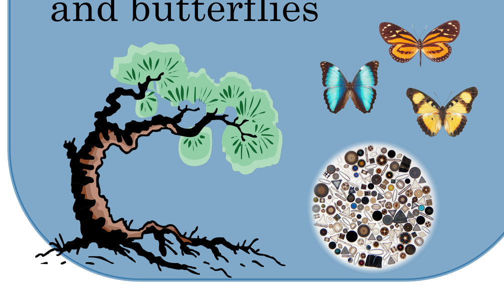
How well do different statistical frameworks predict species- and community-level patterns?

Anna Norberg[1], Guillaume Blanchet, Fred Adler and Otso Ovaskainen[1]
With contributions from the NIMBIOS Biotic Interactions Working Group 2013-2015
[1]Mathematical Biology Group, Metapopulation Research Centre, Department of Biosciences, University of Helsinki, Finland

The idea is to compare various statistical frameworks in terms of their predictive performance

The simulated data are generated by dynamical models, mimicking competitive/mutualistic/trophic interactions, and the real data sets consist of diatoms, trees and butterflies



1 Simulated and real data → **Training data** → **2** Statistical models – SSDMs & JSDMs

1 Simulated and real data → **Validation data**
– interpolation
– extrapolation

2 Statistical models → **3** Predictions → **4** Validation & performance testing

Validation data → **4** Validation & performance testing
– species level
– community level

2 Models

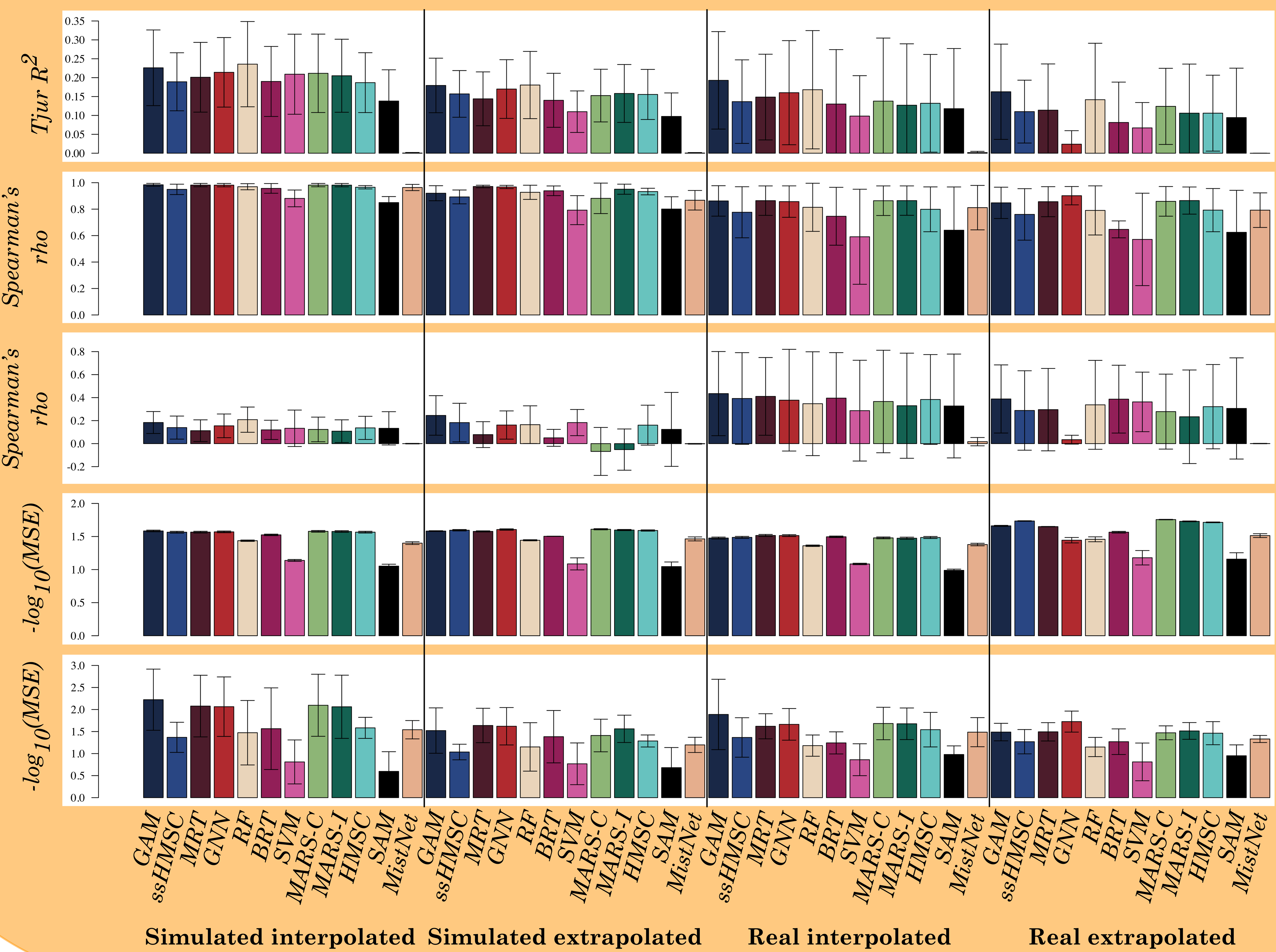
Stacked species distribution models

- Generalized additive models[1]
- Generalized linear models[2-5]
- Multivariate regression trees[6-7]
- Gradient nearest neighbour[8]
- Random forests[9]
- Boosted regression trees[10]
- Support vector machines[11]
- Multivariate adaptive regression splines[12]

Joint species distribution models

- Hierarchical modelling of species communities[2-5]
- Species archetype models[13-15]
- Stochastic feedforward neural network[16]

3 4 Predictions, validation & performance



(a) ***Tjur R^2 values.***
The mean coefficients of discrimination.

(b) ***Species prevalence.***
The mean Spearman's correlations between predicted and true species prevalences.

(c) ***Species richness.***
The mean Spearman's correlations between predicted and true species richnesses.

(d) ***Bray-Curtis dissimilarity.***
The mean squared error of the Bray-Curtis dissimilarity index values.

(e) ***Co-occurrence index.***
The mean squared error of the co-occurrence indices over species pairs.

We **conclude**, that even though there are **differences in the predictive performance** between the models, we so far see **no clear division between stacked and joint models in general.** Most models seem to perform concistently well, but some perform varyingly regarding accuracy and generality (e.g. GNN, MARS, MistNet). Further development in both model fitting procedures and perfomance testing is needed in order to obtain some more conclusive results.

[1]S.N. Wood, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011)
[2]E. Sebastián-González, J.A. Sánchez-Zapata, F. Botella, O. Ovaskainen, Proc. Biol. Sci. 277 (2010)
[3]O. Ovaskainen, J. Soininen, Ecology. 92 (2011)
[4]O. Ovaskainen, J. Hottola, J. Siitonen, Ecology. 91 (2010)
[5]S.M. O'Brien, D.B. Dunson, Biometrics. 60 (2004)
[6]G. De'Ath, Ecology. 83 (2002)
[7]D.R. Larsen, P.L. Speckman, Biometrics. 60 (2004)
[8]N. L. Crookston & A. O. Finley, J. Stat. Soft. 23 (2008)
[9]L.E.O. Breiman,. Machine Learning 45 (2001)
[10]J. Elith, J.R. Leathwick, T. Hastie, A, J. Anim. Ecol. 77 (2008)
[11]J.M. Drake, C. Randin, A. Guisan, J. Appl. Ecol. 43 (2006)
[12]J.R. Leathwick, D. Rowe, J. Richardson, J. Elith, T. Hastie, Freshw. Biol. 50 (2005)
[13]F.K.C. Hui, D.I. Warton, S.D. Foster, P. Dunstan, Ecology. 94 (2013)
[14]P.K. Dunstan, S.D. Foster, R. Darnell, Ecol. Modell. 222 (2011)
[15]P.K. Dunstan, S.D. Foster, F.K.C. Hui, D.I. Warton, J. Agric. Biol. Environ. Stat. 18 (2013)
[16]D.J. Harris, Methods Ecol. Evol. (2015)

anna.norberg@helsinki.fi
@AnnaNorb