Biogeography of crop progenitors and wild plant resources during the transition to agriculture in West Asia, 21–8.3 ka

Joe Roe[a,b,*], Amaia Arranz-Otaegui[c]

[a] *University of Bern, Switzerland,*

[b] *University of Copenhagen, Denmark,*

[c] *University of the Basque Country, Spain,*

Abstract

This paper presents the first comprehensive reconstructions of the palaeodistributions of [X] plant species known to have been gathered or cultivated by early agricultural societies in Southwest Asia, including the progenitors of the first crops. We used machine learning to train an ecological niche model (ENM) of each species based on its present-day distribution in relation to climate and environmental variables. Predictions of the potential ranges of these species at key stages of the Pleistocene–Holocene transition could then be derived from these models using hindcast data from palaeoclimate simulations. Species ranges were on average [X%] [larger|smaller] in the Early Holocene compared to today, indicating [...]. The modelled ranges predict the observed occurrence of these species on archaeological sites with [low|medium|high] accuracy. The regional ubiquity of species in the archaeological record is [not] correlated with the predicted size of its range and the diversity of archaeobotanical assemblages is [not] correlated with the predicted diversity of its environs. This indicates that trends in taxonomic composition of the archaeobotanical record is [not] likely to be influenced

*Corresponding author
Email address: joeroe@hey.com (Joe Roe)

by environmental change and species turnover, not just, as is often assumed, human economic choices.

- ☐ Finish analysis:
  - ☐ Add terrain, soil
  - ☐ Find a resolution that's workable
  - ☐ Do thresholding?
- ☐ First draft:
  - ☐ Introduction
  - ☐ Background
    - ☐ Biogeography
    - ☐ ENM
  - ☒ Methods & Materials
    - ☒ Occurrence data
    - ☒ Predictor data
    - ☒ Random Forest model
  - ☐ Results
    - ☐ Model assessment
    - ☐ Hindcasting
  - ☐ Discussion
    - ☐ General trends
    - ☐ Case studies
  - ☐ Conclusion
- ☐ Figures
  - ☐ tbl-predictors
  - ☐ tbl-climate-periods
- ☐ Appendix with all hindcast predictions
- ☐ References & copyediting
- ☐ Final proofread

1. Introduction

- The first farming societies had an ecological context
- Subsistence is understood (largely) through archaeobotany and zooarchaeology; ecological context from environmental archaeology, palynology, palaeoclimate records, etc.
  - But these have a variety of biases (human selection, taphonomy, etc.)
  - And at the end of the day only represent specific places – interpolation to the entire region is not straightforward
  - Overlap makes it difficult to see where human choices depart from the environmental background (cf. Martin et al., 2016)
  - Or to fully contextualise subsistence strategies and shifts in strategies in response to environmental shifts (e.g. Yaworsky et al., 2023).
- Here we present an alternative approach using ENM
  - Whole region, at multiple climate snapshots
  - Independent of archaeobot. and pal. clim. data, so can verify and compare

2. Background

- The transition to agriculture in West Asia was…

2.1. Biogeography and agricultural origins

- Has always been important in study of agricultural origins
  - Historically: Vavilov, Pumpelly & Childe
  - Genetic studies tell us origin points, but not ranges
- Important to e.g.
  - Distinguish environmental from potentially anthropogenic change (**MartinEtAl2017?**; **MartinEtAl2025?**)
  - Reconstruct sequences of domestication (**YeomansEtAl2017?**)
- Epipal./Neo. plant-based economies were diverse
  - More than the "founder crops";

  - More than food
  - (In archaebot., not all intentionally collected)
  - Regionally and temporally diverse
  - …so we model lots of species!
- Regional ecological reconstructions generally rely on the 'expert interpolation' (or what do they call it with isoscapes?) method
  - See CSEAS (AEA-prep) presentation
  - Figure: comparisons

2.2. Ecological niche modelling in archaeology

Ecological niche modelling (ENM) or species distribution modelling (SDM) is widely used by ecologists to predict the geographic range of a species based on a set of environmental predictors. Essentially, it involves combining records of where an organism has been observed with environmental data (climate, topography, etc.) for those locations to model the range of environmental values at which that species – its environmental niche. This model can then be used to predict the range of the organism in question either in the same or a different environment. (**CITE?**) suggests reserving the term 'species distribution modelling' for when the method is used to recover the verifiable range of a species in a real and existing environment, and using 'ecological niche modelling' as the broader term covering hypothetical or predictive applications – a convention we follow here when referring to predictive or 'hindcast' models of past ranges. Within this overarching framework, ecological niche modelling encompasses a wide range of applications and a variety of potential environmental predictors, modelling approaches, and methodologies, which we will not attempt to review here.

Ecological niche modelling has long been of interest to archaeologists as both a means of exploring the biological niche of humans and for reconstructing the past environments they inhabited (Franklin et al., 2015; **DavidPollyEronen2011?**). In the first sense, it has been used most extensively to model

the range of humans and other hominin species (e.g. **BenitoEtAl2017?**; **YousefiEtAl2020?**; **BanksEtAl2021?**; **YaworskyEtAl2024a?**; **YaworskyEtAl2024b?**; **GuranEtAl2024?**), especially in the Palaeolithic. This overlaps with what archaeologists usually call generically 'predictive modelling' (**VerhagenWhitley2020?**)—more precisely 'site distribution modelling'—which is essentially the same approach as (and often borrows methodologies from) ecological niche modelling but applied to the occurrence of archaeological sites. Here what is modelled is not strictly a biological niche alone, but also aspects of human geography, taphonomy, and archaeological visibility. These applications can be distinguihed from 'palaeoecological niche modelling', where the object of model remains, as in ecology, a non-human biological niche.

(Franklin et al., 2015) review palaeoecological niche modelling and advocate for its greater adoption in environmental archaeology. One relevant early example is (Conolly et al., 2012) used the occurrence of wild and domestic *Bos* remains at prehistoric archaeological sites in Europe and West Asia to map the evolving niche of cattle over the Pleistocene–Holocene transition. It has been used to model the availability of fauna exploited by humans at wider scales (e.g. **deAndresHerreroEtAl2018?**; **YaworksyEtAl2023?**) and, in a West Asian context, of foraged plant resources in the landscape around the Neolithic site of XX (Collins et al., 2018). Modelling the spread of crops has been another significant archaeological application (**CremaEtAl?**).

In the majority of studies to date (palaeo)ecological niche modelling has been applied to archaeological data in an 'inductive' fashion, i.e. faunal and botanical remains from ancient sites are used as the occurrence dataset for training a model using either past or present environmental data. However, both the zooarchaeological and archaeobotanical records are sparse and subject to a complex array of depositional, taphonomic and recovery biases factors that , many of which are not fully understood and/or cannot be corrected for. This means that while the archaeological attestation of the presence of a species might generally be relied upon, it is highly unlikely that its absence is representative of

true past distributions.

The alternative approach is to train the model using contemporary occurrence and environmental data and then use palaeoenvironmental data to 'hindcast' its predictions backwards in time. Like (Franklin et al., 2015), we view the hindcasting approach as more promising, because training datasets for both occurrences and environment are far more readily available, complete and reliable for the present than the past. There is some scepticism in the ecological niche modelling literature about the ability of such models to make accurate predictions in unknown environments (like the past) (**CITES?**), but here the hindcasting approach also presents an opportunity: it reserves archaeological occurrence data as an independent dataset that can be used to assess the retrodictive performance of the model. This possibily was suggested by (Franklin et al., 2015) but to our knowledge our study represents the first attempt to actually do so.

The major practical limitation of the hindcasting approach is that it relies on spatially explicit, high resolution palaeoenvironmental surfaces with continuous coverage of the region and periods of interest. Until recently, this has not been widely available for most applications, which is perhaps why only a minority of studies use it (cf. Yaworsky et al., 2023). In this study, we are able to take advantage of the increasing availability of high resolution, global palaeoclimate data derived from simulation experiments with general circulation models of climate (Brown et al., 2018; **BrownEtAl2020?**; **KargerEtAl2023?**).

3. Methods and materials

3.1. Occurrence data

We consider X distinct taxa (Table **??**) - all the identifiable species known to be present at more than three Neolithic sites in West Asia, according to our previous study (Arranz-Otaegui and Roe, 2023).

Taxonomic names were resolves to the canonical form specified in the GBIF Backbone Taxonomy (**GBIFSecretariat2023?**). So for example occurrences for "Bolboschoenus" included all species and subspecies, including specimens described as *Bolboschoenus sp.*, *Bolboschoenus maritimus* and *Scirpus maritimus*. Domestic species meeting our inclusion criteria were substituted with their wild progenitor(s), where known.

3.2. Occurrence data

Occurrence data was obtained from the Global Biodiversity Information Facility (GBIF) using via its application programming interface and the R package rgbif (Chamberlain et al., 2024; Chamberlain and Boettiger, 2017). Although ENMs have reasonable predictive power even with small training samples (Hernandez et al., 2006; Stockwell and Peterson, 2002; Wisz et al., 2008), we excluded X taxa with less than fifty recorded occurrences in our study region. We also excluded one taxon (*Avena sterilis*) with over 47,000 occurrences, as this would have been computationally prohibitive and we were uncertain what account for such a disproportionately high number of records.

Occurrence data only tells us where a species is present; there is rarely definitive information on where the species is *not* found. We therefore need to generate random background points or "pseudo-absences" to feed to the model. There are several ways to do this. We follow the advice of (Barbet-Massin et al., 2012) for regression-based species distribution models and use a large (:10000) random sample of points, weighted equally against the presences in the regression. (Valavi et al., 2022) also recommend using a very large background sample for random forest models.

3.3. Predictor data

We modelled the occurrence of species as a function of X spatial predictor variables (**?@tbl-predictors**). These included:

- Sixteen 'bioclimatic' variables derived from monthly temperature and precipitation values, following standard practice for species distribution models (Hijmans et al., 2005). Contemporary bioclimatic predictor data for West Asia was extracted from the global CHELSA dataset (Karger et al., 2017), which predicts temperature and precipitation from downscaled general circulation model output at 1 km resolution.

- Terrain aspect and slope, which at high resolution perform well as proxies for solar radiation when modelling plant occurrence (Austin and Van Niel, 2011; Leempoel et al., 2015); and the topographic wetness index (TWI), which serves as a proxy for soil moisture and is particularly important in modelling arid environments (Campos et al., 2016; Di Virgilio et al., 2018; Kopecký and Čížková, 2010). All three were derived from the SRTM15+ digital elevation model using algorithms from WhiteboxTools (Lindsay, 2016).

- Edaphic data from SoilGrids (Hengl et al., 2017, 2014), which improves model performance for plants (Dubuis et al., 2013; Mod et al., 2016; Velazco et al., 2017). Based on a recent assessment of the reliability of SoilGrids data for species distribution modelling (Miller et al., 2024), we used a subset of four variables relating to soil texture (clay, silt, sand) and pH at the surface (0-5 cm depth).

Predictor data was transformed to the same projection system (WGS84 / UTM 37 N) and a common resolution of X km.

For hindcasting, we used reconstructed bioclimatic data for key periods (**?@tbl-climate-periods**) generated from downscaled paleoclimate simulations from the HadCM3 general circulation model (Brown et al., 2018). Terrain and soil predictors were held constant, since reconstructions of these variables in the past are not available at sufficient scale. It is not likely that either macroscale topography or soil characteristics have altered significantly over the period of time considered here, so we assume that this does not degrade model

performance, and may in fact benefit it by providing 'anchoring' predictors that are independent of climate change.

3.4. Random Forest

Ecological niche modelling is a classification problem that can be approached with a wide range of statistical methods. A substantial literature exists on the relatively performance of these approaches and their respective parameterisations (reviewed in Valavi et al., 2022). Random Forest, a widely-used machine learning algorithm, is amongst the best performing approaches for presence-only species distribution models, providing it is appropriately parameterised to account for the class imbalance between presence and background samples (Valavi et al., 2022, 2021). For our application, it also has the advantage of requiring little to no manual parameter tuning to achieve good predictive results, which makes it easier to model a larger numbers of taxa.

For each taxon we trained a classification model to predict occurrence (presence or absence/background) based on our X predictor variables (**?@tbl-predictors**). We used the Random Forest algorithm implemented in the R package 'ranger' (Wright and Ziegler, 2017) and the 'tidymodels' (**tidymodels?**) framework for data preprocessing and model selection. To avoid overfitting, we follow (Valavi et al., 2021) in their recommended hyperparameters and use of down-sampling to balance presence and background samples. Models for each taxon were fit independently, with redundant zero-variance predictors excluded, and assessed based on balanced training (3/4) and test (1/4) partitions.

The output of the model is probabilistic. However, this should not be understood as an actual probability of occurrence (**CITE?**), but more akin to an estimate of habitat suitability. To simplify interpretation, we can convert these predictions into binary presence/absence maps, a process called "thresholding". We select the threshold value for each model individually, using MaxSSS (as recommended by Liu et al., 2013). This also makes it possible to analyses the

predictions together as an assemblage.

4. Results

4.1. Model assessment

- Modelled ecological niches on current data

4.2. Hindcasting

Sensitivity to climate fluctuations?

- Comparison to archaeological occurrences

5. Discussion

- General trends:
  - Quantified sensitivity of plant ranges to climate change
  - Crop progenitors saw range contractions just before the onset of agriculture? (Moreso than other wild resources??)
  - How could are reconstructed ranges at predicting archaeological assemblages? (+Implications)
- Interesting individual case studies:
  - Wheat progenitors
  - ???

6. Conclusion

References

Arranz-Otaegui, A., Roe, J., 2023. Revisiting the concept of the "Neolithic Founder Crops" in southwest Asia. Vegetation History and Archaeobotany. https://doi.org/10.1007/s00334-023-00917-1

Austin, M.P., Van Niel, K.P., 2011. Improving species distribution models for climate change studies: Variable selection and scale. Journal of Biogeography 38, 1–8. https://doi.org/10.1111/j.1365-2699.2010.02416.x

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: How, where and how many?: How to use pseudo-absences in niche modelling? Methods Ecol. Evol. 3, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Brown, J.L., Hill, D.J., Dolan, A.M., Carnaval, A.C., Haywood, A.M., 2018. PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. Sci Data 5, 180254. https://doi.org/10.1038/sdata.2018.254

Campos, V.E., Cappa, F.M., Viviana, F.M., Giannoni, S.M., 2016. Using remotely sensed data to model suitable habitats for tree species in a desert environment. Journal of Vegetation Science 27, 200–210. https://doi.org/10.1111/jvs.12328

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., Ram, K., 2024. Rgbif: Interface to the global biodiversity information facility API.

Chamberlain, S., Boettiger, C., 2017. R python, and ruby clients for GBIF species occurrence data. PeerJ PrePrints.

Collins, C., Asouti, E., Grove, M., Kabukcu, C., Bradley, L., Chiverrell, R., 2018. Understanding resource choice at the transition from foraging to farming: An application of palaeodistribution modelling to the Neolithic of the Konya Plain, south-central Anatolia, Turkey. J. Archaeol. Sci. 96, 57–72. https://doi.org/10.1016/j.jas.2018.02.003

Conolly, J., Manning, K., Colledge, S., Dobney, K., Shennan, S., 2012. Species distribution modelling of ancient cattle from early Neolithic sites in SW Asia and Europe. Holocene 22, 997–1010. https://doi.org/10.1177/0959683612437871

Di Virgilio, G., Wardell-Johnson, G.W., Robinson, T.P., Temple-Smith, D., Hesford, J., 2018. Characterising fine-scale variation in plant species richness and endemism across topographically complex, semi-arid landscapes. Journal of Arid Environments 156, 59–68. https://doi.org/10.1016/j.jaridenv.2018.04.005

Dubuis, A., Giovanettina, S., Pellissier, L., Pottier, J., Vittoz, P., Guisan, A.,

2013. Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. Journal of Vegetation Science 24, 593–606. https://doi.org/10.1111/jvs.12002

Franklin, J., Potts, A.J., Fisher, E.C., Cowling, R.M., Marean, C.W., 2015. Paleodistribution modeling in archaeology and paleoanthropology. Quat. Sci. Rev. 110, 1–14. https://doi.org/10.1016/j.quascirev.2014.12.015

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km–global soil information based on automated mapping. PLoS One 9, e105992. https://doi.org/10.1371/journal.pone.0105992

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Sci Data 4, 170122. https://doi.org/10.1038/sdata.2017.122

Kopecký, M., Čížková, Š., 2010. Using topographic wetness index in vegetation ecology: Does the algorithm matter? Appl. Veg. Sci. 13, 450–459. https:

//doi.org/10.1111/j.1654-109X.2010.01083.x

Leempoel, K., Parisod, C., Geiser, C., Daprà, L., Vittoz, P., Joost, S., 2015. Very high-resolution digital elevation models: Are multi-scale derived variables ecologically relevant? Methods in Ecology and Evolution 6, 1373–1383. https://doi.org/10.1111/2041-210X.12427

Lindsay, J.B., 2016. Whitebox GAT: A case study in geomorphometric analysis. Comput. Geosci. 95, 75–84. https://doi.org/10.1016/j.cageo.2016.07.003

Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of Biogeography 40, 778–789. https://doi.org/10.1111/jbi.12058

Martin, L., Edwards, Y.H., Roe, J., Garrard, A., 2016. Faunal turnover in the Azraq Basin, eastern Jordan 28,000 to 9,000 cal BP, signalling climate change and human impact. Quaternary Research 86, 200–219. https://doi.org/10.1016/j.yqres.2016.07.001

Miller, T., Blackwood, C.B., Case, A.L., 2024. Assessing the utility of SoilGrids250 for biogeographic inference of plant populations. Ecology and Evolution 14, e10986. https://doi.org/10.1002/ece3.10986

Mod, H.K., Scherrer, D., Luoto, M., Guisan, A., 2016. What we use is not what we know: Environmental predictors in plant distribution models. Journal of Vegetation Science 27, 1308–1322. https://doi.org/10.1111/jvs.12444

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

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

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2022. Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. Ecological Monographs 92, e01486. https://doi.org/10.1002/ecm.1486

Velazco, S.J.E., Galvão, F., Villalobos, F., De Marco Júnior, P., 2017. Using

worldwide edaphic data to model plant species niches: An assessment at a continental extent. PLoS One 12, e0186025. https://doi.org/10.1371/journal.pone.0186025

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Group, N.P.S.D.W., 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

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, 1–17. https://doi.org/10.18637/jss.v077.i01

Yaworsky, P.M., Hussain, S.T., Riede, F., 2023. Climate-driven habitat shifts of high-ranked prey species structure Late Upper Paleolithic hunting. Scientific Reports 13, 4238. https://doi.org/10.1038/s41598-023-31085-x