Explaining PaleoIndian settlement in the Intermountain West with comparative machine learning

Kenneth B. Vernon[1,2,3], Kate E. Magargal[3,4], Paul Allgaier[3,5], David Zeanah[6], D. Craig Young[5], Robert G. Elston[7], Simon Brewer[1,2], Brian F. Codding[1,3]

[1] *Scientific Computing and Imaging Institute, University of Utah*
[2] *School of Environment, Society, and Sustainability, University of Utah*
[3] *Department of Anthropology, University of Utah*
[4] *Honors College, University of Utah*
[5] *Far Western Anthropological Research Group*
[6] *Department of Anthropology, California State University, Sacramento*
[7] *Department of Anthropology, University of Nevada, Reno*

Last updated: 2026-04-06

Abstract
Paleoindian sites are exceedingly rare, which limits our ability to train predictive models capable of accurately identifying unknown site locations, or inferential models capable of accounting for the environmental factors that drove past settlement decisions. To overcome these limitations and advance our understanding of Paleoindian settlement, we propose a comparative machine learning approximation to mixed effects models, specifically a Random Forest classification with an interaction term to account for differences in the spatial patterning of PaleoIndian and Archaic sites and down-sampling of background locations to account for class imbalance. As a test case, we apply this framework to the distribution of PaleoIndian sites in Grass Valley, Nevada, a well-studied locale in the Central Great Basin. The input feature set includes travel time estimates to resource rich areas and a spatial basis function to account for residual spatial autocorrelation. Versions of the model are built with increasing complexity and tested using leave-one-out cross validation. Model results indicate that Paleoindian settlements focused on minimizing distance to pluvial lake habitats, especially at the intersection of waterways, while later Archaic settlements were more dispersed across diverse environments. In addition to having decent predictive power, results point to additional research that promises to open up previously untapped data and resources for modeling the initial settlement of North America.

Keywords: Comparative Analysis, Random Forest, Ideal Free Distribution, Great Basin, Western Stemmed Tradition, Clovis

Target Journal: Journal of Archaeological Science

Correspondence: k.vernon@sci.utah.edu (Blake Vernon)

1. Introduction

Research into PaleoIndian settlement in the Great Basin of Western North America has broad implications for the initial settlement of North America more generally (Beck & Jones, 1997; Bradley et al., 2022; Grayson, 2011; Smith et al., 2020). Unfortunately, the PaleoIndian record of the region is extremely patchy and incomplete, leaving scant evidence available to adjudicate between competing hypotheses of early settlement (Grayson, 2011; Jazwa et al., 2021). To address these statistical challenges, we recommend a comparative approach that could, in principle, leverage all of the archaeological sites within a study area to model the subset of them associated with PaleoIndian populations - or, for that matter, any archaeological population of interest. In effect, we propose to model *differences* in the distribution of archaeological sites associated with different archaeological populations using a small set of environmental covariates.

A mixed effects model is preferable for handling small sample sizes as it allows groups with very small samples - even singleton groups - to borrow strength from larger groups, a process known as partial-pooling (Gelman & Hill, 2007). However, this typically requires a large number of groups to estimate mean and variance in the parameters, which, unfortunately, is not available to us in this instance, and also requires strong assumptions about the underlying distributions, so we choose instead to implement an approximation to the mixed effects model using a well-known machine learning algorithm known as Random Forest (Breiman, 2001). We also derive a simple, intuitive justification for the method from the Ideal Free Distribution (IFD) model from ecology (Codding & Bird, 2015; Fretwell & Lucas, 1969; Weitzel & Codding, 2022; Winterhalder et al., 2010), with particular attention given to differences in subsistence strategy and their consequences for settlement. Together, these ideas build on the approach recommended by (Vernon et al., 2021).

As a case study, we model PaleoIndian and Archaic settlement reflected in the distribution of archaeological sites found in Grass Valley, Nevada, an area that has been the focus of considerable archaeological attention and empirical investigation (Brugger & Rhode, 2020; Clewlow et al., 1978; Elston et al., 2025; Wells et al., 2013). For comparative purposes, we choose to focus on sites associated with Archaic hunter-gatherers as they provide an important contrast for PaleoIndian sites in terms of their settlement patterning and subsistence. Following (Elston et al., 2014; Elston & Zeanah, 2002), we expect PaleoIndian sites to occur at greater densities in lower elevation areas that would have been productive marsh habitats during the Pleistocene-Holocene Transition (PHT), from roughly 11,000 to 9,000 years BP. Owing to climate-induced reductions to those marsh habitats (Brugger & Rhode, 2020; Duke & King, 2014), Archaic sites should occur at slightly higher elevations along the slopes of ranges, where habitats would have been more productive during the Middle Holocene (MH), from roughly 7,000 to 4,500 years BP.

2. Background

2.1. Subsistence constraints on settlement

The IFD model (Fretwell & Lucas, 1969) is typically applied to situations in which individuals compete for a finite share of a habitat's limited resources, a situation otherwise known as scramble competition (Nicholson, 1954; Parker, 2000). The habitat's suitability will, thus, exhibit negative

density dependence, declining as more individuals compete for its limited resources. If we assume that the choice of habitats is both *ideal* and *free*, meaning individuals have sufficient knowledge to accurately rank habitats according to their suitability (the ideal condition), and they can settle habitats without restriction or cost (the free condition), then individuals seeking to maximize their food intake (their habitat suitability) will distribute themselves across habitats so as to achieve an equilibrium in food intake per capita. That is, they will achieve an ideal free distribution. Empirical evaluations with ethnographic and historic evidence suggest human settlement decisions follow an ideal free distribution across diverse subsistence strategies (Codding & Jones, 2013; Disma et al., 2011; Einarsson, 2015; Moritz et al., 2015; Yaworsky & Codding, 2018).

While the IFD is a static model, behavioral ecologists will often evaluate changes to the equilibrium under different conditions, a mode of reasoning known in economics as comparative statics. More often than not, that involves consideration of climate induced changes to habitat suitability, with individuals expected to shuffle between habitats in response. Environmental changes also affect decisions related to subsistence, however, and those choices can have additional consequences for how individuals estimate the quality of habitats and, thus, how they choose to distribute themselves across a landscape (Codding et al., 2021; Magargal et al., 2017; Vernon et al., 2021). Suppose, for example, that we have two bundles of subsistence resources, A and B, and two habitats, H1 and H2, and suppose further that bundle A occurs with greater frequency in H1 and bundle B with greater frequency in H2. All else being equal, individuals who rely for their subsistence on bundle A should locate with greater frequency in H1, and individuals who rely on B should locate with greater frequency in H2. Similarly, a population that responds to environmental change by shifting its diet from bundle A to B, should also shift their settlement pattern from H1 to H2. In the language of IFD, the suitability of H1 will be greater for those who rely on bundle A, and the suitability of H2 will be greater for those who rely on bundle B.

The crucial thing to note here is that this logic also informs us as to how two populations reliant on different subsistence goods will locate in relation to each other. This, it must be emphasized, is not the same as asking why the spatial distribution of one population might affect the distribution of another. Rather, the suggestion is that their spatial covariance might in fact be a consequence of spatial covariance in their different subsistence bases. This will especially be the case when evaluating different populations in the same area at different times, with direct causal interaction being impossible. But in either case, if we have that prior knowledge about their diets, and we know where the one population is or was located, we can leverage that information to fill in the gaps in our knowledge regarding the spatial distribution of the other population.

2.2. Paleoenvironmental change and subsistence intensification

Grass Valley, Nevada, lies very near to the geographic center of the North American Great Basin, where it forms one part of a vast network of broad, north-south running valleys separated by steep mountain ranges. While the paleo-environmental response of this region to climate change exhibits a great deal of spatial variability, as one would expect given its rugged topography, the overarching story at the end of the Pleistocene is largely one of retreat, of retreating glaciers, retreating pluvial lakes, and retreating habitats (Grayson, 2011). The cool and wet conditions that characterized the Younger Dryas transitioned gradually into the warmer and drier conditions characteristic of the

middle and later Holocene, with productive habitats slowly marching out of the valley bottoms, up the slopes, and out onto the ridges where cooler and wetter conditions to this day still linger year round (Duke & King, 2014). One of the primary results of all this change is that habitats have become more patchy and unevenly distributed across the landscape.

The most glaring exception to this story of retreat, of course, is the steady growth of Indigenous populations across the same period, from small groups of highly mobile PaleoIndians during the PHT to larger Archaic groups of semi-sedentary hunter-gatherers during the MH. Archaeologists generally agree that both populations had broad diets, including everything from large ungulates to small mammals, birds, and fish, and a wide variety of plants and seeds. However, disagreement surrounds the extent to which low ranked plant and animal resources actually contributed to PaleoIndian diets. Elston et al. (2014) provide a thorough summary of the current state of the evidence relevant to this question, which we will not review here. For our purposes, it is sufficient to note that their summary points to a general process of intensification on low ranked plant and animal resources from the PHT to the MH, which in turn suggests differences in the distributions of PaleoIndian and Archaic sites across the landscape, with PaleoIndian sites expected to occur at greater frequencies closer to basin bottoms where they would have encountered highly productive marsh habitats.

3. Materials and Methods

3.1. Project area

The rough outline of Grass Valley is defined by the Watershed Boundary Dataset (WBD) developed by the US Geological Survey and the Natural Resources Conservation Service within the US Department of Agriculture (USGS & NRCS, 2013). Here, we limit the scope of the project area to the portion of the Grass Valley watershed north of Highway 50 near Austin, Nevada, at roughly the same latitude as Mt. Callaghan near the southern tip of the remnant playa. The project area then extends north all the way to its juncture with Crescent Valley and the Barrick Gold Mine (see Figure 1). If the valley were perfectly rectangular, the dimensions of the project area would be roughly 40 x 60 kilometers, but given that the valley has a certain tilt from the northeast to the southwest, the total project area is closer to 2000 square kilometers. There are also four primary entry points into the valley, one coming up from Austin to the south, and the other three at the northern periphery, including one from Antelope Valley to the west, one from Crescent Valley to the north, and one from Pine Valley to the east, which suggests that most east-west inter-valley travel would have been concentrated at the northern end of the project area.

3.2. Site distribution

To identify Archaic and PaleoIndian sites in the project area, we draw on site records and cultural resource reports hosted by the Nevada Cultural Resource Information System (NVCRIS) with permission from the State Historic Preservation Office. We then supplement the resulting dataset with additional PaleoIndian and Archaic sites identified during National Science Foundation funded fieldwork from 2016 to 2018. This fieldwork was guided by a predictive model based on soil and geomorphological data. The model did not directly predict PaleoIndian site occurrences, but

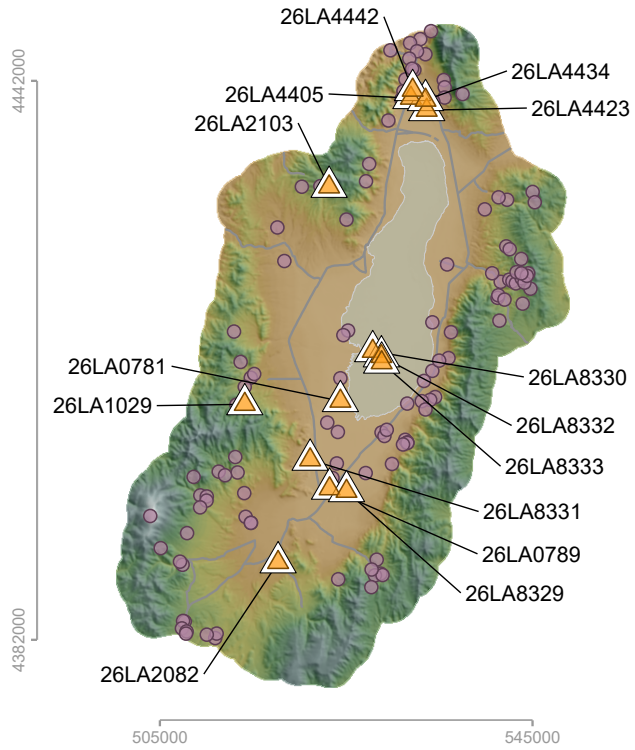

Figure 1: Overview map showing the sites of Archaic (pink circles) and PaleoIndian sites (orange triangles) in Grass Valley, NV. All points are randomly jittered to mask true locations.

rather identified landforms dating to the Terminal Pleistocene and Early Holocene that would have been near wetland habitats. Field research involved systematic survey of targeted locations, followed by testing of selected sites. In total, four new PaleoIndian sites were identified. Two PaleoIndian sites are excluded from this analysis, however, as they occur at the extreme south end of the watershed and thus fall outside the project area. The resulting dataset contains 14 PaleoIndian sites and 123 Archaic sites, as shown in Figure 1.

3.3. Environmental covariates

Limiting the scope of the project area comes with one critically important disadvantage. The more localized the analysis, the less variability will be observed in critical ecological variables, making it harder to tease out a meaningful signal of the presence of archaeological sites, even with really large samples. To address this issue, we rely on topographic variation, estimating cost-distance to critical water resources, namely perennial streams and the hypothesized shoreline of the Pleistocene lake (both measured in hours).

To calculate cost-distance, we use the R package terra (Hijmans, 2025), applying Campbell's hiking function (Campbell et al., 2019; 2022) to slope estimates derived from a digital elevation model (DEM), which we acquired from the United States National Elevation Dataset. Specifically, we apply Campbell's function with coefficients used to estimate hiking speed for the median individual in Campbell's 2022 sample. This allows us to generate estimates of travel time between

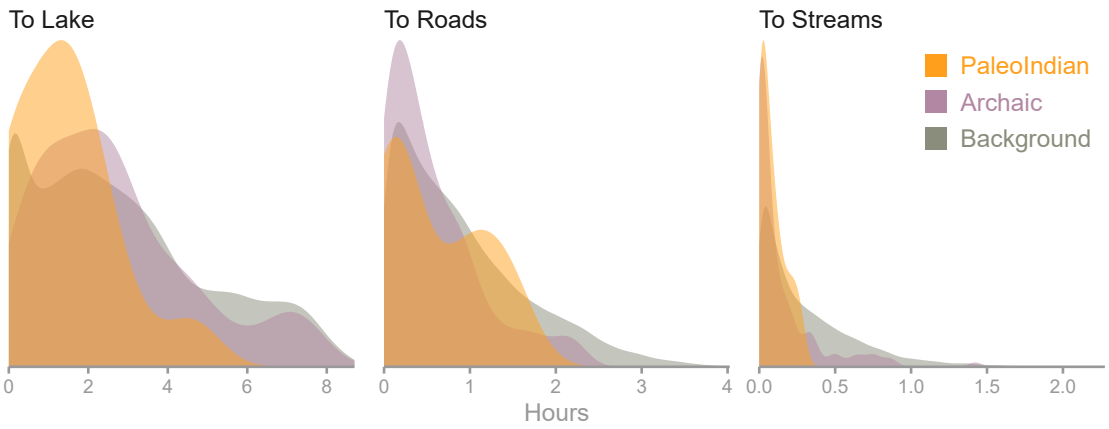

Figure 2: Distribution of covariates in parameter space for Archaic, PaleoIndian, and background sites.

grid cells in the DEM and then calculate the accumulated cost of travel from each perennial stream and the Pleistocene lake shoreline to any grid cell within the valley. To account for potential sampling bias, we also include cost-distance to roads, on the assumption that areas closer to roads are more intensely surveyed than areas farther away from roads.

To estimate the background distribution of these covariates, we use the R package sf (Pebesma, 2018) to generate a spatially uniform grid of 10,000 points, otherwise known as a quadrature scheme, and extract the values of covariates at those sample locations using terra. The distribution of these estimates in parameter space is shown in Figure 2 for PaleoIndian and Archaic presence locations and all background locations.

3.4. Statistical model

We define our target outcome $P(Y(s) = 1 \mid X(s))$ to be the probability that a site is present at a location $s \in \mathcal{D}$ given covariates X observed at s, with $\mathcal{D} \subset \mathbb{R}^2$ being the spatial extent of our project area, Grass Valley. This probability is assumed to be some unknown function $f(X(s))$ of those spatially-resolved covariates. To estimate f, we use the classification Random Forest algorithm implemented in the R package ranger (Wright & Ziegler, 2017), in this case fitting 1000 individual trees to bootstrap subsamples of the data, then averaging their probability estimates to obtain:

$$P(Y(s) = 1 \mid X(s)) = f(X(s)) + e(s)$$

with $e(s)$ being the spatial error.

Unfortunately, our ability to fit this model is hampered by the nature of our site occurrence data, which are presence-only, meaning we know where sites are located, but not necessarily where they are *not* located. To address this limitation, it is common in ecology to fill in for unobserved absence locations by sampling a very large number of randomly chosen locations and then reinterpreting the classification problem as an attempt to distinguish presence locations from those random background - as opposed to observed absence - locations (Elith et al., 2006; 2011; Elith & Leathwick,

2009). This strategy has a decent statistical and ecological justification, but in the context of Random Forest, an additional problem arises from the fact that the algorithm cannot differentiate true absences from random background locations - all it sees are zeroes. Because the sample of background locations is usually orders of magnitude larger than the number of presence locations (in this case, approximately 100 times larger since we use 10,000 background locations), the model's best strategy for achieving high predictive accuracy is to embrace unqualified pessimism: ignore the known presence locations and just assume that everything is a random background location. In machine learning, this is known as class imbalance.

To account for class imbalance, Valavi et al. (2021) recommend a two-part down-sampling strategy. The down-sampling part itself is straightforward. We simply instruct Random Forest to use all of the presence locations and an equal number of regularly sampled background locations in each individual tree, so that if there are N presence locations, the sample size for each tree should be $2N$ (or N presence locations and N background locations). The other part of the strategy is to substantially increase the default number of trees fit to the data (from 500 in ranger to 1000). This has the effect of repeatedly sampling with replacement from the background locations, so that our use of classification Random Forest, in effect, serves to approximate a probability density proportional to the count of presence locations in a localized area. This makes model predictions from the Random Forest analogous to what would be obtained by applying a 2D kernel density smooth to the presence locations, also to the raw outputs of the now quite popular Maximum Entropy approach to species distribution modeling in ecology (Fithian & Hastie, 2013; Merow et al., 2013; Renner et al., 2015). In fact, Valavi *et al* insist that the probability estimates from the classification Random Forest should be interpreted as likelihoods for this very reason.

As this is a comparative analysis, we include an interaction term in X for the time periods, PaleoIndian and Archaic. In the context of Random Forest, this is roughly equivalent to introducing fixed effects for the intercept and slope of each input feature. Introducing this interaction, however, comes with an important implementation cost, namely, that the background locations must be repeated for each level of the interaction term, so that our full model actually uses 20,000 background samples, 2 samples from each of the 10,000 unique background locations. This also requires that our down-sampling strategy for each tree include $4N$ observations, with $N = 14$ being the number of PaleoIndian sites, so that we are also down-sampling the larger contrast group, or sites associated with the Archaic, along with their background sample. As noted, the model is fit using the R package ranger (Wright & Ziegler, 2017), which provides for parallel computing and very fast computation of large datasets.

3.5. Spatial covariance

In addition to drawing statistical power from the distribution of Archaic sites, we can also encourage our model of PaleoIndian settlement to leverage spatial covariance between PaleoIndian sites themselves. That is, the model can take advantage of the fact that the presence of a PaleoIndian site at some location may, in fact, make it more likely that other PaleoIndian sites will be found in nearby locations. This is achieved in the current analysis using a method known as Fixed Rank Kriging (FRK) (Cressie & Johannesson, 2008; Sekulić et al., 2020), which allows us to decompose the error term in the above model into:

$$e(s) = \sum_{k=1}^{K} \phi_k(s)\eta_k + \xi(s)$$

with $\phi_k(s)$ being a spatial basis function evaluated over K spatial knots, η_k a spatial weight (also estimated with Random Forest), and $\xi(s)$ a random noise term. In this case, we set $K = 129$ for knots at two spatial scales (117 at the fine-grain scale and 12 at the coarse-grain scale) over the extent of $\mathcal{D}$ and implement $\phi_k(s)$ using the Matérn32 radial basis function. In previous experiments with spatial bases in Random Forest, we found that Random Forest was extremely sensitive to κ, the maximum range of support in discontinuous functions like bi-square, which return zeroes at distances greater than κ, thus introducing circular artifacts into geographic probability estimates. Unlike bi-square, Matérn32 is smooth and continuous with infinite support. To fit the Random Forest model with FRK, the values of $\phi_k(s)$ are appended to the covariate matrix $X(s)$ as additional features, hence the model estimating the weights η_k.

3.6. Model evaluation

For this modeling exercise, we only care about a model's ability to predict PaleoIndian locations. This leaves us with an extremely small sample, so we choose to evaluate the model using leave-one-out cross-validation (LOO-CV), which allows us to interrogate model behavior on a per site basis. The process involves fitting a model to 13 of the 14 PaleoIndian sites and then estimating the probability of presence of the left out PaleoIndian site. We then calculate the average difference between the probability assigned to the left out PaleoIndian site s^* and the probabilities assigned to all $i \in 1, ..., M$ background points using the formula:

$$\Delta P = P\big(Y\big(s^*\big) = 1\big) - \frac{1}{M}\sum_{i=1}^{M} P(Y(i) = 1)$$

The measure ΔP ranges over the interval $[-1, 1]$ and is roughly analogous to the area under the receiver operating characteristic (ROC) curve, a common summary statistic used for evaluating the predictive power of SDMs like our Random Forest. In the technical jargon, the ROC is the ratio of the true positive to the false positive rate at a given probability threshold, but, intuitively, it is just the idea that increased optimism comes at a price - lowering the probability threshold required to predict that a site is present will definitely net more true positives, but it will catch more false positives, too. A model can minimize the potential for false positives across all thresholds by maximizing the separation between the probability of presence at known presence and background locations, hence our statistic ΔP.

Our model is also quite complex, however, with a fairly large parameter set, especially with the addition of FRK. We want to know if all that additional complexity is actually worth it. Does adding an interaction term so that the model tries to account for differences between PaleoIndian and Archaic sites actually make the model better at estimating PaleoIndian sites? Does incorporating FRK help, too? To answer these questions, we fit four versions of the model: (1) a base model with just the three travel time covariates, (2) a comparative model that extends the base model to include the interaction term, (3) a spatial model that extends the base model to include FRK, and (4) a full model that is both comparative and spatial. We refer to these as Base, Comparative, Spatial,

and Full models, respectively. Each of these models is fit during each iteration of LOO-CV. We, thus, fit 56 (4 models x 14 PaleoIndian sites) versions of the Random Forest model.

All analyses are conducted in the R programming language and environment (R Core Team, 2025). For more details of our methods, please see the supplement.

4. Results

Results of LOO-CV are shown in Figure 3 for each of the four models (Base, Comparative, Spatial, and Full) and each of the 14 PaleoIndian sites. Panel A shows the distribution of ΔP values across models for each individual PaleoIndian site. Panel B shows the distribution of ΔP values across sites for each individual model. These results reveal that all models are generally capable of distinguishing PaleoIndian presence locations from regularly sampled background locations, though there is no statistically significant difference in the median ability of different models to predict the locations of PaleoIndian sites. On the one hand, some sites resist estimation by any configuration of the model. As shown in Panel A of Figure 3, no model can reliably differentiate site 26LA2082 from the background distribution. The models also struggle with 26LA8329 and 26LA1029. On the other hand, all models are proficient at estimating the locations of 26LA4442 and 26LA4405. With the remaining sites, it appears that different versions of the model have variable levels of success, with spatial versions doing better with some sites, and a-spatial versions better with others. To give just two examples, the a-spatial models did equally well at predicting 26LA0789, at least relative to the spatial models. Conversely, the spatial models did equally well at predicting 26LA8332 relative to the a-spatial models.

In general, Figure 4 suggests that the expectation that PaleoIndian sites will be found with greater probability in proximity to the Pleistocene lake shoreline is correct. However, the trend is not very strong and is complicated by interaction with cost-distance to streams, as it appears that the highest probability areas tend to be where streams intersect the modeled shoreline, as shown in Figure 5. One additional result should be emphasized, namely that the spatial process tends to swamp the functional response when included, hence the flat responses in Figure 4 and the smooth estimates in the geographic predictions in Figure 5.

5. Discussion

Our results confirm our expectation that Grass Valley's PaleoIndian population tended to concentrate nearer to the pluvial lake shoreline and that the later Archaic population spread farther afield, mostly on the slopes of adjacent ranges. We interpret this difference as an adaptive response to climate change which reduced the most suitable habitats for later Archaic populations, thus corroborating a widely held view in Great Basin archaeology regarding PaleoIndian settlement (Duke & King, 2014; Grayson, 2011) and adding further support to nearby analyses in Railroad Valley (Elston et al., 2014) and Long Valley (Huckleberry et al., 2001; Jones et al., 1996). It also adds to the finding in Long Valley that streams feeding into wetlands were an important draw for PaleoIndian populations.

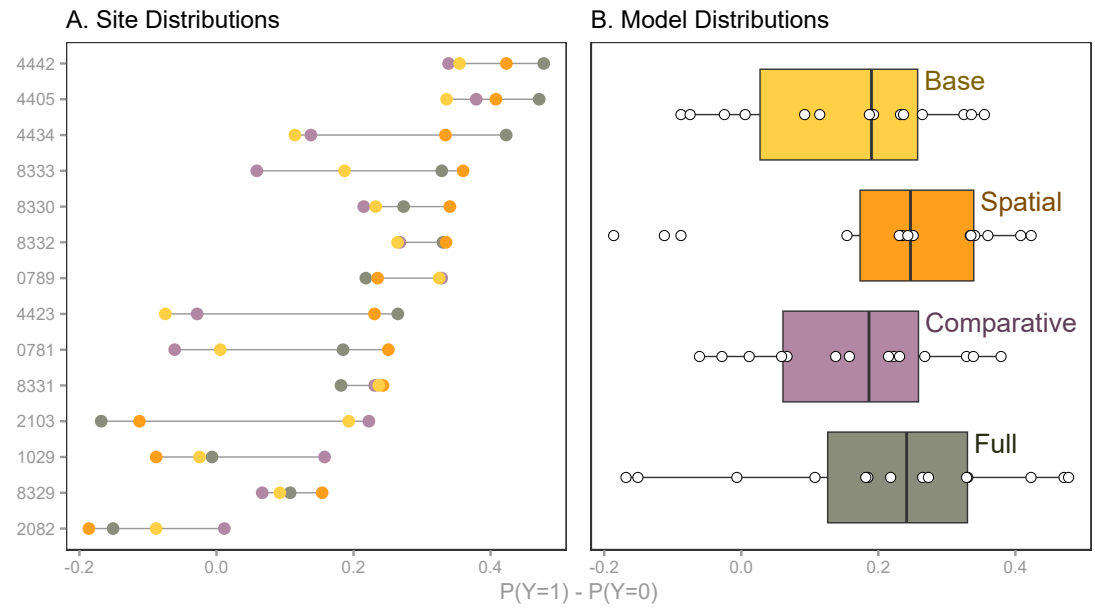

Figure 3: Results of leave-one-out cross validation for each of the 4 models (Base, Comparative, Spatial, and Full) and each of the 14 PaleoIndian sites. (A) Distribution of ΔP values across models for each PaleoIndian site. (B) Distribution of ΔP values across PaleoIndian sites for each model. In the y-axis labels, "26LA" (for Lander County, Nevada) is omitted from each Smithsonian Trinomial

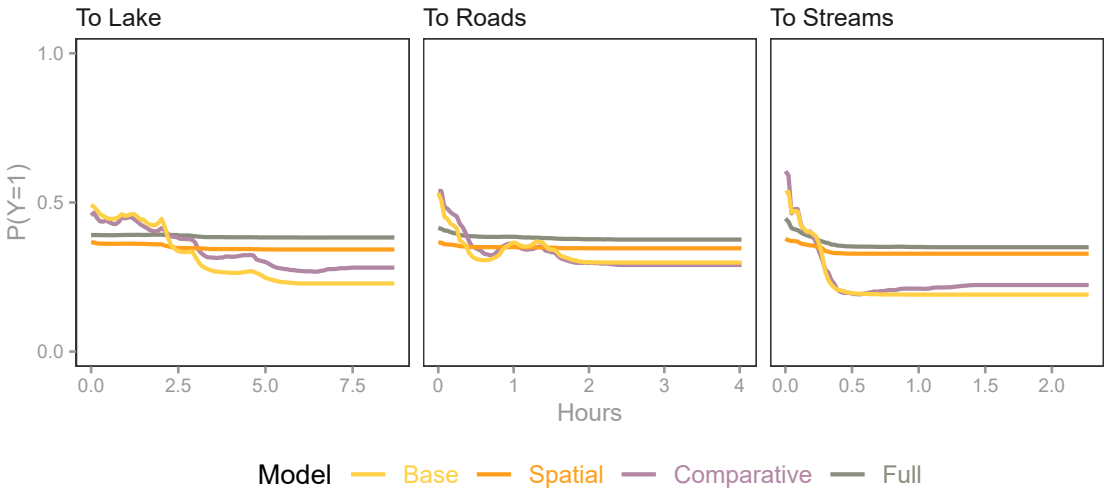

Figure 4: Partial dependence plots. Response is technically the relative likelihood of occurrence of a PaleoIndian site, not a strict probability.

In fact, we believe that the most productive marshes were probably at the inlets on the north and south sides of the Pleistocene lake, where most of the known sites tend to cluster. For instance, the Callahan Creek delta in the south of the valley was most likely the largest and best watered during the terminal Pleistocene, recieving water from both the Toiyabe range to the west and the Simpson Park range to the east. If we had a better way of estimating this, it would likely improve the

performance of our models. Future research should focus on increasing the resolution of marsh habitat reconstructions.

5.1. Model interpretation

While the models have only modest predictive power, variation in the ability of each version of the model to predict each left out PaleoIndian site is actually quite informative. The explanation for these differences is probably owing to the fact that different parts of the model are meant to solve slightly different problems. While the interaction term encourages the model to learn the locations of PaleoIndian sites in relation to nearby Archaic sites, the spatial basis encourages the model to learn PaleoIndian locations in relation to other PaleoIndian sites. And what we see is that the spatial models do better at predicting the location of a PaleoIndian site in areas of obvious spatial clustering, with multiple PaleoIndian sites in proximity to each other. That includes the cluster of four sites to the north, including 26LA4434, and the cluster near the southern end of the hypothesized Pleistocene lake shoreline, including 26LA8332.

We note next that the Comparative model, which is a-spatial, performed substantially better when predicting 3 PaleoIndian sites that are more or less spatially isolated from the main PaleoIndian sites at the center of the valley, including 26LA2082 to the far south, 26LA1029 in the far southwest, and 26LA2103 in the northwest. Adding to this, 26LA2082 looks utterly anomalous, which is probably why all of the models struggle to predict it. Conversely, the Comparative model does a decent job predicting 26LA1029 and 26LA2103, both of which are surrounded by Archaic sites but isolated from other PaleoIndian sites.

Given all of this, should it not be the case that the Full model outperforms the others in every case? To answer this, we need to elaborate on some limitations of the current implementation. It is common in geostatistics to fit two separate models, the first to evaluate the mean trend and the second to evaluate the spatial process over the first model's residuals (Cressie, 1993). To get spatially corrected estimates, predictions from the two models are then summed together. This is fine as long as (a) the outcome is unbounded and continuous like a normally distributed variable, a common assumption in geostatistics, and (b) all effects are additive. If the outcome is not unbounded and continuous, summing in this way risks violating the bounds constraint, for example, by pushing probability estimates outside the unit interval $[0, 1]$.

The obvious solution to this is to use a model that constrains the response to probabilities and then incorporate the spatial component directly, like we have done here. However, doing so increases model complexity by adding a large number of spatial features alongside the explanatory covariates. As a consequence, the model will struggle to disambiguate the effects of the explanatory covariates from the spatial features, potentially attributing variance in the response to the spatial component that would otherwise be absorbed by the covariates, especially in the context of small samples. This is certainly the case with Random Forest, which introduces complex interactions through its bootstrap subsampling and tree splitting procedures, leading the spatial component to swamp the mean effect of the explanatory covariates, as shown in Figure 4. What this means, in effect, is that the full model, while it is sensitive to interaction with Archaic sites, ends up focusing on high density PaleoIndian areas, and largely ignoring the outliers. This can be seen in the

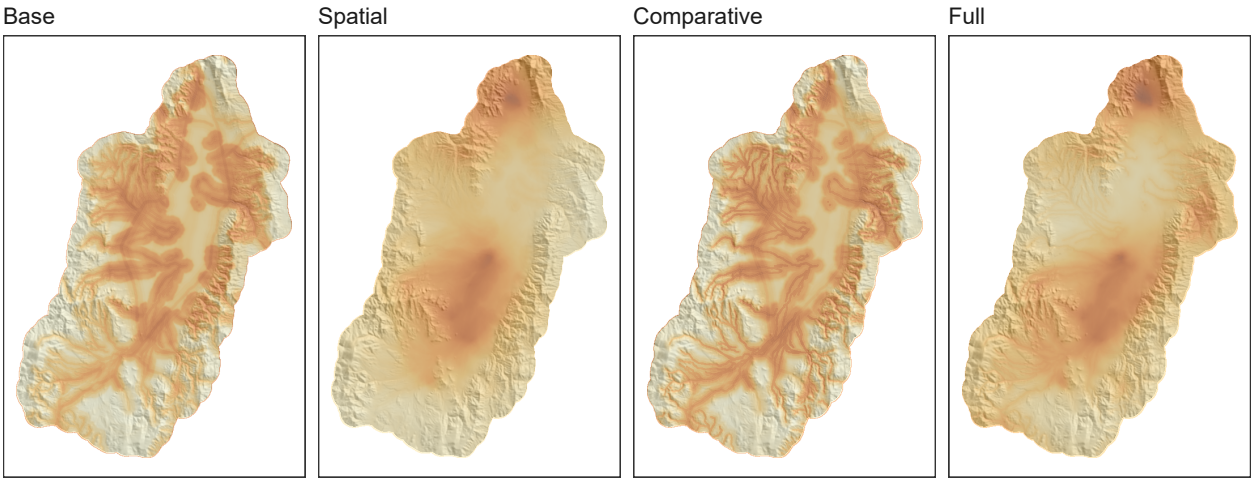

Figure 5: PaleoIndian response in geographic space.

prediction maps in Figure 5, particularly by noting the subtle differences between the Spatial and Full model predictions.

To handle this trade-off, one can either choose to avoid it entirely by simply adopting a different modeling technique, like Joint Species Distribution Models (Ovaskainen et al., 2016; Pollock et al., 2014; Wilkinson et al., 2021), or one can choose to put more thought into the larger modeling objective. If the goal is simply to predict the locations of PaleoIndian sites, then embrace the Full model. With all its limitations, it will still do a pretty good job of predicting PaleoIndian site locations, and with Random Forest, it will do so with minimal work. If the goal is to evaluate hypotheses about the functional response of a population to ecological variation, then the simpler Comparative model will be more appropriate.

5.2. Implications for PaleoIndian research

These findings and, in particular, the comparative, spatial methods used to arrive at them have important implications for the initial settlement of North America, especially for debates surrounding when and by what routes people colonized the interior (Beck & Jones, 1997; Grayson, 2011; Smith et al., 2020) and subsequent settlement patterning over time (Baka et al., 2025; Bradley et al., 2022). In the Great Basin, this debate centers on the chronological relationship between distinct lithic technologies, specifically Western Stemmed Tradition (WST) and Great Basin Fluted (GBF) points (Beck & Jones, 2010; Grayson, 2011; Jazwa et al., 2021).

GBF points are not technically "Clovis" points, at least not in their morphometric properties (Beck et al., 2019; Beck & Jones, 2007). Still, they are Clovis-like in their manufacture and likely function, so they are thought to represent the introduction of Clovis people or Clovis technology (or something descended from those) into the Great Basin, probably from the Great Plains or Southwest, where classic Clovis-type points are more often found (Beck & Jones, 2024; Beck et al., 2019; Jones & Beck, 2024). WST points are harder to pin down. While they are often found co-occurring with GBF points around Pleistocene marshes, their spatial distribution is much wider

than that of GBF points (Basgall, 2000; Eerkens et al., 2007). Recent evidence from sites like Paisley Caves (Jenkins et al., 2012) also suggests that WST points entered the Great Basin around the same time Clovis technology emerged on the Great Plains (Beck et al., 2019; Smith et al., 2020).

On the basis of this evidence, Beck & Jones (2010) have argued that individuals manufacturing and using WST points in the Great Basin likely descended from populations that moved inland from the Pacific coast. Later Clovis populations then encountered these people as they moved into the region, introducing them to fluted point technology. The adoption of GBF points led, in turn, to minor adjustments to the functional role WST played in the Great Basin tool kit, making them multi-purpose, which would possibly explain their wider distribution across the landscape (Beck & Jones, 2013). This is an intriguing hypothesis; however, as Jazwa et al. (2021) note, doubts remain concerning the reliability of WST dates and the paucity of GBF dates.

While our results may not inform on this question directly, instead providing further evidence for widely accepted ideas, our comparative method does offer a promising avenue for advancing this debate. The basic idea here is the same one that motivated this paper, namely, small samples. In this case, of course, we are dealing with small samples of dated WST and GBF points associated with one or more PaleoIndian populations. Following our strategy, we would simply expand on these samples by including all manner of dated projectile point types from across the American West, perhaps focusing on Archaic points like Northern Side-notched, Gatecliff, and Elko points (Thomas, 2013; 1981). We would, in effect, follow the spread of these types back through time to form expectations about the timing and location of older WST and GBF points based on their distributions across various environmental parameters like cost-distance to Pleistocene lakes within the Great Basin. In theory, this would allow us to reconstruct the directions of movement of WST and GBF points and to answer questions about the origins of those points, or at the very least, to determine from what direction they entered the Great Basin.

This could likely be achieved with existing data, too, particularly with the databases established by Thomas (2013) and cultural resource companies like Far Western. In fact, those datasets would seemingly allow us to dispense with categories altogether, a somewhat tedious though often times useful simplifying assumption, and instead look at morphological variation through time and space (Stevens, 2026). In so far as that variation reflects functional differences that entail differences in subsistence and mobility, we could use the method outlined here to go a step further and explain changes in those behaviors over time.

6. Conclusion

Understanding PaleoIndian settlement dynamics is challenging given the small and biased sample of sites available. Here we propose and demonstrate an approach that helps elucidate past settlement and land use decisions within a theoretically-informed machine learning framework. This advances both predictive modeling in archaeology (Kohler & Parker, 1986; Kvamme, 2005; Vaughn & Crawford, 2009; Vernon et al., 2022; Yaworsky et al., 2020) and our ability to evaluate hypotheses of PaleoIndian settlement and subsistence (Elston et al., 2014) during a period of environmental change with no ethnographic analog (Zeanah et al., 2026). Future work can expand

this application across greater spatial extents to identify variation in the drivers of PaleoIndian lifeways across North America.

Bibliography

Baka, A. S., Vernon, K. B., Mackie, M. E., Brewer, S., Spangler, J., Codding, B. F., Louderback, L. A., Flanigan, T. H., & Greenwald, A. M. (2025). Targeted explorations of Pleistocene-Holocene transition archaeology on the Colorado Plateau in southern Utah. *Quaternary International*, *746*, 109992. https://doi.org/10.1016/j.quaint.2025.109992

Basgall, M. E. (2000). *The Structure of Archaeological Landscapes in the North-Central Mojave Desert* (J. S. Schneider, R. M. Yohe, & J. K. Gardner, Eds.; pp. 123–138). Western Center for Archaeology, Paleontology.

Beck, C., & Jones, G. T. (2024). *Cultural transmission and the interaction of two cultural traditions* (K. N. McDonough, R. L. Rosencrance, & J. E. Pratt, Eds.; pp. 240–261). University of Utah Press.

Beck, C., & Jones, G. T. (1997). The Terminal Pleistocene/Early Holocene Archaeology of the Great Basin. *Journal of World Prehistory*, *11*(2), 161–236. http://www.jstor.org/stable/25801110

Beck, C., & Jones, G. T. (2007). *Early Paleoarchaic Point Morphology and Chronology* (K. E. Graf & D. N. Schmitt, Eds.; pp. 23–41). University of Utah Press.

Beck, C., & Jones, G. T. (2010). Clovis and Western Stemmed: Population Migration and the Meeting of Two Technologies in the Intermountain West. *American Antiquity*, *75*(1), 81–116. http://www.jstor.org/stable/20622483

Beck, C., & Jones, G. T. (2013). *Complexities of the Colonization Process: A View from the North American West* (K. E. Graf, C. V. Ketron, & M. R. Waters, Eds.). Center for the Study of the First Americans, Texas A&M University.

Beck, C., Jones, G. T., & Taylor, A. K. (2019). What's Not Clovis? An Examination of Fluted Points in the Far West. *Paleoamerica*, *5*(2), 109–120. https://doi.org/10.1080/20555563.2019.1613145

Bradley, E. J., Smith, G. M., & Nussear, K. E. (2022). Ecological niche modeling and diachronic change in Paleoindian land use in the northwestern Great Basin, USA. *Journal of Archaeological Science: Reports*, *45*, 103564. https://doi.org/10.1016/j.jasrep.2022.103564

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

Brugger, S. O., & Rhode, D. (2020). Impact of Pleistocene–Holocene climate shifts on vegetation and fire dynamics and its implications for Prearchaic humans in the central Great Basin, USA. *Journal of Quaternary Science*, *35*(8), 987–993. https://doi.org/10.1002/jqs.3248

Campbell, M. J., Dennison, P. E., & Thompson, M. P. (2022). Predicting the variability in pedestrian travel rates and times using crowdsourced GPS data. *Computers, Environment and Urban Systems*, *97*, 101866. https://doi.org/10.1016/j.compenvurbsys.2022.101866

Campbell, M. J., Dennison, P. E., Butler, B. W., & Page, W. G. (2019). Using crowdsourced fitness tracker data to model the relationship between slope and travel rates. *Applied Geography*, *106*, 93–107. https://doi.org/https://doi.org/10.1016/j.apgeog.2019.03.008

Clewlow, C. W., Wells, H. F., & Ambro, R. D. (1978). *History and prehistory at Grass Valley, Nevada* (Issue 7). University of California Institute of Archaeology Monograph.

Codding, B. F., & Bird, D. W. (2015). Behavioral ecology and the future of archaeological science. *Journal of Archaeological Science*, *56*, 9–20.

Codding, B. F., & Jones, T. L. (2013). Environmental productivity predicts migration, demographic, and linguistic patterns in prehistoric California. *Proceedings of the National Academy of Sciences*, *110*(36), 14569–14573. https://doi.org/10.1073/pnas.1302008110

Codding, B. F., Brenner Coltrain, J., Louderback, L., Vernon, K. B., Magargal, K. E., Yaworsky, P. M., Robinson, E., Brewer, S. C., & Spangler, J. D. (2021). Socioecological Dynamics Structuring the Spread of Farming in the North American Basin-Plateau Region. *Environmental Archaeology*, 1–13. https://doi.org/10.1080/14614103.2021.1927480

Cressie, N. A. (1993). *Statistics for spatial data* (Revised). John Wiley & Sons, Inc.

Cressie, N. A., & Johannesson, G. (2008). Fixed Rank Kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 209–226. https://doi.org/10.1111/j.1467-9868.2007.00633.x

Disma, G., Sokolowski, M. B., & Tonneau, F. (2011). Children's competition in a natural setting: Evidence for the ideal free distribution. *Evolution and Human Behavior*, *32*(6), 373–379. https://doi.org/10.1016/j.evolhumbehav.2010.11.007

Duke, D., & King, J. (2014). A GIS model for predicting wetland habitat in the Great Basin at the Pleistocene–Holocene transition and implications for Paleoindian archaeology. *Journal of Archaeological Science*, *49*, 276–291. https://doi.org/10.1016/j.jas.2014.05.012

Eerkens, J. W., Rosenthal, J. S., Young, D. C., & King, J. (2007). Early Holocene Landscape Archaeology in the Coso Basin, Northwestern Mojave Desert, California. *North American Archaeologist*, *28*(2), 87–112. https://doi.org/10.2190/NA.28.2.a

Einarsson, Á. (2015). Viking age fences and early settlement dynamics in Iceland. *Journal of the North Atlantic*, *2015*(27), 1–21. https://doi.org/10.3721/037.006.2703

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, And Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Elston, R., Zeanah, D., & Codding, B. (2014). Living outside the box: An updated perspective on diet breadth and sexual division of labor in the Prearchaic Great Basin. *Quaternary Investigations from Antarctica across South America to the North Atlantic*, *352*, 200–211. https://doi.org/10.1016/j.quaint.2014.09.064

Elston, R. G., & Zeanah, D. W. (2002). Thinking outside the box: A new perspective on diet breadth and sexual division of labor in the Prearchaic Great Basin. *World Archaeology*, *34*(1), 103–130. https://doi.org/10.1080/00438240220134287

Elston, R. G., Codding, B. F., Craig Young, D., & Zeanah, D. W. (2025). A new open-air PaleoIndian site in the Central Great Basin of North America. *Paleoamerica*, *11*(2), 179–182. https://doi.org/10.1080/20555563.2025.2553469

Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, *7*(4), 1917–1939. https://doi.org/10.1214/13-AOAS667

Fretwell, S. D., & Lucas, H. L. (1969). On territorial behavior and other factors influencing habitat distribution in birds I. Theoretical development. *Acta Biotheoretica*, *19*, 16–36.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Grayson, D. K. (2011). *The Great Basin: A Natural Prehistory*. University of California Press.

Hijmans, R. J. (2025). *terra: Spatial data analysis*. https://rspatial.org/

Huckleberry, G., Beck, C., Jones, G. T., Holmes, A., Cannon, M., Livingston, S., & Broughton, J. M. (2001). Terminal Pleistocene/Early Holocene Environmental Change at the Sunshine Locality, North-Central Nevada, U.S.A. *Quaternary Research*, *55*(3), 303–312. https://doi.org/10.1006/qres.2001.2217

Jazwa, C. S., Smith, G. M., Rosencrance, R. L., Duke, D. G., & Stueber, D. (2021). Reassessing the Radiocarbon Date from the Buhl Burial from South-Central Idaho and Its Relevance to the Western Stemmed Tradition–Clovis Debate in the Intermountain West. *American Antiquity*, *86*(1), 173–182. https://doi.org/10.1017/aaq.2020.36

Jenkins, D. L., Davis, L. G., Stafford, T. W., Campos, P. F., Hockett, B., Jones, G. T., Cummings, L. S., Yost, C., Connolly, T. J., Yohe, R. M., Gibbons, S. C., Raghavan, M., Rasmussen, M., Paijmans, J. L. A., Hofreiter, M., Kemp, B. M., Barta, J. L., Monroe, C., Gilbert, M. T. P., & Willerslev, E. (2012). Clovis Age Western Stemmed Projectile Points and Human Coprolites at the Paisley Caves. *Science*, *337*(6091), 223–228. https://doi.org/10.1126/science.1218443

Jones, G. T., & Beck, C. (2024). *The Clovis record of the Far West* (K. N. McDonough, R. L. Rosencrance, & J. E. Pratt, Eds.; pp. 163–183). University of Utah Press.

Jones, G. T., Beck, C., Nials, F. L., Neudorfer, J. J., Brownholtz, B. J., & Gilbert, H. B. (1996). Recent Archaeological and Geological Investigations at the Sunshine Locality, Long Valley, Nevada. *Journal of California and Great Basin Anthropology*, *18*(1), 48–63. http://www.jstor.org/stable/27825597

Kohler, T. A., & Parker, S. C. (1986). *Predictive models for archaeological resource location* (M. B. Schiffer, Ed.; pp. 397–452). Academic Press. https://doi.org/10.1016/B978-0-12-003109-2.50011-8

Kvamme, K. L. (2005). *There and back again: Revisiting archaeological locational modeling* (M. W. Mehrer & K. L. Wescott, Eds.; pp. 23–55). CRC Press.

Magargal, K. E., Parker, A. K., Vernon, K. B., Rath, W., & Codding, B. F. (2017). The ecology of population dispersal: Modeling alternative basin-plateau foraging strategies to explain the Numic expansion. *American Journal of Human Biology*, *29*(4), e23000–n/a. https://doi.org/10.1002/ajhb.23000

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Moritz, M., Hamilton, I. M., Yoak, A. J., Scholte, P., Cronley, J., Maddock, P., & Pi, H. (2015). Simple movement rules result in ideal free distribution of mobile pastoralists. *Ecological Modelling*, *305*, 54–63. https://doi.org/10.1016/j.ecolmodel.2015.03.010

Nicholson, A. (1954). An outline of the dynamics of animal populations. *Australian Journal of Zoology*, *2*(1), 9–65. https://doi.org/10.1071/ZO9540009

Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit Joint Species Distribution Models. *Methods in Ecology and Evolution*, *7*(4), 428–436. https://doi.org/10.1111/2041-210X.12502

Parker, G. A. (2000). Scramble in behaviour and ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *355*(1403), 1637–1645. https://doi.org/10.1098/rstb.2000.0726

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*(5), 397–406. https://doi.org/10.1111/2041-210X.12180

R Core Team. (2025). *R: A language and environment for statistical computing*. https://www.r-project.org/

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379. https://doi.org/10.1111/2041-210X.12352

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., & Bajat, B. (2020). Random Forest spatial interpolation. *Remote Sensing*, *12*(10), 1687. https://doi.org/10.3390/rs12101687

Smith, G. M., Duke, D., Jenkins, D. L., Goebel, T., Davis, L. G., O'Grady, P., Stueber, D., Pratt, J. E., & Smith, H. L. (2020). The Western Stemmed Tradition: Problems and Prospects in Paleoindian Archaeology in the Intermountain West. *Paleoamerica*, *6*(1), 23–42. https://doi.org/10.1080/20555563.2019.1653153

Stevens, N. E. (2026). Investigating technology change without typology: The spread of the bow and arrow in California. *American Antiquity*, *91*(1), 117–134. https://doi.org/10.1017/aaq.2025.10118

Thomas, D. H. (2013). Great Basin Projectile Point Chronology: Still Relevant?. *Journal of California and Great Basin Anthropology*, *33*, 133–152.

Thomas, D. H. (1981). How to Classify the Projectile Points from Monitor Valley, Nevada. *Journal of California and Great Basin Anthropology*, *3*(1), 7–43. http://www.jstor.org/stable/27825055

USGS, & NRCS. (2013). *Federal Standards and Procedures for the National Watershed Boundary Dataset (WBD) (4 ed.): Techniques and Methods* (p. 63). https://pubs.usgs.gov/tm/11/a3/

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742. https://doi.org/10.1111/ecog.05615

Vaughn, S., & Crawford, T. (2009). A predictive model of archaeological potential: An example from northwestern Belize. *Applied Geography*, *29*(4), 542–555. https://doi.org/10.1016/j.apgeog.2009.01.001

Vernon, K. B., Spangler, J. D., McCool, W. C., Yaworsky, P. M., Brewer, S. C., & Codding, B. F. (2022). *The Fremont frontier: Living at the margins of maize farming*. https://kbvernon.github.io/saa_2022-western_fremont/#1

Vernon, K. B., Yaworsky, P. M., Spangler, J. D., Brewer, S., & Codding, B. F. (2021). Decomposing habitat suitability across the forager to farmer transition. *Environmental Archaeology*. https://doi.org/10.1080/14614103.2020.1746880

Weitzel, E. M., & Codding, B. F. (2022). The ideal distribution model and archaeological settlement patterning. *Environmental Archaeology*, *27*(4), 349–356. https://doi.org/10.1080/14614103.2020.1803015

Wells, H. F., Seelinger, E., & Ambro, R. D. (2013). The Grass Valley Archaeological Project: Looking back and looking forward. *Journal of California and Great Basin Anthropology*, *33*(2), 153–165. http://www.jstor.org.ezproxy.lib.utah.edu/stable/24644391

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2021). Defining and evaluating predictions of Joint Species Distribution Models. *Methods in Ecology and Evolution*, *12*(3), 394–404. https://doi.org/10.1111/2041-210X.13518

Winterhalder, B., Kennett, D. J., Grote, M. N., & Bartruff, J. (2010). Ideal free settlement of California's Northern Channel Islands. *Journal of Anthropological Archaeology*, *29*, 469–490.

Wright, M. N., & Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yaworsky, P. M., & Codding, B. F. (2018). The ideal distribution of farmers: Explaining the Euro-American settlement of Utah. *American Antiquity*, *83*(1), 75–90. https://doi.org/10.1017/aaq.2017.46

Yaworsky, P. M., Vernon, K. B., Spangler, J. D., Brewer, S. C., & Codding, B. F. (2020). Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument. *PLOS ONE*, *15*(10), e239424. https://doi.org/10.1371/journal.pone.0239424

Zeanah, D. W., Elston, R. G., & Codding, B. F. (2026). Gender convergent labor and technological change at the Pleistocene to Holocene Transition in the Great Basin. *Journal of Anthropological Archaeology*, *81*, 101746. https://doi.org/10.1016/j.jaa.2025.101746