given the difficulties we encountered with RTA in the extant dataset, however, we believe that it is safe to dismiss these particularly low estimates from RTA as spurious.

7. Conclusions

Using South American and Australian mammalian communities as data, our results suggest that ecological information from the mammalian fauna can be successfully used to retrodict (or reconstruct) a number of bioclimatic variables in the Miocene of South America, not just precipitation as has been previously demonstrated Among the six bioclimatic variables included in this study, MAT, MAP, temperature seasonality, and canopy height were the most accurately predicted by our regression models. Precipitation seasonality and NPP are much more difficult to reconstruct accurately, presumably because both of these variables have much more complex relationships with the local geography and/or other historical factors.

Additionally, of the regression models that we include in this study, we demonstrate that two machine-learning approaches—random forests and Gaussian process regression—consistently show high levels of accuracy in predicting our bioclimatic variables in a test dataset. The tradeoff that comes with this accuracy is the lack of interpretability and clarity of the model(s). As a way of accommodating these issues with the fact that machine-learning methods do seem to be more accurate, we suggest including a more conventional approach along with RF or GPR. This can allow readers to appreciate the general pattern/direction of the relationships in the data, while also presenting results from RF or GPR.

Acknowledgments

We would like to thank Drs. JD Pampush, Blythe Williams, Doug Boyer, and Louise Roth for helpful comments and suggestions regarding this paper. We would also like to thank Drs. Sergio Vizcaíno and Susanna Bargo for formative discussions and helpful comments on the manuscript. This research is funded by the National Science Foundation, grant numbers EAR 1338694 and EAR 1349741.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.palaeo.2019.01.014.

References

Abello, M.A., Ortiz-Jaureguizar, E., Candela, A.M., 2012. Paleoecology of the Paucituberculata and Microbiotheria (Mammalia, Marsupialia) from the late Early Miocene of Patagonia. In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 156–172.

Aizen, M.A., Ezcurra, C., 1998. High incidence of plant-animal mutualisms in the woody flora of the temperate forest of southern South America: biogeographical origin and present ecological significance. Ecol. Austral 8, 217–236.

Andrews, P., O'Brien, E.M., 2000. Climate, vegetation, and predictable gradients in mammal species richness in southern Africa. J. Zool. 251, 205–231.

Andrews, P., Van Couvering, J.A.H., 1975. Palaeoenvironments in the East African Miocene. In: Approaches to Primate Paleobiology. 5. pp. 62–103.

Andrews, P., Lord, J.M., Nesbit-Evans, E.M., 1979. Patterns of ecological diversity in fossil and modern mammalian faunas. Biol. J. Linn. Soc. 11, 177–205.

Anthony, M.R.L., Kay, R.F., 1993. Tooth form and diet in ateline and alouattine primates: reflections on the comparative method. Am. J. Sci. 283A, 356–382.

Bargo, M.S., Toledo, N., Vizcaíno, S.F., 2012. Paleobiology of the Santacrucian sloth and anteaters (Xenarthra, Pilosa). In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 216–242.

Barke, R., Lamb, S., 2006. Late Cenozoic uplift of the Eastern Cordillera, Bolivian Andes. Earth Planet. Sci. Lett. 249, 350–367.

Behrensmeyer, A.K., Dechant-Boaz, D.E., 1980. The recent bones of Amboseli Park, Kenya, in relation to East African paleoecology. In: Fossils in the Making: Vertebrate Taphonomy and Paleoecology. University of Chicago Press, Chicago, pp. 72–92.

Boyer, D.M., 2008. Relief Index of second mandibular molars is a correlate of diet among prosimian primates and other euarchontan mammals. J. Hum. Evol. 55, 1118–1137.

Boyer, D.M., Seiffert, E.R., Gladman, J.T., Bloch, J.I., 2013. Evolution and allometry of calcaneal elongation in living and extinct primates. PLoS One 8 (7), e67792.

Brea, M., Zucol, A.F., Iglesias, A., 2012. Fossil plant studies from late Early Miocene of the Santa Cruz formation: paleoecology and paleoclimatology at the passive margin of Patagonia, Argentina. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 104–128.
Brieman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Candela, A.M., Rasia, L.L., Pérez, ME., 2012. Paleobiology of Santacrucian caviomorph rodents: a morphofunctional approach. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 287–305.
Cassini, G.H., Vizcaíno, S.F., 2006. Paleoecology of the large carnivore guild from the late Pleistocene of Argentina. Acta Palaeontol. Pol. 51 (3), 407–422.
Cassini, G.H., Mendoza, M., Vizcaíno, S.F., Bargo, M.S., 2011. Inferring habitat and feeding behaviour of early Miocene notoungulates from Patagonia. Lethaia 44, 153–165.
Cassini, G.H., Cerdeño, E., Villafañe, A.L., Muñoz, N.A., 2012. Paleobiology of Santacrucian native ungulates (Meridiungulata: Astraopotheria, Litopterna and Notoungulata). In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 243–286.
Cerling, T.E., Wynn, J.G., Andanje, S.A., Bird, M.I., Korir, D.K., Levin, N.E., Mace, W., Macharia, A.N., Quade, J., Remien, C.H., 2011. Woody cover and hominin environments in the past 6 million years. Nature 476, 51–56.
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250.
Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J., 2010. Training and testing low-degree polynomial data mappings via linear SVM. J. Mach. Learn. Res. 1471–1490.
Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V., Wikle, C.K., 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol. Appl. 19 (3), 553–570.
Croft, D.A., 2001. Cenozoic environmental change in South America as indicated by mammalian body size distributions (cenograms). Divers. Distrib. 7, 271–287.
Croft, D.A., 2006. Do marsupials make good predators? Insights from predator-prey diversity ratios. Evol. Ecol. Res. 8 (7), 1193–1214.
Croft, D.A., 2013. What constitutes a fossil mammal community in the early Miocene Santa Cruz Formation? J. Vertebr. Paleontol. 33 (2), 401–409.
Crompton, A.W., Lumsden, A.G.S., 1970. Functional significance of the therian molar pattern. Nature 227, 197–199.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.
Damuth, J., 1982. Analysis of the preservation of community structure in assemblages of fossil mammals. Paleobiology 8 (4), 434–446.
Danell, K., Lundberg, P., Niemela, P., 1996. Species richness in mammalian herbivores: patterns in the boreal zone. Ecography 19, 404–409.
Darwin, C., 1859. The Origin of Species, Sixth edition. 1872 D. Appleton and Company, London (Reprinted by) (1859).
Degrange, F.J., Noriega, J.I., Areta, JI., 2012. Diversity and paleobiology of the Santacrucian birds. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 138–155.
De'ath, G., 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83 (4), 1105–1117.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.
Di Giacomo, M., Fariña, R.A., 2017. Allometric models in paleoecology: trophic relationships among Pleistocene mammals. Palaeogeogr. Palaeoclimatol. Palaeoecol. 471, 15–30.
Dunning, J., Stewart, D., Danielson, B., Noon, B., Root, T., Lamberson, R., Stevens, E., 1995. Spatially explicit population models: current forms and future uses. Ecol. Appl. 5, 3–11.
Duque-Caro, H., 1990. Neogene stratigraphy, paleoceanography and paleobiogeography in northwest South America and the evolution of the Panama Seaway. Palaeogeogr. Palaeoclimatol. Palaeoecol. 77, 203–234.
Efron, B., Tibshirani, R., 1991. Statistical data analysis in the computer age. Science 253, 390–395.
Fariña, R.A., 1996. Trophic relationship among Lujanian mammals. Evol. Theory 11, 125–134.
Fernicola, J.C., Albino, A., 2012. Amphibians and squamate reptiles from the Santa Cruz Formation (late Early Miocene), Santa Cruz Province, Argentina: paleoenvironmental and paleobiological considerations. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 129–137.
Fleming, T.H., 1973. Numbers of mammalian species in north and central american forest communities. Ecology 54, 555–563.
Flynn, J.J., Guerrero, J., Swisher, C.C., 1996. Geochronology of the Honda group. In: Kay, R.F., Madden, R.H., Cifelli, R.L., Flynn, J.J. (Eds.), Vertebrate Paleontology in the Neotropics: The Miocene Fauna of La Venta, Colombia. Smithsonian Institution Press, Washington, D.C..
Gingerich, P.D., 1989. New Earliest Wasatchian Mammalian Fauna From the Eocene of Northwestern Wyoming: Composition and Diversity in a Rarely Sampled High-Floodplain Assemblage. No. 56 GIN.
Golding, N., Purse, B.V., 2016. Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods Ecol. Evol. 7, 598–608.
Gould, S.J., 1982. The stinkstones of oeningen. Nat. Hist. 91 (6), 6–13.
Gould, S.J., Lewontin, R.C., 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B Biol. Sci. 205 (1161), 581–598.
Guerrero, J., 1996. Stratigraphy, sedimentary environments, and the Miocene uplift of the Colombian Andes. In: Kay, R.F., Madden, R.H., Cifelli, R.L., Flynn, J.J. (Eds.), Vertebrate Paleontology in the Neotropics: The Miocene Fauna of La Venta, Colombia. Smithsonian Institution Press, Washington, D.C..
Harrison, J.L., 1962. The distribution of feeding habits among animals in a tropical rain forest. J. Anim. Ecol. 34, 53–64.
Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guegan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M., Porter, E.E., 2003. Energy, water, and broad-scale geographic patterns of species richness. Ecology 84 (12), 3105–3117.
Hayes, J.M., Freeman, K.H., Popp, B.N., Hoham, C.H., 1990. Compound-specific isotopic analyses: a novel tool for reconstruction of ancient biogeochemical processes. Org. Geochem. 16 (4), 1115–1128.
Hoffman, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. Ann. Stat. 1171–1220.
Hoorn, C., Guerrero, J., Sarmiento, G.A., Lorente, M.A., 1995. Andean tectonics as a cause for changing drainage patterns in Miocene northern South America. Geology 23 (5), 237–240.
Horton, B.K., Parra, M., Saylor, J.E., Nie, J., Mora, A., Torres, V., Stockli, D.F., Strecker, M.R., 2010. Resolving uplift of the northern Andes using detrital zircon age signatures. GSA Today 20 (7), 4–10.
Hurlbert, A.H., Haskell, J.P., 2003. The effect of energy and seasonality on avian species richness and community composition. Am. Nat. 161, 83–97.
Hutchinson, G.E., 1957. Concluding remarks. Cold Spring Harb. Symp. Quant. Biol. 22, 415–427.
Hutcinson, G.E., 1965. The Ecological Theater and the Evolutionary Play. Yale University Press.
Iverson, L., Prasad, A., 1998. Estimating regional plant biodiversity with GIS modelling. Divers. Distrib. 4 (2), 49–61.
Kay, R.F., Cartmill, M., 1977. Cranial morphology and adaptations of *Palaechthon nacimienti* and other Paromomyidae (Plesiadapoidea, ? primates), with a description of a new genus and species. J. Hum. Evol. 6 (1), 19IN533-3253.
Kay, R.F., Madden, R.H., 1996. Paleogeography and paleoecology. In: Kay, R.F., Madden, R.H., Cifelli, R.L., Flynn, J.J. (Eds.), Vertebrate Paleontology in the Neotropics: The Miocene Fauna of La Venta, Colombia. Smithsonian Institution Press, Washington, D.C..
Kay, R.F., Madden, R.H., 1997. Mammals and rainfall: paleoecology of the middle Miocene at La Venta (Colombia, South America). J. Hum. Evol. 32, 161–199.
Kay, R.F., Madden, R.H., Van Schaik, C., Higdon, D., 1997. Primate species richness is determined by plant productivity: implications for conservation. Proc. Natl. Acad. Sci. U. S. A. 94, 13023–13027.
Kay, R.F., Schmitt, D., Vinyard, C.J., Perry, J.M., Shigehara, N., Takai, M., Egi, N., 2004. The paleobiology of Amphipithecidae, south Asian late Ecocene priamtes. J. Hum. Evol. 46 (1), 3–25.
Kay, R.F., Vizcaíno, S.F., Bargo, M.S., 2012a. The paleoenviroinment and paleoecology of the coastal Miocene Santa Cruz Formation (late Early Miocene, Argentina). J. Vertebr. Paleontol. 32 (6), 119.
Kay, R.F., Vizcaíno, S.F., Bargo, M.S., 2012b. A review of the paleoenvironment and paleoecology of the Miocene Santa Cruz Formation. In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 331–364.
Kidwell, S.M., Flessa, K.W., 1996. The quality of the fossil record: populations, species and communities 1. Annu. Rev. Earth Planet. Sci. 24 (1), 433–464.
Kohn, M.J., 2010. Carbon isotope compositions of terrestrial C3 plants as indicators of (paleo) ecology and (paleo) climate. Proc. Natl. Acad. Sci. 107 (46), 19691–19695.
Kowaleski, M., 1996. Time-averaging, overcompleteness, and the geological record. J. Geol. 317–326.
Lenfant, C., 1973. High altitude adaptation in mammals. Am. Zool. 13 (2), 447–456.
Liaw, A., Wiener, M., 2002. Classification and regresion by random forest. R News 2 (3), 18–22.
Louys, J., Meloro, C., Elton, S., Ditchfield, P., Bishop, L.C., 2015. Analytical framework for reconstructing environmental variables from mammal community structure. J. Hum. Evol. 78, 1–11.
Madden, R.H., Guerrero, J., Kay, R.F., Flynn, J.J., Swisher, C.C., Walton, A.H., 1996. The Laventan stage and age. In: Kay, R.F., Madden, R.H., Cifelli, R.L., Flynn, J.J. (Eds.), Vertebrate Paleontology in the Neotropics: The Miocene Fauna of La Venta, Colombia. Smithsonian Institution Press, Washington, D.C..
Matheos, S.D., Raigemborn, MS., 2012. Sedimentology and paleoenvironment of the Santa Cruz formation. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 59–82.
Mendoza, M., Palmqvist, P., 2008. Hypsodonty in ungulates: an adaptation for grass consumption or for foraging in open habitat? J. Zool. 274, 134–142.
Mercer, J., 1909. Functions of positive and negative type and their connection with the theory of integral equations. Phil. Trans. R. Soc. A 209 (441–458), 415–446.
Michaelson, J., Schimel, D.S., Friedl, M.A., Davis, F.W., Dubayah, R.C., 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J. Veg. Sci. 5 (5), 673–686.
Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I., Gough, L., 2001. What is the observed relationship between species richness and productivity? Ecology 82 (9), 2381–2396.
Moisen, G.G., 2008. Classification and regression trees. 2008. In: Jørgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology. vol. 1. Elsevier, Oxford, UK, pp. 582–588.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J.P., Guinet, C., 2005. Comparison of model based geostatistical methods in ecology: application to fin whale spatial distribution in northwestern Mediterranean Sea. In: Geostatistics Banff 2004. Springer, Netherlands, pp. 777–786.
Monge, C., Leon-Velarde, F., 1991. Physiological adaptation to high altitude: oxygen transport in mammals and birds. Physiol. Rev. 71 (4), 1135–1172.
Moore, D.E., Lees, B.G., Davey, S.M., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. J. Environ. Manag. 15, 59–71.
Ogle, K., 2009. Hierarchical Bayesian statistics: mergin experimental and modeling approaches in ecology. Ecol. Appl. 19 (3), 577–581.
Okubo, A., Levin, S.A., 1980. Diffusion and Ecological Problems. Springer, New York.
Pausas, J.G., Austin, M.P., 2001. Patterns of plant species richness in relation to different environments: an appraisal. J. Veg. Sci. 12 (2), 153–166.
Perkins, M.E., Fleagle, J.G., Heizler, M.T., Nash, B., Bown, T.M., Tauber, A., Dozo, MT., 2012. Tephrochronology of the Miocene Santa Cruz and Pinturas formations, Argentina. In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 23–40.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
Prevosti, F.J., Forasiepi, A.M., Ercoli, M.D., Turazzini, GF., 2012. Paleoecology of the mammalian carnivores (Methatheria, Sparassodonta) of the Santa Cruz Formation (late Early Miocene). In: Vizcaíno, SF, Kay, RF, Bargo, MS (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, pp. 173–193.
QGIS, 2017. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning.
Reed, K.E., 1998. Using large mammal communities to examine ecological and taxonomic structure and predict vegetation in extant and extinct assemblages. Paleobiology 24 (3), 384–408.
Rigby, P., Pizarro, O., Williams, S.B., 2010. Towards adaptive benthic habitat mapping using gaussian process classification. J. Field Rob. 27 (6), 741–758.
Robinson, J.R., Rowan, J., Campisano, C.J., Wynn, J.G., Reed, K.E., 2017. Late Pliocene environmental change during the transition from *Australopithecus* to *Homo*. Nat. Ecol. Evol. 1 (6), 0159.
Ross, C.F., Kirk, E.C., 2007. Evolution of eye size and shape in primates. J. Hum. Evol. 52, 294–313.
Rowan, J., Kamilar, J.M., Beaudrot, L., Reed, K.E., 2016. Strong influence of palaeoclimate on the structure of modern African mammal communities. Proc. R. Soc. London 283 (1840), 1207.
Royle, J.A., Dorazio, R.M., 2008. Hierarchical Modeling and Inference in Ecology: The Analysis of Data From Populations, Metapopulations and Communities. Elsevier.
Sandom, C., Dalby, L., Fløjgaard, C., Kissling, W.D., Lenoir, J., Sandel, B., Trøjelsgaard, K., Ejrnæs, Svenning, J.-C., 2013. Mammal predator and prey species richness are strongly linked at macroscales. Ecology 94 (5), 1112–1122.
Sawada, Y., Suwa, R., Jindo, K., Endo, T., Oki, K., Sawada, H., Arai, E., Shimabukuro, Y.E., Celes, C.H.S., Campos, M.A.A., Higuchi, F.C., Lima, A.J.N., Higuchi, N., Kajimoto, T., Ishizuka, M., 2015. A new 500-m resolution map of canopy height for Amazon forest using spaceborne LiDAR and cloud-free MODIS imagery. Int. J. Appl. Earth Obs. Geoinf. 43, 92–101.
Seiffert, E.R., Perry, J.M., Simons, E.L., Boyer, D.M., 2009. Convergent evolution of anthropoid-like adaptations in Eocene adapiform primates. Nature 461 (7267), 1118–1121.
Sept, J.M., 1994. Bone distribution in a semi-arid riverine habitat in eastern Zaire: implications for the interpretation of faunal assemblages at early archaeological sites. J. Archaeol. Sci. 21 (2), 217–235.
Spradley, J.P., Williams, B.A., Kay, R.F., 2015. Environmental variables affecting primate richness in the Neotropics. Am. J. Phys. Anthropol. 156 (S60), 294.
Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., Wetson, J., 1999. Support vector regression with ANOVA decomposition kernels. In: Advances in Kernel Methods—Support Vector Learning, pp. 285–292.
Tauber, A., 1994. Estratigrafía y vertebrados fósiles de la Formacíon Santa Cruz (Mioceno inferior) en la costa atlántica entre las rías del Coyle y Río Gallegos, Provincia de Santa Cruz, República Argentina. In: Facultdad de Ciencias Exactas, Físicas y Naturales. Córdoba, Universidad Nacional de Córdoba, República Argentina, pp. 422.
Therneau, T., Atkinson, B., Ripley, B., 2015. Package 'rpart'. 3-1.
Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J., Moreno, J., 2012. Retrieval of vegetation biophysical parameters using Gaussian process techniques. IEEE Trans. Geosci. Remote Sens. 50 (5), 1832–1843.
Vizcaíno, S.F., Bargo, M.S., Kay, R.F., Fariña, R.A., Di Giacomo, M., Perry, J.M., Prevosti, F.J., Toledo, N., Cassini, G.H., Fernicola, J.C., 2010. A baseline paleoecological study for the Santa Cruz Formation (late-early Miocene) at the Atlantic coast of Patagonia, Argentina. Palaeogeogr. Palaeoclimatol. Palaeoecol. 292 (3), 507–519.
Vizcaíno, S.F., Kay, R.F., Bargo, M.S., 2012a. Background for a paleoecological study of the Santa Cruz Formation (late Early Miocene) on the Atlantic Coast of Patagonia. In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 1–22.
Vizcaíno, S.F., Fernicola, J.C., Bargo, M.S., 2012b. Paleobiology of Santacrucian glyptodonts and armadillos (Xenarthra, Cingulata). In: Vizcaíno, S.F., Kay, R.F., Bargo, M.S. (Eds.), Early Miocene Paleobiology in Patagonia: High-Latitude Paleocommunities of the Santa Cruz Formation. Cambridge University Press, Cambridge, UK, pp. 194–215.
Vizcaíno, S.F., Toledo, N., Bargo, M.S., 2017. Advantages and limitations in the use of extant xenarthrans (Mammalia) as morphological models for paleobiological reconstruction. J. Mamm. Evol. https://doi.org/10.1007/s10914-017-9400-2.
Walker, K.R., Bambach, R.K., 1971. The significance of fossil assemblages from fine-grained sediments: time-averaged communities. Geol. Soc. Am. Abstr. Programs 3, 783–784.
Welling, S.H., Refsgaard, H.H., Brockhoff, P.B., Clemmensen, L.H., 2016. Forest Floor Visualizations of Random Forests. arXiv (preprint arXiv:1605.09196).
Whitmore, F.C., Stewart, R.H., 1965. Miocene mammals and central American seaways. Science 148, 180–185.
Williams, C.K., 1998. Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In: Learning in Graphical Models. Springer, Netherlands, pp. 599–621.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.
Willmott, C.J., Matsurra, K., Robeson, S.M., 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. 43, 749–752.
Wilson, J.L., 1974. Characteristics of carbonate platform margins. Am. Assoc. Pet. Geol. Bull. 58, 810–824.