Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data – Supporting Information

David W. Redding[1*], Tim C.D. Lucas[1,2], Tim Blackburn[1,3], and Kate E. Jones[1,3*]

[1]Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom.

[2]Big Data Institute, University of Oxford, Peter Medawar Building, South Parks Road, Oxford, OX1 3SY, United Kingdom.

[3]Institute of Zoology, Zoological Society of London, Regent's Park, London, NW1 4RY, United Kingdom.

*Corresponding authors: dwredding@gmail.com and kate.e.jones@ucl.ac.uk (Tel: +44 (0)20 17 31084230)

Running title: Bayesian SDM methods for non-random data

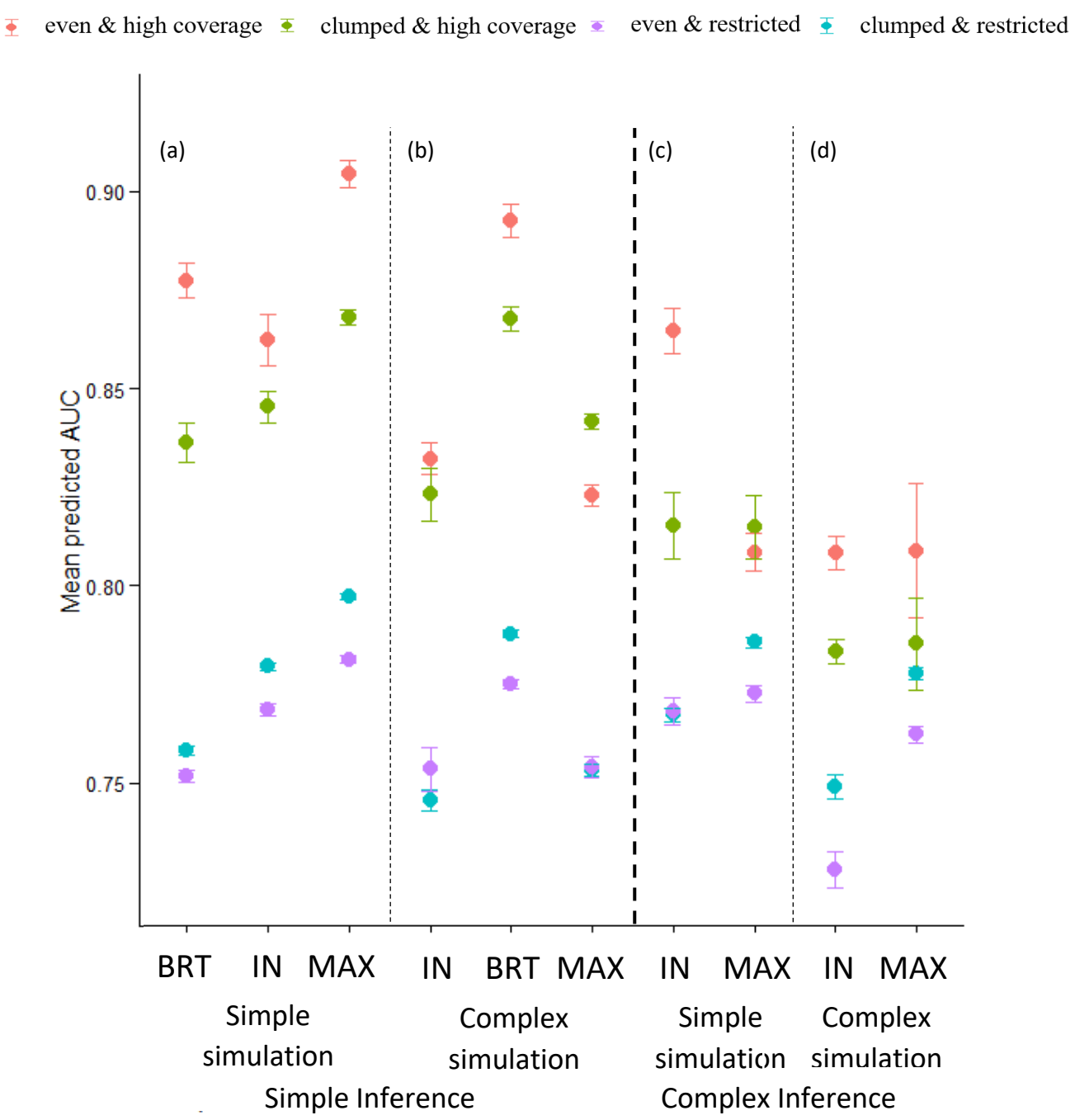
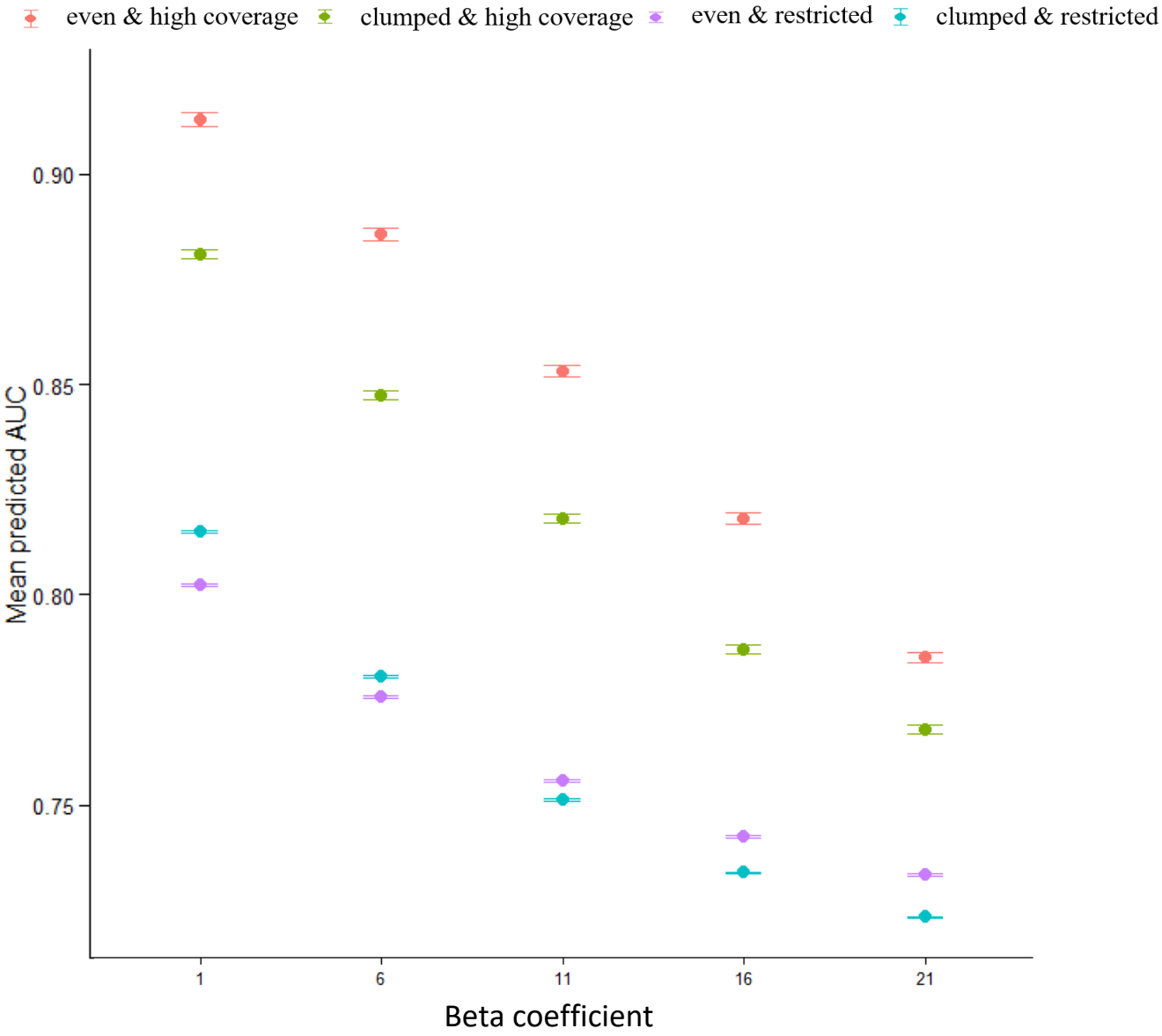


Figure S1. Predictive accuracy of three species distribution modelling methods (IN – spatial INLA, MAX – MAXENT, BRT – Boosted Regression trees) to infer 5000 simulated species' ranges that were generated using either (a) just additive or (b) additive and interaction terms in formula to determine the relationship between species' presence and a set of simulated covariates. A repeated set of comparisons (c-d) is made for SDM methods where interactions can also be specified for the inference formulae (i.e. INLA & MAXENT). Points represent mean AUC score over all simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.



Figures S2. Predictive accuracy of the MAXENT species distribution modelling method when varying the complexity of the inference models using the beta, or "regularisation", coefficient. Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.

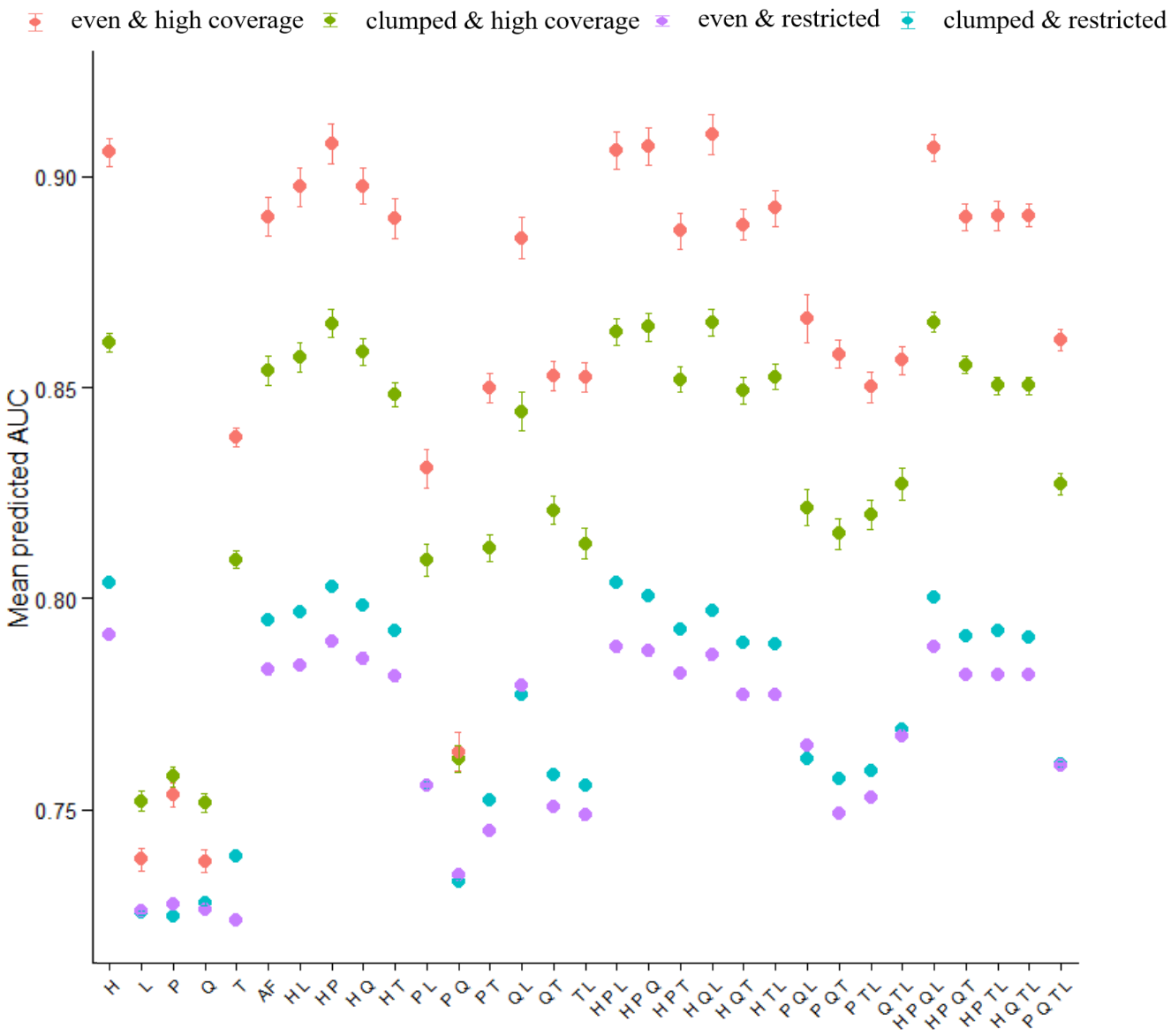


Figure S3. Predictive accuracy of the MAXENT species distribution modelling method when varying what class of terms are include in the inference model (Hinge – H, Product – P, Quadratic – Q, Threshold – T, Linear – L, Auto Feature – AF). Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.

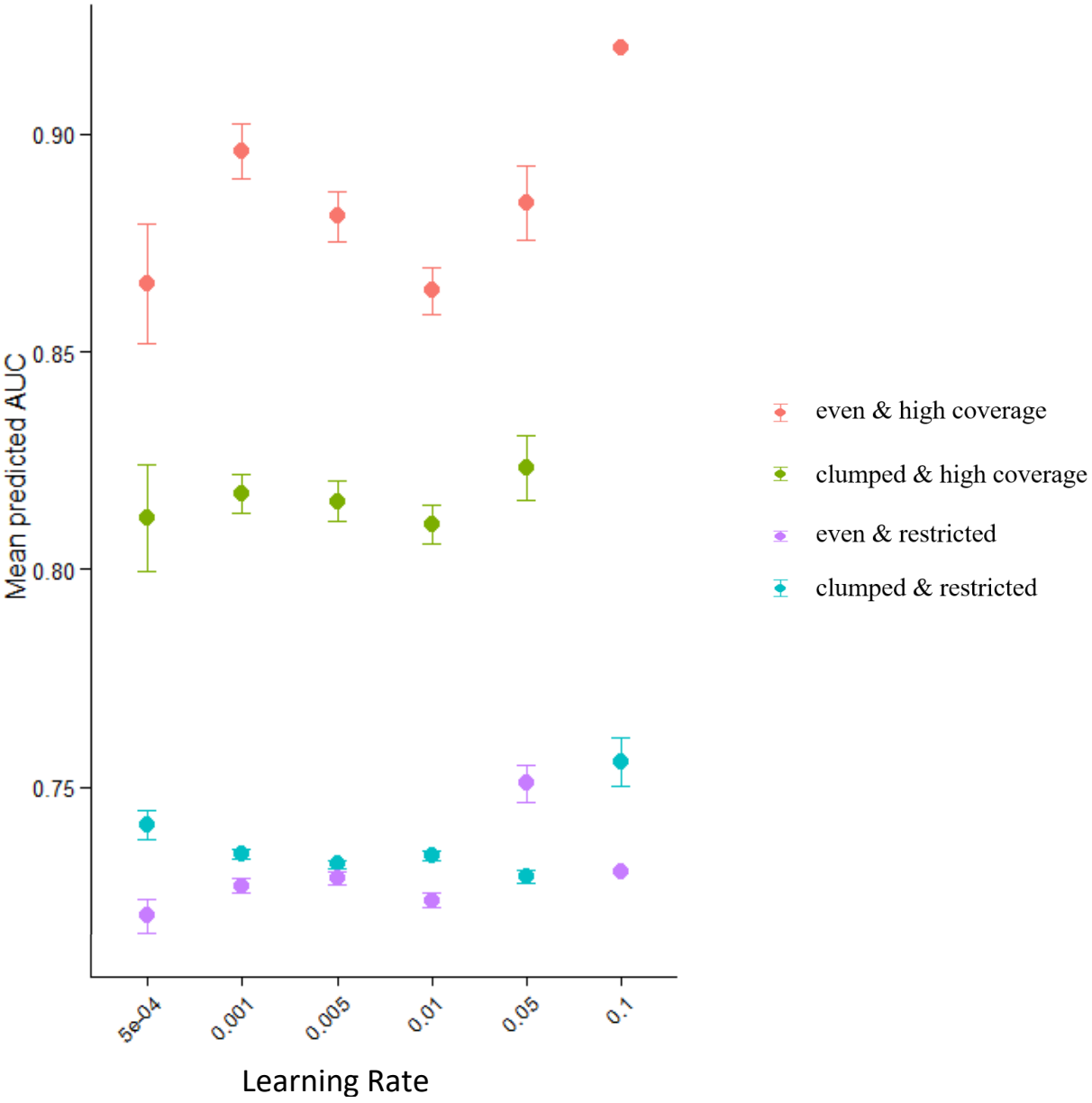


Figure S4. Predictive accuracy of Boosted-Regression Trees (BRT) species distribution modelling method when varying the learning rate of tree inference algorithm. Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.

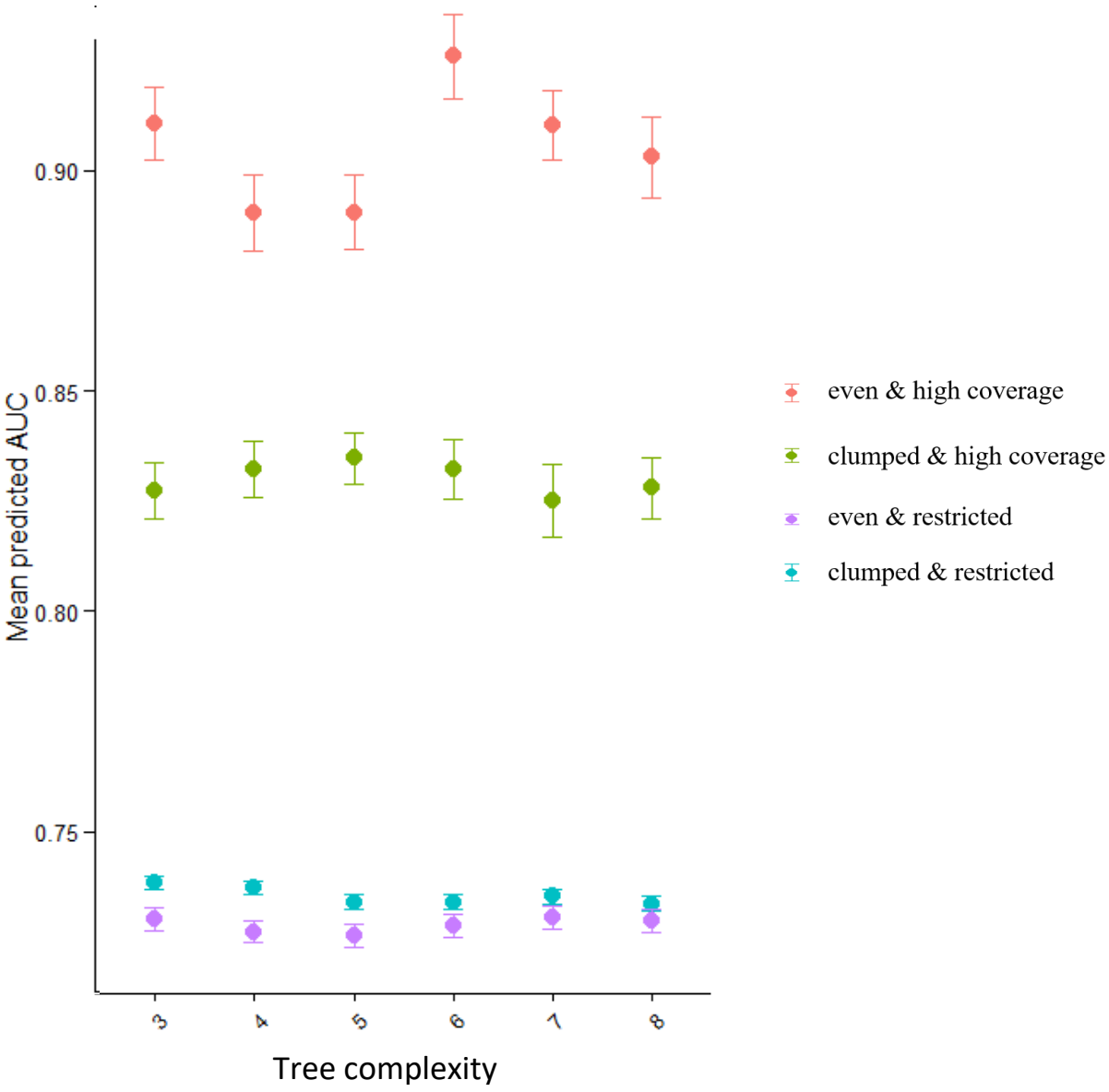


Figure S5 Predictive accuracy of Boosted-Regression Trees (BRT) species distribution modelling method when varying the complexity of the underlying regression trees during inference. Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.

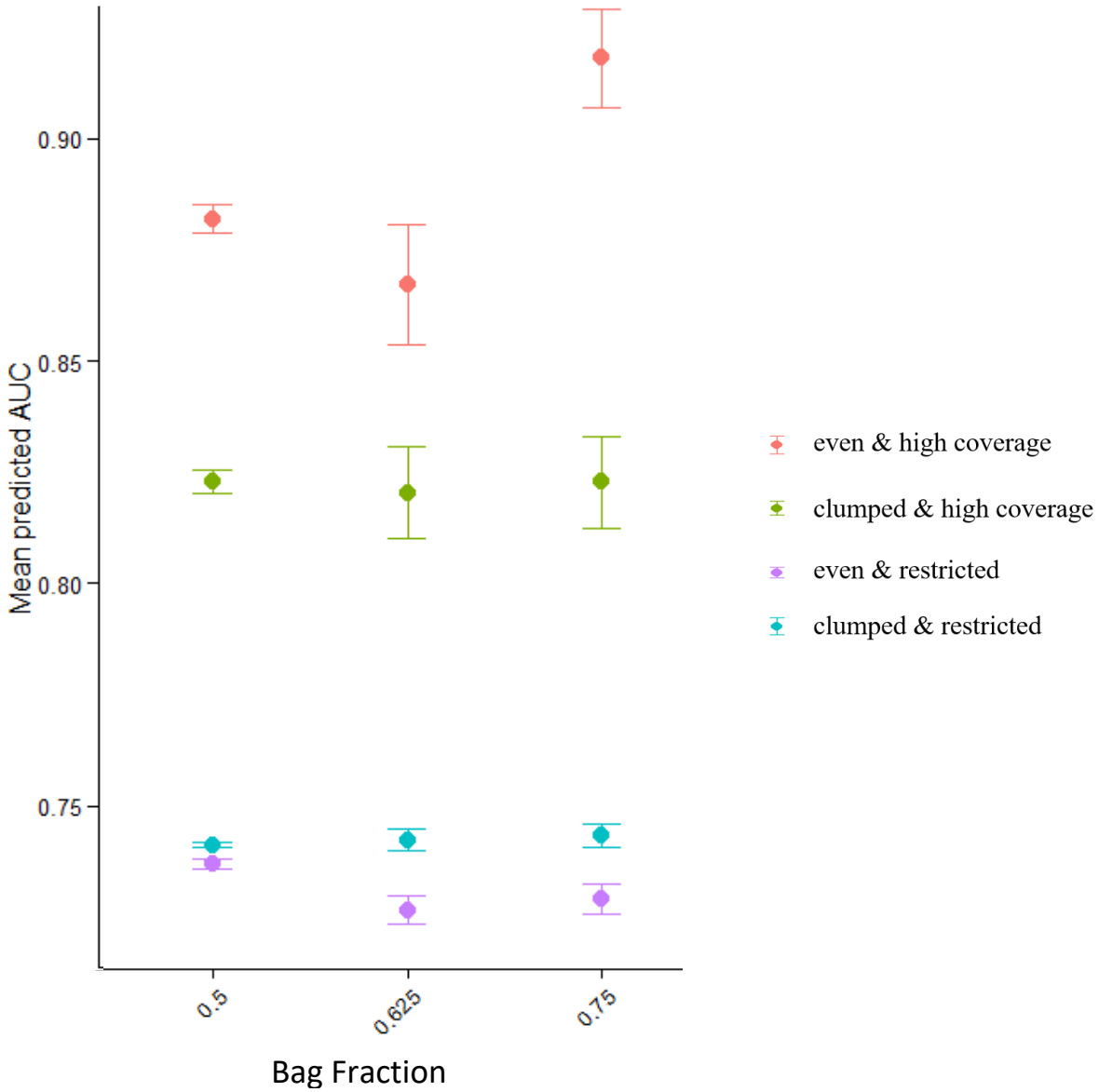


Figure S6. Predictive accuracy of Boosted-Regression Trees (BRT) species distribution modelling method when varying the bag fraction used to hold-back parts of data for internal validation. Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.

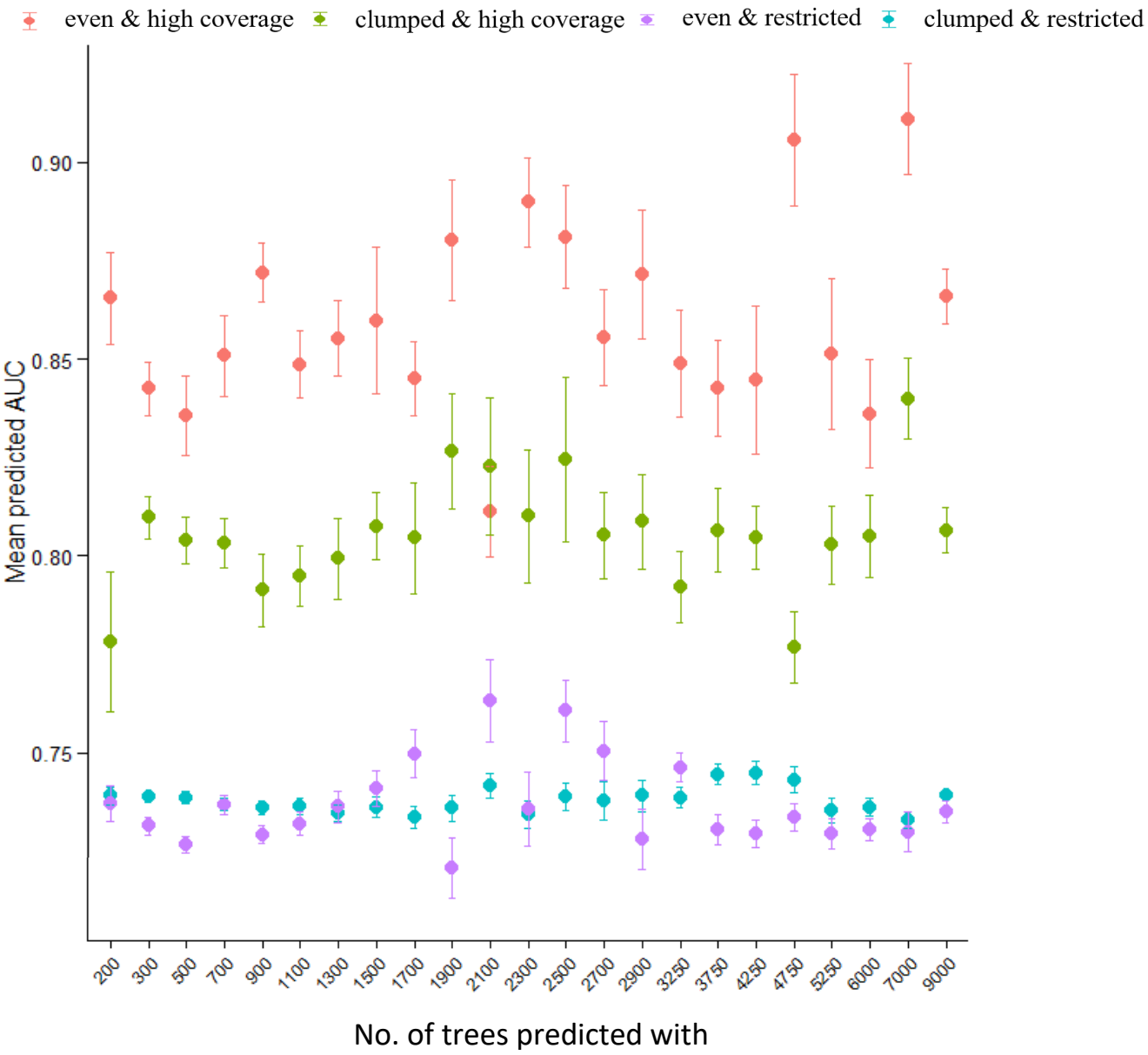


Figure S7. Predictive accuracy of Boosted-Regression Trees (BRT) species distribution modelling method when varying the number of regression trees retained in the final modelling set. Points represent mean AUC score over a set of 5000 simulated species where a prediction of the true range is attempted using a set of simulated sampling points, with whiskers showing the 95% confidence intervals. Different colours show the predictive accuracy of subsets of the 5000 datasets when binning the input samples from each dataset into either high or low clumping and high or low coverage of the simulated "true" range.