WALLACE: A FLEXIBLE PLATFORM FOR REPRODUCIBLE MODELING OF SPECIES NICHES AND DISTRIBUTIONS BUILT FOR COMMUNITY EXPANSION

JAMIE M. KASS, BRUNO VILELA, MATTHEW E. AIELLO-LAMMENS, ROBERT MUSCARELLA, CORY MEROW AND ROBERT P. ANDERSON

ABTRACT

1) Scientific research increasingly calls for open-source software that is flexible, interactive, and expandable, while providing methodological guidance and reproducibility. Currently, many analyses in ecology are implemented with "black box" graphical user interfaces that lack flexibility or command-line interfaces that are infrequently used by non-specialists.

2) To help remedy this situation in the context of species distribution modeling, we created *Wallace*, an open and modular application with a richly documented graphical user interface to underlying R scripts that is flexible and highly interactive.

3) *Wallace* guides users from acquiring and processing data to building models and examining predictions. Additionally, it is designed to grow via community contributions of new modules to expand functionality. All results are downloadable, along with code to reproduce the analysis.

4) *Wallace* provides an example of an innovative platform to increase access to cutting-edge methods and encourage plurality in science and collaboration in software development.

INTRODUCTION

Ecological and evolutionary studies have shifted over the past 20 years towards increasingly complex analyses (Bolker 2008). This has been enabled, in part, by a rise in computing power and the increasing openness of data and software (Gimenez *et al.* 2014). As a result, most current methods are accessed as either: 1) programming-language scripts run in command-line interfaces (CLIs; e.g. R and Python), or 2) software with graphical user interfaces (GUIs). On one hand, programming scripts provide flexibility, but custom code is often poorly documented and tailored to specific analyses (Mislan *et al.* 2016). GUIs, on the other hand, are easy to navigate and extend accessibility of analyses to more users, but are less flexible than custom code and often necessitate using multiple software packages to complete a study. This exacerbates a problem with GUIs: lack of reproducibility (Hampton *et al.* 2015). Additionally, GUI implementations of methods often lag behind the cutting-edge analyses enabled by the frequent release of scripts with new publications. Hence, tools that combine the positive aspects of CLI and GUI methods can help advance ecological research.

We developed *Wallace* to address these issues specifically for user communities in ecology and the environmental sciences. *Wallace* is an open-source GUI application that offers user-friendly access to R-scripted modern workflows. It is available as the R package "wallace" on CRAN, with a development version on Github (see Data Accessibility for links). *Wallace* currently focuses on a workflow for modeling species niches and geographic distributions (Fig. 1; Guisan & Thuiller 2005; Peterson *et al.* 2011), but we anticipate that future versions will expand the analyses offered for

biogeographical and macroecological modeling. *Wallace* is written for R (R Core Team 2017) using `shiny` (a package for developing interactive applications; Chang *et al.* 2017), and can thus leverage the rapidly expanding suite of R packages authored by the scientific community. Six main qualities of *Wallace* distinguish it as a model for providing access and guidance for advanced methodologies (Table 1).

Figure 1

Below, we present several important issues in niche/distribution modeling and explain how we address them with *Wallace*, first conceived as a response to the Global Biodiversity Information Facility's 2015 Ebbe Nielsen Challenge (https://devpost.com/software/wallace-round-2). We then provide a walkthrough of the application and conclude by discussing the general utility of *Wallace*'s framework for disseminating scientific methods and encouraging community-wide innovation.

Table 1

CURRENT ISSUES IN NICHE/DISTRIBUTION MODELING

Wallace currently implements analyses for species niche/distribution modeling (hereafter "distribution modeling"). These correlative models estimate the response of a species to the environment and with clear assumptions can be used to infer (or hypothesize) geographic ranges, environmental suitability across a landscape, or niche requirements (Franklin 2010a; Peterson *et al.* 2011). Distribution modeling is used in many disciplines,

such as phylogeography (Alvarado-Serrano & Knowles 2014), community ecology (Guisan & Rahbek 2011), evolutionary biology (McCormack *et al.* 2010), and conservation (Franklin 2010b). At a minimum, it requires georeferenced occurrence records of the study species (e.g., from field surveys, museum collections, citizen science) and environmental predictors (e.g., climate, land cover, topography). Occurrence data generally represent the primary ecological information available for the vast majority of species. "Presence-only" distribution models use environmental values at occurrences, typically contrasting them with those available in the study region ("background" or pseudoabsence samples; Elith *et al.* 2011). Since reliable absence data are unavailable for most species, research focusing on presence-only models has grown tremendously over the past two decades. The current implementation of *Wallace* concentrates on these models, highlighting two algorithms with differing complexity: BIOCLIM (Booth *et al.* 2014) and Maxent (Phillips *et al.* 2006). Many approaches exist for making such models, and comparing them conveys to users that the utility of a model does not necessarily improve with complexity (Jiménez-Valverde *et al.* 2008).

Confusion abounds regarding how best to choose and implement presence-only distribution modeling methods and interpret their outputs (Joppa *et al.* 2013). There have been numerous calls to address a range of complicating issues (Elith *et al.* 2011; Merow *et al.* 2013), among them sampling bias (Bean *et al.* 2012), selection of study extent (VanDerWal *et al.* 2009), model evaluation (Radosavljevic & Anderson 2014), model selection (Warren & Seifert 2011), and considering key assumptions (Yackulic *et al.* 2013). *Wallace* provides extensive guidance text and enables user experimentation with a

variety of modules (see *Walkthrough*), directly addressing some of these issues and encouraging the use of a diversity of methods (Fig. 2).

Figure 2

Difficulties For Programmers and Non-Programmers Alike

Wallace combines the strengths of GUI and CLI approaches to enable research in distribution modeling for a broad audience. A number of GUI-based applications have been widely used for distribution modeling analyses (e.g., maxent.jar (Phillips *et al.* 2006); DesktopGARP (Scachetti-Pereira 2002); openModeller (de Souza Muñoz *et al.* 2011)), but an ongoing problem is that many researchers treat them like "black boxes", even though documentation exists in the literature (Joppa *et al.* 2013). In addition to the shortfalls mentioned above, these GUIs lack adequate guidance within the software. Further, relying on CLIs for distribution modeling can be challenging even for specialists because it involves a combination of map inspection, spatial analysis, and statistical modeling.

Very recently, a number of new distribution modeling applications present exciting developments in reproducible science and indicate the demand for software that advances accessibility and collaboration (e.g. De Giovanni *et al.* 2016; Hallgren *et al.* 2016; Hardisty *et al.* 2016; Naimi *et al.* 2016; Golding *et al.* Accepted). However, those that highlight customizability and modularity require programming skills and currently lack integrated guidance on methods, while others that feature user-friendly interfaces and

extensive educational resources are less flexible and have fewer opportunities for user contributions. *Wallace* aims to provide a wide variety of advantages by having an easily navigable interface featuring advanced and expandable modeling tools, guidance on theory and methods, and access to the underlying code. Further development of these innovative applications and cross-collaboration among them—including *Wallace*—would benefit the field greatly. Clearly no single lab or research group can address all the needs of biogeography and related fields, and *Wallace* is designed to expand in an agile fashion as the field advances and new demands arise.

WALKTHROUGH

We present a brief walkthrough of *Wallace* v1.0, which is divided into a series of components that each feature one or more modules. Module authors and featured R packages are documented in each module (Fig. 2). All major modules have associated unit tests, and these will be standard with module submission going forward.

1) Obtain Occurrence Data: Species occurrence records can be obtained from online databases or supplied by the user. *Wallace* currently accesses GBIF, VertNet, and BISON, removes duplicate coordinates, plots localities on a map, and populates a data table.

2) Process Occurrence Data: The user chooses which localities to include in the analysis and can address sampling bias by selecting localities on a map or using a spatial-thinning algorithm (Aiello-Lammens *et al*. 2015).

3) Obtain Environmental Data: For gridded predictor variables to characterize the species' response to the environment, *Wallace* currently offers WorldClim bioclimatic rasters (Hijmans *et al.* 2005) or allows user-input rasters.

4) Process Environmental Data: The user delineates a study extent to crop the predictor grids and draw background samples, as required by most presence-only models. *Wallace* offers four alternatives, with optional buffering: bounding box, minimum convex polygon, buffers around occurrence points, and user input.

5) Partition Occurrence Data: To evaluate models, the user chooses among (spatial and non-spatial) methods to partition occurrence localities into groups for *k*-fold cross-validation.

6) Build and Evaluate Niche Models: To examine model complexity, users can fit multiple models and use evaluation statistics to identify optimal settings (e.g., regularization multipliers, feature classes; Hijmans et al. 2017; Muscarella et al. 2014).

7) Visualize Model Results: The user can pan around the map to explore suitability predictions for the study extent, examine response curves for predictor variables, and view evaluation plots.

8) Project Model: The user can project models to other areas or time periods. *Wallace* currently allows future projections based on the estimates of different global circulation models (Hijmans *et al*. 2005). Critically, users can view the magnitude of environmental novelty between the study extent and the projected area/time, which can highlight areas to exercise caution in interpretation (Elith *et al.* 2010).

9) Session Code: The user can download an R Markdown script that reproduces the analysis undertaken during the *Wallace* session.

TARGET AUDIENCES

We developed *Wallace* with a wide range of audiences in mind. Graduate students interested in distribution modeling and coding but who are not yet advanced programmers should benefit from learning interactively using *Wallace*. Conservation practitioners and natural resource managers may want to assess data availability and quality for a study species, learn about methods, run analyses, and share results with colleagues. Experienced programmers can run models, download the session code, and customize it to modify or extend the analysis. Those developing new methods may also want to disseminate their products by contributing new modules to *Wallace*. Lastly, educators can use *Wallace* to teach interactive lessons about ecology, programming, and scientific best practices.

CONCLUSIONS AND FUTURE DIRECTIONS

Wallace demonstrates an innovative, open platform for rapid dissemination of scientific methods to a broad audience—specifically encouraging plurality in methodology and ongoing community development. Over the next three years under funding from the U.S. National Science Foundation, we plan to work closely with a cadry of international research groups to integrate new modules, both expanding available options within the existing scope of *Wallace* and broadening the scope of its capabilities. Some plans for the future include providing more environmental datasets and modeling algorithms, measuring prediction uncertainty, integrating analyses that use distribution models as inputs (e.g. measuring biodiversity, conservation planning), and model comparison tools.

Above all, our vision for an expandable software like *Wallace* is that users decide what needs to be added and become contributors themselves.

Although *Wallace* is currently focused on distribution modeling, other fields may benefit from adopting a similar framework for software development. Like distribution models, many complex analyses can often be broken down into components and assembled into teachable workflows for disseminating methods to a broad audience. Furthermore, science advances most quickly when researchers share advancements and build tools together, which *Wallace* enables. Finally, *Wallace*'s interactive nature demonstrates an alternative to static manuals, tutorials, or vignettes for presenting new methods. The next generation of scientific software will benefit from these ideas, which could lead to more individuals learning, contributing to, and engaged in a dynamic process of creative collaboration.

AUTHORS' CONTRIBUTIONS

J.M.K., R.P.A., M.A.-L., B.V., and R.M. conceived of the original genesis of *Wallace*. J.M.K led code development, and B.V., M.A.-L., R.M., and C.M. were co-developers. J.M.K. drafted the manuscript, with major input from R.P.A., and all other coauthors provided revisions. R.P.A. led overall project design and drafted guidance text, on which J.M.K. provided revisions.

ACKNOWLEDGMENTS

This work was made possible via an award from the Global Biodiversity Information Facility (Finalist, 2015 Ebbe Nielsen Challenge, led by J.M.K.), a CUNY Graduate

Center Provost Digital Innovation Grant to J.M.K., and grants from the U.S. National Science Foundation (DBI-1650241, DBI-1661510, and DEB-1119915 to R.P.A.). B.V. was supported by a CAPES doctoral grant (Brazil), M.A.-L. by NSF DEB-1046328, and R.M. by NSF DBI-1401312. We appreciate helpful comments and suggestions from the Anderson lab and other members of the CUNY Ecology, Evolution, and Behavior group; Mary E. Blair; and Morgan W. Tingley.

DATA ACCESSIBILITY

No data were included in this article. The R package is freely available under the GPL-3 license at Github (development version: https://github.com/wallaceEcoMod/wallace) and CRAN (stable version: https://cran.r-project.org/package=wallace). There is also a project webpage that will be updated with ongoing development (https://wallaceecomod.github.io).

REFERENCES CITED

1. Aiello-Lammens M.E., Boria R.A., Radosavljevic A., Vilela B. & Anderson R.P. (2015) spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38, 541-545.
2. Alvarado-Serrano D.F. & Knowles L.L. (2014) Ecological niche models in phylogeographic studies: applications, advances and precautions. *Molecular Ecology Resources*, 14, 233-248.
3. Bean W.T., Stafford R. & Brashares J.S. (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, 35, 250-258.
4. Bolker B.M. (2008) *Ecological models and data in R*, Princeton University Press, Princeton.
5. Booth T.H., Nix H.A., Busby J.R. & Hutchinson M.F. (2014) BIOCLIM: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, 20, 1-9.
6. Chang W., Cheng J., Allaire J.J., Xie Y. & McPherson J. (2017) shiny: Web Application Framework for R. R package version 1.0.3, http://cran.r-project.org/package=shiny.
7. De Giovanni R., Williams A.R., Ernst V.H., Kulawik R., Fernandez F.Q. & Hardisty A.R. (2015) ENM Components: a new set of web service-based workflow components for ecological niche modelling. *Ecography*, 39, 376-383.
8. Elith J., Kearney M. & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 4, 330-342.

9. 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, 43-57.
10. Franklin J. (2010a) *Mapping species distributions: spatial inference and prediction*, Cambridge University Press, New York.
11. Franklin J. (2010b) Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, 16, 321-330.
12. Gimenez O., *et al.* (2014) Statistical ecology comes of age. *Biology Letters*, 10, 20140698.
13. Golding N., August T.A., Lucas T.C.D., Gavaghan D.J., van Loon E.E. & McInerny G. The zoon R package for reproducible and shareable species distribution modelling. *Methods in Ecology and Evolution*, Accepted Author Manuscript, doi:10.1111/2041-210X.12858.
14. Guisan A. & Thuiller W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8, 993-1009.
15. Guisan A. & Rahbek C. (2011) SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 38, 1433-1444.
16. Hallgren W., *et al.* (2016) The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data. *Environmental Modelling & Software*, 76, 182-186.
17. Hardisty A.R., *et al.* (2016) BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. *BMC Ecology*, 16, 49.

18. Hampton S.E., *et al.* (2015) The Tao of open science for ecology. *Ecosphere*, 6, 120.
19. 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. *International Journal of Climatology*, 25, 1965-1978.
20. Hijmans R.J., Phillips S., Leathwick J. & Elith J. (2017) dismo: Species Distribution Modeling. R package version 1.1-4, http://cran.r-project.org/package=dismo.
21. Jiménez-Valverde A., Lobo J.M. & Hortal J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14, 885-890.
22. Joppa L.N., McInerny G., Harper R., Salido L., Takeda K., O'Hara K., Gavaghan D. & Emmott S. (2013) Troubling trends in scientific software use. *Science*, 340, 814-815.
23. McCormack J.E., Zellmer A.J. & Knowles L.L. (2010) Does niche divergence accompany allopatric divergence in *Aphelocoma* jays as predicted under ecological speciation?: Insights from tests with niche models. *Evolution*, 64, 1231-1244.
24. Merow C., Smith M.J. & Silander J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36, 1058-1069.
25. Mislan K.A.S., Heer J.M. & White E.P. (2016) Elevating the status of code in ecology. *Trends in Ecology & Evolution*, 31, 4-7.

26. Muscarella R., Galante P.J., Soley-Guardia M., Boria R.A., Kass J.M., Uriarte M. & Anderson R.P. (2014) ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, 5, 1198-1205.
27. Naimi B. & Araújo M.B. (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, 39, 368-375.
28. Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martinez-Meyer E., Nakamura M. & Araújo M.B. (2011) *Ecological niches and geographic distributions*, Princeton University Press.
29. Phillips S.J., Anderson R.P. & Schapire R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231-259.
30. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL: http://www.R-project.org [accessed 02 February 2017]
31. Radosavljevic A. & Anderson R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, 41, 629-643.
32. Scachetti-Pereira R. (2002) DesktopGarp: a software package for biodiversity and ecologic research. United States: The University of Kansas Biodiversity Research Center. Available online at http://www.nhm.ku.edu/desktopgarp/
33. de Souza Muñoz M.E. *et al.* (2011) openModeller: a generic approach to species' potential distribution modelling. *GeoInformatica*, 15, 111-135.

34. VanDerWal J., Shoo L.P., Graham C. & Williams S.E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220, 589-594.

35. Warren D.L & Seifert S.N. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 21, 335-342.

36. Yackulic C.B., Chandler R., Zipkin E.F., Royle J.A., Nichols J.D., Campbell Grant E.H. & Veran S. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, 4, 236-243.

318 Table 1. Advantages of the *Wallace* framework.

OPEN	- code is free and open-source (GNU GPL 3.0) - users can access data from online databases
EXPANDABLE	- modules (discrete methodological options) can be contributed by the community
FLEXIBLE	- multiple options exist for user data uploads and downloads of results
INTERACTIVE	- sessions are participatory and encourage experimentation - a variety of visualizations are provided (maps, tables, figures)
INSTRUCTIVE	- guidance text (theoretical and methodological) is included for all components and modules
REPRODUCIBLE	- an annotated and executable R Markdown file is produced for rerunning analyses, sharing results, providing supplemental information / educational resources

319