Validating SDMs

Bob O'Hara

```
Loading required package: future
```

Abstract

Introduction

- SDMs often used. Validation, i.e. how good are they, is an issue. (A. Lee-Yaw et al. (2022)).
- Current validation often on same data, which can lead to over-fitting. Also, metrics often wrong: because it is the same data, will replicate biases. Thus TSS should be FSS in this context.
- A better approach is to validate on external data. Could calculate AUC, TSS etc. but teh validation daat will be imperfect too. Here we take a slightly different approach that can be adapted to different data types.

Smith and Levine (2025) showed that Yackulic et al. (2013) were right: you shouldn't think at all species are equal. But…

The same data was used by Valavi et al. (2022) to compare the perform,ance of different models. Here our focus is on the comparisons can and should be made, and what the results tell us.

Theory

Most data is modelled as presence-only, so we will follow that here but indicate the adjustments needed if the data are of a different type. For presence-only data the underlying theory can be derived from point processes (e.g. Fithian and Hastie (2013), Renner and Warton (2013), Warton and Shepherd (2010), Aarts, Fieberg, and Matthiopoulos (2012)). The actual distribution is viewed as an intensity surface, with a higher intensity where a species is more likely to be present. The data is made up of locations of observations of the species of interest, along with locations in the same region where the species was not observed. These latter points are usually called "pseudo-absences" in the SDM world. In a point process approach they are

seen as integration points - to calculate the likelihood for the presences we should integrate the intensity over the full space.

Of concern here is the model for the intensity surface. It is assumed that it is affected by several environmental covariates in an additive way. Specifically, of $\lambda(\mathbf{s})$ is the intensity at point $\mathbf{s}$, the assumption is that $\log \lambda(\mathbf{s}) = \alpha + \sum_i \beta_i X_i(\mathbf{s}) = \alpha + \eta(\mathbf{s})$ where $X_i(\mathbf{s})$ is a *feature*, i.e. a function of the environmental covariates. For simplicity we can think of this as an environmental covariate, but it could also be a non-linear term, such as a quadratic or the product of two environmental covariates (i.e. an interaction). This same approach is used in GLMs, MaxEnt, GAMs and many other methods. The β_is are standard regression coefficients, and α is an intercept. This is not normally of interest for presence-only data, as it is determined by the amount of data, i.e. the sampling effort. Thus the focus is on $\eta(\mathbf{s})$.

If the model is correct, then the number of individuals of a species in an area A would follow a Poisson distribution with mean $A\lambda(\mathbf{s})$, where A is a constant that will depend on sampling effort. The probability of an absence is then the probability of zero individuals, i.e. $e^{-A\lambda(\mathbf{s})}$. This leads to a model for presence/absence where $P(Z = 1) = 1 - e^{e^{-\alpha_c - \eta(\mathbf{s})}}$, which is equivalent to a GLM with a binomial response and a cloglog link function ($\log(-\log(1 - P(Z = 1))) = \alpha_c + \eta(\mathbf{s})$). An alternative is to use a logistic regression (Elith et al. (2011)), i.e. $\log P(Z = 1)/(1 - P(Z = 1)) = \alpha_l + \eta(\mathbf{s})$, with a derivation based on averaging over possible distributions of the covariates and response (Phillips and Dudík (2008)).

The purpose of laying this out is to suggest an approach to model validation. If we fit a model to presence-only data, we get estimates for $\eta(\mathbf{s}) = \sum_i \beta_i X_i(\mathbf{s})$. We can then use this to calculate predicted probabilities (up to an intercept) for a new presence/absence data set. We can then use these predictions in a GLM with a cloglog or logit link. If the model is correct, the slope should equal 1.

Although we are not primarily interested in the intercept, except to note that it should not equal 0 (as pointed out by both Yackulic et al. (2013) and Smith and Levine (2025)), we might expect that if we look at the estimates for the fitting and validation models across species from the same data sets, they should be correlated. i.e. a species that is more common in the presence only data set may also be more common in the presence-absence data, because it may be more common or be easier to spot.

Why, though, might a regression coefficient not equal 1? Aside from random chance, it may be because the model is wrong. Indeed, as all models are wrong this is likely. But it may also be because, even if the fitted model is correct, it is fitted to finite data, so there is uncertainty in the parameter estimates. This uncertainty will feed through to validation model, where $\eta(\mathbf{s})$ has been estimates with error. The overall effect of this is not just to increase uncertainty, but to bias the estimate of the slope towards zero (Carroll et al. (1984)). This could be mitigated by incorporating this uncertainty into the second model (REF).

Methods

Data

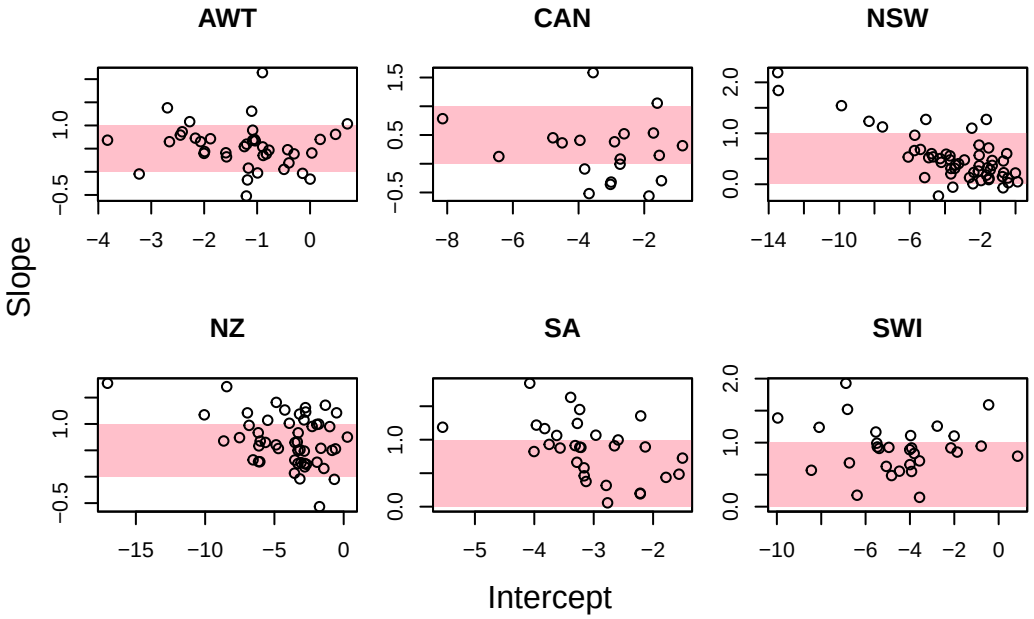
From Elith et al. (2020)

Modelling

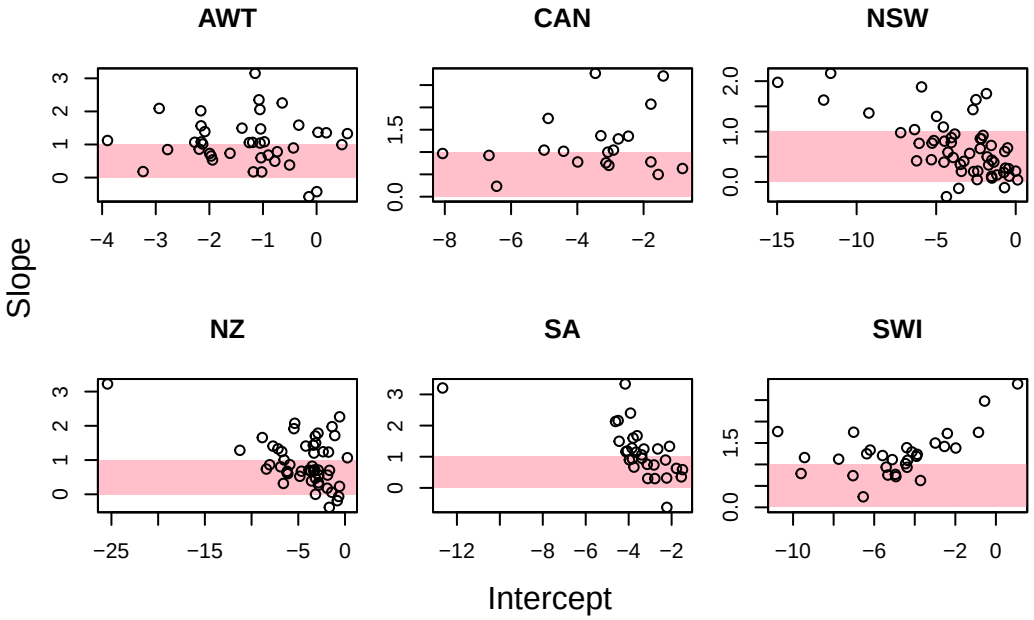
```
plan(multisession, workers = 4) ## Run in parallel on local computer, with 4 nodes (cores?)
AllCoefs.l <- sapply(c("AWT", "CAN", "NSW", "NZ", "SA", "SWI"), JustMaxEnt,
                     remove=RemoveNames, classes="l", valid = TRUE, pred=TRUE)
AllCoefs.l.sp <- sapply(c("AWT", "CAN", "NSW", "NZ", "SA", "SWI"), JustMaxEnt,
                     remove=RemoveNames, classes="l", valid = TRUE, otherSpBG=TRUE, pred=TRUE)
```

Results

Plot of models with linear features



Plot of models with linear features, using presences of other species as background points

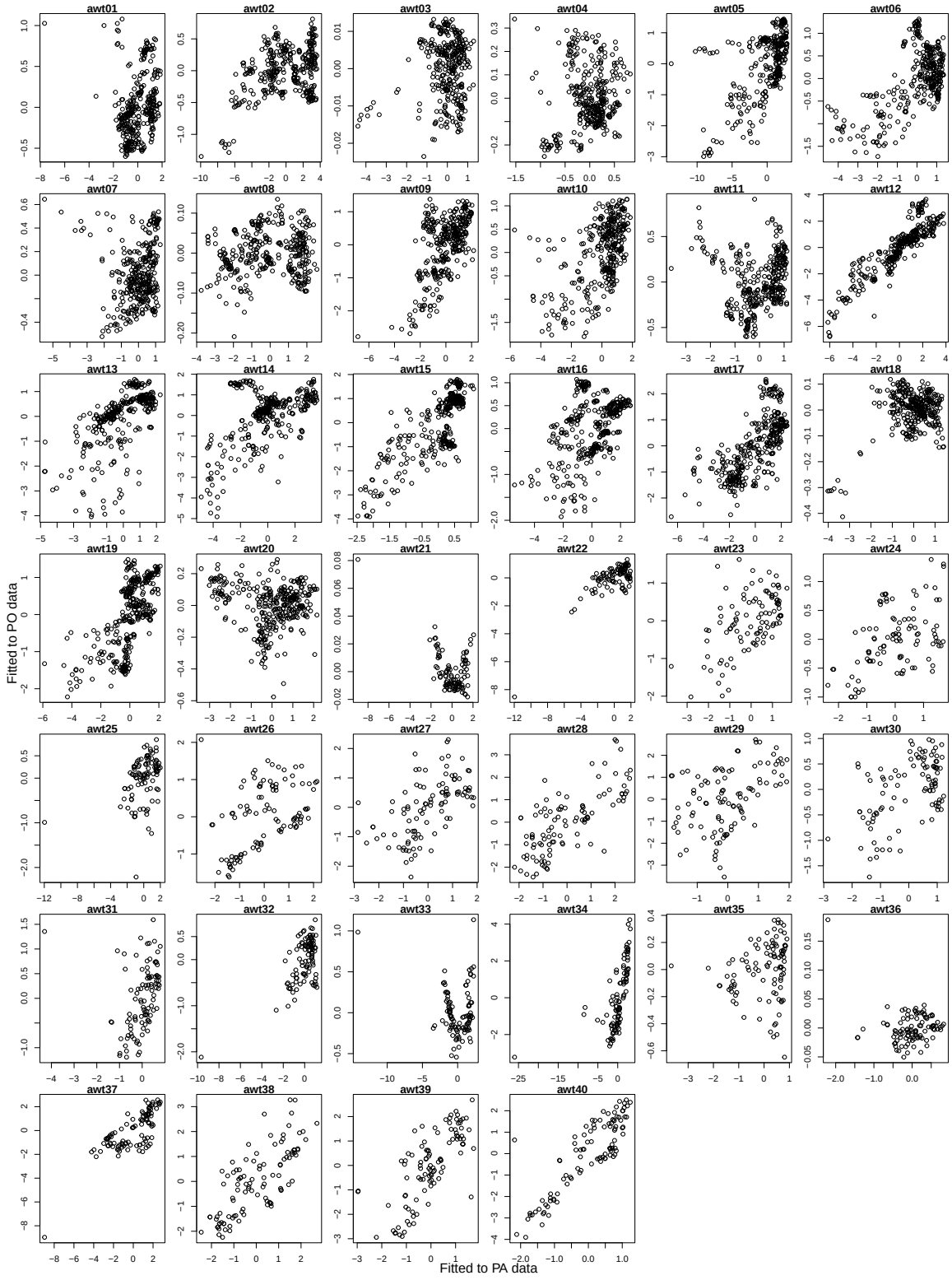


Plot of models with all features except hinges

Plots of predictions from the MaxEnt model on the validation data model plotted against predictions for a model with MaxEnt features fitted to the presence-absence data.

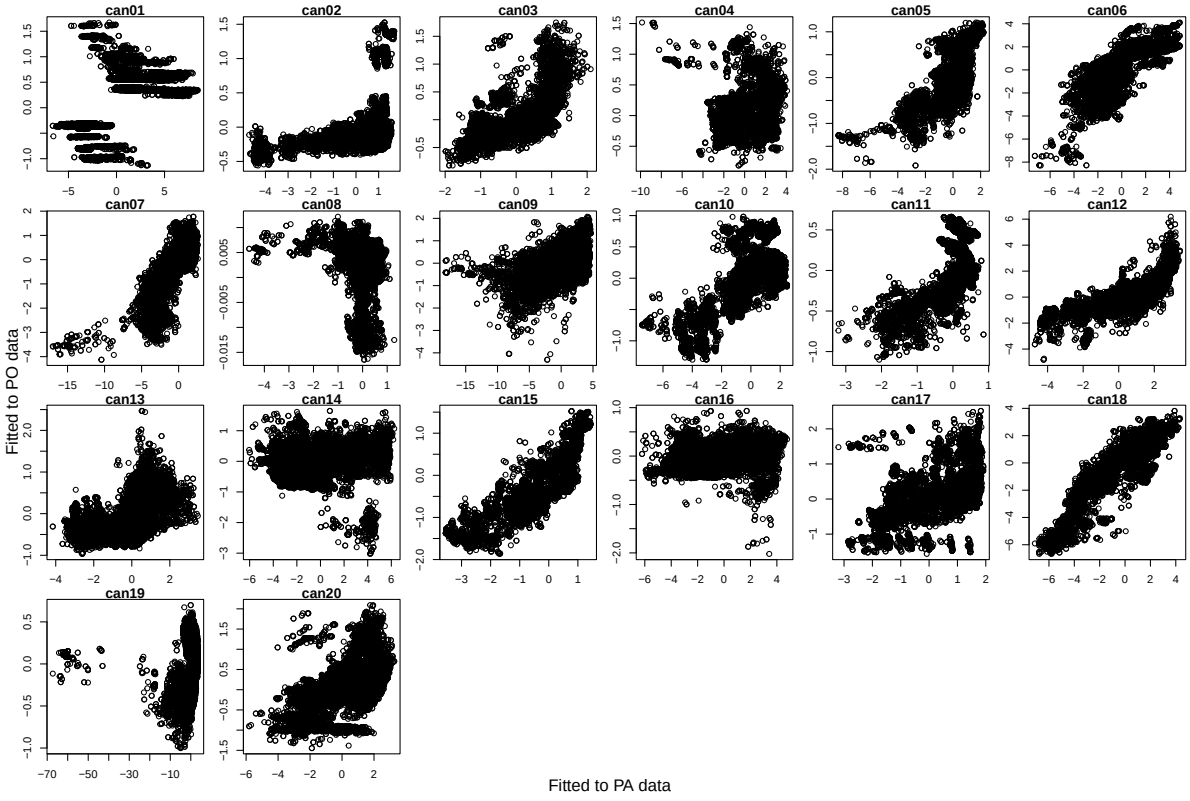
AWT

```
par(mfrow=c(ceiling(40/6), 6), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.lqpt$AWT), function(nm, allC) {
  lst <- allC[[nm]]
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.lqpt$AWT)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```


awt01
awt02
awt03
awt04
awt05
awt06
awt07
awt08
awt09
awt10
awt11
awt12
awt13
awt14
awt15
awt16
awt17
awt18
awt19
awt20
awt21
awt22
awt23
awt24
awt25
awt26
awt27
awt28
awt29
awt30
awt31
awt32
awt33
awt34
awt35
awt36
awt37
awt38
awt39
awt40
Fitted to PO data
Fitted to PA data

CAN

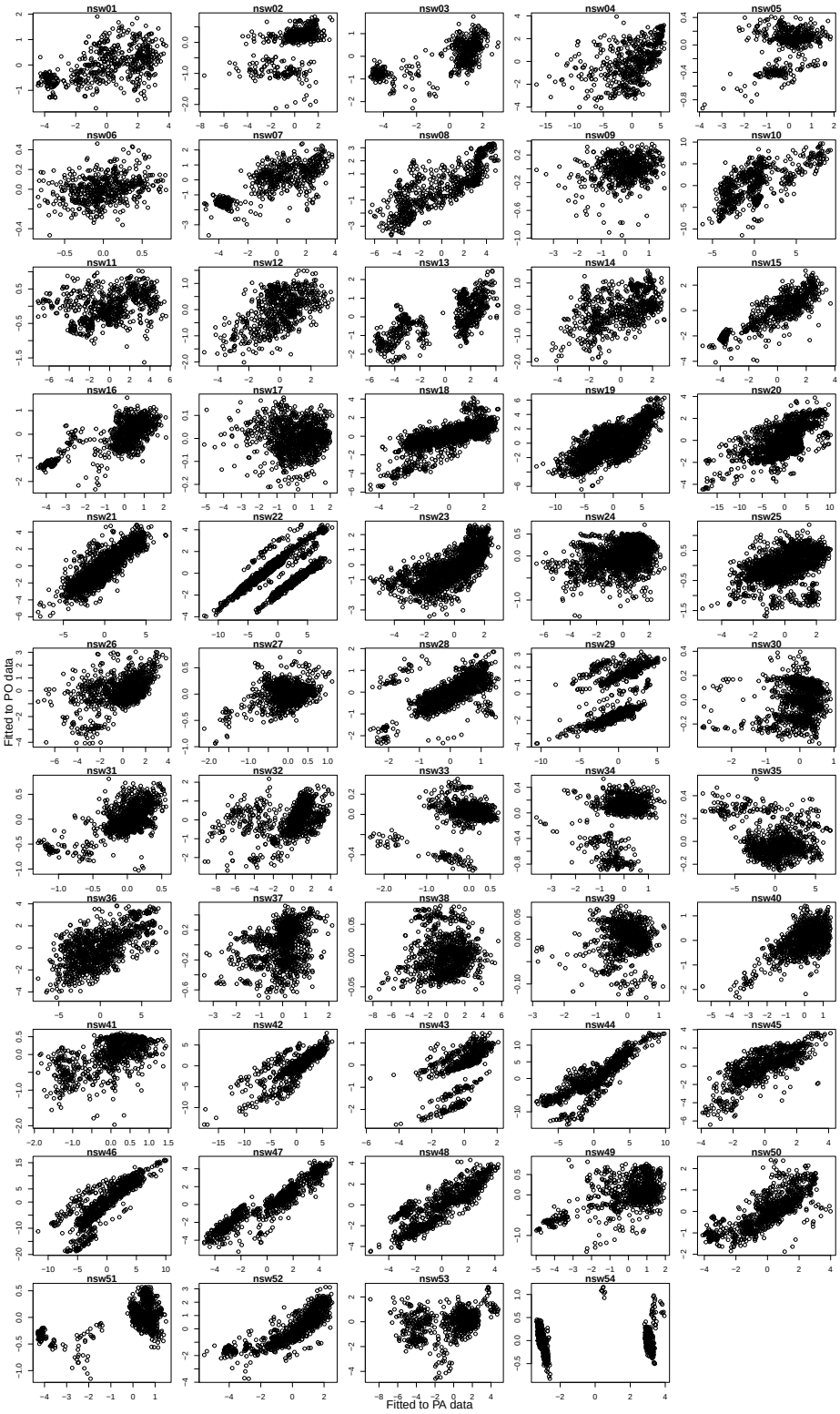
```
par(mfrow=c(ceiling(20/6), 6), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.lqpt$CAN), function(nm, allC) {
  lst <- allC[[nm]]
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.lqpt$CAN)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```



NSW

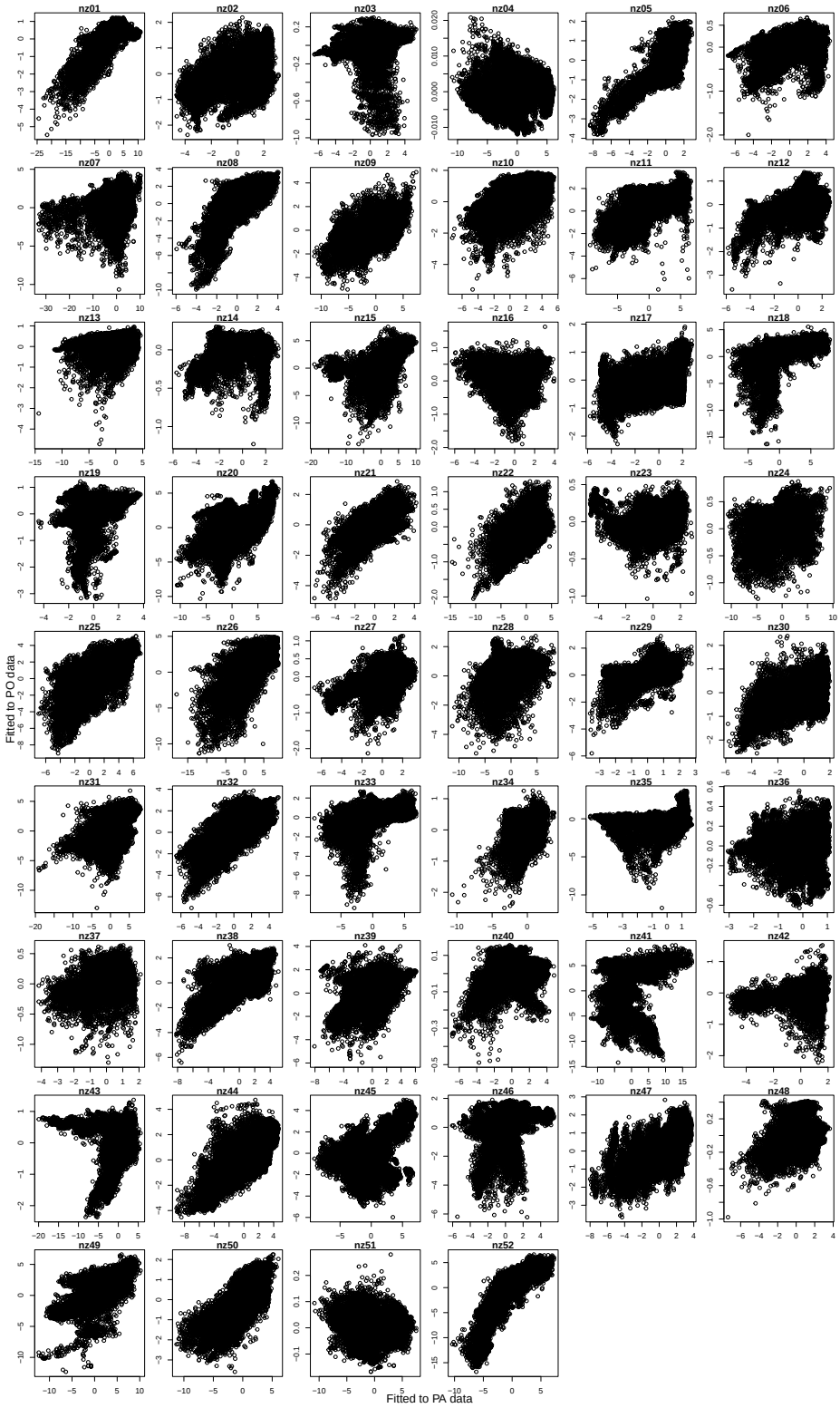
```
par(mfrow=c(11,5), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.l$NSW), function(nm, allC) {
  lst <- allC[[nm]]
```

```
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.l$NSW)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```


nsw01
nsw02
nsw03
nsw04
nsw05
nsw06
nsw07
nsw08
nsw09
nsw10
nsw11
nsw12
nsw13
nsw14
nsw15
nsw16
nsw17
nsw18
nsw19
nsw20
nsw21
nsw22
nsw23
nsw24
nsw25
nsw26
nsw27
nsw28
nsw29
nsw30
nsw31
nsw32
nsw33
nsw34
nsw35
nsw36
nsw37
nsw38
nsw39
nsw40
nsw41
nsw42
nsw43
nsw44
nsw45
nsw46
nsw47
nsw48
nsw49
nsw50
nsw51
nsw52
nsw53
nsw54
Fitted to PO data
Fitted to PA data

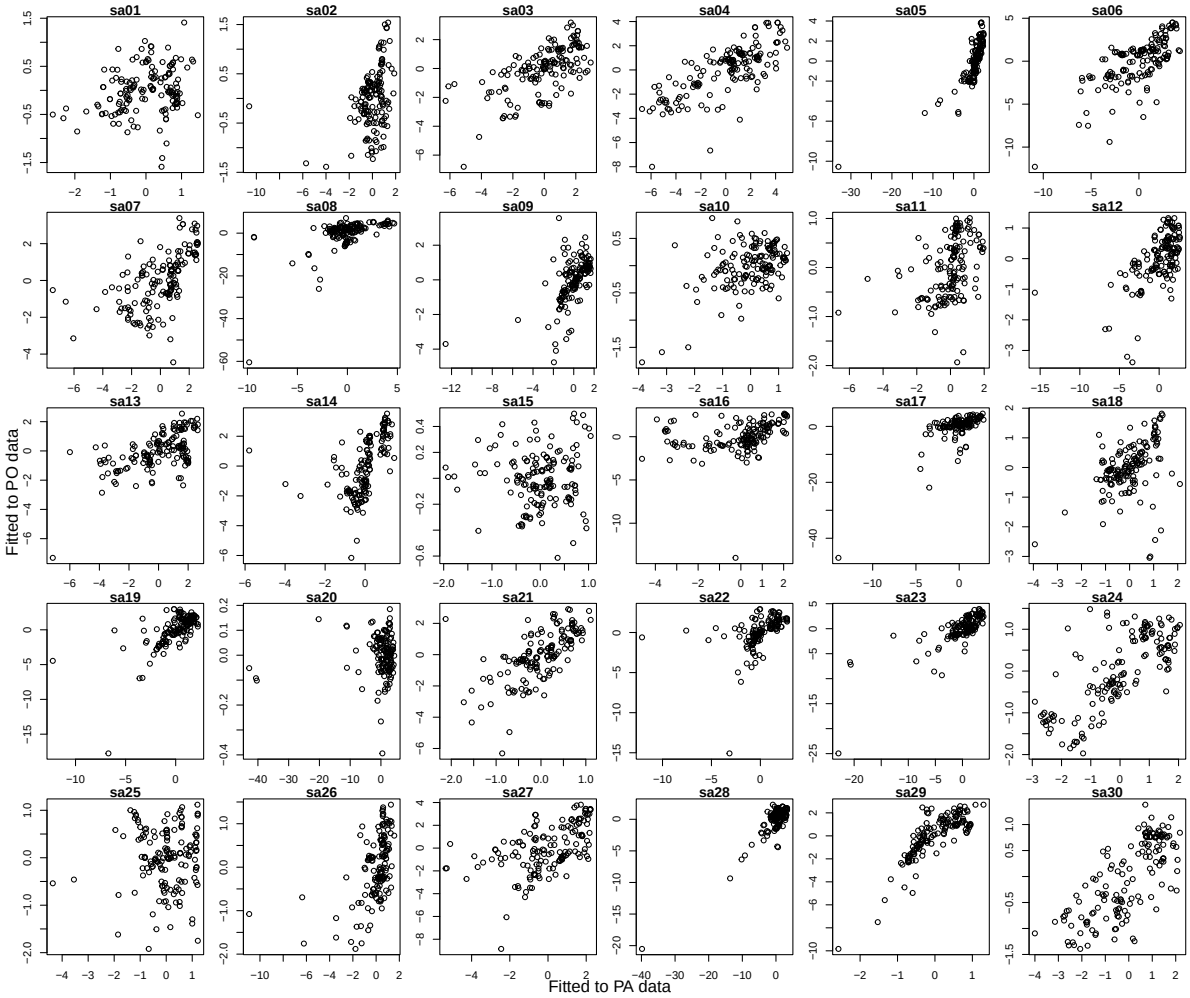
NZ

```
par(mfrow=c(ceiling(52/6),6), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.lqpt$NZ), function(nm, allC) {
  lst <- allC[[nm]]
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.lqpt$NZ)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```


nz01
nz02
nz03
nz04
nz05
nz06
nz07
nz08
nz09
nz10
nz11
nz12
nz13
nz14
nz15
nz16
nz17
nz18
nz19
nz20
nz21
nz22
nz23
nz24
nz25
nz26
nz27
nz28
nz29
nz30
nz31
nz32
nz33
nz34
nz35
nz36
nz37
nz38
nz39
nz40
nz41
nz42
nz43
nz44
nz45
nz46
nz47
nz48
nz49
nz50
nz51
nz52
Fitted to PO data
Fitted to PA data

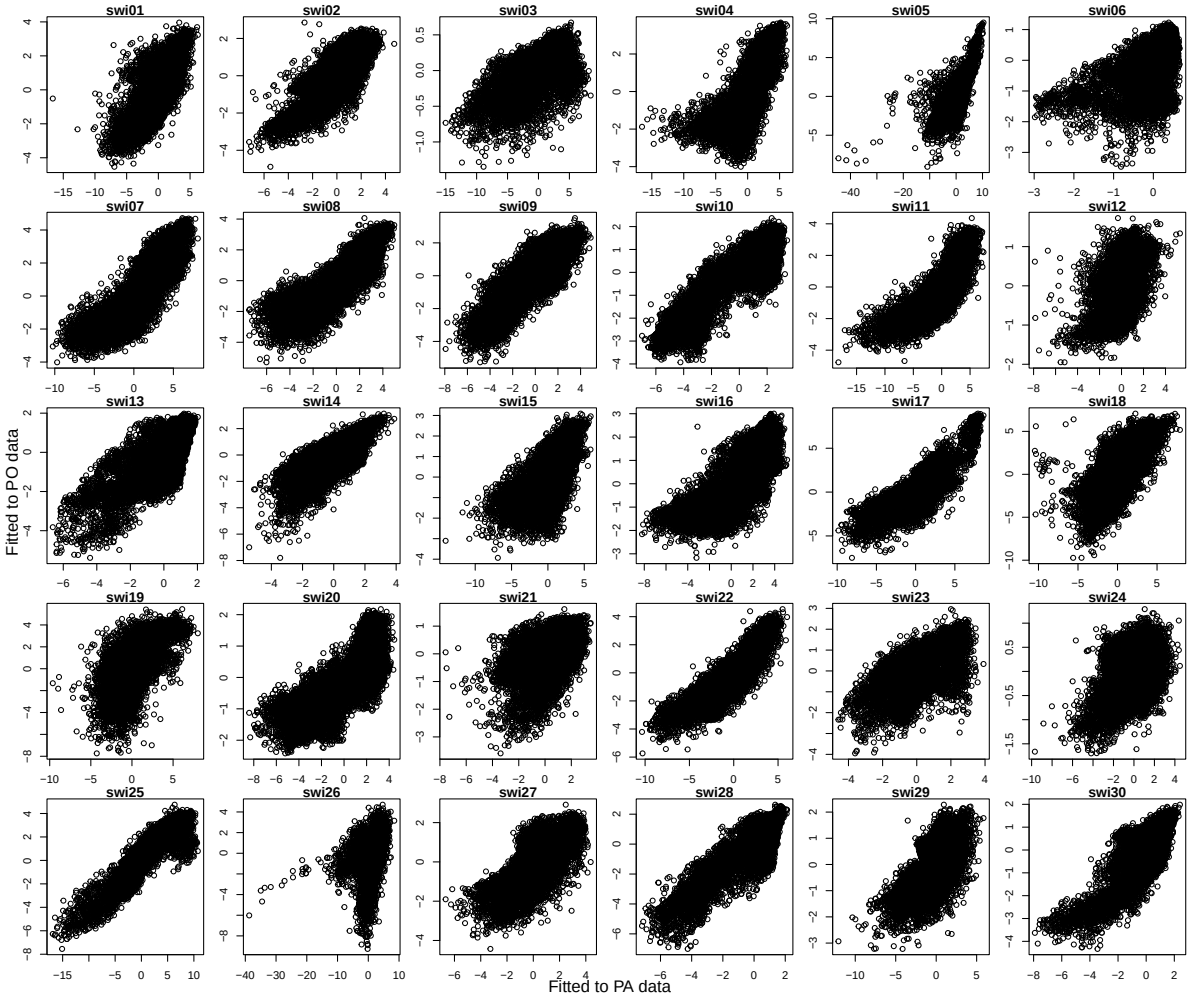
SA

```
par(mfrow=c(ceiling(30/6),6), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.lqpt$SA), function(nm, allC) {
  lst <- allC[[nm]]
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.lqpt$SA)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```



SWI

```
par(mfrow=c(ceiling(30/6),6), mar=c(2,2,1,1), oma=c(2,2,0,0))
sapply(names(AllCoefs.lqpt$SWI), function(nm, allC) {
  lst <- allC[[nm]]
  plot(lst$pred[,"PA"], lst$pred[,"valid"], xlab="", ylab="",
       main=nm)
}, allC=AllCoefs.lqpt$SWI)
mtext("Fitted to PA data", 1, outer=TRUE)
mtext("Fitted to PO data", 2, outer=TRUE)
```



References

A. Lee-Yaw, Julie, Jenny L. McCune, Samuel Pironon, and Seema N. Sheth. 2022. "Species Distribution Models Rarely Predict the Biology of Real Populations." *Ecography* 2022 (6): e05877. https://doi.org/https://doi.org/10.1111/ecog.05877.

Aarts, Geert, John Fieberg, and Jason Matthiopoulos. 2012. "Comparative Interpretation of Count, Presence–Absence and Point Methods for Species Distribution Models." *Methods in Ecology and Evolution* 3 (1): 177–87. https://doi.org/https://doi.org/10.1111/j.2041-210X.2011.00141.x.

Carroll, Raymond J., Clifford H. Spiegelman, K. K. Gordon Lan, Kent T. Bailey, and Robert D. Abbott. 1984. "On Errors-in-Variables for Binary Regression Models." *Biometrika* 71 (1): 19–25. http://www.jstor.org/stable/2336392.

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. 2011. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions* 17 (1): 43–57. https://doi.org/https://doi.org/10.1111/j.1472-4642.2010.00725.x.

Elith, J., Graham, C.H., Valavi, R., Abegg, et al. 2020. "Presence-Only and Presence-Absence Data for Comparing Species Distribution Modeling Methods." *Biodiversity Informatics* 15: 69–80. https://journals.ku.edu/jbi.

Fithian, William, and Trevor Hastie. 2013. "Finite-sample equivalence in statistical models for presence-only data." *The Annals of Applied Statistics* 7 (4): 1917–39. https://doi.org/10.1214/13-AOAS667.

Phillips, Steven, and Miroslav Dudík. 2008. "Generative and Discriminative Learning with Unknown Labeling Bias." In *Advances in Neural Information Processing Systems*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2008/file/9cf81d8026a9018052c429cc4e56739b-Paper.pdf.

Renner, Ian W., and David I. Warton. 2013. "Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology." *Biometrics* 69 (1): 274–81. https://doi.org/https://doi.org/10.1111/j.1541-0420.2012.01824.x.

Smith, Jeffrey R., and Jonathan M. Levine. 2025. "Linking Relative Suitability to Probability of Occurrence in Presence-Only Species Distribution Models: Implications for Global Change Projections." *Methods in Ecology and Evolution* 16 (4): 854–65. https://doi.org/https://doi.org/10.1111/2041-210X.70003.

Valavi, Roozbeh, Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, and Jane Elith. 2022. "Predictive Performance of Presence-Only Species Distribution Models: A Benchmark Study with Reproducible Code." *Ecological Monographs* 92 (1): e01486. https://doi.org/https://doi.org/10.1002/ecm.1486.

Warton, David I., and Leah C. Shepherd. 2010. "Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology." *The Annals of Applied Statistics* 4 (3): 1383–1402. https://doi.org/10.1214/10-AOAS331.

Yackulic, Charles B., Richard Chandler, Elise F. Zipkin, J. Andrew Royle, James D. Nichols, Evan H. Campbell Grant, and Sophie Veran. 2013. "Presence-Only Modelling Using

MAXENT: When Can We Trust the Inferences?" *Methods in Ecology and Evolution* 4 (3): 236–43. https://doi.org/https://doi.org/10.1111/2041-210x.12004.