Received: 18 May 2020 | Accepted: 6 August 2020

DOI: 10.1111/2041-210X.13469

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

dsmextra: Extrapolation assessment tools for density surface models

Phil J. Bouchet[1,2] | David L. Miller[1,2] | Jason J. Roberts[3] | Laura Mannocci[4] | Catriona M. Harris[1,5] | Len Thomas[1,2]

[1]Centre for Research into Ecological and Environmental Modelling (CREEM), University of St Andrews, St Andrews, UK

[2]School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

[3]Marine Geospatial Ecology Lab, Duke University, Durham, NC, USA

[4]MARBEC (Marine Biodiversity, Exploitation and Conservation), University of Montpellier, CNRS, IFREMER, IRD, Montpellier, France

[5]School of Biology, University of St Andrews, St Andrews, UK

Correspondence
Phil J. Bouchet
Email: pb282@st-andrews.ac.uk

Funding information
OPNAV N45 and the SURTASS LFA Settlement Agreement; U.S. Navy's Living Marine Resources program, Grant/Award Number: N39430-17-C-1982

Handling Editor: Laura Graham

Abstract

1. Forecasting the responses of biodiversity to global change has never been more important. However, many ecologists faced with limited sample sizes and shoestring budgets often resort to extrapolating predictive models beyond the range of their data to support management actions in data-deficient contexts. This can lead to error-prone inference that has the potential to misdirect conservation interventions and undermine decision-making. Despite the perils associated with extrapolation, little guidance exists on the best way to identify it when it occurs, leaving users questioning how much credence they should place in model outputs. To address this, we present dsmextra, a new R package for measuring, summarizing and visualizing extrapolation in multivariate environmental space.
2. dsmextra automates the process of conducting quantitative, spatially explicit assessments of extrapolation on the basis of two established metrics: the Extrapolation Detection (ExDet) tool and the percentage of data nearby (%N). The package provides user-friendly functions to (a) calculate these metrics, (b) create tabular and graphical summaries, (c) explore combinations of covariate sets as a means of informing covariate selection and (d) produce visual displays in the form of interactive html maps.
3. dsmextra implements a model-agnostic approach to extrapolation detection that is applicable across taxonomic groups, modelling techniques and datasets. We present a case study fitting a density surface model to visual detections of pantropical spotted dolphins *Stenella attenuata* in the Gulf of Mexico.
4. Predictive modelling seeks to deliver actionable information about the states and trajectories of ecological systems, yet model performance can be strongly impaired out of sample. By assessing conditions under which models are likely to fail or succeed in extrapolating, ecologists may gain a better understanding of biological patterns and their underlying drivers. Critical to this is a concerted effort to standardize best practice in model evaluation, with an emphasis on extrapolative capacity.

KEYWORDS
cetaceans, distance sampling, ecological predictions, extrapolation, model transferability, R package, spatial modelling, wildlife surveys

1 | INTRODUCTION

The expanding footprint of human activities is rapidly creating novel challenges for the conservation of biodiversity. Maps of wildlife density patterns are fundamental to mitigating the impacts of anthropogenic pressures, but reliable estimates remain elusive for many animal populations inhabiting areas subject to limited sampling effort (e.g. Kaschner, Quick, Jewell, Williams, & Harris, 2012). Density surface modelling (DSM) (Hedley & Buckland, 2004; Miller, Burt, Rexstad, & Thomas, 2013) is a popular framework for estimating species abundance, particularly in support of spatial planning and decision-making (Becker et al., 2018; Mannocci, Roberts, Miller, & Halpin, 2017; Redfern et al., 2017). However, demands for solutions to large-scale management problems mean that predictions are increasingly being made far beyond the boundaries of the study regions where the data used to fit DSMs were originally collected (e.g. Mannocci, Monestiez, Spitz, & Ridoux, 2015). Statisticians usually issue strong warnings against such practice, as extrapolating relies heavily on assumptions that are unlikely to hold outside the range of sampled conditions (Conn, Johnson, & Boveng, 2015). Consequently, most extrapolations are considered prone to errors, the magnitude of which can vary substantially across taxonomic groups, habitats, study systems and/or modelling techniques (Yates et al., 2018).

Despite these difficulties, untested predictions derived from the best available science are still viewed by many as more desirable than proceeding blindly. In particular, projecting models into novel contexts (e.g. new geographical areas and/or future time periods) remains a practical necessity in many risk assessments, making the ability to evaluate the extent and magnitude of extrapolation a pivotal component of research agendas in applied ecology (Sequeira, Bouchet, Yates, Mengersen, & Caley, 2018). Although several extrapolation metrics have been proposed to address this (Table S1), there is little consensus on which works best for a given ecological scenario, and scant consideration of how results are sensitive to metric choice (Grenier, Parent, Huard, Anctil, & Chaumont, 2013). Standard protocols for supporting consistent assessments of extrapolation are therefore lacking (Sequeira et al., 2018; Yates et al., 2018), leaving end users deprived of both the guidelines and the tools they require to objectively audit models projected into novel and/or data-deficient contexts.

We introduce the dsmextra R package, a user-friendly toolkit for quantitatively assessing extrapolation. The package was built as a companion to dsm, a sibling package focused on the analysis of distance sampling survey data using generalized additive models (Miller et al., 2013). However, dsmextra only requires knowledge of covariate values at sampled and prediction locations, meaning that it is not restricted to line/strip-transect data and that its functions are readily applicable to a wide variety of datasets, irrespective of how they were collected or of the statistical frameworks in which they are analysed. Here, we describe the package, explain the concepts underpinning its use and illustrate some of its features using a survey dataset on pantropical spotted dolphins *Stenella attenuata* in the Gulf of Mexico.

2 | SOFTWARE AND EXAMPLE

dsmextra provides functions (Table S2) designed to assist users in measuring, summarizing and visualizing extrapolation, as precursory steps in the traditional modelling workflow (Figure 1). A number of

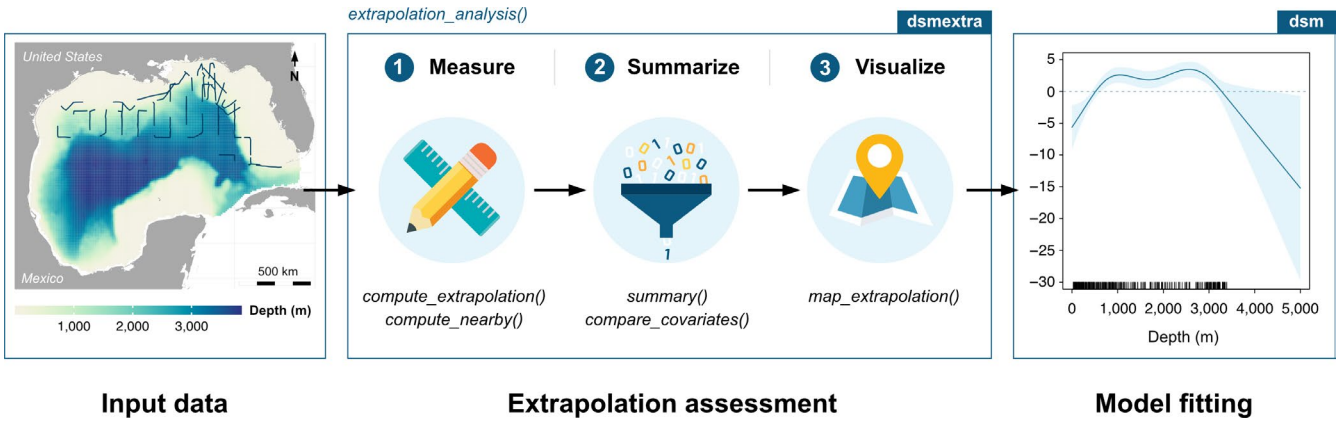

FIGURE 1 Typical workflow for conducting an a priori assessment of extrapolation within the dsmextra R package, using line transect surveys (solid lines, left panel) of pantropical spotted dolphins *Stenella attenuata* in the Gulf of Mexico as an example (see Supporting Information for details). Package functions are shown in italics. A wrapper, extrapolation_analysis, allows all three steps (middle panel) to be performed within a single function call. Results can then be used to inform model development and interpretation, e.g. in package dsm (right panel)

these functions are adapted from existing—but scattered—code released on CRAN (https://cran.r-project.org/) or available as ad hoc scripts in the supplemental materials of peer-reviewed articles (e.g. Mannocci et al., 2018). dsmextra represents the first attempt to consolidate these methods while extending their usability and functionality.

Attempts at characterizing extrapolation are typically made following model calibration, as a test of model adequacy, realism and predictive performance (Robinson, Nelson, Costello, Sutherland, & Lundquist, 2017). We argue that this is unproductive, in part because model fitting can be time-consuming (especially for large datasets), meaning that there is little to gain from spending hours (or more) building a model that is unlikely to return credible outputs in the first place. Additionally, a number of comparative studies have demonstrated mixed extrapolation success among different models (or model types) fitted to the same datasets (e.g. Shabani, Kumar, & Ahmadi, 2016), suggesting that idiosyncrasies in model structures and parameterizations may result in inconsistent out-of-sample behaviour that may compromise the interpretation of model outputs. Lastly, the independent data required for rigorous model validation are often lacking, precluding any opportunities to objectively measure the quality of model predictions under novel conditions (Sequeira et al., 2018). For these reasons, we instead suggest that extrapolation assessments take place a priori, ensuring that models that do perform poorly due to extrapolative issues outside the sampled covariate space can be identified early on and avoided (Figure 1). Doing so requires that we determine the degree to which the sampled environmental conditions are comparable to those prevailing in the system to which the model is being projected (Guisan, Thuiller, & Zimmermann, 2017). In the past, many ecologists have typically taken a largely qualitative approach to drawing this comparison, adopting a binary view of extrapolation that only considers the ranges of each individual covariate (Figure S1). With the compute family of functions, dsmextra improves on this by making available two quantitative metrics that capture complementary dimensions of extrapolation related to environmental analogy (i.e. conditions existing in one system but not the other) and availability (combinations of conditions and their frequency of occurrence; Guisan et al., 2017). Together, these yield a more nuanced picture of departures from sampled conditions.

compute_extrapolation is an R implementation of the Extrapolation Detection (ExDet) tool proposed by Mesgaran, Cousens, and Webber (2014), which uses Euclidean and Mahalanobis distances to characterize both univariate and combinatorial extrapolation (Bouchet et al., 2019). Detecting the latter is especially important for identifying changes in collinearity patterns among covariates, which may undermine model transferability (Rödder & Engler, 2012). dsmextra also makes it possible to identify the covariates making the largest contributions to extrapolation (the 'most influential covariates' or MIC) (Mesgaran et al., 2014). compute_nearby returns the percentage of data nearby (%N), as described by King and Zeng

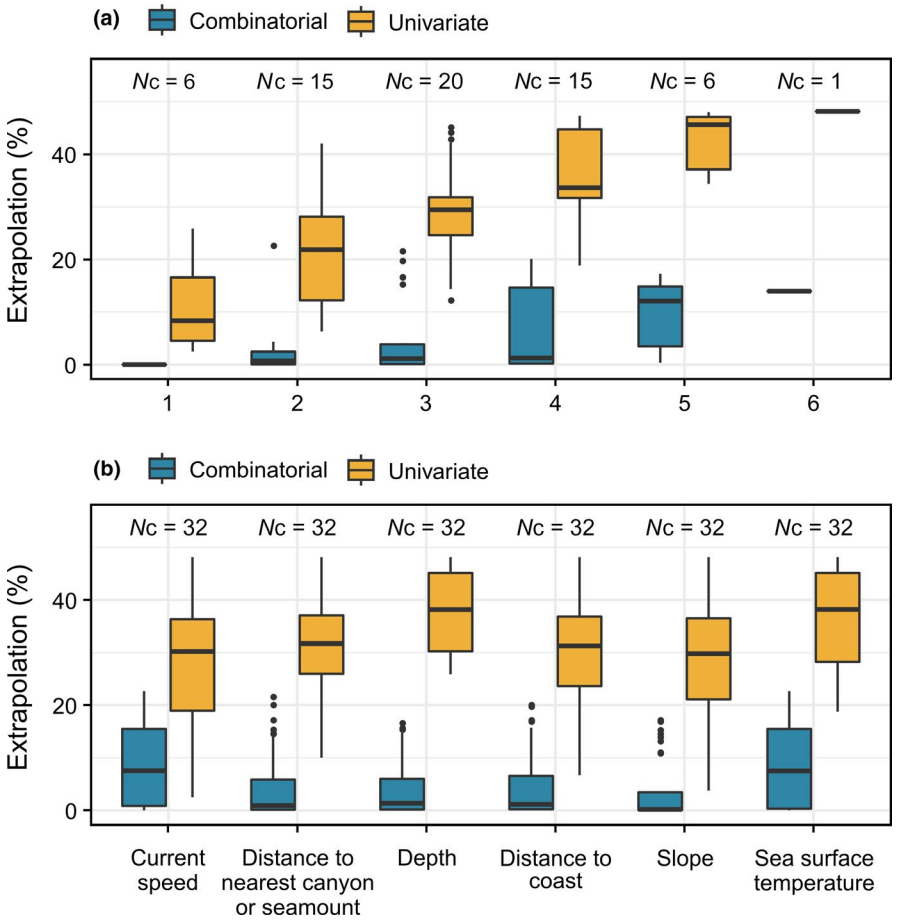

FIGURE 2 Graphical comparison of the magnitude of univariate and combinatorial extrapolation associated with different sets of six input covariates. Results are summarised for (a) sets of increasing size (all combinations of one to six covariates), and (b) covariate type, and expressed as the proportion of locations (i.e. grid cells) subject to extrapolation in the prediction area. In (a, b) *Nc* respectively indicates the number of possible covariate combinations and the number of combinations in which each covariate is found

(2007). %N relies on the Gower's distance to calculate the proportion of reference data lying within a given radius of any prediction point, with the expectation that extrapolations made in proximity to a larger amount of sample data will be better 'informed'—and thus more reliable—all else being equal (Figure S1). As a rule, this radius is set to the mean Gower's distance between all pairs of reference points (King & Zeng, 2007); however, the nearby argument allows users to change this value. Both functions require a minimum of four inputs, some of which are similar to dsm: (a) a samples data.frame object containing the survey data used for model building; (b) a prediction.grid data.frame object containing the geographic coordinates (labelled *x*, *y*) of the target locations where model predictions are desired, and the values of covariates of interest at these locations; (c) a coordinate reference system, supplied either as an object of class CRS or as simple character vector and (d) a vector of covariate names, as follows:

```
stenella.exdet <- compute_extrapolation(samples = survey.segs,
  covariate.names = stenella.covariates, prediction.grid = pred.grid,
  coordinate.system = GulfMexico_proj)
```

```
stenella.nearby <- compute_nearby(samples = survey.segs, covariate.
  nnames = stenella.covariates, prediction.grid = pred.grid, nearby =
  1, coordinate.system = GulfMexico_proj)
```

Extrapolated predictions should always be checked for biological plausibility, yet model users are frequently tempted to include a kaleidoscope of predictors with the expectation that the true ecological drivers among them will naturally transpire (Guisan et al., 2017). Covariate selection thus remains a prominent issue in statistical ecology, particularly as several model formulations may fit the sample data equally well but lead to vastly diverging predictions (Dormann et al., 2012). Furthermore, the more covariates are considered, the more combinations of their values there can be and the higher the risk of extrapolation (Mahony, Cannon, Wang, & Aitken, 2017). Failure to filter variables may thus lead to unreasonable predictions of species' responses to novel conditions (Petitpierre, Broennimann, Kueffer, Daehler, & Guisan, 2017). compare_covariates is a wrapper around compute_extrapolation designed to run iteratively for all combinations of n.covariates, helping

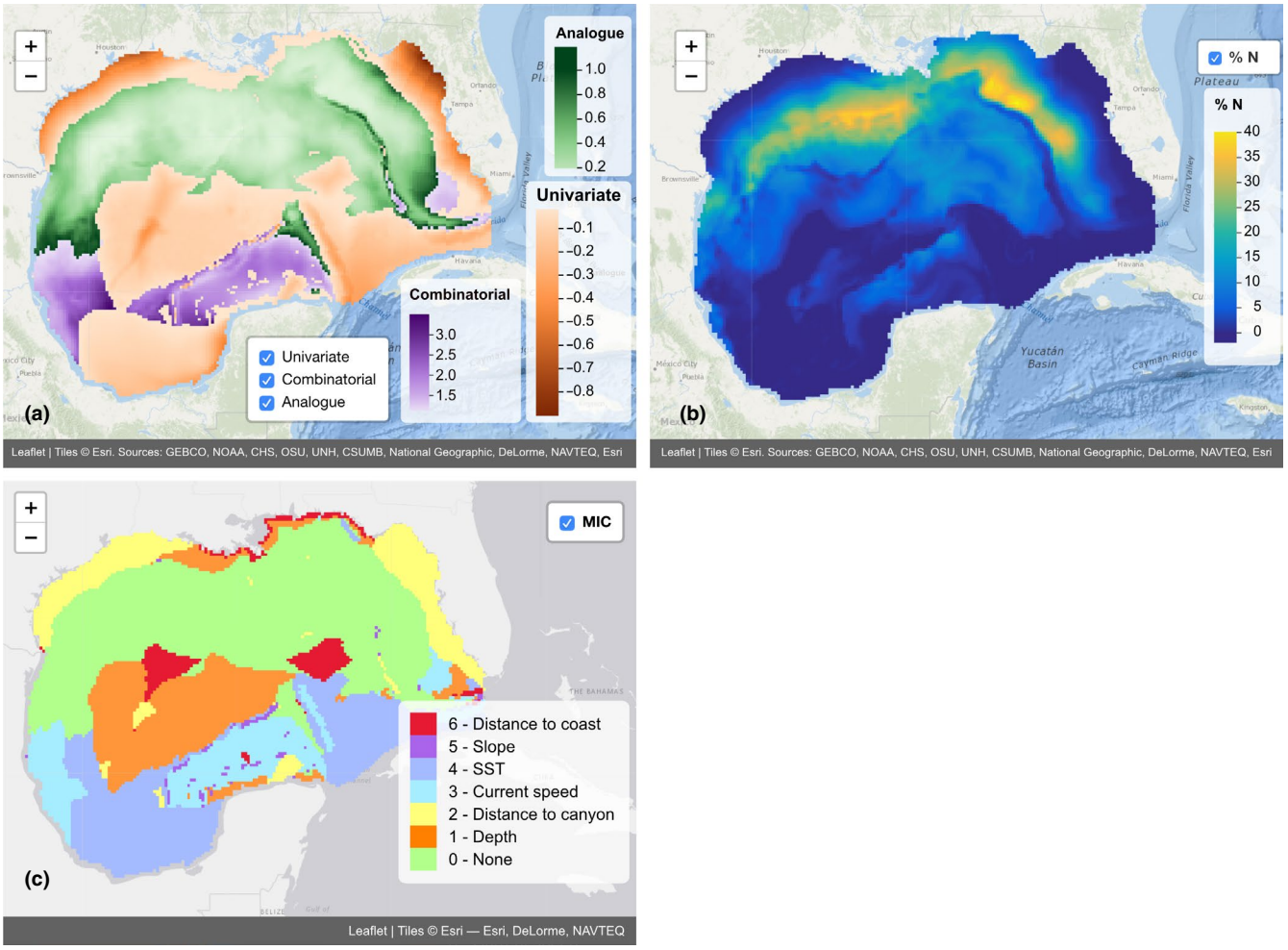

FIGURE 3 Visualisation of extrapolation using the map_extrapolation function in dsmextra. Maps display the extent of extrapolation in the multivariate space defined by six input covariates, and respectively measured by (a) the Extrapolation Detection tool (ExDet) and (b) the percentage of data nearby (%N). Panel (c) shows the distribution of the most influential covariates (MIC), i.e. those making the largest contributions to univariate and combinatorial extrapolation

to guide this process. The function returns graphical and tabular overviews of the covariates sets that minimize and maximize univariate and combinatorial extrapolation, measured as the number of prediction.grid cells subject to either (Figure 2). n.covariates can be supplied as any vector of integers from 1 to p, where p is the total number of covariates available. For instance, Figure 2 was produced with the command:

```
compare_covariates(extrapolation.type = "both", samples = survey.
  segs, n.covariates = 5, covariate.names = stenella.covariates,
  prediction.grid = pred.grid, coordinate.system = GulfMexico_proj,
  create.plots = TRUE)
```

Map production is the final step in the dsmextra workflow. map_extrapolation generates interactive visualizations of extrapolation using leaflet (https://rstudio.github.io/leaflet/), giving users the capacity to pan/zoom in on areas of interest and toggle map layers on/off as desired (Figure 3). Species sightings and survey tracks can also be overlaid, if available. Figure 3b was obtained as follows:

```
dsmextra::map_extrapolation(map.type = "nearby", gower.values =
  stenella.nearby, covariate.names = stenella.covariates, prediction.
  grid = pred.grid, coordinate.system = GulfMexico_proj)
```

3 | DISCUSSION

There is still much to learn about extrapolation and its implications for predictive inference. By extrapolating, we are without a doubt using models in risky ways and making implicit assumptions about the generality of ecological processes, which we often have no data to validate empirically. The potential for errors is therefore non-negligible, with growing evidence that predictive performance may be impaired when a model is transferred to novel contexts, especially when relationships vary geographically and/or temporally (Mannocci, Roberts, Pedersen, & Halpin, 2020) (Figure S2). Worryingly, general understanding of extrapolation and its consequences seems to be lacking in many disciplines, often misleading practitioners and policy-makers into taking extrapolated predictions at face value, irrespective of their uncertainties and biases. Developing more rigorous extrapolation practice thus hinges on raising awareness of possible shortcomings as well as harmonizing strategies for its detection and reporting. dsmextra was conceived with this goal in mind, and offers an intuitive, model-agnostic toolkit for diagnosing extrapolation that accounts for various types of departures from reference conditions, and enables the geographic and temporal distribution of extrapolation to be easily displayed on a map. The latter is especially important for alleviating pervasive scepticism in conservation decisions, and guiding survey design by delineating areas where future sampling would likely improve model predictions.

ACKNOWLEDGEMENTS

We would like to thank Natalie Kelly and one anonymous reviewer for their thoughtful feedback on earlier versions of this manuscript. This work was funded by OPNAV N45 and the SURTASS LFA Settlement Agreement, managed by the U.S. Navy's Living Marine Resources program under Contract No. N39430-17-C-1982. The icons used in Figure 1 were made by macrovector from www.Freepik.com and Vectors market from www.flaticon.com.

AUTHORS' CONTRIBUTIONS

P.J.B. and D.L.M. conceived and designed the package. All authors contributed to developing R functions. P.J.B. led the writing of the manuscript, with critical inputs from all authors.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13469.

DATA AVAILABILITY STATEMENT

The package website (https://densitymodelling.github.io/dsmextra/) provides full documentation, including a comprehensive tutorial for novice users. Package files and code are hosted on Github at https://github.com/densitymodelling/dsmextra and archived on Zenodo at https://doi.org/10.5281/zenodo.3529465 (Bouchet et al., 2020). The dolphin dataset used in the case study is shipped with dsm and described at http://seamap.env.duke.edu/dataset/25.

ORCID

Phil J. Bouchet https://orcid.org/0000-0002-2144-2049
David L. Miller https://orcid.org/0000-0002-9640-6755
Laura Mannocci https://orcid.org/0000-0001-8147-8644
Catriona M. Harris https://orcid.org/0000-0001-9198-2414
Len Thomas https://orcid.org/0000-0002-7436-067X

REFERENCES

Becker, E. A., Forney, K. A., Redfern, J. V., Barlow, J., Jacox, M. G., Roberts, J. J., & Palacios, D. M. (2018). Predicting cetacean abundance and distribution in a changing climate. *Diversity & Distributions*, *43*, 459.

Bouchet, P. J., Miller, D. L., Roberts, J. J., Mannocci, L., Harris, C. M., & Thomas, L. (2019). From here and now to there and then: Practical recommendations for extrapolating cetacean density surface models to novel conditions (CREEM technical report No. 2019–01). University of St Andrews.

Bouchet, P. J., Miller, D. L., Roberts, J. J., Mannocci, L., Harris, C. M., & Thomas, L. (2020). Data from: dsmextra: Extrapolation assessment tools for density surface models. *Zenodo*, https://doi.org/10.5281/zenodo.3529465

Conn, P. B., Johnson, D. S., & Boveng, P. L. (2015). On extrapolating past the range of observed data when making statistical predictions in ecology. *PLoS ONE*, *10*(10), e0141416. https://doi.org/10.1371/journal.pone.0141416

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., … Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, *39*(12), 2119–2131. https://doi.org/10.1111/j.1365-2699.2011.02659.x

Grenier, P., Parent, A.-C., Huard, D., Anctil, F., & Chaumont, D. (2013). An assessment of six dissimilarity metrics for climate analogs. *Journal*

of Applied Meteorology and Climatology, *52*(4), 733–752. https://doi.org/10.1175/JAMC-D-12-0170.1

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge, UK: Cambridge University Press.

Hedley, S. L., & Buckland, S. T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, *9*(2), 181–199. https://doi.org/10.1198/1085711043578

Kaschner, K., Quick, N. J., Jewell, R., Williams, R., & Harris, C. M. (2012). Global coverage of cetacean line-transect surveys: Status quo, data gaps and future challenges. *PLoS ONE*, *7*(9), e44075. https://doi.org/10.1371/journal.pone.0044075

King, G., & Zeng, L. (2007). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly*, *51*(1), 183–210. https://doi.org/10.1111/j.1468-2478.2007.00445.x

Mahony, C. R., Cannon, A. J., Wang, T., & Aitken, S. N. (2017). A closer look at novel climates: New methods and insights at continental to landscape scales. *Global Change Biology*, *23*(9), 3934–3955. https://doi.org/10.1111/gcb.13645

Mannocci, L., Monestiez, P., Spitz, J., & Ridoux, V. (2015). Extrapolating cetacean densities beyond surveyed regions: Habitat-based predictions in the circumtropical belt. *Journal of Biogeography*, *42*(7), 1267–1280. https://doi.org/10.1111/jbi.12530

Mannocci, L., Roberts, J. J., Halpin, P. N., Authier, M., Boisseau, O., Bradai, M. N., … Vella, J. (2018). Assessing cetacean surveys throughout the Mediterranean Sea: A gap analysis in environmental space. *Scientific Reports*, *8*(1), 3126. https://doi.org/10.1038/s41598-018-19842-9

Mannocci, L., Roberts, J. J., Miller, D. L., & Halpin, P. N. (2017). Extrapolating cetacean densities to quantitatively assess human impacts on populations in the high seas. *Conservation Biology*, *31*(3), 601–614. https://doi.org/10.1111/cobi.12856

Mannocci, L., Roberts, J. J., Pedersen, E. J., & Halpin, P. N. (2020). Geographical differences in habitat relationships of cetaceans across an ocean basin. *Ecography*, *43*(8), 1250–1259. https://doi.org/10.1111/ecog.04979

Mesgaran, M. B., Cousens, R. D., & Webber, B. L. (2014). Here be dragons: A tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity & Distributions*, *20*(10), 1147–1159. https://doi.org/10.1111/ddi.12209

Miller, D. L., Burt, M. L., Rexstad, E. A., & Thomas, L. (2013). Spatial models for distance sampling data: Recent developments and future directions. *Methods in Ecology and Evolution*, *4*(11), 1001–1010. https://doi.org/10.1111/2041-210X.12105

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, *26*(3), 275–287. https://doi.org/10.1111/geb.12530

Redfern, J. V., Moore, T. J., Fiedler, P. C., de Vos, A., Brownell, R. L., Forney, K. A., … Ballance, L. T. (2017). Predicting cetacean distributions in data-poor marine ecosystems. *Diversity & Distributions*, *23*(4), 394–408. https://doi.org/10.1111/ddi.12537

Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E., & Lundquist, C. J. (2017). A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, *4*, art421. https://doi.org/10.3389/fmars.2017.00421

Rödder, D., & Engler, J. O. (2012). Disentangling interpolation and extrapolation uncertainties in species distribution models: A novel visualization technique for the spatial variation of predictor variable collinearity. *Biodiversity Informatics*, *8*(1), 30–40.

Sequeira, A. M. M., Bouchet, P. J., Yates, K. L., Mengersen, K., & Caley, M. J. (2018). Transferring biodiversity models for conservation: Opportunities and challenges. *Methods in Ecology and Evolution*, *9*(5), 1250–1264. https://doi.org/10.1111/2041-210X.12998

Shabani, F., Kumar, L., & Ahmadi, M. (2016). A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecology and Evolution*, *6*(16), 5973–5986. https://doi.org/10.1002/ece3.2332

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., … Sequeira, A. M. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Bouchet PJ, Miller DL, Roberts JJ, Mannocci L, Harris CM, Thomas L. dsmextra: Extrapolation assessment tools for density surface models. *Methods Ecol Evol*. 2020;11:1464–1469. https://doi.org/10.1111/2041-210X.13469