Ecography

EGOG-02881

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. - Ecography doi: 10.1111/ecog-02881.

Appendix 1

Supplementary tables

List of tables

A1.1 Overview of model validation methods in ecology. 2
A1.2 Model validation methods from the ecology literature 4
A1.3 Summary of results and claims for non-random cross-validation 8

Table A1.1. Overview of model validation methods and blocking options for each, where applicable.

Method	Definition	Blocking examples
Data splitting	Generally refers to performing a 2-fold cross-validation but in only one direction. Data is split into two groups: one for model training and one for model testing.	
k-fold	A preliminary partition of data are done into k samples (folds), where $k \geq 2$. k models are then built, each using all but one of the k samples. Each model is validated against the withheld sample (Verbyla & Litvaitis, 1989; Fielding & Bell, 1997; Arlot & Celisse, 2010)	Random spatial: folds are defined as spatially contiguous, but arranged randomly within the data coverage (Lieske & Bender, 2011) Stratified spatial: folds are defined along a geographic spatial gradient (e.g. elevation, latitude) (Vaughan & Ormerod, 2005) Temporal: folds are defined along a temporal gradient (e.g. different observation periods, different portions of a time series) (Vaughan & Ormerod, 2005) Environmental: folds are defined along a measured environmental gradient (e.g. similar temperature, similar soil condition) (Vaughan & Ormerod, 2005) Unit: folds are defined by individual experimental units or subjects (e.g. single radio-collared animal) (Koper & Manseau, 2009; Buston & Elith, 2011) Observer: folds are defined as those portions of data observed or recorded by different individuals. Deployment factor: folds are by any factor that may affect the performance of a model, such as different users, data
Leave-n-out	Every possible subset of n is left out of the sample and used for validation (Fielding & Bell, 1997)	
Leave-one-out	A special case of leave-n-out where each record is used individually for the validation. With n data points, the model is built with n-1 records and validated on the withheld n^{th} record (resulting in 0 or 100% validation accuracy). The process is repeated for each record and the average validation percentage is calculated from the n runs (Fielding & Bell, 1997)	Buffered-leave-one-out: a buffer of predetermined geographic distance around the withheld point is removed from the model training data (Bahn, 2009; Telford & Birks, 2009; Le Rest *et al.*, 2014). In temporally sequential observations, h observations preceding and following the observation in the test data are removed before validation (h-block) (Burman *et al.*, 1994)

Continued on next page

Table A1.1 – continued from previous page

Method	Definition	Blocking examples
Bootstrap resampling	Similar to "leave-one-out" but with replacement: the validation is still run n times for a dataset of n points, but individual points are permitted to be randomly selected more than once (Verbyla & Litvaitis, 1989; Arlot & Celisse, 2010)	
Bootstrapping	Similar to "bootstrap resampling" but where the validation samples (with replacement) can be larger than n=1 (Fielding & Bell, 1997; Verbyla & Litvaitis, 1989)	
Randomization	Equivalent to "bootstrapping" without replacement (Fielding & Bell, 1997)	

Table A1.2. Model validation methods from the ecology literature, used in presence-absence and presence-only (presence-available) applications. We included all presence-only / presence-available papers we found that described CV procedures, but we focused only on presence-absence papers that incorporated a non-random CV or a comparison of CV approaches. For presence-only papers, we tried to complete an exhaustive review. Because of the extremely large number of presence-absence papers, our search is not exhaustive, though it should be representative of the relative commonness of non-random CV methods. Papers only incorporating a random CV are not included in this table. Note that a single paper incorporating multiple methods is listed in multiple groups.

Method	Presence-Absence	Presence-Only
Goodness-of-fit		
Resubstitution	Bahn & McGill (2013) Barbosa *et al.* (2009) Edwards *et al.* (2006) Fielding & Haworth (1995) Graf *et al.* (2006) Hartley *et al.* (2006) Lieske & Bender (2011) Littlewood & Young (2008) Murray *et al.* (2011) Olden & Jackson (2000) Olden *et al.* (2002) Pearson *et al.* (2013) Roberts & Hamann (2012) Rodriguez-Rey *et al.* (2013) Seoane *et al.* (2005) Sundblad *et al.* (2009) Wenger & Olden (2012) Whittingham *et al.* (2003)	Merckx *et al.* (2011)
Prospective sampling	-	-
Exhaustive		
Leave-one-out	Anderson *et al.* (2008) Littlewood & Young (2008) Olden & Jackson (2000) Olden *et al.* (2002) Telford & Birks (2009)	Chow *et al.* (2005)
Leave-n-out	-	-

Continued on next page

Table A1.2 – continued from previous page

Method	Presence-absence	Presence-only
Random		
Data splitting	Araújo *et al.* (2005) Bahn & McGill (2013) Crimmins *et al.* (2013) Duncan *et al.* (2009) Ervin & Holly (2011) Fielding & Haworth (1995) Heikkinen *et al.* (2012) Heinnen & von Numers (2009) Heinänen *et al.* (2012) Olden & Jackson (2000) Peterson *et al.* (2007) Roberts & Hamann (2012) Steyerberg *et al.* (2001) Torres *et al.* (2015) Vanreusel *et al.* (2007)	Hijmans (2012) Veloz (2009) Zharikov *et al.* (2007)
k-fold	Chee & Elith (2012) Edwards *et al.* (2006) Graf *et al.* (2006) McAlpine *et al.* (2008) Newbold *et al.* (2015) Randin *et al.* (2006) Seoane *et al.* (2005) Steyerberg *et al.* (2001) Sundblad *et al.* (2009) Telford & Birks (2005) Wenger *et al.* (2013)	Boyce *et al.* (2002) Brakker *et al.* (2014) Fortin *et al.* (2009) Hirzel *et al.* (2006) Johnson *et al.* (2006) Latombe *et al.* (2014) Merckx *et al.* (2011) Olivier & Wotherspoon (2005) Radosavljevic & Anderson (2014) Wiens *et al.* (2008) Zielinski *et al.* (2010)
Bootstrapping	Steyerberg *et al.* (2001) Vernier *et al.* (2008)	-

Continued on next page

Table A1.2 – continued from previous page

Method	Presence-absence	Presence-only
Blocked		
Spatial	Anderson & Gonzalez (2011) Bahn & McGill (2013) Bulluck *et al.* (2006) Fielding (1994) Fielding & Haworth (1995) Fithian *et al.* (2014) Fløjgaard *et al.* (2009) Gavin & Hu (2006) Graf *et al.* (2006) Hartley *et al.* (2006) Heikkinen *et al.* (2012) Heinänen *et al.* (2012) Lieske & Bender (2011) McAlpine *et al.* (2008) Murray *et al.* (2011) Pearson *et al.* (2013) Peterson *et al.* (2007) Roberts & Hamann (2012) Seoane *et al.* (2005) Sundblad *et al.* (2009) Vanreusel *et al.* (2007) Wenger & Olden (2012) Wenger *et al.* (2013) Whittingham *et al.* (2007)	Irwin *et al.* (2012) Hijmans (2012) Radosavljevic & Anderson (2014) Wiens *et al.* (2008)
Temporal	Bulluck *et al.* (2006) Burman *et al.* (1994)	Anderson *et al.* (2005) Wiens *et al.* (2008)
Spatio-temporal	Harris (2015)	Johnson *et al.* (2000)
Environmental	Fløjgaard *et al.* (2009) Newbold *et al.* (2015) Roberts & Hamann (2012) Wenger & Olden (2012)	-
Unit	Buston & Elith (2011) Torres *et al.* (2015)	Aarts *et al.* (2008) Anderson *et al.* (2005) Koper & Manseau (2009) Long *et al.* (2009)
Observer	-	-
Deployment factor	-	-

Continued on next page

Table A1.2 – continued from previous page

Method	Presence-absence	Presence-only
Independent		
Spatial	Barbosa *et al.* (2009) Bozek & Rahel (1992) Chee & Elith (2012) Duncan *et al.* (2009) Freeman *et al.* (1997) Lawler & Edwards (2002) Littlewood & Young (2008) Schröder & Richter (2000) Telford & Birks (2005) Teresa & Casatti (2013) Thomas & Bovee (1993) Torres *et al.* (2015) Verbruggen *et al.* (2013) Vernier *et al.* (2008) Zharikov *et al.* (2007)	Olivier & Wotherspoon (2005) Skarin (2007)
Temporal	Araújo *et al.* (2005) Bozek & Rahel (1992) Crimmins *et al.* (2013) Dobrowski *et al.* (2011) Fouquet *et al.* (2010) Lawler & Edwards (2002) Martinez-Meyer *et al.* (2004) Newbold *et al.* (2015) Pearman *et al.* (2008) Rapacciuolo *et al.* (2012) Roberts & Hamann (2012) Rodriguez-Rey *et al.* (2013) Schröder & Richter (2000)	Garaffoa *et al.* (2010) Rodrìguez-Sánchez & Arroyo (2008) Waltari *et al.* (2007)
Spatiotemporal	Dennis & Eales (1999) Heinnen & von Numers (2009) Leftwich *et al.* (1997) Randin *et al.* (2006)	Coe *et al.* (2011)
Individual or group	-	Aarts *et al.* (2008) Polfus *et al.* (2014)
Observer or collection effort	-	Edwards *et al.* (2006)

Table A1.3. Summary of results and claims for non-random cross-validation (CV) from various ecological papers. We include 1) stated motivations for non-random CV, 2) comparisons between random and non-random validations or CVs, and 3) comparisons between non-random CVs and independent validations with new data. Papers are separated into those dealing with presence-absence data (typically in a species distribution modelling context), and those dealing with presence-only data (typically in a resource selection function context). In many cases, authors did not distinguish between a "non-random" or "block" cross-validation and an "independent" validation.

Presence-Absence	Presence-Only
1) Motivations for including a non-random validation or cross-validation:	
Model predictions improve when non-random CV is used in model building.	
Fløjgaard *et al.* (2009)	Aarts *et al.* (2008)
Heikkinen *et al.* (2012)	Wiens *et al.* (2008)
Non-random CV increases independence of validation data (reduces autocorrelation).	
Araújo *et al.* (2005)	Aarts *et al.* (2008)
Bahn & McGill (2013)	Hijmans (2012)
Fithian *et al.* (2014)	Merckx *et al.* (2011)
Rodriguez-Rey *et al.* (2013)	Radosavljevic & Anderson (2014)
Telford & Birks (2005)	Veloz (2009)
Non-random CV helps avoid overfitting.	
Barbosa *et al.* (2009)	Aarts *et al.* (2008)
Bahn & McGill (2013)	Anderson & Gonzalez (2011)
Chee & Elith (2012)	Radosavljevic & Anderson (2014)
Fløjgaard *et al.* (2009)	
Heikkinen *et al.* (2012)	
Heinänen *et al.* (2012)	
Merckx *et al.* (2011)	
Seoane *et al.* (2005)	
Sundblad *et al.* (2009)	
Wenger & Olden (2012)	
Non-random CV produces a better error estimate.	
Fløjgaard *et al.* (2009)	-
Hartley *et al.* (2006)	
Non-random CV produces unbiased parameter estimates.	
-	-

Continued on next page

Table A1.3 – continued from previous page

Presence-Absence	Presence-Only
Non-random CV can identify non-transferability or a general inability to extrapolate.	
Barbosa *et al.* (2009)	Anderson & Gonzalez (2011)
Duncan *et al.* (2009)	
Fielding & Haworth (1995)	
Graf *et al.* (2006)	
Heikkinen *et al.* (2012)	
Heinänen *et al.* (2012)	
Lieske & Bender (2011)	
McAlpine *et al.* (2008)	
Murray *et al.* (2011)	
Peterson *et al.* (2007)	
Randin *et al.* (2006)	
Rapacciuolo *et al.* (2012)	
Rodriguez-Rey *et al.* (2013)	
Schröder & Richter (2000)	
Sundblad *et al.* (2009)	
Torres *et al.* (2015)	
Vanreusel *et al.* (2007)	
Vernier *et al.* (2008)	
Wenger & Olden (2012)	
Wenger *et al.* (2013)	
Whittingham *et al.* (2003)	
2) Comparisons between random and non-random validations or CVs:	
Non-random CV demonstrated that random CV or validation is overly optimistic.	
Bahn & McGill (2013)	Koper & Manseau (2009)
Graf *et al.* (2006)	Hijmans (2012)
Heikkinen *et al.* (2012)	Radosavljevic & Anderson (2014)
Heinänen *et al.* (2012)	Veloz (2009)
Lieske & Bender (2011)	Wiens *et al.* (2008)
McAlpine *et al.* (2008)	
Merckx *et al.* (2011)	
Murray *et al.* (2011)	
Roberts & Hamann (2012)	
Seoane *et al.* (2005)	
Sundblad *et al.* (2009)	
Vanreusel *et al.* (2007)	

Continued on next page

Table A1.3 – continued from previous page

Presence-Absence	Presence-Only
Non-random CV demonstrated non-transferability or non-stationarity of models.	
Fielding & Haworth (1995)	Wiens *et al.* (2008)
Graf *et al.* (2006)	
Heikkinen *et al.* (2012)	
Lieske & Bender (2011)	
McAlpine *et al.* (2008)	
Murray *et al.* (2011)	
Peterson *et al.* (2007)	
Seoane *et al.* (2005)	
Wenger & Olden (2012)	
Vanreusel *et al.* (2007)	
Whittingham *et al.* (2003)	
Non-random CV provided evidence of overfit of at least some models.	
Bahn & McGill (2013)	Anderson & Gonzalez (2011)
Hartley *et al.* (2006)	Radosavljevic & Anderson (2014)
Heikkinen *et al.* (2012)	
Heinänen *et al.* (2012)	
Merckx *et al.* (2011)	
Wenger & Olden (2012)	
3) Validations with truly independent data:	
Independent validations demonstrated that random CV or validation is overly optimistic.	
Araújo *et al.* (2005)	-
Barbosa *et al.* (2009)	
Chee & Elith (2012)	
Duncan *et al.* (2009)	
Edwards *et al.* (2006)	
Ervin & Holly (2011)	
Littlewood & Young (2008)	
Randin *et al.* (2006)	
Roberts & Hamann (2012)	
Rodriguez-Rey *et al.* (2013)	
Telford & Birks (2005)	
Vernier *et al.* (2008)	
Independent validations showed that blocked CV can produce better error estimates.	
Roberts & Hamann (2012)	-

Continued on next page

Table A1.3 – continued from previous page

Presence-Absence	Presence-Only
Independent validations demonstrated non-transferability or non-stationarity in models.	
Barbosa *et al.* (2009)	-
Duncan *et al.* (2009)	
Ervin & Holly (2011)	
Leftwich *et al.* (1997)	
Randin *et al.* (2006)	
Rodriguez-Rey *et al.* (2013)	
Telford & Birks (2005)	
Teresa & Casatti (2013)	
Vernier *et al.* (2008)	
Independent validations provided evidence of overfit of at least some models.	
Araújo *et al.* (2005)	Zharikov *et al.* (2007)
Randin *et al.* (2006)	
Telford & Birks (2005)	

References

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M., & Matthiopoulos, J. 2008. Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, **31**(1), 140–160.

Anderson, D. P., Turner, M. G., Forester, J. D., Zhu, J., Boyce, M. S., Beyer, H., & Stowell, L. 2005. Scale-dependent summer resource selection by reintroduced elk in Wisconsin, USA. *Journal of Wildlife Management*, **69**(1), 298–310.

Anderson, D. P., Forester, J. D., & Turner, M. G. 2008. When to slow down: Elk residency rates on a heterogeneous landscape. *Journal of Mammalogy*, **89**(1), 105–114.

Anderson, R. P., & Gonzalez, I. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, **222**(15), 2796–2811.

Araújo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. 2005. Validation of species-climate impact models under climate change. *Global Change Biology*, **11**(9), 1504–1513.

Arlot, S., & Celisse, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.

Bahn, V. 2009. A new method for evaluating species distribution models. *In: 94th Ecological Society of America Annual Meeting.*

Bahn, V., & McGill, B. J. 2013. Testing the predictive performance of distribution models. *Oikos*, **122**(3), 321–331.

Barbosa, A. M., Real, R., & Vargas, J. M. 2009. Transferability of environmental favourability models in geographic space: The case of the Iberian desman (Galemys pyrenaicus) in Portugal and Spain. *Ecological Modelling*, **220**(5), 747–754.

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. 2002. Evaluating resource selection functions. *Ecological Modelling*, **157**(2-3), 281–300.

Bozek, M. A., & Rahel, F. J. 1992. Generality of Microhabitat Suitability Models for Young Colorado River Cutthroat Trout (Oncorhynchus-Clarki-Pleuriticus) across Sites and among Years in Wyoming Streams. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**(3), 552–564.

Brakker, S., Moretti, M., Boesch, R., Ghazoul, J., Obrist, M. K., & Bontadina, F. 2014. Assessing habitat connectivity for ground-dwelling animals in an urban environment. *Ecological Applications*, **24**, 1583–1595.

Bulluck, L., Fleishman, E., Betrus, C., & Blair, R. 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. *Global Ecology and Biogeography*, **15**(1), 27–38.

Burman, P., Chow, E., & Nolan, D. 1994. A Cross-Validatory Method for Dependent Data. *Biometrika*, **81**(2), 351–358.

Buston, P. M., & Elith, J. 2011. Determinants of reproductive success in dominant pairs of clownfish: a boosted regression tree analysis. *Journal of Animal Ecology*, **80**(3), 528–538.

Chee, Y. E., & Elith, J. 2012. Spatial data for modelling and management of freshwater ecosystems. *International Journal of Geographical Information Science*, **26**(11), 2123–2140.

Chow, T. E., Gaines, K. F., Hodgson, M. E., & Wilson, M. D. 2005. Habitat and exposure modelling for ecological risk assessment: A case study for the raccoon on the Savannah River Site. *Ecological Modelling*, **189**, 151–167.

Coe, P. K., Johnson, B. K., Wisdonm, M. J., Cook, J. G., Vavra, M., & MNielson, R. M. 2011. Validation of elk resource selection models with spatially independent data. *Journal of Wildlife Management*, **75**(1), 159–170.

Crimmins, S. M., Dobrowski, S. Z., & Mynsberge, A. R. 2013. Evaluating ensemble forecasts of plant species distributions under climate change. *Ecological Modelling*, **266**, 126–130.

Dennis, R. L. H., & Eales, H. T. 1999. Probability of site occupancy in the large heath butterfly Coenonympha tullia determined from geographical and ecological data. *Biological Conservation*, **87**(3), 295–301.

Dobrowski, S. Z., Thorne, J. H., Greenberg, J. A., Safford, H. D., Mynsberge, A. R., Crimmins, S. M., & Swanson, A. K. 2011. Modeling plant ranges over 75 years of climate change in California, USA: temporal transferability and species traits. *Ecological Monographs*, **81**(2), 241–257.

Duncan, R. P., Cassey, P., & Blackburn, T. M. 2009. Do climate envelope models transfer? A manipulative test using dung beetle introductions. *Proceedings of the Royal Society B-Biological Sciences*, **276**(1661), 1449–1457.

Edwards, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., & Moisen, G. G. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**(2), 132–141.

Ervin, G. N., & Holly, D. C. 2011. Examining Local Transferability of Predictive Species Distribution Models for Invasive Plants: An Example with Cogongrass (Imperata cylindrica). *Invasive Plant Science and Management*, **4**(4), 390–401.

Fielding, A. 1994. A review of methods used to investigate bird habitat associations. *Scottish Natural Heritage Review*, **No. 4**.

Fielding, A. H., & Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**(01), 38–49.

Fielding, A. H., & Haworth, P. F. 1995. Testing the generality of bird-habitat models. *Conservation Biology*, **9**(6), 1466–1481.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. 2014. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, In press.

Fløjgaard, C., Normand, S., Skov, F., & Svenning, J. C. 2009. Ice age distributions of European small mammals: insights from species distribution modelling. *Journal of Biogeography*, **36**(6), 1152–1163.

Fortin, D., Fortin, M. E., Beyer, H. L., Duchesne, T., Courant, S., & Dancose, K. 2009. Group-size-mediated habitat selection and group fusion-fission dynamics of bison under predation risk. *Ecology*, **90**(9), 2480–2490.

Fouquet, A., Ficetola, G. F., Haigh, A., & Gemmell, N. 2010. Using ecological niche modelling to infer past, present and future environmental suitability for Leiopelma hochstetteri, an endangered New Zealand native frog. *Biological Conservation*, **143**(6), 1375–1384.

Freeman, M. C., Bowen, Z. H., & Crance, J. H. 1997. Transferability of habitat suitability criteria for fishes in warmwater streams. *North American Journal of Fisheries Management*, **17**, 20–31.

Garaffoa, G. V., Dansa, S. L., Crespoa, E. A., Degratia, M., Giudicia, P., & Gagliardinia, D. A. 2010. Dusky dolphin: modeling habitat selection. *Journal of Mammalogy*, **91**(1), 54–65.

Gavin, D. G., & Hu, F. S. 2006. Spatial variation of climatic and non-climatic controls on species distribution: the range limit of Tsuga heterophylla. *Journal of Biogeography*, **33**(8), 1384–1396.

Graf, R. F., Bollmann, K., Sachot, S., Suter, W., & Bugmann, H. 2006. On the generality of habitat distribution models: a case study of capercaillie in three Swiss regions. *Ecography*, **29**(3), 319–328.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, **6**(4), 465–473.

Hartley, S., Harris, R., & Lester, P. J. 2006. Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology Letters*, **9**(9), 1068–1079.

Heikkinen, R. K., Marmion, M., & Luoto, M. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography*, **35**(3), 276–288.

Heinänen, S., Erola, J., & von Numers, M. 2012. High resolution species distribution models of two nesting water bird species: a study of transferability and predictive performance. *Landscape Ecology*, **27**(4), 545–555.

Heinnen, S., & von Numers, M. 2009. Modelling species distribution in complex environments: an evaluation of predictive ability and reliability in five shorebird species. *Diversity and Distributions*, **15**(2), 266–279.

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**(3), 679–688.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **119**, 142–152.

Irwin, L. L., Rock, D. F., & Rock, S. C. 2012. Habitat selection by northern spotted owls in mixed-coniferous forests. *Journal of Wildlife Management*, **76**(1), 200–213.

Johnson, B. K., Kern, J. W., Wisdon, M. J., Findholt, S. L., & Kie, J. G. 2000. Resource Selection and Spatial Separation of Mule Deer and Elk during Spring. *Journal of Wildlife Management*, **64**(3), 685–697.

Johnson, C. J., Nielsen, S. E., Merrill, E. H., Mcdonald, T. L., & Boyce, M. S. 2006. Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. *Journal of Wildlife Management*, **70**(2), 347–357.

Koper, N., & Manseau, M. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. *Journal of Applied Ecology*, **46**(3), 590–599.

Latombe, G., Parrott, L., Basille, M., & Fortin, D. 2014. Uniting Statistical and Individual-Based Approaches for Animal Movement Modelling. *Plos One*, **9**(6), e99938.

Lawler, J. J., & Edwards, T. C. 2002. Landscape patterns as habitat predictors: building and testing models for cavity-nesting birds in the Uinta Mountains of Utah, USA. *Landscape Ecology*, **17**(3), 233–245.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, **23**(7), 811–820.

Leftwich, K. N., Angermeier, P. L., & Dolloff, C. A. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. *Transactions of the American Fisheries Society*, **126**(5), 725–734.

Lieske, D. J., & Bender, D. J. 2011. A Robust Test of Spatial Predictive Models: Geographic Cross-Validation. *Journal of Environmental Informatics*, **17**(2), 91–101.

Littlewood, N. A., & Young, M. R. 2008. A habitat suitability model for the narrow-headed ant, Formica exsecta, evaluated against independent data. *Insect Conservation and Diversity*, **1**(2), 108–113.

Long, R. A., Muir, J. D., Rachlow, J. L., & Kie, J. G. 2009. A Comparison of Two Modeling Approaches for Evaluating Wildlife-Habitat Relationships. *Journal of Wildlife Management*, **73**(2), 294–302.

Martinez-Meyer, E., Townsend Peterson, A., & Hargrove, W. W. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography*, **13**(4), 305–314.

McAlpine, C. A., Rhodes, J. R., Bowen, M. E., Lunney, D., Callaghan, J. G., Mitchell, D. L., & Possingham, H. P. 2008. Can multiscale models of species' distribution be generalized from region to region? A case study of the koala. *Journal of Applied Ecology*, **45**(2), 558–567.

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., & Vanaverbeke, J. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, **222**(3), 588–597.

Murray, J. V., Choy, S. L., McAlpine, C. A., Possingham, H. P., & Goldizen, A. W. 2011. Evaluating model transferability for a threatened species to adjacent areas: Implications for rock-wallaby conservation. *Austral Ecology*, **36**(1), 76–89.

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Brger, L., Bennett, D. J., C., Argyrios, Collen, B., Day, J., De Palma, A., Daz, S., Echeverria-Londoo, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., Ingram, D. J., Itescu, Y., Kattge, J., Kemp, V., Kirkpatrick, L., Kleyer, M., Correia, D. L. P., Martin, C. D., Meiri, S., Novosolov, M., Pan, Y., Phillips, H. R. P., Purves, D. W., Robinson, A., Simpson, J., Tuck, S. L., Weiher, E., White, H. J., Ewers, R. M., Mace, G. M., Scharlemann, J. P. W., & Purvis, A. 2015. Global effects of land use on local terrestrial biodiversity. *Nature*, **520**, 45–50.

Olden, J. D., & Jackson, D. A. 2000. Torturing data for the sake of generality: How valid are our regression models? *Ecoscience*, **7**(4), 501–510.

Olden, J. D., Jackson, D. A., & Peres-Neto, P. R. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**(2), 329–336.

Olivier, F., & Wotherspoon, S. J. 2005. GIS-based application of resource selection functions to the prediction of snow petrel distribution and abundance in East Antarctica: Comparing models at multiple scales. *Ecological Modelling*, **189**, 105–129.

Pearman, P. B., Randin, C. F., Broennimann, O., Vittoz, P., van der Knaap, W. O., Engler, R., Le Lay, G., Zimmermann, N. E., & Guisan, A. 2008. Prediction of plant species distributions across six millennia. *Ecology Letters*, **11**(4), 357–369.

Pearson, R. G., Phillips, S. J., Loranty, M. M., Beck, P. S. A., Damoulas, T., Knight, S. J., & Goetz, S. J. 2013. Shifts in Arctic vegetation and associated feedbacks under climate change. *Nature Climate Change*, **3**(7), 673–677.

Peterson, A. T., Papes, M., & Eaton, M. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30**(4), 550–560.

Polfus, J. L., Heinemeyer, K., Hebblewhite, M., & Nation, The Taku River Tlingit First. 2014. Comparing Traditional Ecological Knowledge and Western Science Woodland Caribou Habitat Models. *Journal of Wildlife Management*, **78**(1), 112–121.

Radosavljevic, A., & Anderson, R. P. 2014. Making better MAXENT models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, **41**(4), 629–643.

Randin, C. F., Dirnbock, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. 2006. Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**(10), 1689–1703.

Rapacciuolo, G., Roy, D. B., Gillings, S., Fox, R., Walker, K., & Purvis, A. 2012. Climatic Associations of British Species Distributions Show Good Transferability in Time but Low Predictive Accuracy for Range Change. *Plos One*, **7**(7).

Roberts, D. R., & Hamann, A. 2012. Method selection for species distribution modelling: are temporally or spatially independent evaluations necessary? *Ecography*, **35**(9), 792–802.

Rodriguez-Rey, M., Jimenez-Valverde, A., & Acevedo, P. 2013. Species distribution models predict range expansion better than chance but not better than a simple dispersal model. *Ecological Modelling*, **256**, 1–5.

Rodrìguez-Sánchez, F., & Arroyo, J. 2008. Reconstructing the demise of Tethyan plants: climate-driven range dynamics of Laurus since the Pliocene. *Global Ecology and Biogeography*, **17**(6), 685–695.

Schröder, B., & Richter, O. 2000. Are habitat models transferable in space and time? *Journal for Nature Conservation*, **8**, 195–205.

Seoane, J., Bustamante, J., & Diaz-Delgado, R. 2005. Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**(2), 512–522.

Skarin, A. 2007. Habitat use by semi-domesticated reindeer, estimated with pellet-group counts. *Rangifer*, **27**(2), 121–132.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. 2001. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, **54**(8), 774–781.

Sundblad, G., Harma, M., Lappalainen, A., Urho, L., & Bergstrom, U. 2009. Transferability of predictive fish distribution models in two coastal systems. *Estuarine Coastal and Shelf Science*, **83**(1), 90–96.

Telford, R. J., & Birks, H. J. B. 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews*, **24**(20-21), 2173–2179.

Telford, R. J., & Birks, H. J. B. 2009. Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, **28**(13-14), 1309–1316.

Teresa, F. B., & Casatti, L. 2013. Development of habitat suitability criteria for Neotropical stream fishes and an assessment of their transferability to streams with different conservation status. *Neotropical Ichthyology*, **11**(2), 395–402.

Thomas, J. A., & Bovee, K. D. 1993. Application and Testing of a Procedure to Evaluate Transferability of Habitat Suitability Criteria. *Regulated Rivers-Research & Management*, **8**(3), 285–294.

Torres, L. G., Sutton, P. J. H., Thompson, D. R., Delord, K., Weimerskirch, H., Sagar, P. M., Sommer, E., Dilley, B. J., Ryan, P. G., & Phillips, R. A. 2015. Poor Transferability of Species Distribution Models for a Pelagic Predator, the Grey Petrel, Indicates Contrasting Habitat Preferences across Ocean Basins. *Plos One*, **10**(3).

Vanreusel, W., Maes, D., & Van Dyck, H. 2007. Transferability of species distribution models: a functional habitat approach for two regionally threatened butterflies. *Conservation Biology*, **21**(1), 201–212.

Vaughan, I. P., & Ormerod, S. J. 2005. The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**(4), 720–730.

Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, **36**(12), 2290–2299.

Verbruggen, H., Tyberghein, L., Belton, G. S., Mineur, F., Jueterbock, A., Hoarau, G., Gurgel, C. F. D., & De Clerck, O. 2013. Improving transferability of introduced species distribution models: new tools to forecast the spread of a highly invasive seaweed. *PLoS One*, **8**(6), e68337.

Verbyla, D. L., & Litvaitis, J. A. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, **13**(6), 783–787.

Vernier, P. R., Schmiegelow, F. K. A., Hannon, S., & Cumming, S. G. 2008. Generalizability of songbird habitat models in boreal mixedwood forests of Alberta. *Ecological Modelling*, **211**(1-2), 191–201.

Waltari, E., Hijmans, R. J., Peterson, A. T., Nyari, A. S., Perkins, S. L., & Guralnick, R. P. 2007. Locating Pleistocene refugia: Comparing phylogeographic and ecological niche model predictions. *PLoS One*, **2**(7), –.

Wenger, S. J., & Olden, J. D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**(2), 260–267.

Wenger, S. J., Som, N. A., Dauwalter, D. C., Isaak, D. J., Neville, H. M., Luce, C. H., Dunham, J. B., Young, M. K., Fausch, K. D., & Rieman, B. E. 2013. Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biology*, **19**(11), 3343–3354.

Whittingham, M. J., Wilson, J. D., & Donald, P. F. 2003. Do habitat association models have any generality? Predicting skylark Alauda arvensis abundance in different regions of southern England. *Ecography*, **26**(4), 521–531.

Whittingham, M. J., Krebs, J. R., Swetnam, R. D., Vickery, J. A., Wilson, J. D., & Freckleton, R. P. 2007. Should conservation strategies consider spatial generality? Farmland birds show regional not national patterns of habitat association. *Ecology Letters*, **10**(1), 25–35.

Wiens, T. S., Dale, B. C., Boyce, M. S., & Kershaw, G. P. 2008. Three way k-fold cross-validation of resource selection functions. *Ecological Modelling*, **212**(3-4), 244–255.

Zharikov, Y., Lank, D. B., & Cooke, F. 2007. Influence of landscape pattern on breeding distribution and success in a threatened Alcid, the marbled murrelet: model transferability and management implications. *Journal of Applied Ecology*, **44**(4), 748–759.

Zielinski, W. J., Hunter, J. E., Hamlin, R., Slauson, K. M., & Mazurek, M. J. 2010. Habitat Characteristics at Den Sites of the Point Arena Mountain Beaver (Aplodontia rufa nigra). *Northwest Science*, **84**(2), 119–130.

Appendix 2

Spatial blocking (Box 1)

Complete R scripts for this simulation are provided in Supplementary materials Appendix 6.

Introduction

Evaluation methods need to address dependence between training and test data to supply realistic error estimates. However, such strategies may come at a cost. Here, using a simulation model, we specifically test the performance of block cross-validation and leave-one-out cross-validation with spatial buffers in the presence of spatial autocorrelation. Since geographic and environmental space are closely linked via distribution and spatial autocorrelation, breaking the dependence between training and test data due to spatial autocorrelation may have the side effect of forcing environmental extrapolation when predicting to test locations for evaluation. The latter may lead to an overestimation of model error, while residual dependence between training and test data may lead to an underestimation of error. When the two balance each other we might find the right error for the wrong reason.

Here, we investigate the effects of residual dependence between training and test data on error estimates. By using a simulation model, we can determine the expected error estimate from a reliable evaluation. We also consider the effect of environmental extrapolations introduced in cross-validation procedures.

Methods

All analysis was performed within the R framework for statistical computing (R Core Team, 2015).

Simulated data

We simulated 100 landscapes on a 50×50 grid creating 13 "environmental" variables. Eight of the variables were created directly from random Gaussian fields to introduce spatial structure, using the *randomFields* package for R (Schlather *et al.*, 2015), while five were derivatives of the direct variables creating complicated interactions (for specifics see Table A2.1, for examples see Figure A2.1). Species abundance depended on hypothetical environmental data in complex and indirect ways including interactions, non-linear combinations, limiting effects, and exclusion by hypothetical disease (Table A2.1). Four of the eight direct variables (x.7 - x.10) were not used in the creation of the virtual species, but are included in the model to allow for overfitting.

Species Distribution Models

We modelled species abundances using the *randomForest* package for R (Breiman, 2001). Only three of the four direct variables and none of the derived, indirect variables used in the virtual species creation were included in the Random Forest to challenge the models. Furthermore, we supplied four additional variables that were not used for simulating species to allow for overfitting. Variables included in the model were x.2, x.3, and x.6 to x.10 (see Table A2.1).

Methods of error estimates

We calculated root mean squared errors (RMSE) for comparing several evaluation methods. In the most basic version training and test data are identical (resubstitution). This would, for example, be the basis for an R^2 in a regular regression. A more advanced

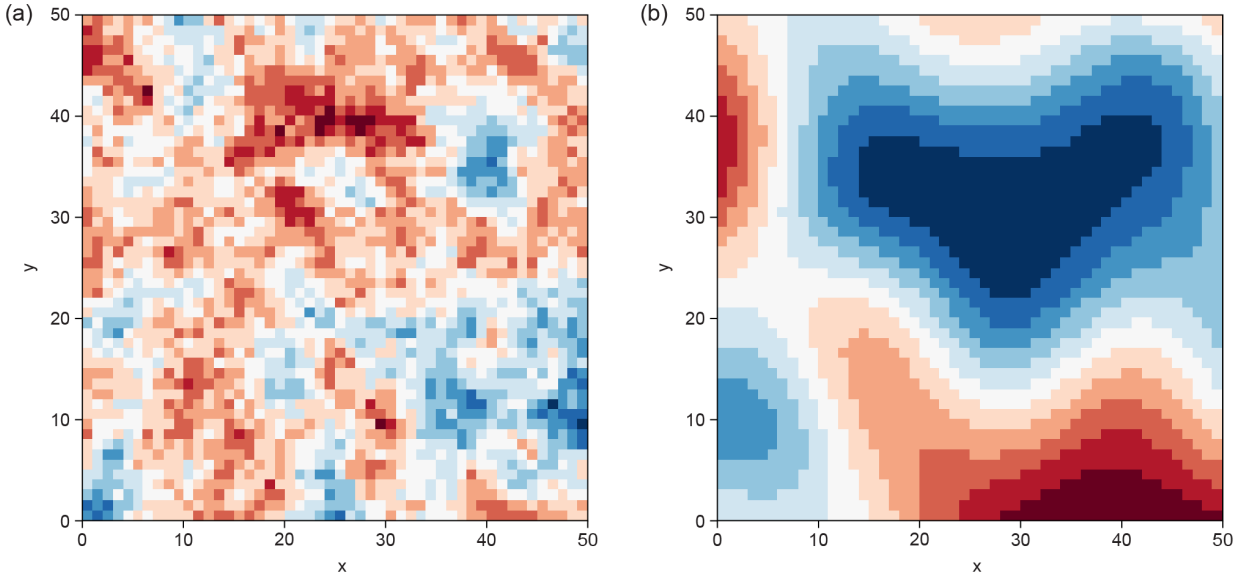

Figure A2.1. Example of simulated environmental variables created with a random Gaussian field with **(a)** an exponential covariance model, variance = 0.1, scale = 0.1, and **(b)** with a Gaussian covariance model, variance = 0.1, scale = 0.3.

Table A2.1. Details on the simulation of environmental variables and a species that depends on them. Note that some variables are not used in the simulation of the species, and others are only used indirectly.

Variable	Generation	Interpretation
x.1	Random Gaussian field with exponential covariance model (variance = 0.1, scale = 0.1)	A standard environmental variable with a medium-ranged spatial autocorrelation and normal variance such as soil or topography
x.2	Random Gaussian field, with exponential covariance model (variance = 0.3, scale = 0.1)	An environmental variable such as precipitation with medium-ranged spatial autocorrelation and high variance
x.3	Random Gaussian field, with Gaussian covariance model (variance = 0.1, scale = 0.3)	An environmental variable such as temperature with long-ranged spatial autocorrelation and normal variance
x.4	Binary (0 or 1) variable based on the ratio x.2/x.3. Values above a specified quantile (here 0.95) of this ratio led to a zero in x.4.	A disease that excludes the species at the "warmest" and "wettest" locations
x.5	x.1 + x.2 + x.3 + x.2 × x.3	Linear combination and interaction of previous variables here symbolizing "food"
x.6 - x.8	Random Gaussian field with exponential covariance model (variance = 0.1, scale = 0.1)	More standard environmental variables with a medium-ranged spatial autocorrelation and normal variance
x.9 - x.10	Random Gaussian field, with Gaussian covariance model (variance = 0.1, scale = 0.3)	More long-ranged spatial autocorrelation and normal variance environmental variables such as temperature
x.11	x.2/x.3	"Water availability"
x.12	$1/(\text{sqrt}(2\times\pi))\times\text{exp-}(x.3^2/4)$	Gaussian unimodal dependence on "temperature"
x.13	$1/(\text{sqrt}(2\times\pi))\times\text{exp-}(x.2^2/4)$	Gaussian unimodal dependence on "precipitation"
y	x.1 + x.5 + x.6 + x.12 + x.13; and x.4 is excluding (i.e. if x.4 = 0, y = min(y)); and x.11 is limiting (i.e. when y >x.11, y is set to x.11)	Dependent variable: "species abundance"
$\hat{y}$	x.2 + x.3 + x.6 + x.7 + x.8 + x.9 + x.10	Model formulation [a]

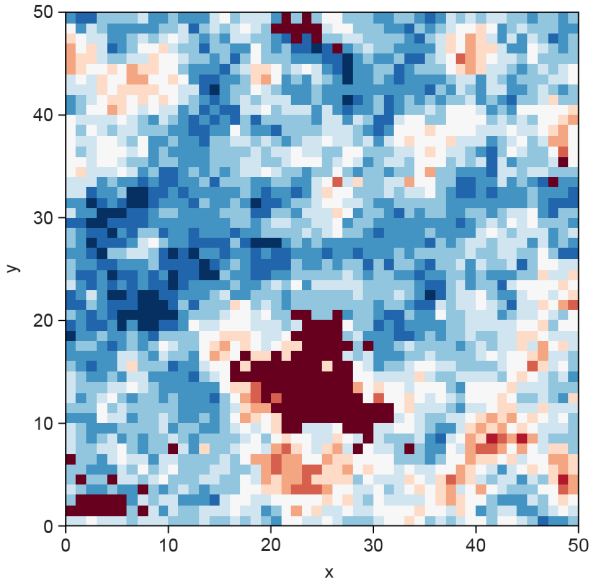

Figure A2.2. Example of a virtual species. See Table A2.1 for details. Note the dark areas where simulated disease has excluded the species.

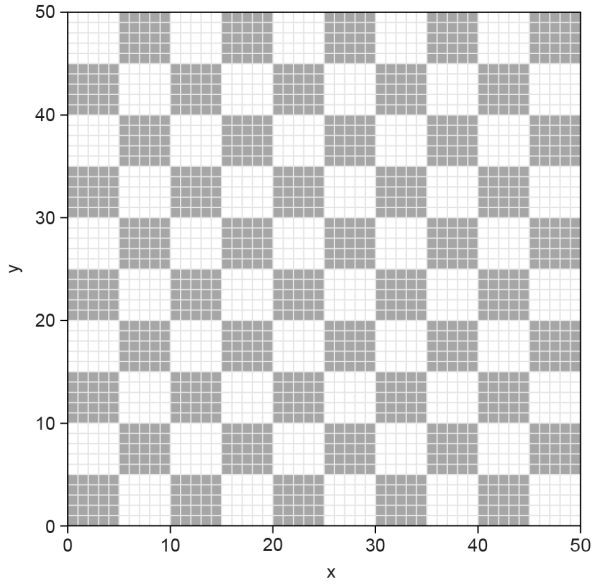

Figure A2.3. Setup for block cross validation with block size of five cells along each edge. Grey cells ini-tially supply training data and white cells test data and then roles are reversed for a two-fold cross validation.

method is randomly splitting the cells into training and test data and conducting a two-fold cross validation, which means that one half is used for training and the other for testing and then the roles are switched and the final error is calculated as the average of the two errors. This method separates training and test data and therefore fully addresses overfitting if it provides for independence between training and test data. However, here spatial autocorrelation will prevent them to be fully independent.

An alternative to splitting the cells randomly is using a systematic spatial pattern to increase the average distance between training and test cells and thereby reducing dependence (block cross validation). We used regular, square blocks to separate training and test data spatially, investigating a range of block sizes. Within a single evaluation the blocks were of uniform size and distributed in a checkerboard fashion between training and test sites (Figure A2.3). After one evaluation with half of the blocks being training data and the other half being test data, we reversed the roles and averaged the errors for a two-fold cross validation.

For assuring a given minimum distance between training and test data, we did a leave-one-out cross-validation with spatial buffers. In this approach only one cell is held out for evaluation per model run, which we buffered by omitting neighbouring cells up to a certain distance from training data (Figure A2.4).

To calculate the RMSE, we used 100 separate runs with different individual test cells, which were drawn from a regular pattern (Figure A2.4). Finally, we determined our "ideal" error rate by projecting upon new landscapes that were produced in an identical way. We produced a model on each of the 100 landscapes and projected to the remaining 99 landscapes, averaging the errors for a single error estimate per landscape.

Analogue vs. non-analogue conditions

To find out whether environmental extrapolation had occurred, we tested whether test locations were within conditions covered by training locations. The values of x.2, x.3, and x.6 (see Table A2.1) at a test location had to be within the ranges of these vari-

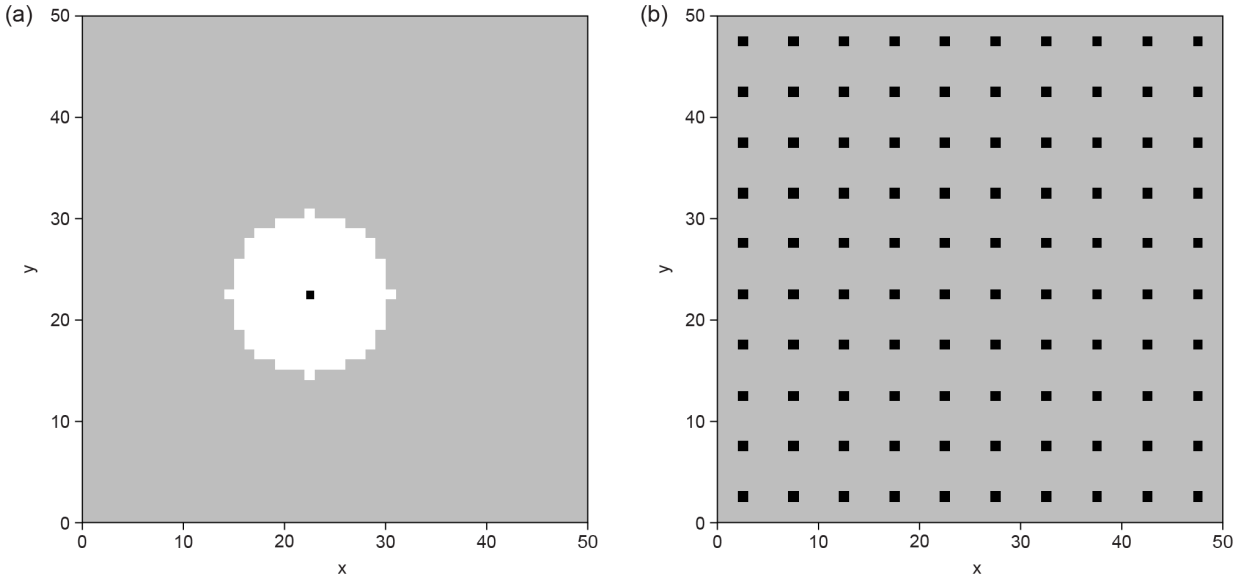

Figure A2.4. The setup for spatially independent leave-one-out cross validation with a buffer size of eight cells, showing **(a)** an individual validation with grey cells supplying the training data, the black cell as the test datum surrounded by the white buffer of radius 8. As the leave-one-out is extremely computationally intensive, it was not run exhaustively (i.e. not all 2,500 cells were withheld in turn). Rather 100 equally spaced test cells were used for each leave-one-out evaluation, as shown in **(b)**.

ables at training locations for the test location to be deemed "analogous". We did not check variables x.7 to x.10 because they were not used in simulating the species nor did we check x.1, x.4, or x.5 because they were not available to the models. We determined error estimates using the above methods based on all test locations (non-analogue) and on analogue test locations only.

Results and discussion

While blocking increased the average distance between training and test locations, it was not able to move all test cells out of autocorrelation range of training cells (determined to be 10 on average in our residuals; Figure A2.5). However, the distance between training and test cells was increased, particularly for the larger block sizes (up to half of the whole grid). The price to address the autocorrelation between training and test data was an increase in environmental extrapolation (Figure A2.6). However, we were able to address the problem of envi-

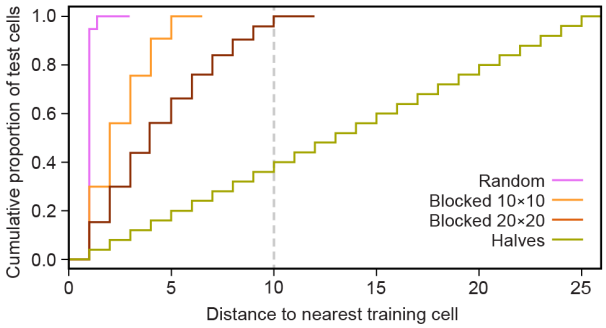

Figure A2.5. Cumulative proportion of test sites that are within a given minimum distance to a training site. In a random hold-out almost all test sites have a neighbouring training site, well within the range of spatial autocorrelation (distance = 10; shown as a dashed grey line), while a split of the area in half leads to approximately 40% of the test sites having a training site within the spatial autocorrelation range of 10.

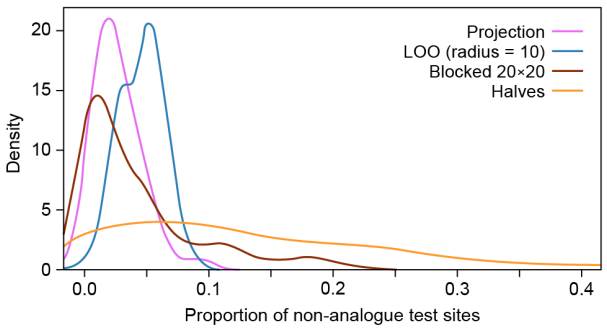

Figure A2.6. Distribution of the proportion of non-analogue test sites (extrapolating in environmental space) across the 100 simulated landscapes. Splitting the range in halves for training and testing data led to highly variable but sometimes substantial extrapolation.

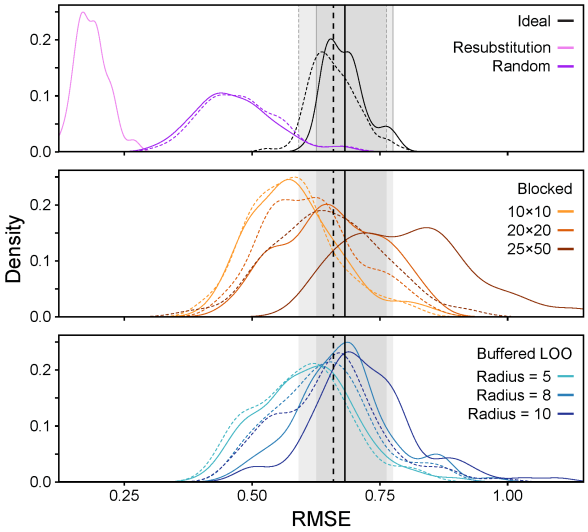

Figure A2.7. Model error estimate distributions and means over 100 landscape simulations. When training and test sites were identical (Resubstitution) or when they were split randomly (Random) only one distribution is shown in black. For the "ideal" error of projecting to a new landscape (Ideal), and the blocked cross validation schemes (10×10 blocks, 20×20 blocks, or 25×50 blocks for the area split in half), results for all test sites are shown as solid lines while evaluation results using only test sites found to be analogous are shown as dashed lines.

ronmental extrapolation by excluding non-analogue test sites, leading to good error estimates for the largest block size (Figure A2.7, middle). Error rates are substantially and systematically underestimated by resubstitution (training data = test data) indicating strong overfitting of the models (Figure A2.7, top). Random two-fold cross validation also underestimates errors but less so than resubstitution, suggesting that some overfitting was non-spatial.

The remainder of the error underestimation by the random cross-validation can reasonably be attributed to the proximity and resulting correlation between training and test sites caused by spatial autocorrelation (Figure A2.5). This problem is further alleviated by blocking in space for the cross-validation. While small blocks (size 10×10) lead to virtually no environmental extrapolation (dashed and solid lines are almost identical in Figure A2.7, middle, 10×10), they still consistently underestimate error, because they still only weakly spatially separate training and test data (Figure A2.5).

In contrast, splitting the range in half for the largest block size overestimates the error due to substantial environmental extrapolation when all the test sites are included, though omitting the non-analogue test sites provides an error estimate close to the ideal error (Figures A2.7, middle, solid line, 25×50). Finally, the block size 20×20 initially yields the most reliable error estimate (Figure A2.7, middle, solid line, 20×20). However, it arrives at this error estimate for the wrong reason, balancing an error underestimation due to correlation between training and test data with an error overestimation due to environmental extrapolation. This is revealed when considering only the analogue sites: a 20×20 block still leads to a noticeable underestimation of the model error (Figure A2.7, middle, dashed line, 20×20).

The buffered leave-one-out provides error estimates close to the desirable error when the buffer is selected to cover the range of spatial autocorrelation, which was about 10 cells, and when non-analogue test sites were omitted (Figure A2.7, bottom, dashed line, radius = 10). A smaller buffer led to slight underestimation of the error (Figure A2.7, bottom, dashed line, radius = 5).

Conclusions

When model overfitting and spatial autocorrelation are present, block cross-validation provides an estimate of model error closer to the true error than either resubstitution or random hold out. Further, the size of the blocks may need to be substantially larger than the measured range of spatial autocorrelation to provide reliable error estimates. The buffered leave-one-out cross-validation consistently provides error estimates close to the ideal when the buffer size is adjusted to spatial autocorrelation range. However, this method comes at an extremely high computational cost.

References

Breiman, L. 2001. Random forests. *Machine Learning*, **45**(1), 5–32.

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., & Strokorb, K. 2015. Analysis, Simulation and Prediction of Multivariate Random Fields with Package Random Fields. *Journal of Statistical Software*, **63**(8), 1–25.

Appendix 3

Blocking by individual or group (Box 2)

Complete R scripts and data for this case study are provided in Supplementary material Appendix 6.

Introduction

We evaluate resource selection functions (RSFs) for elk (*Cervus elaphus*) in the Canadian Rocky Mountains using k-fold cross validation with several different blocking approaches.

Methods

All analysis was performed within the R framework for statistical computing (R Core Team, 2015). Spatial data visualization was done in ESRI ArcMap 10.3 (ESRI, 2011).

Study area

The study took place in southwest Alberta (AB, Canada) and extended into northwest Montana (USA) and southeast British Columbia (BC, Canada; Figure A3.1). The core of the study area is a montane ecosystem along the eastern slopes of the Canadian Rocky Mountains, along the AB-BC border. The majority of our study area within Alberta is comprised of provincial forest reserve with mixed livestock ranching with cropland on its eastern boundary. This boundary is a transition zone from prairie grassland into Rocky Mountain montaine, and several elk subpopulations are present here. Natural predators of elk are wolf (*Canis lupus*), cougar (*Puma concolor*), and grizzly bear (*Ursus arctos;* Muhly *et al.*, 2011). There is also extensive human presence in the area (Ciuti *et al.*, 2012b), including industrial activities such as forestry and natural gas extraction, as well as recreational activities, especially during summer and the autumn hunting season (Ciuti *et al.*, 2012b).

Elk satellite telemetry data

Data used in this study have been collected by one of the largest satellite telemetry studies of a large herbivore and its main predators (http://montaneelk.com/). For more study details, see Muhly *et al.* (2011); Ciuti *et al.* (2012a,b).

In the original elk study, 98 adult female elk between 2 and 19 years of age were captured during winters from 2007–2012 (capture site locations: Figure A3.1, left panel) using helicopter net-gunning (animal care protocol no. 536-1003 AR University of Alberta). Elk were fitted with Lotek GPS-4400 radiotelemetry collars (Lotek wireless Inc., Ontario, Canada; Figure A3.2). All collars were programmed with a two-hour relocation schedule and data were remotely downloaded in the field.

For this case study, we limit the data to a subsample of 43 female elk monitored during summer 2008, a period in which the largest number of individuals was tracked simultaneously (27,148 GPS relocations). Summer ranges were defined as all telemetry locations between July 1st through August 31st, 2008 (100% minimum convex polygon summer home ranges shown in Figure A3.1, right panel).

Environmental predictors

We selected four environmental predictors known to be key drivers of elk resource selection in the region (Ciuti *et al.*, 2012a; Killeen *et al.*, 2014): Terrain Ruggedness Index (TRI), Normalized Differenced

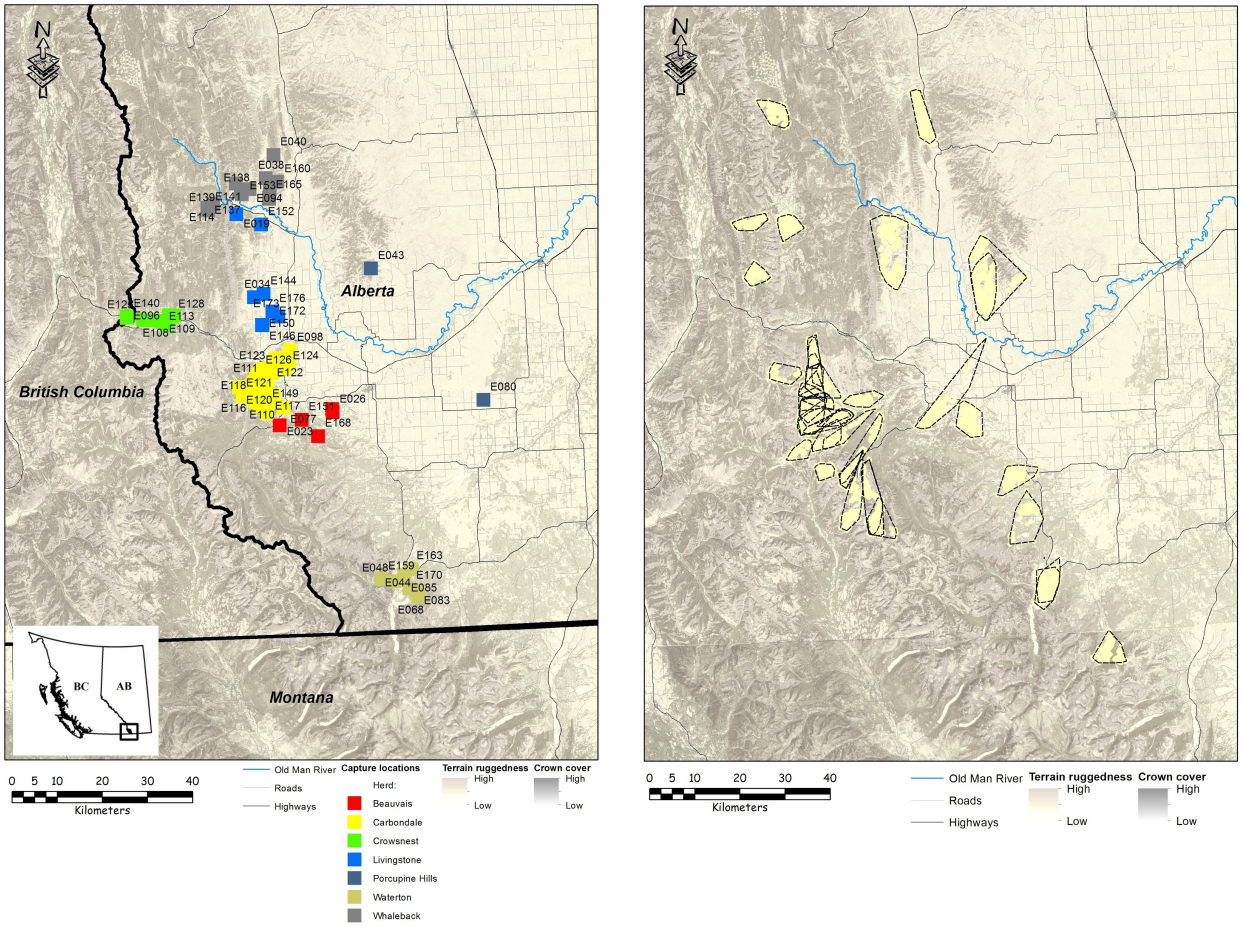

Figure A3.1. Location of sites where female elk were captured from 2007 to 2012 by the montane elk research program (http://montaneelk.com/) on the eastern slope of the Rocky Mountains, Alberta, Canada (left panel). Elk capture sites are coloured-coded by herd (n = 7) using the traditional name assigned by local managers based on overwintering areas. Summer 2008 satellite telemetry relocations (data used in this case study) and the corresponding individual home ranges (minimum convex polygons 100%, n = 43 females) are shown in the right panel.

Figure A3.2. Female elk E053 (on left) taken by a camera trap on July 2008 in southwest Alberta, Canada.

Vegetation Index ($NDVI$), the distance to the closest road (d_{road}), and 30-metre resolution land cover map (lc30). Terrain ruggedness is an effective proxy for elevation (Pearson correlation coefficient $r_p >$ 0.6), whereas 30-metre resolution land cover map is an effective proxy for canopy cover ($r_p <$ -0.9). The RSF was assumed to take the exponential form (Lele *et al.*, 2013, see below for more details), which assumes normally distributed predictors. We thus transformed continuous predictors to approximate a normal distribution (log-transformation for TRI, square-root transformation for d_{road}, and square-transformation for $NDVI$). Land cover was reclassified into four main habitat types: coniferous forest, deciduous forest, mixed forest, and open areas. A final data screening revealed that the four predictors were not collinear ($r_p \ll 0.7$) and could be included in the same model structure.

Random availability and sensitivity analysis

We followed the RSF 'sampling protocol A' (Manly *et al.*, 2002) in a use-available design (a.k.a. presence-available) with sampled used (individual elk telemetry relocations collected during summer) and available resource units (random points drawn within individual summer home ranges). Individual summer home ranges were computed using the 100% minimum convex polygon (MCP) estimator (Figure A3.1, right panel).

We ran a sensitivity analysis to estimate the minimum sample size of available random points needed to achieve stable parameter estimates in our generalized linear mixed models (GLMMs; see below for full details on model class and structure). We created scenarios with varying ratios of used:available points as follows: 1:0.05, 1:0.1, 1:0.5, 1:1, 1:2, ... 1:15. We fitted 100 models for each scenario and investigated the trend of parameter estimates depending on the number of random available points per used point. We stopped the simulation when used:available ratios resulted in stable parameter estimates. Our sensitivity analysis suggested that parameter estimate tends to be stable with at least 6-7 random locations per used point (see Figure A3.3 for an example output). Based on the time required by each GLMM to run, and considering the number of models expected to be fit depending on the block-cross validation schemes, we decided we could afford to sample 10 random available points per used location. We arranged our use-available dataset accordingly.

RSF model class and structure

We fitted a generalized mixed linear model (GLMM) with presence (1's) availability (0's) as a binary response variable and the four environmental variables (TRI, d_{road}, $NDVI$, and lc30) as predictors. We used the *glmer* function from the *lme4* package for R (Bates *et al.*, 2015) with binomial distribution of errors and logit link function. Predictors were scaled as $x–\bar{x}/\sigma_x$ prior to fitting the model. We included a quadratic term for TRI, d_{road}, and $NDVI$ to allow for non linear effects. Because elk resource selection is expected to vary depending on the distance to roads, i.e., a proxy for human presence and disturbance, we also included two interaction terms ($TRI \times d_{road}$, $NDVI \times d_{road}$). Elk identity was fitted as a random intercept with the underlying assumption that random effects are normally distributed (Figure A3.4, Figure A3.5). Beta coefficients estimated by the GLMM were plugged in the resource selection function to obtain RSF scores, which are proportional to the probability of selection. We assumed

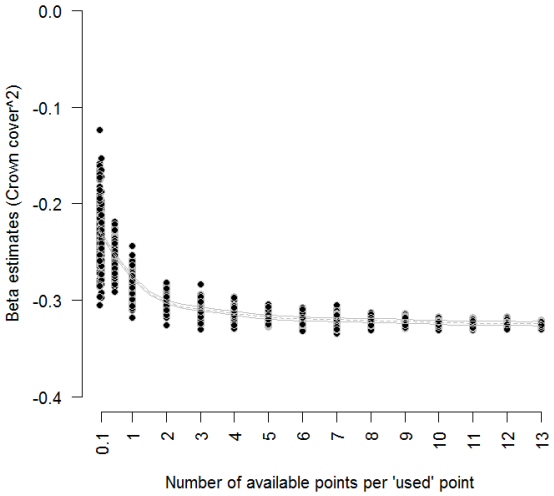

Figure A3.3. Output example of the sensitivity anal-ysis performed to estimate the minimum number of random 'available' points (background availability in presence-availability design) needed to get stable param-eter estimates in generalized linear mixed effect models (GLMMs). The example depicted here refers to the variation of GLMM beta estimates for the quadratic term of crown cover (30-metre resolution, collinear with land cover map 30-metre resolution) depending on the number of available points per used point. Parameters for each use:available ratio have been estimated 100 times, each run with a new selection of random available points. A spline from a generalized additive model has been added to improve readability of model parameter stabilization.

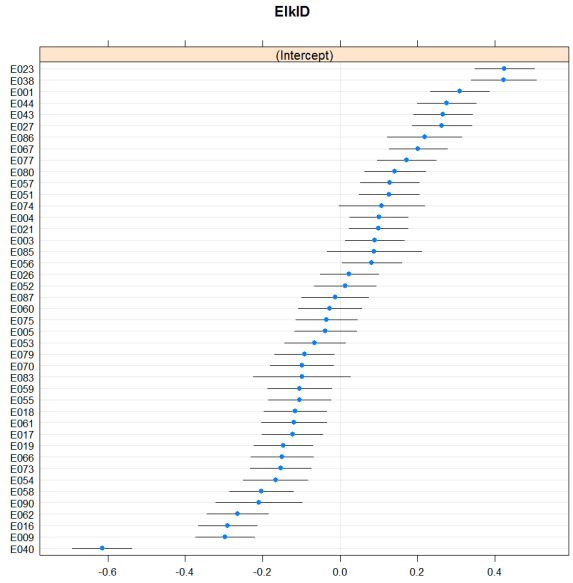

Figure A3.4. Variation of random intercepts in the GLMM fitted to estimate beta coefficients needed to build the elk resource selection function RSF. Individual elk (ElkID, n = 43) were fitted as random intercept (1|ElkID) in the GLMM framework and were correctly assumed to be normally distributed (also see Figure A3.5).

the RSF to take the form:

$$w(x) = exp(\beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n)$$

where β_1 to β_n are coefficients estimated by the GLMM, which are associated with a vector x of environmental variables x_1 to x_n, respectively (Manly *et al.*, 2002; Lele *et al.*, 2013).

RSF evaluation

RSF models estimated from presence-available data create unique problems for evaluating model predictions because presence-available data are not truly binary (presence-absence) data (Boyce *et al.*, 2002; Boyce, 2010). RSFs were thus validated using the method introduced by Boyce *et al.* (2002) developed for presence-available designs. This involves calculating the correlation between RSF ranks and area-adjusted frequencies for a withheld sub-sample of data, e.g. 1/5 of the data in a 5-fold cross-validation scheme. We investigated the pattern of predicted RSF scores for partitioned testing data (presence-only) against categories of RSF scores (10 bins). A Spearman rank correlation between area-adjusted frequency of cross-validation points within individual bins and the bin rank was calculated for each cross-validated model. A model with good predictive performance would be expected to be one with a strong positive correlation, as more use locations (area-adjusted) would progressively fall into higher RSF bins (Boyce *et al.*, 2002; Hirzel *et al.*, 2006; Wiens *et al.*, 2008).

Residual spatial autocorrelation

We computed the variogram depicting the residual spatial autocorrelation in the full model (GLMM) using the *variogram* function from the *gstat* package

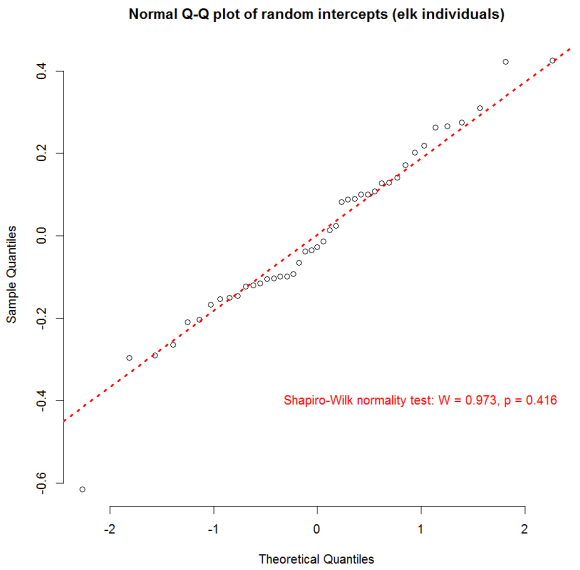

Figure A3.5. Q-Q plot for the random intercepts of the GLMM fitted to estimate beta coefficients needed to build the elk resource selection function RSF. Individ-ual elk (ElkID, n = 43) were fitted as random intercept (1|ElkID) in a GLMM framework and were correctly as-sumed to be normally distributed, as confirmed visually by the Q-Q plot and formally by the Shapiro-Wilk nor-mality test.

for R (Pebesma, 2004). The variogram showed no evidence for spatial autocorrelation suggesting that it was addressed well in the GLMM (Figure A3.6).

However, presence-available data should be treated with great care when dealing with spatial autocorrelation of residuals. Whereas presence data are locations where a given animal (or plant) has been located (e.g. telemetry data, direct observations, museum records; see Manly *et al.*, 2002), available data are randomly drawn locations used to merely characterize resource availability. Many random points per presence location are usually needed to get stable parameter estimates in GLMMs (Figure A3.3). By nature, available points are randomly distributed and reflect the same spatial autocorrelation patterns of environmental predictors. Once we fit a GLMM with proper environmental covariates as predictors, residual autocorrelations for the major-ity of the locations (largely the available locations)

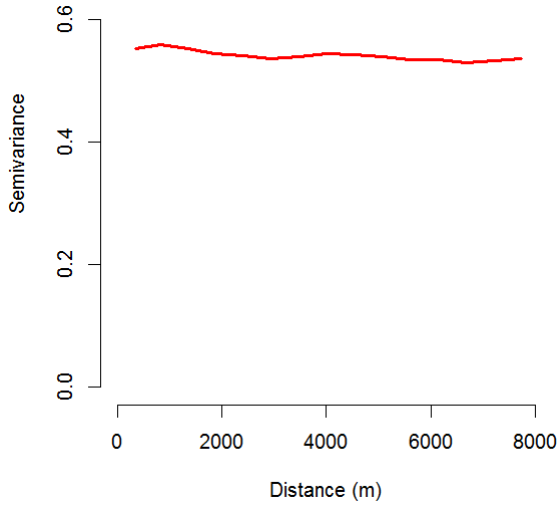

Figure A3.6. Variogram computed with the residuals of the GLMM trained with elk presence-available data.

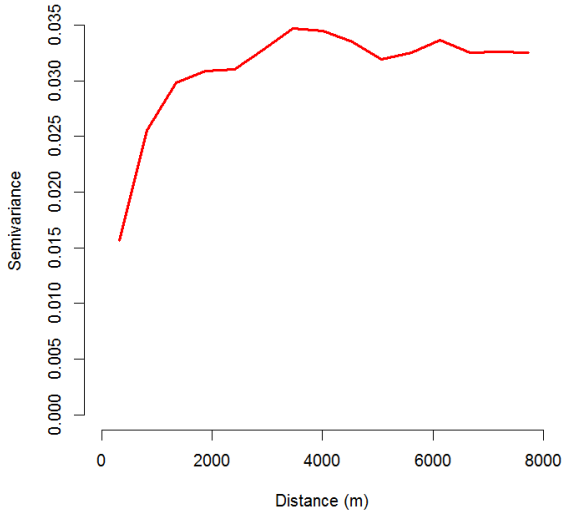

Figure A3.7. Variogram computed with residuals of the GLMM, this time considering only those related to elk presence data (compared to Figure A3.6, which depicts the variogram calculated with all residuals).

are addressed, resulting in a model that appears to have limited residual autocorrelation (Figure A3.6).

In other words, spatial autocorrelation between presence locations, if present in the data, is confounded by the noise produced by the large number of uncorrelated residuals linked to random available points. We can learn very little from this type of variogram. If we compute the variogram of residuals for

the presence data only, however, the picture is different (Figure A3.7). It reveals spatial autocorrelation between elk relocation residuals at a distance lower than 4 km. This is likely related to within-individual correlations (e.g. correlation between relocations of the same individual at distance lower than 4 km), although inter-individual correlations in animals with overlapping home ranges cannot be excluded.

Splitting data for different cross-validation schemes: cluster analysis

Splitting satellite relocations randomly into 5 folds, as usually done in RSF k-fold cross-validation (Boyce *et al.*, 2002) may create non-independent folds because each individual contributes to all 5 folds with some of its (dependent) locations ($\sim$20% of relocations per fold). The alternative is to split data by individuals, and thus avoid that the same individual contributing dependent data to different folds.

In order to visualize the spatial distribution of monitored elk, we built an Euclidean distance matrix with the distances between individual elk home range centres. We fitted a cluster analysis with complete linkage using the *hclust* function of the base *stats* package for R. To block by individuals, we use two approaches. First, we randomly select individuals and allocate them into 5 different folds (*blocking by random individuals*; Figure A3.8a). Based on our residual correlation analysis, this should create independent folds. However, by chance two folds may be not independent if they contain two individual elk with overlapping home ranges (Figure A3.9b). Therefore, we also block individuals by allocating elk into different folds only if their ranges do not overlap, making folds spatially independent. We chose 20 km as the minimum inter-individual distance required for spatial independence, a distance well above the $\sim$4 km extent of residual spatial autocorrelation (Figure A3.7). This 20 km threshold also results in a similar number of animals in each fold (Figure A3.9a). Home range sizes recorded in 2008 ranged from 8 to 140 km^2, with an average size of 34 km^2, or an area of approximately 6$\times$6 km. The largest home range recorded in 2008 (140 km^2) corresponded to an area of 12$\times$12 km (Figure A3.9b).

Depending on the ecology of the target species, inter-individual autocorrelation structures may differ significantly, thus affecting the design of blocks. Social mammals with strong social bonds (e.g. monkeys in the same troop, wolves in the same pack), for instance, may have strongly correlated individuals, which would require greater care in the blocking strategy. The sampling protocol implemented here involved the intentional capture of elk from different social units (i.e. different groups), which reduces the likelihood of inter-individual spatial autocorrelation patterns.

In summary, we:

1. Created a presence-availability dataset using elk telemetry relocations combined with random points drawn within individual home ranges, and associated the values of four spatial environmental predictors;
2. Fit a GLMM to estimate beta coefficients for environmental predictors, and used them to build the RSF, which was assumed to take the exponential form;
3. Showed, based on GLMM residual spatial autocorrelation, that residuals related to elk locations at a distance lower than 4 km were autocorrelated (within-individual correlation, and, possibly, inter-individual correlation);
4. Split the data into 5 folds following different blocking schemes which took care of the correlation between telemetry relocations differently.

RSF evaluation was thus performed using 4 different designs:

1. *Resubstitution.* The model was trained and tested on the same complete data set (i.e. no data splitting).
2. *Random cross-validation.* 5-fold cross-validation with all data (GPS location fixes) assigned to folds randomly (i.e. all elk con-

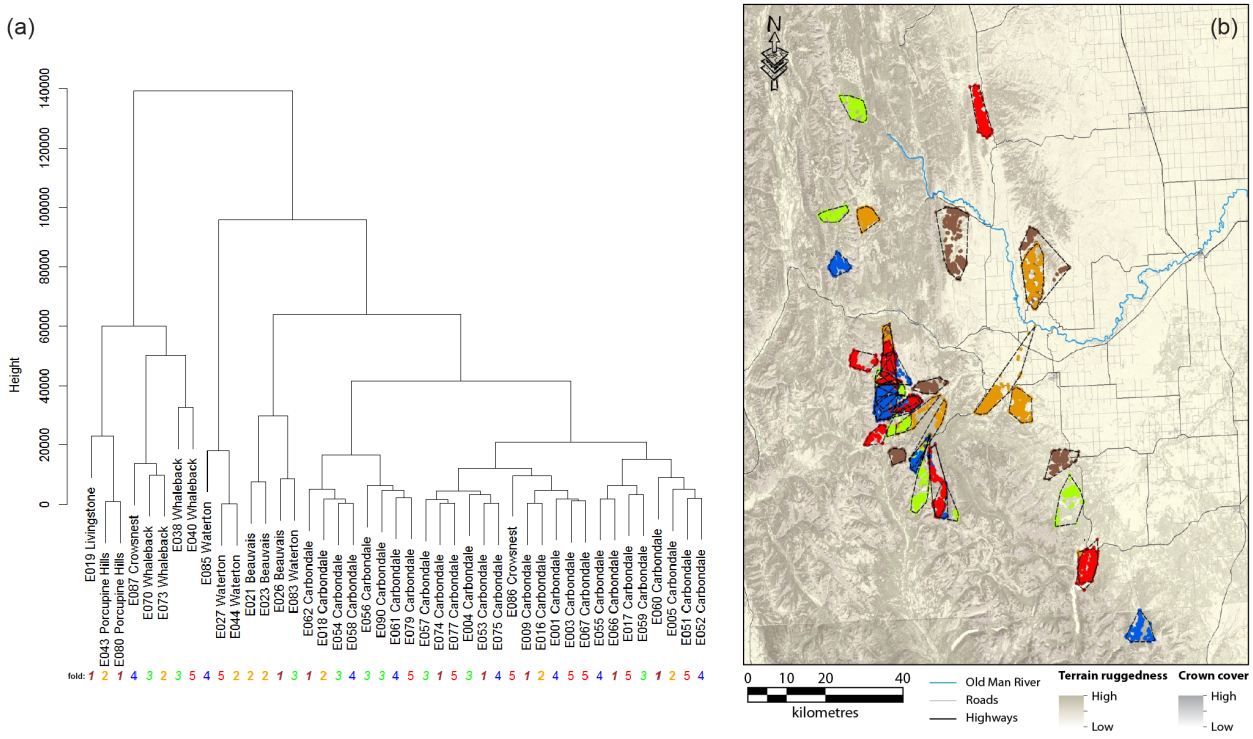

Figure A3.8. **(a)** Cluster analysis performed on the Euclidean distance matrix built with individual home range centres. **(b)** Telemetry relocations (dots) and summer home ranges (polygons) of 43 elk monitored in southwest Alberta, Canada. Elk individuals were randomly selected and assigned to 5 folds. A consistent colour legend for folds is used in the cluster tree and in the map.

tributed with telemetry fixes to all folds), resulting in 5 folds with ∼20% of telemetry relocations each.

3. *individual block cross-validation.* 5-fold cross-validation with data split by randomly assigning all GPS fixes from a single individual to a given fold, resulting in 3 folds with 9 individuals each and 2 folds with 8 individuals each. Home ranges of individuals assigned to different folds may overlap (Figures A3.8).

4. *Spatially independent individual block cross-validation.* 5-fold cross-validation with data split by spatially independent individuals. Data from each individual contributed only to one fold and individuals closer than 20 km were never allocated to the same fold. This resulted in 3 folds with 8 individuals each, 1 fold with 7 individuals, and 1 fold with 12 individuals. Home ranges of individuals assigned to different folds did not overlap (Figures A3.9).

Results

Model performance

Model evaluation for the full resubstitution (train data = test data) resulted in very strong Spearman rank correlation between RSF bin ranks and area-adjusted frequencies ($r_s = 1.00$, $p < 0.001$; Figure A3.10a and b). Evaluations for the random cross-validation was very similar with an average r_s across folds of 0.997 (Table A3.1; Figure A3.10a and b). Both evaluations, resubstitution and random cross-validation, suggest outstanding model performance.

In contrast, both of the block cross-validations, by individual and by spatially independent individual, showed notably lower performance estimates on average and much higher variability in Spear-

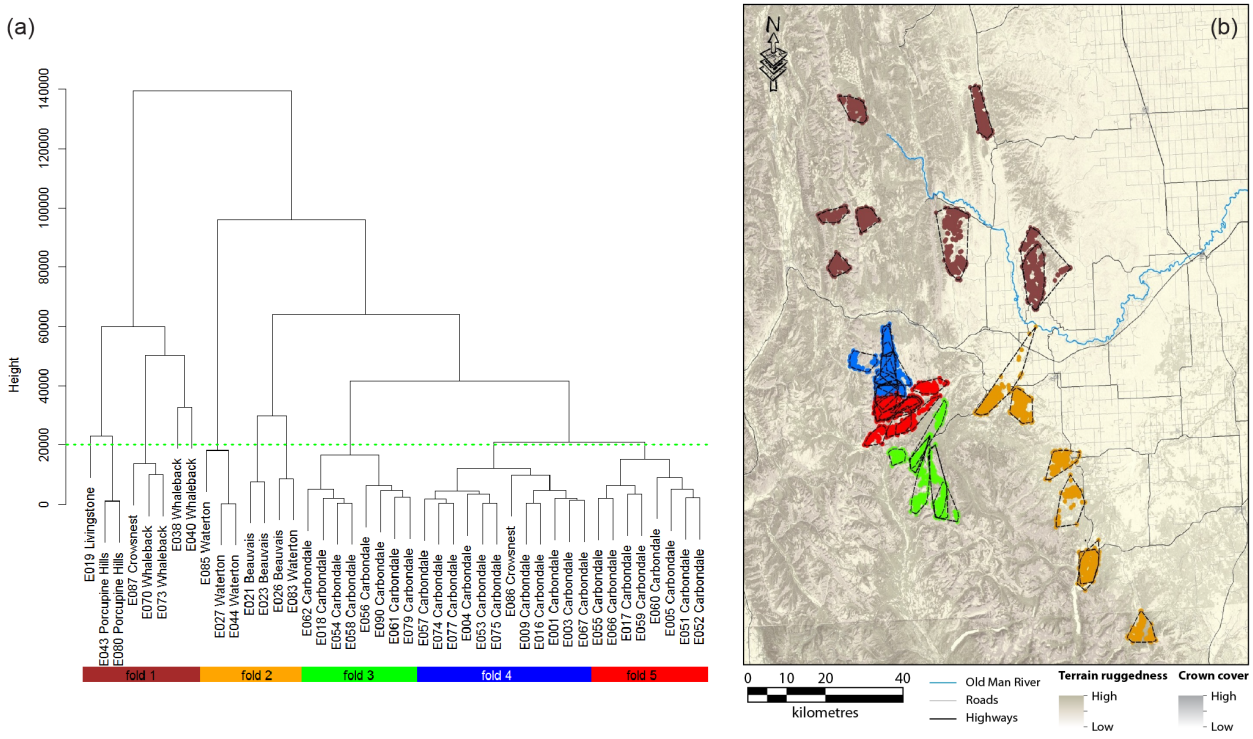

Figure A3.9. **(a)** Cluster analysis performed on the Euclidean distance matrix built with individual home range centres. Spatially-blocked animals were split and allocated into 5 folds. Home range centres between animals belonging to different folds (but not animals within the same fold) were at least 20 km distant (see text for more details). **(b)** Telemetry relocations (dots) and summer home ranges (polygons) of 43 elk monitored in southwest Alberta, Canada, as split into 5 folds. A consistent colour legend for folds is used in the cluster tree and in the map.

man rank correlations across folds (Table A3.1; Figure A3.10a and c). Spearman rank correlations for the cross-validation with randomly blocked individuals and spatially blocked individuals, averaged across all folds, were 0.916 and 0.925, respectively. Given that blocking by spatially independent individuals resulted in no further decrease in model performance, independence between folds was likely achieved at the level of individual animals. In ecological terms: individuals with overlapping home ranges did not behave more similarly than any two random individuals. Thus, while the data contains within-individual but not inter-individual residual spatial autocorrelation.

Parameter estimates

Beta coefficients for environmental predictors estimated by random cross-validations were consistent with those estimated by the full model (Figure A3.10c). Sample sizes in satellite telemetry studies are usually high and a model trained with 4/5 of the data (but including all individuals) performs similarly to a model trained with all data. Thus, it is likely that we can learn very little from the performances of the model by predicting to withheld data.

Beta coefficients showed increasing variability when blocking was introduced (Figure A3.10c). For example, the beta estimate for terrain ruggedness (TRI) by the full model was 0.152 while estimates fell between 0.147 and 0.156 (5.9% of the full model point estimate) in the random cross-validation, between 0.098 to 0.173 (49.3% of the full model point estimate) in the cross-validation blocked by random individuals, and between 0.094 to 0.203 (71.7% of the full model point estimate) for the cross-

Table A3.1. Spearman rank correlations between RSF bin ranks and area-adjusted frequencies for three different cross-validation designs, as well as the average across all folds for each blocking design.

Blocking design	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Random cross-validation	$r_s = 0.987$ $p < 0.001$	$r_s = 1.000$ $p < 0.001$	$r_s = 1.000$ $p < 0.001$	$r_s = 1.000$ $p < 0.001$	$r_s = 1.000$ $p < 0.001$	$r_s = 0.997$
Random individual block cross-validation	$r_s = 0.838$ $p = 0.002$	$r_s = 0.987$ $p < 0.001$	$r_s = 0.927$ $p < 0.001$	$r_s = 0.878$ $p < 0.001$	$r_s = 0.951$ $p < 0.001$	$r_s = 0.916$
Spatially independent individual block cross-validation	$r_s = 0.842$ $p = 0.004$	$r_s = 0.915$ $p < 0.001$	$r_s = 0.951$ $p < 0.001$	$r_s = 0.915$ $p < 0.001$	$r_s = 1.000$ $p < 0.001$	$r_s = 0.925$

validation by spatially independent individuals (Figure A3.10c). Beta estimates of all other environmental predictors showed similar patterns of variation. Despite consistent values on average across methods, parameter estimates covered a wider breadth of values in the blocked cross-validations, providing a measure of uncertainty for their true values.

Conclusions

By only implementing resubstitutions or random cross-validations in RSF approaches, modellers accept an optimism in model error estimates and a precision in parameter estimates due to non-independence of validation data. To date, both practices are prevalent in RSFs studies (Supporting Information Table S2). Block cross-validation can help avoid such overconfidence in model performance and foster greater care in the search for sound model structures.

References

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Boyce, M. S. 2010. Presence-only data, pseudo-absences, and other lies about habitat selection. *Ideas in Ecology and Evolution*, **3**, 26–27.

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. 2002. Evaluating resource selection functions. *Ecological Modelling*, **157**(2-3), 281–300.

Ciuti, S., Muhly, T. B., Paton, D. G., McDevitt, A. D., Musiani, M., & Boyce, M. S. 2012a. Human selection of elk behavioural traits in a landscape of fear. *Proceedings of the Royal Society B-Biological Sciences*, **279**(1746), 4407–4416.

Ciuti, S., Northrup, J. M., Muhly, T. B., Simi, S., Musiani, M., Pitt, J. A., & Boyce, M. S. 2012b. Effects of humans on behaviour of wildlife exceed those of natural predators in a landscape of fear. *Plos One*, **7**(11).

ESRI. 2011. *ArcGIS Desktop v.10*. Redlands, CA, USA: Environmental Systems Research Institute.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **119**, 142–152.

Killeen, J., Thurfjell, H., Ciuti, S., Paton, D., Musiani, M., & Boyce, M. S. 2014. Habitat selection during ungulate dispersal and exploratory movement at broad and fine scale with implications for conservation management. *Movement Ecology*, **2**, 1–13.

Lele, S. R., Merrill, E. H., Keim, J., & Boyce, M. S. 2013. Selection, use, choice and occupancy: clarifying concepts in resource selection studies. *Journal of Animal Ecology*, **82**(6), 1183–1191.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L., & Erickson, W. P. 2002. *Resource Selection by Animals: Statistical Design and Analysis for field studies*. New York, NY, USA: Kluwer Academic Publishers.

Muhly, T. B., Semeniuk, C., Massolo, A., Hickman, L., & Musiani, M. 2011. Human activity helps prey win the predator-prey space race. *Plos One*, **6**(3).

Pebesma, E. J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**(7), 683–691.

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Wiens, T. S., Dale, B. C., Boyce, M. S., & Kershaw, G. P. 2008. Three way k-fold cross-validation of resource selection functions. *Ecological Modelling*, **212**(3-4), 244–255.

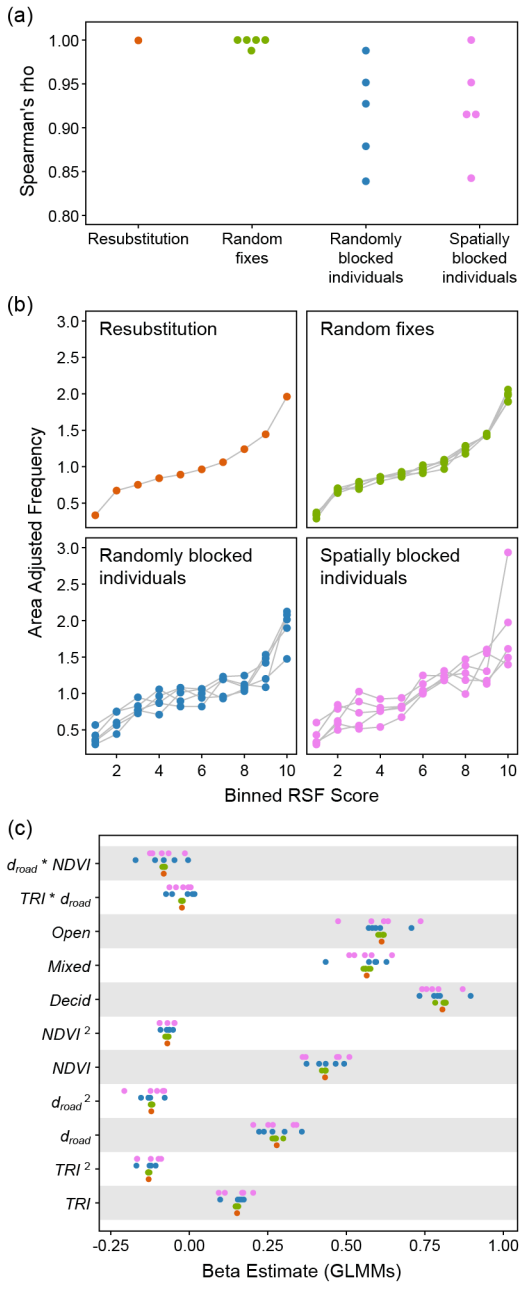

Figure A3.10. Resource selection function (RSF) evaluation depending on different validation designs (resubstitution, cross-validation with random fixes, with randomly blocked individuals, or with spatially blocked individuals). **(a)** Cross-validated Spearman rank correlations between RSF bin ranks and area-adjusted frequencies, which refer to the relationships depicted in the the lower panel plots **(b)** from top-left clockwise: area-adjusted frequency of categories (bins) of RSF scores for full model resubstitution, for withheld elk relocations, for withheld spatially blocked elk individuals, and for randomly withheld elk individuals. **(c)** Point estimates (betas) predicted by GLMMs for each model set (TRI: terrain ruggedness index; d_{road}: distance to the closest road; $NDVI$: normalized differenced vegetation index; *Decid*, *Mixed* and *Open* are deciduous forest, mixed forest, and open areas, respectively, with conifer forest as the reference category).

Appendix 4

Blocking to address phylogenetic correlation (Box 3)

Complete R scripts for this simulation are provided in Supplementary material Appendix 6.

Introduction

We simulate a simple trait-environment relationship (body mass as a function of latitude) for 50 hypothetical species observations to test error estimation in prediction of missing trait values when data are autocorrelated in phylogenetic distance (species relatedness). The simulation was repeated 100 times.

For model selection and error estimation, we test three cross-validation approaches: k-fold with random data splits, blocked k-fold with splits based on phylogenetic distance, and a leave-one-out approach with test points buffered based on phylogenetic distance. We also test parametric methods of model estimation, including a stepwise and a model dredging approach.

We compare selected models to the data generating model structure to assess the extent of overfit in selected models from the various cross-validation approaches. We also compare error distributions across the 100 simulations to assess the reliability of error estimates from the various cross-validation approaches.

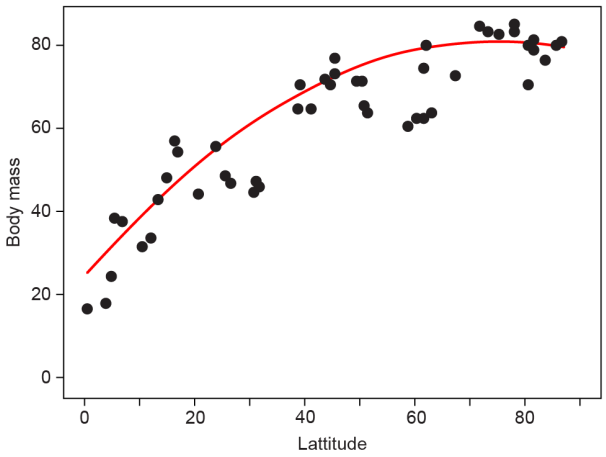

Figure A4.1. Simulated regression showing body mass as a function of latitude, where each point (n=50) represents a single tree tip (hypothetical species). The generating model prior to the introduction of autocorrelated error is shown as a red line.

Methods

All analysis was performed within the R framework for statistical computing (R Core Team, 2015). The following methodology was repeated for 100 simulations.

Simulating the trait / environment relationship

Trait values (simulated body mass) were generated for 50 hypothetical species through a three-step process. First, we generated an ordered set of 50 environmental observations (simulated latitude; one observation for each species) between 0 and 90 from a random distribution (the *runif* command in the base *stats* package for R).

Prior to the calculation of trait values (body mass), we created an error term that ensured model residuals were autocorrelated in phylogenetic distance. To accomplish this, we simulated a phylogenetic tree of 50 terminal nodes (hypothetical species), generated with default settings in the *pbtree* command from the *phytools* package for R (Revell, 2012). We then used a Brownian Motion (BM) simulation, the *fastBM* command from the *phytools* package for R (Revell, 2012), to generate a vector of tip state values (one for each tip) from the given

tree. Tree tip values were generated with a mean of zero and an instantaneous variance of 1.0 in the BM process.

Last, we generated the simulated values of body mass from a quadratic polynomial function with the environmental observations (latitude) as the predictor. We included the tip state values from the BM process as an error term to ensure that residual errors were autocorrelated within the phylogenetic distance structure of the tree (Figure A4.1). A non-zero intercept of 25 was added to ensure positive body mass values. The data generating model is as follows:

$$R = 25 + (1.5 \times E) - (0.01 \times E^2) + (P \times 10)$$

where:

R = the trait value response (body mass),
E = the environmental predictor (latitude), and
P = the tip state values from the BM process.

Phylogenetic autocorrelation

To determine the phylogenetic distance required to achieve independence of the test data, we evaluated residual phylogenetic autocorrelation using both semivariograms as well as correlograms with Moran's *I*. A single plot (semivariogram or correlogram) was made for each of the 100 simulations, then all averaged to create the final plots. Both approaches plot a measure of covariance (semivariance or Moran's *I*) against distance.

A distance matrix between all tree tips was created using the *cophenetic.phylo* command from the *ape* package for R (Paradis *et al.*, 2004) and the *cmdscale* command from the *MASS* package for R (Venables & Ripley, 2002) was used to generate coordinates in xy space for each tree tip. Residuals were from a fitted a quadratic regression (the same order as the true data generating model) of body mass as a function of latitude. Semivariograms were created with the *variog* command from the *geoR* package for R (Ribeiro Jr. & Diggle, 2015). Correlograms with Moran's *I* were generated using the *correlog* command from the *ncf* package for R (Bjørnstad, 2013).

While Moran's *I* decreased to ∼0.0 by phylogenetic distances of only 0.1, the semivariogram indicated that residual autocorrelation extended to ∼0.4 units of phylogenetic distance. Beyond ∼0.5 distance units, semivariance remained consistent or decreased (Figure A4.2).

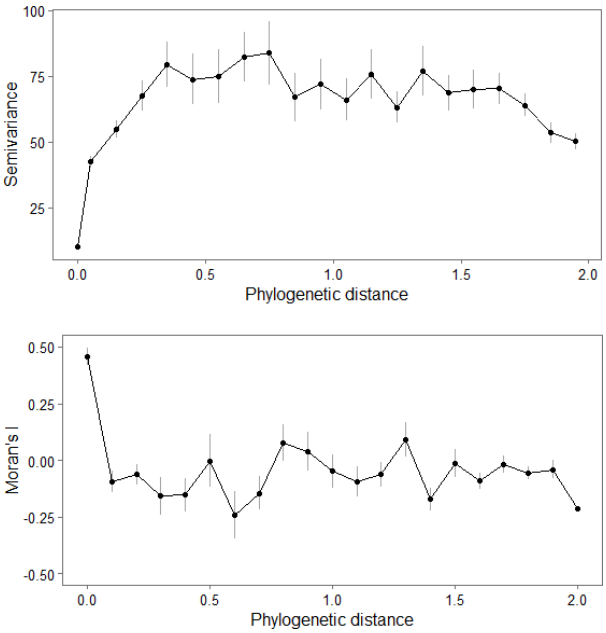

Figure A4.2. **(a)** Semivariogram of phylogenetic distance and residuals of a quadratic regression model fitted with all data. **(b)** Correlogram showing Moran's *I* as a function of phylogenetic distance, based on the residuals from a quadratic regression model fitted with all data. Both plots are averaged across all 100 simulations, with vertical bars showing standard errors.

Model evaluation

We evaluated the ability of several evaluation and cross-validation approaches to 1) select an appropriate model structure similar to the true (i.e. data generating) model and 2) provide reliable error estimates (i.e. similar to the error term included in the data generating model). To test both, we evaluated eight predefined regression models of increasing complexity (i.e. increasing polynomial order) with four cross-validation techniques: two *k*-fold cross-validations, one based on random and one on non-random data splits, and a buffered leave-one-out cross-validation. We also performed a full data re-substitution, where all data are used for both model

Table A4.1. Pre-defined model structures included in the model selection process. Note that the number of parameters includes a model intercept (not shown in formulae).

Parameters	Formula	Relationship to true model
2	$R \sim E$	Underfit
3	$R \sim E + E2$	True model
4	$R \sim E + E2 + E3$	Overfit
5	$R \sim E + E2 + E3 + E4$	Overfit
6	$R \sim E + E2 + E3 + E4 + E5$	Overfit
7	$R \sim E + E2 + E3 + E4 + E5 + E6$	Overfit
8	$R \sim E + E2 + E3 + E4 + E5 + E6 + E7$	Overfit
9	$R \sim E + E2 + E3 + E4 + E5 + E6 + E7 + E8$	Overfit

training and testing, incorporating a stepwise model selection and a model dredging.

The first cross-validation was a k-fold with random data splits. For this approach, each data point was randomly assigned to one of either 5 or 10 folds. As in a typical k-fold approach, each of the k folds was left out in turn and a model trained with data from the remaining folds. The withheld fold was then used for model prediction. This process was run iteratively until all of the k folds had been held-out exactly once.

The second cross-validation was a blocked k-fold, identical in procedure to the random k-fold but incorporating non-random data splits defined based on phylogenetic distance between data points. To define folds, the 50 tree tips were clustered into k groups based on their phylogenetic distance using the pam command from the cluster package for R (Reynolds et al. 1992).

In addition to the k-fold approaches, we also considered a buffered leave-one-out (SLOO) cross-validation. This approach is akin to the spatially independent leave-one-out cross-validation approach in the spatial simulation (Box 1; Appendix Box 1). A standard leave-one-out approach withholds a single data point, trains a model on the remaining points, then tests the prediction to the withheld point. In SLOO approaches, data points within a given distance (in space, time, etc.) the withheld data point are also removed from the training data to achieve independence between training and testing data (Bahn, 2009; Le Rest *et al.*, 2014). We implemented the buffered LOO removing points within phylogenetic distances of 0.00, 0.25, 0.50, 0.75, and 1.00 units. According to the analysis of residual autocorrelation (Figure A4.2), buffer distances of 0.50 and higher achieve phylogenetic independence of the withheld point.

In all implementations of each cross-validation approach, all predictions (for which there was always exactly one for each data point) were then evaluated against the known true value and a root mean squared error (RMSE) of the predictions was calculated.

Model selection

We pre-defined eight model structures of increasing complexity (i.e. increasing polynomial order) from which to select a "best" predictive model (Table A4.1). Each cross-validation approach (random k-folds, blocked k-folds, and buffered LOOs) was run using all 8 model structures. The "best" model from the cross-validations was defined as that with the lowest RMSE across all withheld folds.

We also included a parametric model selection using AIC in which we compared AICs from 1) those calculated for all eight pre-defined model structures, 2) a stepwise model selection using the stepAIC command from the MASS package for R (Venables & Ripley 2002), and 3) a full model dredging using the *dredge* command from the *MuMIn* package for R (Barton 2015). For comparison, we considered

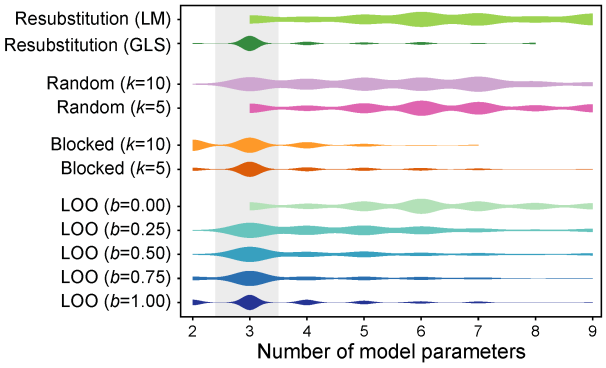

Figure A4.3. Violin plot of the number of model parameters in selected models by each evaluation and cross-validation procedure. The LM and GLS models were selected by minimum AIC while the random, blocked, and LOO cross-validation models were selected by minimum RMSE. Frequencies of each selected model are provided in Table A4.2.

both a standard regression (LM) using the *lm* command from the base *stats* package for R, as well as a geographical least squares regression (GLS; see Dormann *et al.*, 2007, for an ecological explanation and application) using the *gls* command from the *nlme* package for R (Pinheiro *et al.*, 2015), with the correlation structured defined therein by the *corBrownian* command from the *ape* package for R (Paradis *et al.*, 2004). In the parametric selections, the best model was defined as that with the lowest AIC. Because a new model is created for each fold in the cross-validations, values for AIC (and for the number of model parameters in the stepwise and dredging approaches) are averaged for each cross-validation approach for tree.

Results

Due to consistently positively skewed distributions of RMSE, we compare medians and percentiles of RMSE across all simulations, rather than means and standard deviations. Values of AIC and numbers of model parameters are presented as averages across all simulations.

Model selection

Frequencies of parameter counts for the final models as selected by the parametric and cross-validation approaches are listed in Table A4.2 and plotted in Figure A4.3 and Figure A4.4a. Average parameter counts for each method (across the 100 simulations) are provided in Table A4.3.

The GLS offered improvement in model selection over the LM, selecting an average of 3.9 ($\sigma = 1.5$) model parameters, compared with an average of 6.2 ($\sigma = 1.9$) parameters for the LM. Most of the selected models for the GLS were generated via the stepwise or dredging approach, which unsurprisingly tended to optimise AICs by considering model combinations beyond simple additive arrangements (e.g. dredging or stepwise selection could result in a final model with a squared and quartic term but not a cubed term, whereas this combination is not offered in the pre-defined model structures). For the GLS, the correct model complexity was selected in 60% of the simulations (the most common), while in the LM this dropped to 12% (with 6 parameters being the most common).

The random cross-validations and the unbuffered LOO approach performed similarly to the LM resubstitution for model selection, averaging 5.5 ($\sigma = 1.8$) and 6.0 ($\sigma = 1.8$) parameters for the 5- and 10-fold random approaches respectively, and 6.0 ($\sigma = 1.8$) for the unbuffered LOO, with the number of parameters in the final models distributed fairly equally.

When cross-validations were performed using less dependent data, as in the blocked and buffered LOO approaches, model selection more reliably chose structures similar to the true data generating model, with most simulations for these methods resulting in the appropriate model complexities. The 5- and 10-fold block cross-validations selected the appropriate model complexity 42% and 52% of the time, respectively, while the LOO with buffers of 0.25, 0.50, 0.75, and 1.00 selected the appropriate model complexity 30%, 43%, 49%, and 45% of the time, respectively. The most reliable model selection was in the 10-fold block cross-validation (correct complexity in 52% of the simulations), while the 5-fold block cross-

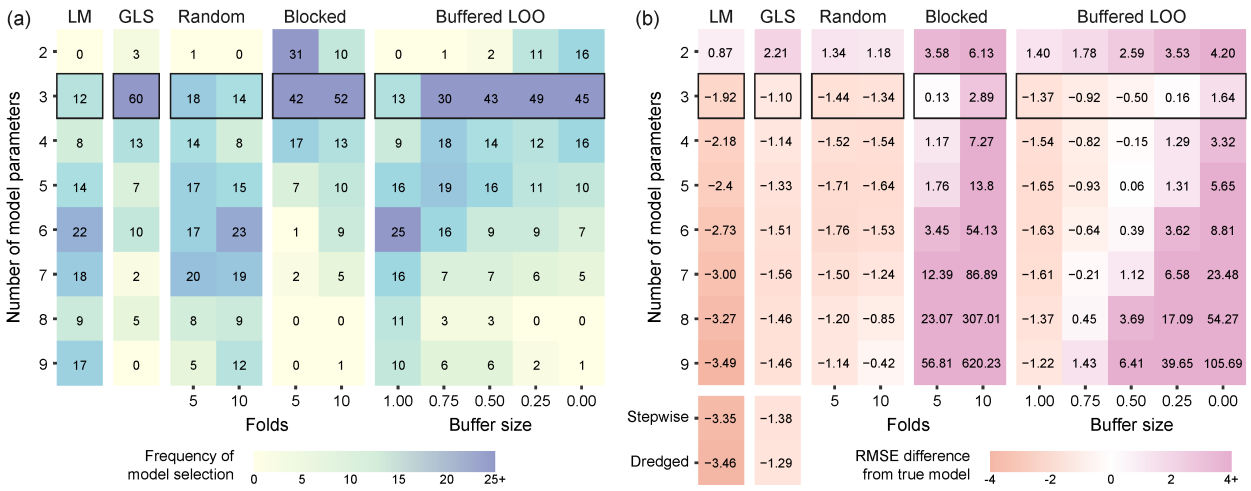

Figure A4.4. (a) The frequency of model selection for each level of model complexity (number of parameters) by each evaluation (LM and GLS; selected by minimum AIC) and cross-validation (Random, Blocked, and Buffered LOO; selected by minimum RMSE) approaches. **(b)** The difference between the median RMSE from the true model (8.91) and the median RMSE from the resubstitutions and cross-validations. Models highlighted in orange are those with optimistic error estimates (RMSE lower than the true model) while models highlighted in purple are those with pessimistic error estimates (RMSE higher than the true model). Models in white have error estimates that closely match the true model RMSE. In both figures, the true data generating model complexity (3 parameters) is outlined in black.

validation was the best general performer, averaging 3.1 ($\sigma = 1.1$) parameters across the 100 simulations.

Error estimates

Percentiles of RMSE for each evaluation and cross-validation approach are provided in Table A4.3. Differences between median RMSE values from each model structure for each evaluation and cross-validation are shown in Figure A4.4b. Plots of RMSE distributions are shown in Figure A4.5.

As expected, RMSE values for the LM resubstitution (median RMSE = 5.51) are the most optimistic relative to the true median RMSE of 8.91. The GLS resubstitution (median RMSE = 7.54), while improving on the LM, still resulted in optimistic error estimates. Among the cross-validations, those which did not adequately address dependence in validation data, including the random cross-validations (RMSE = 6.64 and 6.66 for the 5- and 10-fold validations respectively) and the LOO cross-validations with buffer sizes just at or below that which accounts for autocorrelation (RMSE = 6.62, 7.41, and 7.79 for buffer sizes of 0.00, 0.25, and 0.50, respectively) also resulted in the optimistic error estimates.

The blocked 10-fold cross-validation (median RMSE = 8.56) and the LOO with a buffer of 0.75 units (median RMSE = 8.52) provided the best error estimates relative to those from the true model. Only the 5-fold blocked cross-validation (median RMSE = 9.87) and the leave-one-out with the largest buffer size (median RMSE = 9.05) resulted in RMSE values higher than those of the true model (i.e. overly-pessimistic validations). In both cases, values at the high RMSE end extend well beyond the highest true RMSE values, while values at lower percentiles match those of the true model very closely.

Extrapolation

Phylogenetic distance in the cross-validations (random, blocked, and LOO) was calculated be measuring the minimum distance between all points in the training and testing folds. Phylogenetic dissimilarity for the resubstitutions is consistently 0.0 as all data are able to measure to themselves RMSE followed a

predictable pattern when compared to the phylogenetic distance between testing and training data: as distance increases, RMSE decreases linearly (Figure A4.6).

References

Bahn, V. 2009. A new method for evaluating species distribution models. *In: 94th Ecological Society of America Annual Meeting*.

Bjørnstad, O. N. 2013. *ncf: spatial nonparametric covariance functions. R package version 1.1-5.*

Dormann, C. F., McPherson, J. M., Araujo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kuhn, I., Ohlemuller, R., Peres-Neto, P. R., Reineking, B., Schroder, B., Schurr, F. M., & Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**(5), 609–628.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, **23**(7), 811–820.

Paradis, E., Claude, J., & Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2), 289–290.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Team, R Core. 2015. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-121.

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**(2), 217–223.

Ribeiro Jr., P. J., & Diggle, P. J. 2015. geoR: Analysis of Geostatistical Data. R package version 1.7-5.1.

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S*. 4 edn. New York, NY, USA: Springer-Verlag.

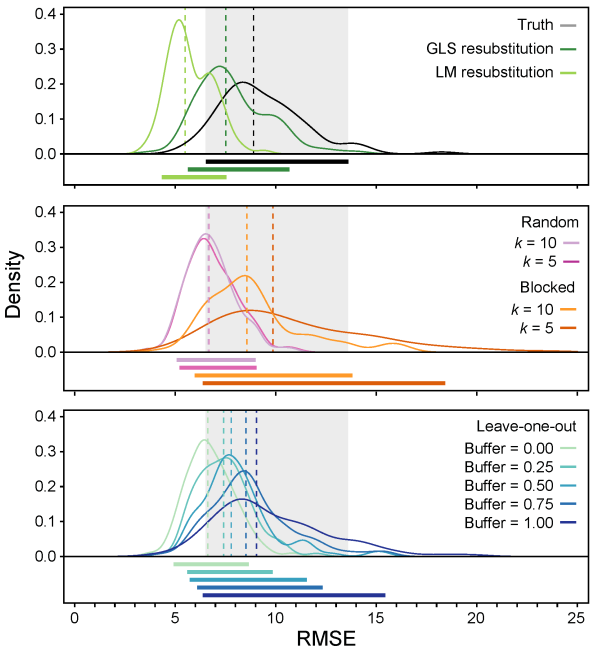

Figure A4.5. Distributions of RMSE across 100 simulations for the resubstitution and for the various implementations of each cross-validation approach, including the random k-fold, the blocked k-fold, and the buffered leave-one-out. A density plot is provided for each number of folds (k) and for each buffer size. To facilitate comparison, coloured horizontal lines represent the 5th to 95th percentile range of each distribution. Dashed vertical lines represent distribution medians. The shaded grey areas shows the RMSE distribution of the true model.

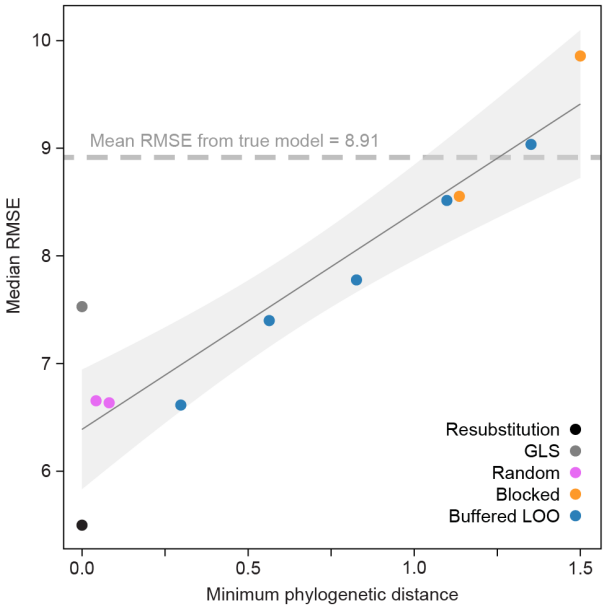

Figure A4.6. The median RMSE of the selected model for each parametric approach (Resubstitution and GLS; based on minimum AIC) and cross-validation (Random, Blocked, and Buffered LOO; based on minimum RMSE), plotted against the minimum phylogenetic distance between training and testing data, averaged across all folds (in the random and blocked k-folds) or across all withheld points (in the LOO), averaged across all 100 simulations. The median RMSE from the true data generating model is shown as a dashed grey line. RMSE values for all models are provided Table A4.3 and Figure A4.3. Phylogenetic distances are also provided in Table A4.3.

Table A4.2. The count of final models from each evaluation or cross-validation approach, including the linear model resubstituion (LM), the GLS resubstitution (GLS), the random cross-validation (Ran), the block cross-validation (Block), and the leave-one-out cross-validation (LOO), falling into each level of model complexity (number of model parameters), of the 100 simulations. The data generating "true" model contains 3 parameters. For each method, the number of parameters of the highest frequency is shown in bold.

	Number of parameters in final model							
	2	3	4	5	6	7	8	9
AIC-selected models:								
LM	0	12	8	14	**22**	18	9	17
GLS	3	**60**	13	7	10	2	5	0
RMSE-selected models:								
Ran (k=10)	0	14	8	15	**23**	19	9	12
Ran (k=5)	1	18	14	17	17	**20**	8	5
Block (k=10)	10	**52**	13	10	9	5	0	1
Block (k=5)	31	**42**	17	7	1	2	0	0
LOO (b=0.00)	0	13	9	16	**25**	16	11	10
LOO (b=0.25)	1	**30**	18	19	16	7	3	6
LOO (b=0.50)	2	**43**	14	16	9	7	3	6
LOO (b=0.75)	11	**49**	12	11	9	6	0	2
LOO (b=1.00)	16	**45**	16	10	7	5	0	1

Table A4.3. The average and standard deviation number of parameters (Parameters) and percentiles of RMSE across all 100 simulations for the selected "best" models based on minimum AIC for the linear model (LM) and GLS resubstitutions or based on minimum RMSE for the random (Ran), blocked (Block), and leave-one-out (LOO) cross-validations. For comparison, values for the true data generating model are also shown. Median RMSE and the difference between model and true median RMSE for each evaluation and cross-validation are also plotted in Figure A4.4. Also noted is the minimum phylogenetic distance (Dist) between the training and validation data in the various cross-validation approaches, averaged across all folds (see also Figure A4.5).

	Parameters		RMSE percentiles					
	Avg	SD	5^{th}	25^{th}	50^{th}	75^{th}	95^{th}	Dist
True model	3.0	2.2	6.54	7.86	8.91	10.55	13.64	
AIC-selected models:								
LM	6.2	1.9	4.35	4.95	5.51	6.66	7.56	0.00
GLS	3.9	1.5	5.64	6.66	7.54	8.97	10.69	0.00
RMSE-selected models:								
Ran (k=10)	6.0	1.8	5.08	5.96	6.64	7.41	8.99	0.04
Ran (k=5)	5.5	1.8	5.21	5.98	6.66	7.65	9.05	0.08
Block (k=10)	3.8	1.4	5.98	7.21	8.56	9.68	13.81	1.10
Block (k=5)	3.1	1.1	6.38	8.04	9.87	12.93	18.43	1.45
LOO (b=0.00)	6.0	1.8	4.93	5.93	6.62	7.45	8.66	0.29
LOO (b=0.25)	4.8	1.8	5.61	6.45	7.41	8.16	9.85	0.54
LOO (b=0.50)	4.5	1.8	5.73	7.00	7.79	8.79	11.55	0.80
LOO (b=0.75)	3.9	1.6	6.10	7.54	8.52	9.75	12.33	1.06
LOO (b=1.00)	3.7	1.4	6.38	7.90	9.05	11.47	15.44	1.31

Appendix 5

Blocking for extrapolation (Box 4)

Complete R scripts and data for this case study are provided in Supplementary materials, Appendix 6.

Introduction

We examined the effect of blocking in environmental space on cross-validation, in a typical species distribution modelling approach for Douglas-fir (*Pseudotsuga menziesii*) habitats in western North America.

Methods

All analysis was performed within the R framework for statistical computing (R Core Team, 2015).

Species and climate data

Douglas-fir presence and absence records were compiled from various forest inventory databases by Rehfeldt *et al.* (2014) and used with permission. We included data from Canada, the continental USA, and Mexico, west of -95° longitude, resulting in 53,293 unique inventory records (18,601 presences; 34,692 absences; 34.9% prevalence). The data covered the entire natural range of Douglas-fir in North America (Figure A5.1a). To maintain a consistent Cartesian grid scale across the entire study area, geographic coordinates of the original data (decimal degrees of latitude and longitude) were re-projected into XY metres in Lambert Conformal Conic projection using ESRI ArcGIS 10.3.

Environmental predictors were chosen from a suite of climate variables at each observed presence and absence location, generated by the publicly available ClimateNA software (http://tinyurl.com/ClimateNA), which topographically downscales observed weather station data at specifically requested lat-long coordinates using the PRISM approach (Wang *et al.*, 2016). Climate data layers are also publicly available at the Adaptwest Databasin (http://adaptwest.databasin.org). Six observed and derived variables were included based on Douglas-fir literature (e.g. Leites *et al.*, 2012; Montwe *et al.*, 2015) or for their more general biological relevance:

1. Average temperature of the coldest month (MCMT),
2. Average temperature of the warmest month (MWMT),
3. Total summer (May to Aug) precipitation (PPT_sm),
4. Total precipitation of the driest month (MDMP),
5. Annual heat-moisture index (AHM), and
6. Growing degree days above 5°C (DD5)

To remove collinearity between variables, climate data were scored with a principal component analysis and all six components (PC1 to PC6) were used in the modelling. Component loadings are listed in Table A5.1.

Species distribution models

Modelling was performed with Random Forest, a bootstrap-aggregation approach to classification trees, using the *randomForest* package for R (Breiman, 2001). Forests were built using 1,000 classification trees with the species presence or absence (1 or 0) defined as a logistic response variable. The Random Forest models created in the

Table A5.1. For the six included climate variables, the principal component loadings (PC1 to PC6), the standard deviation of the loadings (*SD*),and the variance explained, both by each component (*VarEx*) and cumulatively (*CumVarEx*).

Variable	PC1	PC2	PC3	PC4	PC5	PC6
MWMT	-0.49	-0.26	0.25	0.04	-0.52	0.60
MCMT	-0.43	-0.29	-0.60	0.12	0.55	0.23
PPT_sm	0.21	-0.64	0.30	-0.62	0.27	0.05
MDMP	0.35	-0.49	0.22	0.77	0.08	-0.01
DD5	-0.48	-0.35	-0.01	0.02	-0.26	-0.76
AHM	-0.43	0.27	0.66	0.12	0.53	-0.05
SD	*1.81*	*1.29*	*0.66*	*0.57*	*0.52*	*0.19*
VarEx	*0.55*	*0.28*	*0.07*	*0.05*	*0.04*	*0.01*
CumVarEx	*0.55*	*0.82*	*0.90*	*0.95*	*0.99*	*1.00*

cross-validation procedure predict a probability (between 0 and 1) that a given data point will contain a species presence, based on what proportion of the 1,000 classification trees result in a species presence for each data point. Spatial autocorrelation of model residuals was assessed with maps of model residuals as well as with correlograms based on Moran's *I*. Correlograms were implemented with the *correlog* function from the *ncf* package for R (Bjrnstad, 2013) using a random subsample of 5,000 data points with 5 resampling runs.

Cross-validation structures

In addition to a full data resubstitution as a measure of goodness-of-fit, k-fold cross-validations were performed using several data-splitting approaches, including splitting the model training data randomly, in geographic space, and in environmental space (Figure A5.1). Because of the large size of the data in general, sample size differences in training data sets made no appreciable difference to model structure or accuracy (data not shown). Specifically, validations were of the following structures:

Resubstitution. All data used for model training and model evaluation. No data splitting.

Random splits. Data were randomly assigned to one of either two, four, or eight folds.

Spatial splits. Data was divided along x and y coordinates into contiguous blocks with the number of blocks in xy equalling 1×2, 2×2, 2×3, 2×4, 3×6, 4×8, 5×10, 10×20, or 20×40. Blocks were then treated as individual folds for cross-validation to minimise environmental extrapolation between spatial folds.

Environmental splits. Data were assigned to groups based on similar values of environmental variables (PC1 to PC6), using three different methods. First, a k-means cluster analysis (*kmeans* command in the base *stats* package for R) was used to group data into two, four, six, or eight environmentally similar folds. Second, the values of the predictor variables were equally split in multidimensional space using a nearest neighbour approach with reference points placed at the first and third quartile of each variable. Three iterations considering only PC1, PC1 and PC2, or PC1 and PC2 and PC3, resulted in two, four, or eight environmentally similar folds, respectively.

The k-fold cross-validation approach involves iteratively withholding each one of the k folds in turn, training a model on the remaining k-1 folds, then predicting to the withheld k^{th} fold. Because the k-fold is exhaustive, all data are withheld once resulting in a single set of predictions for the entire data set.

We also performed several buffered leave-one-out validations, which are identical to traditional leave-one-out approaches but with points within a given radius around the hold-out point also withheld from the training data (Bahn, 2009; Le Rest *et al.*, 2014). The leave-one-out validations were performed with buffer radii of 100, 500, 1000, and 1500 km. Because leave-one-out processes can be computationally intensive, we ran these validations on a random subsample of 5,000 data points. This sample size was selected as it balances reliability of the error estimates with computational efficiency (from a sensitivity analysis using progressively increasing sample sizes; data not shown but code provided in R scripts).

Model evaluation was undertaken using the area

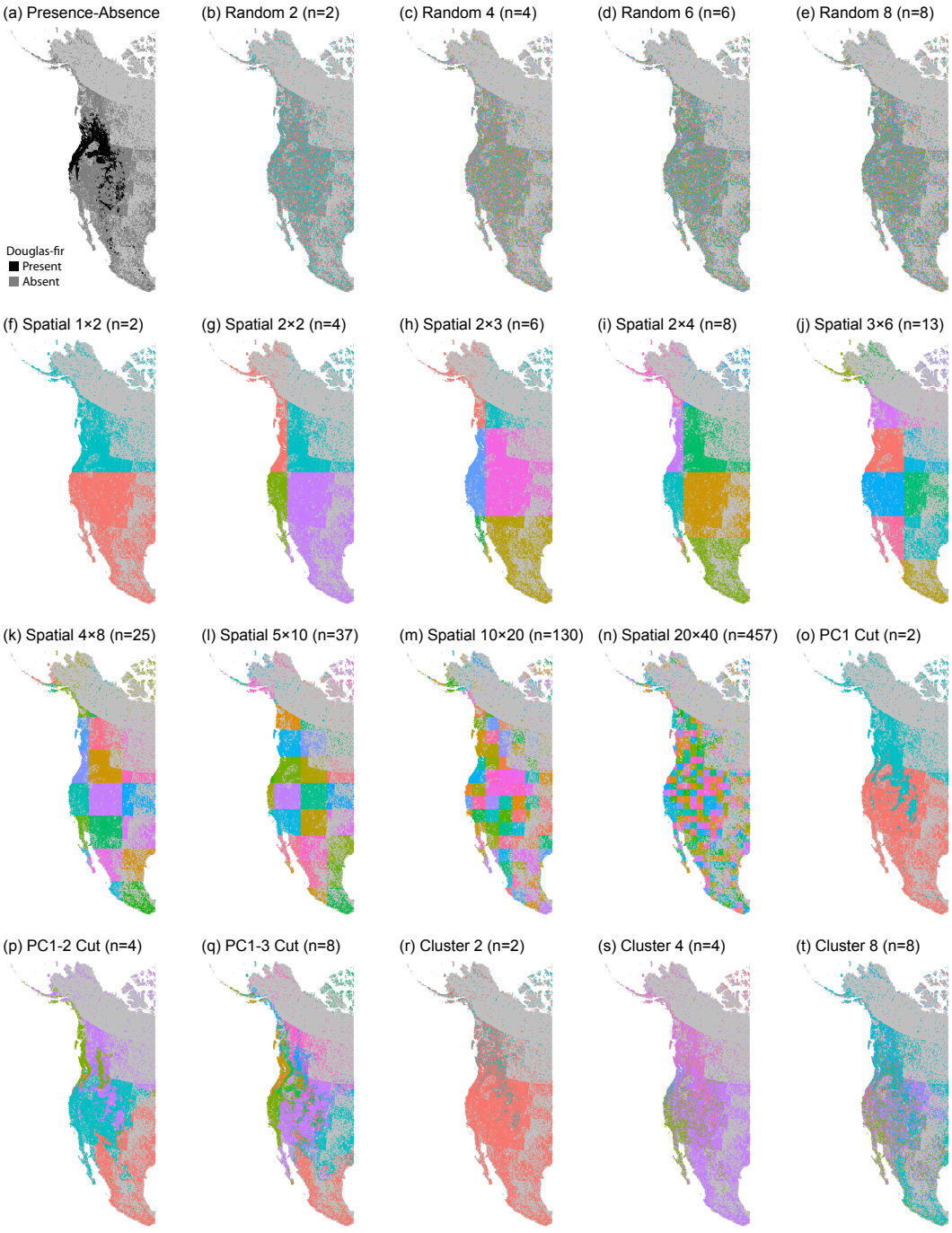

Figure A5.1. Maps of training data and cross-validation blocks showing **(a)** the presence-absence records for reference, **(b-e)** the random data splits, **(f-n)** spatial blocks from divisions in xy coordinates, **(o-q)** environmental blocks from divisions in principal component ranges, and **(r-t)** environmental blocks from the cluster analysis. Some spatial blocking scenarios are omitted for space. The number of blocks created in each splitting approach is shown in parentheses.

under the curve (AUC) of the receiver operating characteristic (Swets, 1988), a common validation statistic for presence-absence models when species non-detections are assumed to be true absences. AUC values were calculated using the *ROCR* package for R (Sing *et al.*, 2005). The AUC statistic can vary between 0 and 1, where 1 represents perfect model performance, 0 represents perfectly incorrect model performance, and 0.5 is the expected value from a random model.

Measuring extrapolation

Environmental extrapolations were quantified using multivariate distances across environmental predictors, a common for measure of environmental dissimilarity (e.g. Williams *et al.*, 2001; Roberts & Hamann, 2012; Eiserhardt *et al.*, 2013; Mesgaran *et al.*, 2014). Because variables are standardised and correlation is removed through the PCA rotation, a simple Euclidean distance between PCA variables can be used to measure dissimilarity. First, each PCA variables was weighted by its corresponding variance explained. Then, for each fold, the Euclidean distance across all PCA environmental variables was measured between the withheld data and the remaining model training data. In the leave-one-out approaches, the distance from each hold-out point was measured back to the training data (with the buffer removed). Reported environmental distances for each k-fold approach are the averaged minimum distances for all hold-out points. For computational efficiency, a consistent random selection of 5,000 training points was used for distance measurements. Spatial distances between folds were calculated in the same way using XY coordinates.

Results

Residual spatial autocorrelation

Autocorrelation in model residuals was assessed with a map of the residuals, a semivariogram, and a correlogram with Moran's *I*. The mapped residuals show spatial structure: residuals through the presence range of the species are of a larger magnitude (in both the positive and negative direction) than those through the absence range (Figure A5.2a). The semivariogram and the correlogram give similar estimates for the range of residual spatial autocorrelation. In the semivariogram, semivariance plateaus at 213 km (semivariance = 0.013) before decreasing consistently (Figure A5.2b). In the correlogram, Moran's I is minimised at 226 km (Moran's I = -0.005) and remains negligible thereafter (Figure A5.2c).

Environmental and spatial distances in data splits

Minimum multivariate distances between validation folds was smallest for the resubstitution and random data splits and increased with increasing size of spatial data splits (Table A5.2, Figure A5.3). Distances between environmental data splits were largest when the fewest folds were used (dividing only along PC1 or in the 2-group cluster analysis) and decreased with increasing number of folds.

Goodness-of-fit and cross-validations

The Random Forest model showed near perfect goodness-of-fit (AUC = 1.00) when evaluated using a complete data resubstitution (all data used for both training and evaluation; Table A5.2). When evaluated using a k-fold with data divided randomly into two, four, six, or eight folds, AUC decreased only slightly and was consistent across the number of folds (AUC = 0.97).

In the spatial blocking approaches, model accuracy decreased relatively consistently as the number of spatial blocks increased (i.e. as the size of blocks decreased). The highest accuracy (AUC = 0.95) was found in the 20×40 arrangement (457 blocks of 193×192 km each) and was only slightly lower than in the random data splits. The lowest accuracy (AUC = 0.68) was found in the 1×2 arrangement (2 blocks of 3862×3834 km each).

In the environmental blocking approaches, accuracies for the eight- and six-group clustering (Cluster-8 and Cluster-6) and the eight-way PCA split (PC1-3) were consistently high (AUC = 0.95 to 0.96). Accuracy dropped slightly for the four-way PCA split

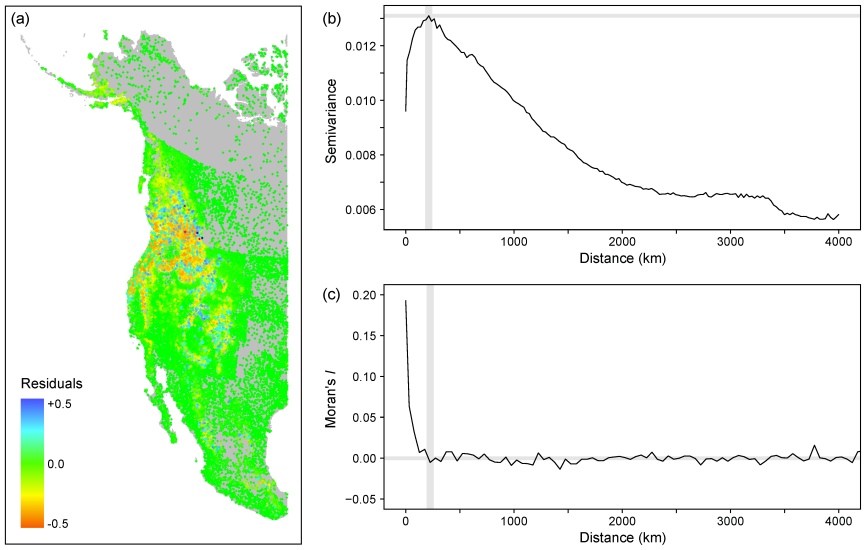

Figure A5.2. Spatial autocorrelation of model residuals as visualised in **(a)** a geographical map of residuals, **(b)** a semivariogram, and **(c)** a correlogram with Moran's *I*. In the semivariogram, the plateau in semivariance is highlighted with a horizontal grey line (semivariance = 0.013) and a vertical grey line (geographic distance = 213 km). In the correlogram, the distance to a minimal value of Moran's *I* is highlighted with a horizontal grey line (Morans *I* = -0.005) and a vertical grey line (geographic distance = 226 km).

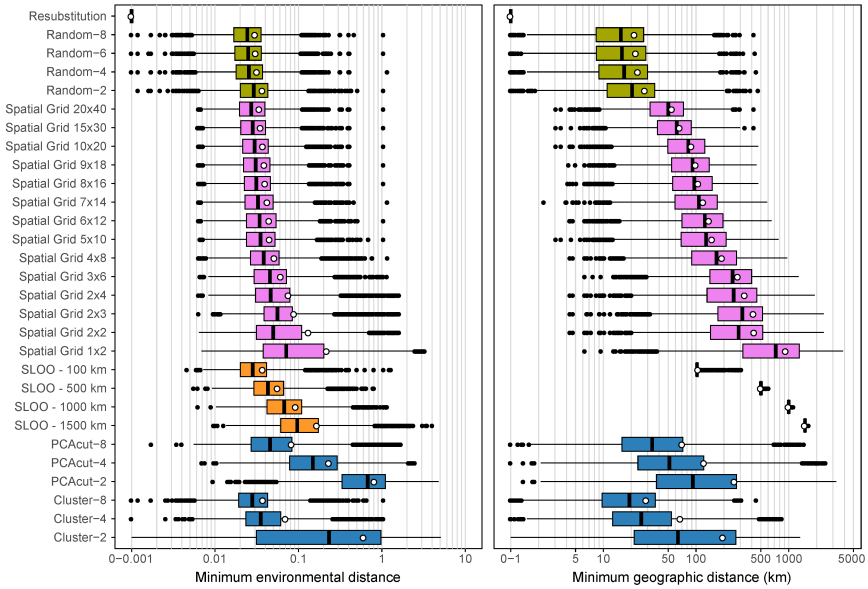

Figure A5.3. Boxplots of minimum environmental and minimum geographic distance of the hold-out data measured back to the model training data. White circles represent group averages, which are reported in Table A5.2.

Table A5.2. Correlation statistics for cross-validation accuracy relative to environmental and geographic distance for each of the cross-validation data splitting approaches, listing the number of cross-validations (n), the Pearson correlation coefficient of AUC and environmental distance (r), the corresponding p-values (p), and the slopes of the corresponding regression lines (*Slope*). Individual values for each approach are listed in Table A5.2.

		Environmental distance			Geographic distance		
Method	n	r	p	slope	r	p	Slope
Random splitting	4	-0.995	0.005	-0.53	-0.997	0.003	-5.79e-4
Spatial blocking	14	-0.949	<0.001	-1.45	-0.978	<0.001	-3.28e-4
Buffered leave-one-out	4	-0.852	0.148	-1.78	-0.945	0.055	-1.88e-4
Environmental blocking	7	-0.996	<0.001	-0.15	-0.976	<0.001	-5.37e-4

(PC1-2; AUC = 0.93). Notably lower accuracies were observed in the two-group clustering (Cluster-2) and the single PCA split (PC1), which had the lowest accuracies of the environmental blocking approaches (AUC = 0.86 and 0.84 respectively; Table A5.2).

Prediction accuracy vs. environmental and geographic distance

In all cross-validations, across the spatial, environmental, random, and leave-one-out approaches, model accuracy was strongly negatively correlated with geographic distance (r = -0.93, p < 0.001) but not with environmental distance (r = -0.31, p = 0.106), though consistently negative. However, within each cross-validation approach (i.e. random, spatial, or environmental), relationships were consistently strongly negative (Table A5.3, Figure A5.4) showing trends, particularly in the environmental distance, unique to each cross-validation approach. In other words, while geographic distance tended to have a consistent response across all blocking approaches, the relationship between model accuracy and environmental distance was strongly dependent on the blocking method implemented.

The decline in predictive accuracy with increasing minimum environmental distance was much stronger in the spatial blocking than in the environmental blocking, despite the coarsest environmental blocking approaches (Cluster-2 and PCA Cut-2) resulting in larger environmental distances overall (Figure A5.4). While the trend of decreasing AUC with increasing environmental distance was very consistent for the environmental blocking approaches (r = -0.996), the slope was the least steep of any approach, indicating that blocking purely in environment provided relatively little challenge for the model prediction. By contrast, spatial blocks or validations with the spatially buffered leave-one-out resulted in very steep decreases in model predictive accuracy, with the spatial blocking and the buffered leave-one-out approach resulted in very similar decreases in model accuracy relative to environmental distance (Figure A5.4a).

The decline in predictive accuracy with increasing minimum geographic distance was more consistent across all methods (Figure A5.4b). The environmental blocking resulted in the steepest slope here, likely due to the combination of spatial and environmental extrapolation required in those cross-validations. Conversely, the leave-one-out approach resulted in the least steep slope here, likely due to the maximisation of the training data for each validation point (i.e. the lack of required environmental extrapolation). That said, while increasing block size could decrease accuracy simply by decreasing the amount of data available for model training, among each fold within each cross-validation approach, accuracy was not generally correlated with the number of points used for training (r = -0.10, p = 0.63).

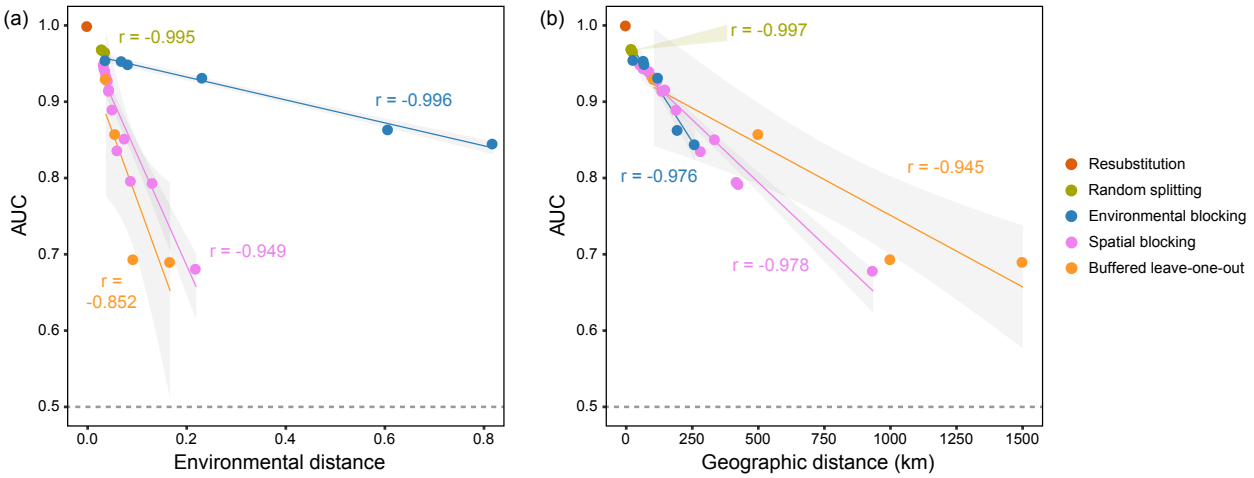

Figure A5.4. Relationship between distance measurements from hold-out to model training data, as measured by **(a)** the average minimum environmental distance and **(b)** the geographic distance. While relationships are drawn as linear, the theoretical minimum AUC is 0.5 (for a random model).

References

Bahn, V. 2009. A new method for evaluating species distribution models. *In: 94th Ecological Society of America Annual Meeting.*

Bjrnstad, O. N. 2013. *ncf: spatial nonparametric covariance functions. R package version 1.1-5.*

Breiman, L. 2001. Random forests. *Machine Learning*, **45**(1), 5–32.

Eiserhardt, W. L., Svenning, J. C., Baker, W. J., Couvreur, T. L. P., & Balslev, H. 2013. Dispersal and niche evolution jointly shape the geographic turnover of phylogenetic clades across continents. *Scientific Reports*, **3**.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, **23**(7), 811–820.

Leites, L. P., Robinson, A. P., Rehfeldt, G. E., Marshall, J. D., & Crookston, N. L. 2012. Height-growth response to climatic changes differs among populations of Douglas-fir: a novel analysis of historic data. *Ecological Applications*, **22**(1), 154–165.

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 and Distributions*, **20**(10), 1147–1159.

Montwe, D., Spiecker, H., & Hamann, A. 2015. Five decades of growth in a genetic field trial of Douglas-fir reveal trade-offs between productivity and drought tolerance. *Tree Genetics & Genomes*, **11**(2).

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rehfeldt, G. E., Jaquish, B. C., Lopez-Upton, J., Saenz-Romero, C., St Clair, J. B., Leites, L. P., & Joyce, D. G. 2014. Comparative genetic responses to climate for the varieties of Pinus ponderosa and Pseudotsuga menziesii: Realized climate niches. *Forest Ecology and Management*, **324**, 126–137.

Roberts, D. R., & Hamann, A. 2012. Predicting potential climate change impacts with bioclimate envelope models: a palaeoecological perspective. *Global Ecology and Biogeography*, **21**(2), 121–133.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941.

Swets, J. A. 1988. Measuring the Accuracy of Diagnostic Systems. *Science*, **240**(4857), 1285–1293.

Wang, T. L., Hamann, A., Spittlehouse, D., & Carroll, C. 2016. Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. *Plos One*, **11**(6).

Williams, J. W., Shuman, B. N., & Webb, T. 2001. Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. *Ecology*, **82**(12), 3346–3362.

Table A5.3. For each of the validation and cross-validation approaches, the average minimum environmental distance in PCA units (*Enviro*) and geographic distance in kilometres (*Geog*) between the hold-out and the model training data, the AUC of the combined folds (AUC), and the geographic dimensions of the spatial blocks in x×y kilometres, where appropriate.

	Distance			
	Env	Geog	AUC	Size
Resubstitution	0.000	0	1.00	
Random splitting				
Random-2	0.036	27	0.97	
Random-4	0.031	23	0.97	
Random-6	0.030	22	0.97	
Random-8	0.029	21	0.97	
Environmental blocking				
Cluster-2	0.607	196	0.86	
Cluster-4	0.070	67	0.95	
Cluster-6	0.042	38	0.96	
Cluster-8	0.037	28	0.95	
PC1	0.818	261	0.84	
PC1-2	0.232	122	0.93	
PC1-3	0.082	70	0.95	
Spatial blocking				
Spatial Grid 1×2	0.219	935	0.68	3862×3834
Spatial Grid 2×2	0.132	426	0.79	1931×3834
Spatial Grid 2×3	0.088	419	0.79	1931×2556
Spatial Grid 2×4	0.075	337	0.85	1931×1917
Spatial Grid 3×6	0.061	284	0.84	1287×1278
Spatial Grid 4×8	0.051	236	0.89	966×958
Spatial Grid 5×10	0.044	149	0.92	772×767
Spatial Grid 6×12	0.044	139	0.91	644×639
Spatial Grid 7×14	0.042	120	0.93	552×548
Spatial Grid 8×16	0.039	106	0.93	483×479
Spatial Grid 9×18	0.038	99	0.93	429×426
Spatial Grid 10×20	0.037	89	0.94	386×383
Spatial Grid 15×30	0.034	67	0.94	257×256
Spatial Grid 20×40	0.033	55	0.95	193×192
Buffered leave-one-out				
Buffer = 100 km	0.037	105		
Buffer = 500 km	0.056	501		
Buffer = 1000 km	0.092	1002		
Buffer = 1500 km	0.167	1501		