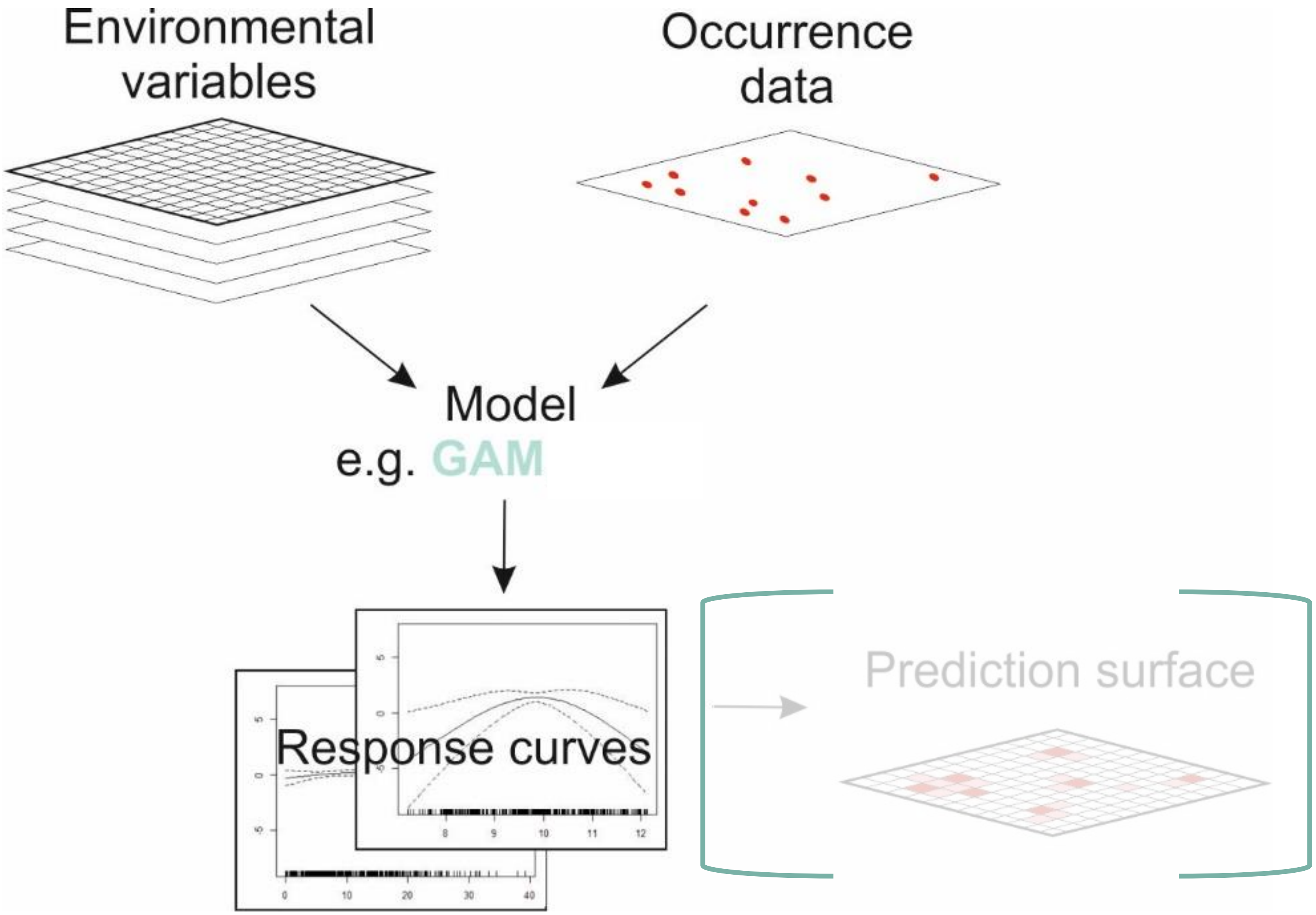
Generalized additive models

Modelling in physical geography, 5cr, 30.10.-30.11.2017



Environmental variables

Occurrence data

Model

e.g. GAM

Response curves

Prediction surface

GAM = generalized additive model

1. Theory

Semi-parametric versions of GLMs

- GAMs relate a response variable to predictor variables, but these variables don’t have to be related linearly

Splines or **smooth functions** of predictor variables (vs. higher order terms in GLMs) relax assumption of linearity

- investigate influence of predictor along range of its possible values
- each predictor is separated into sections (delimited by 'knots’)
- polynomial functions are fitted to each section separately

Compared to GLMs:

+ no need to identify appropriate polynomial terms and predictor transformations to improve model fit
+ model flexibility (to approximate true response)
+ relaxed assumptions on response-predictor relationship
+ potential for better fit to data (vs. purely parametric models)
+ distinguish weak/indirect/complex responses and thresholds
– some loss of interpretability, more complex to realize
– restricted to be additive (may miss important interactions)

$$g(\mu) = \beta_0 + s_1(x_1) + s_2(x_2) + \ldots + s_k(x_k),$$

Monographs on Statistics and Applied Probability 43
Generalized Additive Models
T.J. Hastie and R.J. Tibshirani
CHAPMAN & HALL/CRC
Texts in Statistical Science
Generalized Additive Models
An Introduction with R
SECOND EDITION
Simon N. Wood
CRC Press
A CHAPMAN & HALL BOOK

2. Applications

Used to understand the effect of covariates and test hypotheses about effects

The regression form is not defined via a function but found from the data

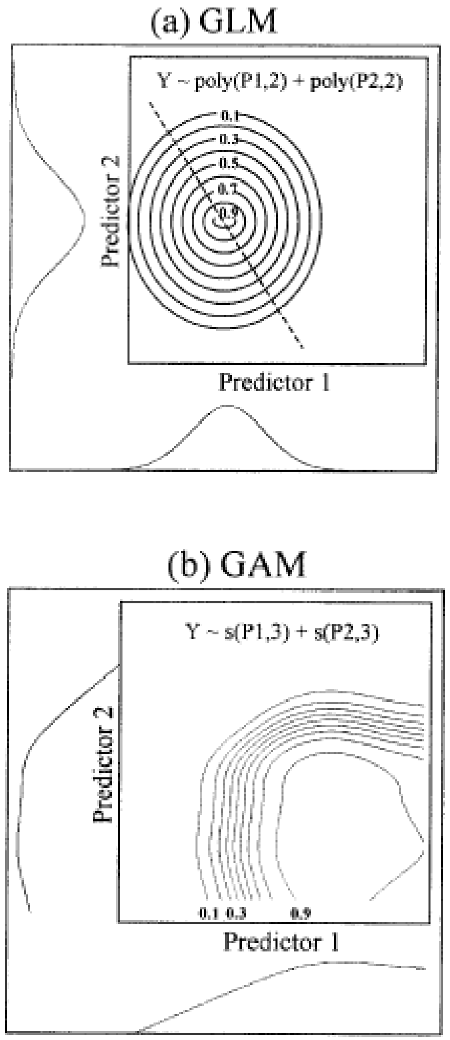
- Letting the data 'speak for themselves’ (vs GLM, which is model-driven)
- The sensitivity of this regression to local components (knots) can be defined by the modeller (similar to choosing number of parameters in *lm*)
- Can bring to light nonlinear dependency structures in the data

Easy-to-read, easy-to-interpret. Useful for:

- When no *a priori* reason for choosing a particular response function
- Spatial prediction/interpolation
- Exploratory analyses regarding the functional nature of a response

Greater flexibility than GLMs
→ GAMs more capable of modelling complex ecological response shapes
→ Possible to examine the shape of curves using GAMs, then reconstruct the curve shape parametrically with GLM (conservative approach)

-- Non-parametric regression is computationally heavy → can be slow



Guisan et al. 2000

3. Examples in R – GAM model building

Very similar to building GLMs!

> ***library(mgcv)*** #opens gam library

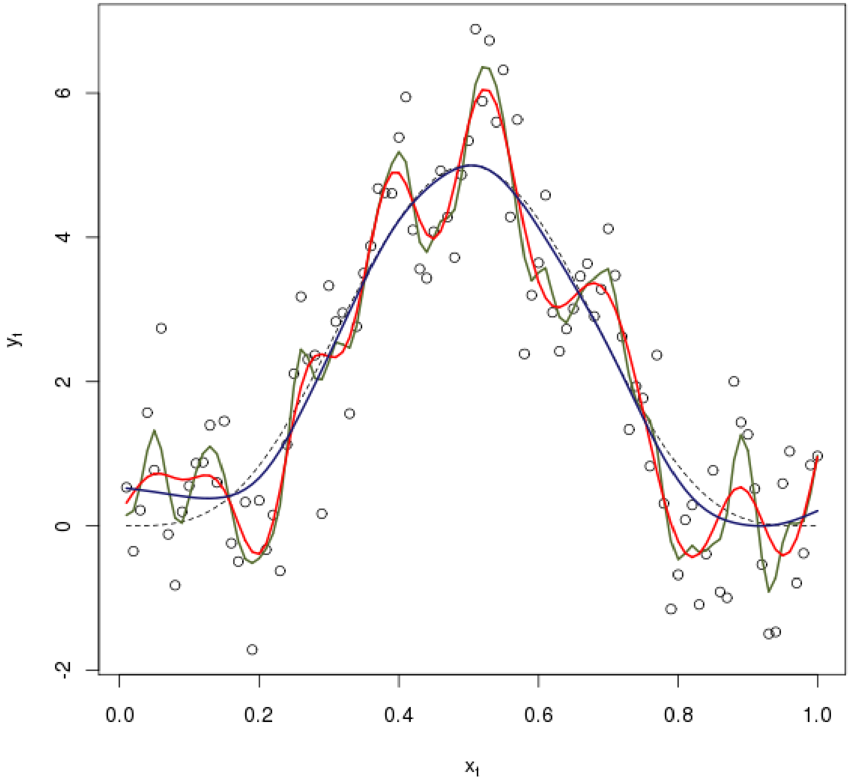
Input: like most R modelling functions, ***gam()*** expects a model formula to be supplied, specifying model structure to fit

> ***gam1 <- gam(Y~s(Xi, k=3), family="binomial")***

where: Y = response variable
Xi = linear predictor variable
s = fitting the smoothing spline, estimated from data
k = number of knots/degrees of freedom for spline (model flexibility)
family = same as GLM; Gaussian, binomial, poisson

Output: a fitted gam object to be analyzed via functions
> ***summary()*** # view model parameters
> ***plot()*** # view response curves

Appropriate smoothing helps to correctly understand and predict data: fit to data is best possible, and is as smooth as possible



```
> gam1 <- gam(empher~s(altitude, k=3), family="binomial")
> gam2 <- gam(totalspr~s(altitude, k=3), family=”poisson")
> gam3 <- gam(gdd~s(altitude, k=3), family=”gaussian")
```

3. Examples in R – GAM model summary

The reported statistics of the model summary show:

The estimated degrees of freedom (edf)

A test of whether the smoothed functions significantly reduce model deviance

How much of the variation in the response variable the predictor variable/s explain:

- R-sq.(adj) – depends on variables; their significance and number (pseudo-R^2)
- Deviance explained: 1 – (residual deviance/null deviance)

Un-Biased Risk Estimator (UBRE) or Minimised generalised cross-validation (GCV) score

- smoothing parameter estimation
- similar to AIC, smaller values indicate better fit

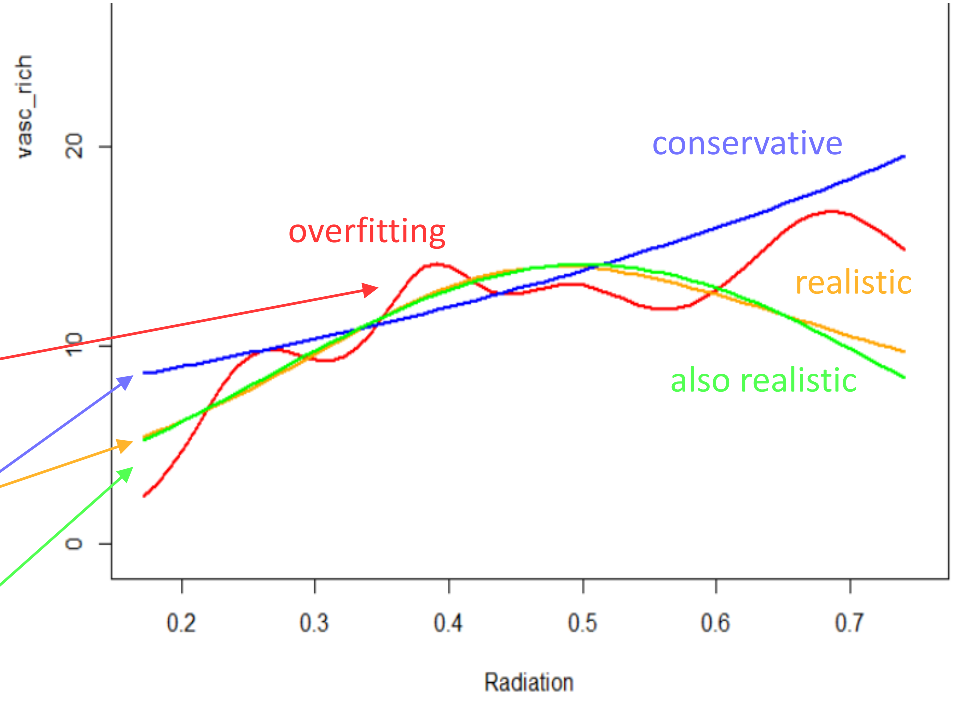
```
> gam1 <- gam(empher~s(altitude, k=3), family="binomial")

> summary(gam1)

Family: binomial
Link function: logit

Formula:
empher ~ s(altitude, k = 3)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.82844    0.06864   12.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(altitude) 1.995      2    309  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.338   Deviance explained = 27.5%
UBRE = -0.078376  Scale est. = 1         n = 1451
```

3. Examples in R – response curves

- The data determines shape of response curve (rather than being limited by shapes available in parametric GLM)

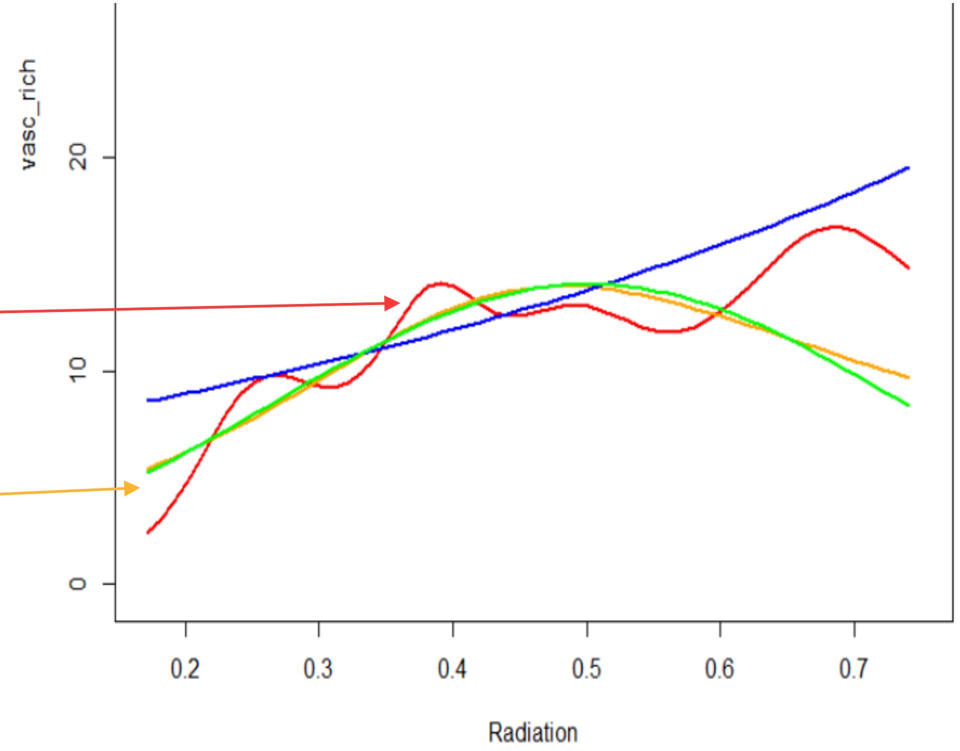
The effect of one variable, here solar radiation, on vascular plant richness

```
> #GAM with freedom term set to 9
> gam9 <- gam(vasc_rich~s(Radiation, k=9), family="poisson")
> Radiation2 <- seq(min(Radiation), max(Radiation), 0.01)
> newdata <- data.frame(Radiation=Radiation2)
> pred.gam9 <- predict.gam(gam9, newdata, type="response")
> plot(Radiation, vasc_rich, pch=19, cex=0.2, col="grey",type="n")
> lines(Radiation2, pred.gam9, lty=1,lwd= 2,col="red")

> # GAM 3
> gam3 <- gam(vasc_rich~s(Radiation, k=3), family="poisson")
> Radiation2 <- seq(min(Radiation), max(Radiation), 0.01)
> pred.gam3 <- predict.gam(gam3, newdata, type="response")
> lines(Radiation2, pred.gam3, lty=1,lwd= 2,col="orange")

> # GLM
> glm1 <- glm(vasc_rich~Radiation, family="poisson")
> pred.glm1 <- predict.glm(glm1, newdata, type="response")
> lines(Radiation2, pred.glm1, lty=1,lwd= 2,col="blue")

> # GLM 2nd order polynomial function
> glm2 <- glm(vasc_rich~Radiation+I(Radiation^2), family="poisson")
> pred.glm2 <- predict.glm(glm2, newdata, type="response")
> lines(Radiation2, pred.glm2, lty=1,lwd= 2,col="green")
```



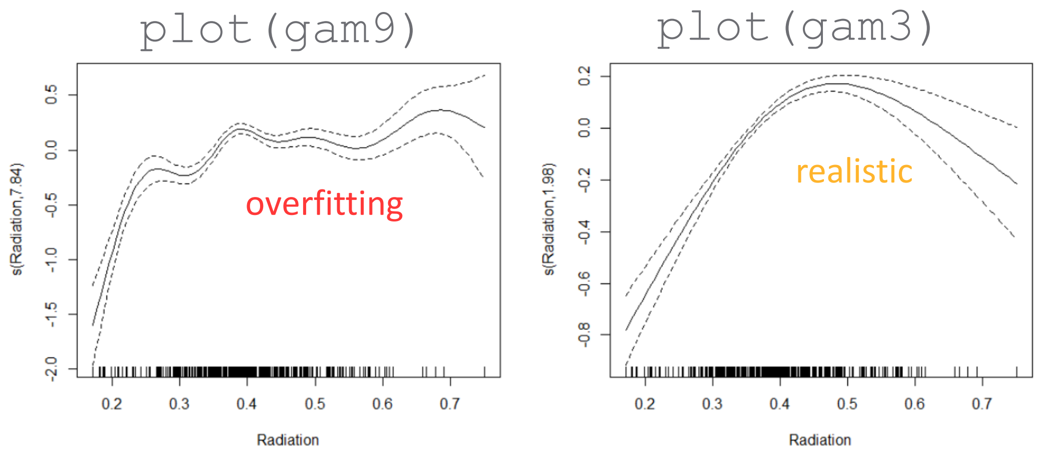
3. Examples in R – response curves

```
> #GAM with freedom term set to 9
> gam9 <- gam(vasc_rich~s(Radiation, k=9), family="poisson")
> Radiation2 <- seq(min(Radiation), max(Radiation), 0.01)
> newdata <- data.frame(Radiation=Radiation2)
> pred.gam9 <- predict.gam(gam9, newdata, type="response")
> plot(Radiation, vasc_rich, pch=19, cex=0.2, col="grey",type="n")
> lines(Radiation2, pred.gam9, lty=1,lwd= 2,col="red")

> # GAM 3
> gam3 <- gam(vasc_rich~s(Radiation, k=3), family="poisson")
> Radiation2 <- seq(min(Radiation), max(Radiation), 0.01)
> pred.gam3 <- predict.gam(gam3, newdata, type="response")
> lines(Radiation2, pred.gam3, lty=1,lwd= 2,col="orange")
```



Notice:

- More degrees of freedom → increased flexibility of response → models explanatory power increases (but danger of overfitting!)
- Appropriate df → variation can be captured while avoiding overfitting
- Precise freedom terms given on y-axis (7.84 ja 1.98)
- Y-axis = strength of the smoothing function (s)
- Fewer observations (see dashes on x-axis) increases confidence interval at both ends of the gradient


plot(gam9)

plot(gam3)

Modelled effect of radiation on species richness

3. Examples in R – variable selection

GAMs can be made more realistic by increasing model complexity, i.e. using multiple predictors

When choosing variables consider:

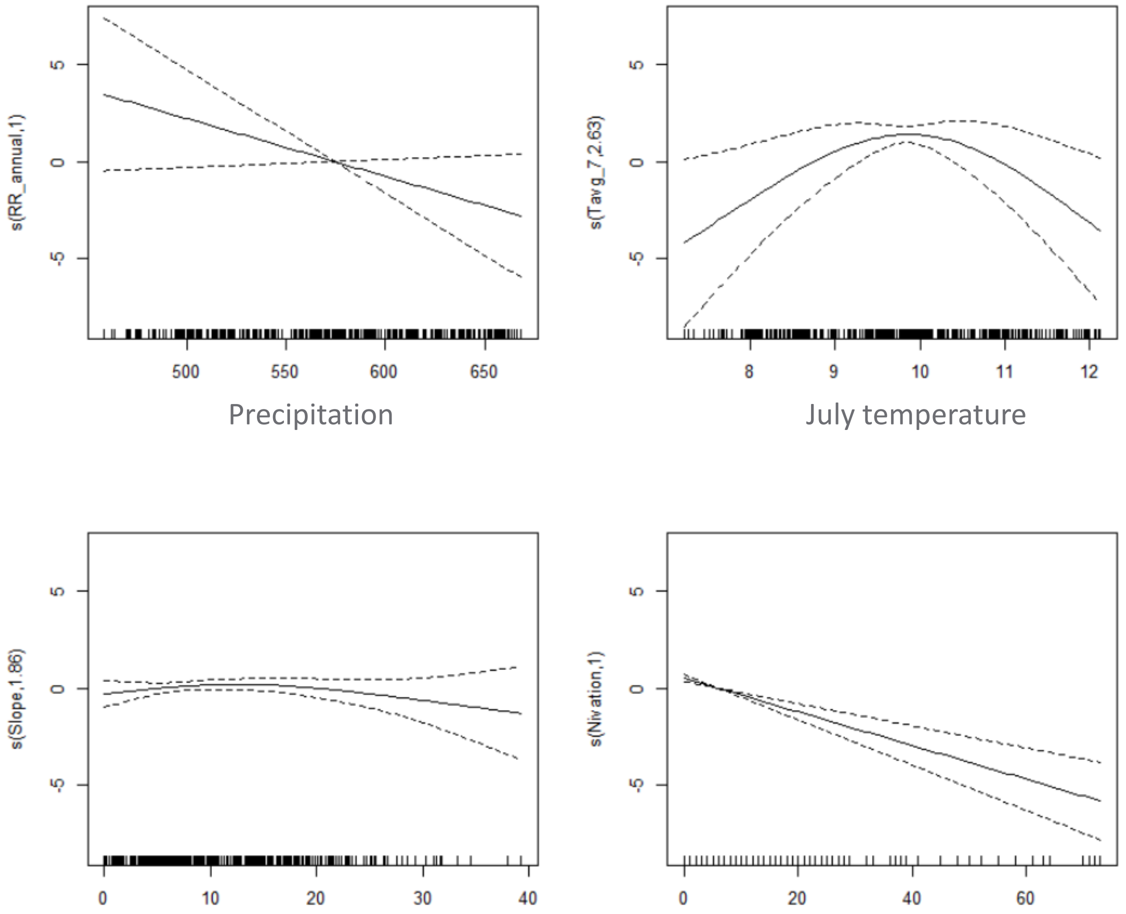
- statistical significance
- but also: theoretical importance

In the ***mgcv*** package, the stat. significance of model terms can be measured through the χ2 and p values

- ***> summary()*** # summary of one model
- ***> anova()*** # compare models

Visualizing the response curves also helps choose important predictors →

The distribution of *Empetrum hermaphroditum* (crowberry) in relation to different environmental parameters



Precipitation

July temperature

Slope steepness

Nivation (snow patches)

3. Examples in R – variable selection

- Statistical significance!

Consider removing X (even if significant) when:

- Confidence interval is large in relation to response curve (no variation is explained)
- No effect on adj.R^2 (see model summary) when X is removed

Consider including X as only a linear predictors when:

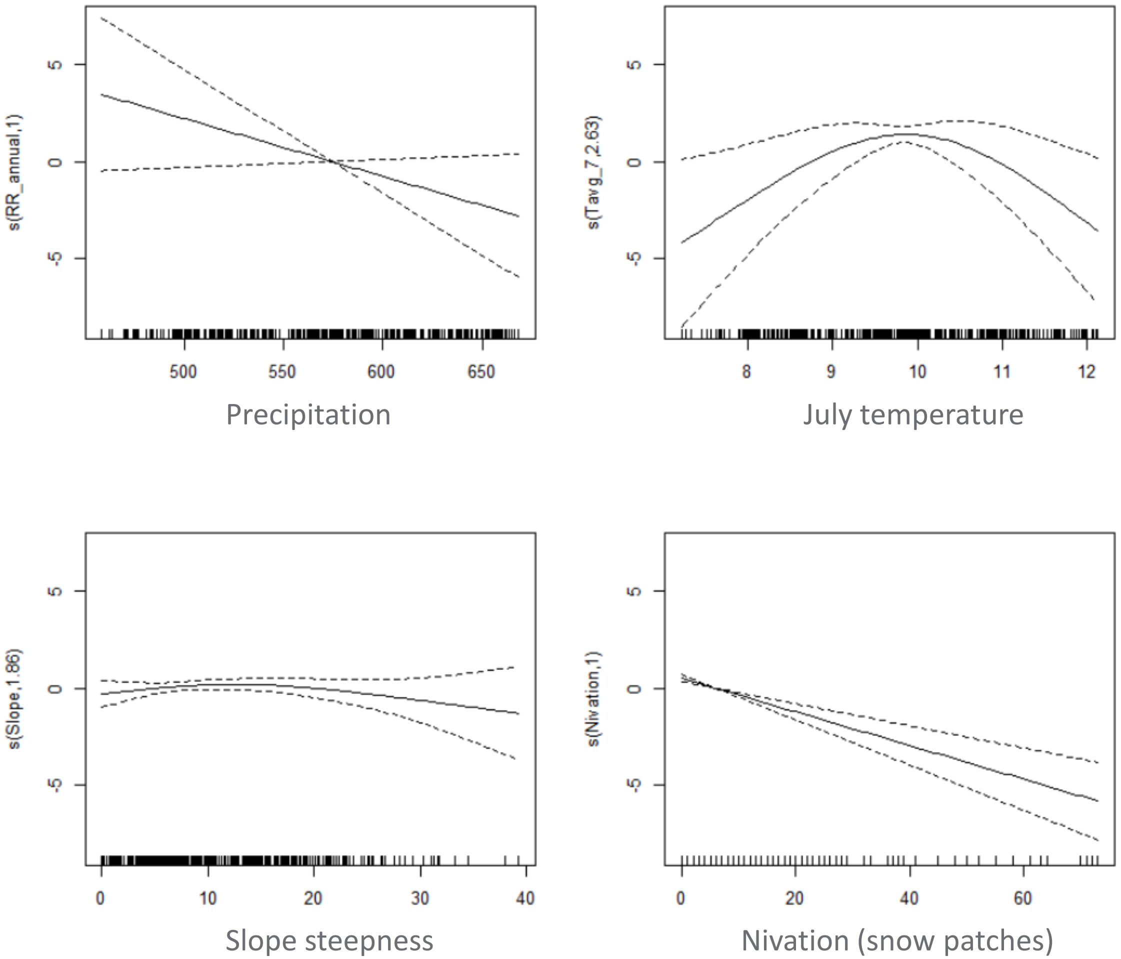
- Actual k is near 1 (see y-axis or model summary)
- e.g. ***s(Precipitation, k =3)*** → just ***Precipitation***

Consider decreasing degrees of freedom (k) when:

  - The response curve is very wiggly (over-fitting),
  - **e.g. k = 3 or 4 is usually a safe bet!**

- But: statistical significance is not enough, remember theoretical importance!
- Don't get too greedy: too many predictors (Xs) → model may not work!

The distribution of *Empetrum hermaphroditum* (crowberry) in relation to different environmental parameters



Precipitation

July temperature

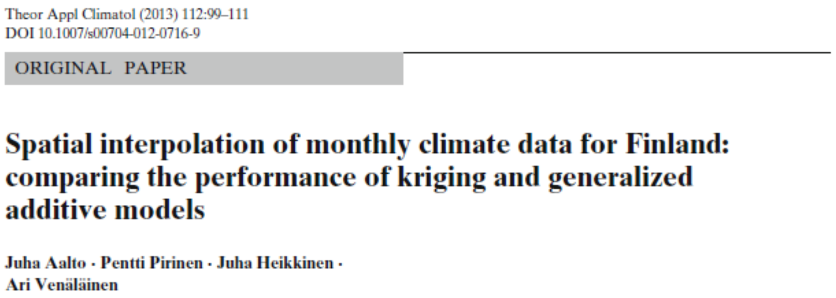
Slope steepness

Nivation (snow patches)

GAMs in R – main functions

- *?gam =* get help on gam
- *gam() =* for fitting the model
- *summary() =* for extracting model results (estimated coefficients and their p-values)
- *plot() =* for viewing response curves
- *predict.gam() =* use the fitted model to predict *Yi* at *Xi ,*
  remember → *(type=”response”)* inside the function!
- *anova() =* significance of the model terms, model inter-comparison; lower is better
- *AIC()* = model inter-comparison; lower AIC is better

4. Examples from the literature

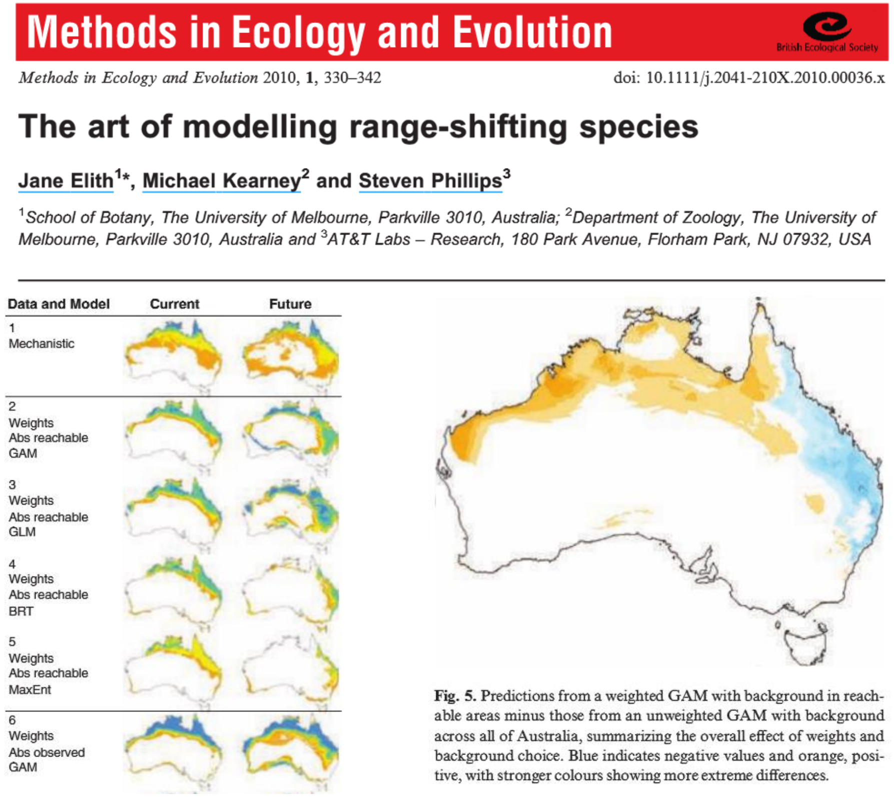


Theor Appl Climatol (2013) 112:99–111
DOI 10.1007/s00704-012-0716-9

ORIGINAL PAPER

Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models

Juha Aalto · Pentti Pirinen · Juha Heikkinen · Ari Venäläinen



Methods in Ecology and Evolution

British Ecological Society

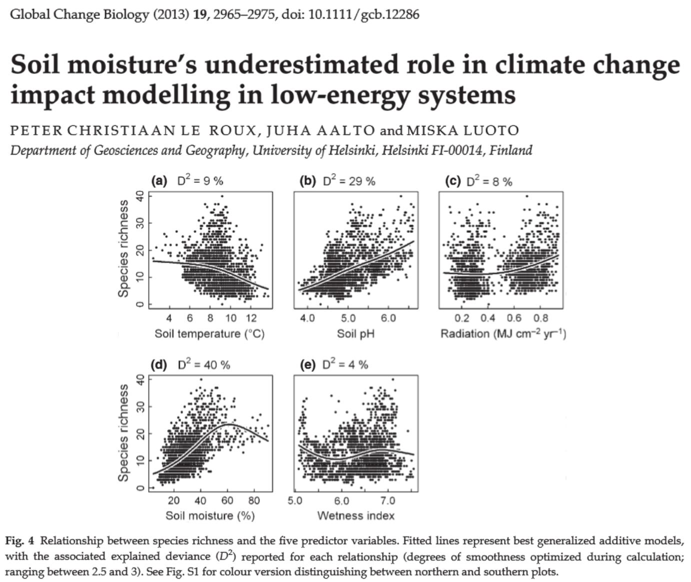
Methods in Ecology and Evolution 2010, **1**, 330–342 doi: 10.1111/j.2041-210X.2010.00036.x

The art of modelling range-shifting species

Jane Elith[1]*, Michael Kearney[2] and Steven Phillips[3]

[1]School of Botany, The University of Melbourne, Parkville 3010, Australia; [2]Department of Zoology, The University of Melbourne, Parkville 3010, Australia and [3]AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA

Fig. 5. Predictions from a weighted GAM with background in reachable areas minus those from an unweighted GAM with background across all of Australia, summarizing the overall effect of weights and background choice. Blue indicates negative values and orange, positive, with stronger colours showing more extreme differences.

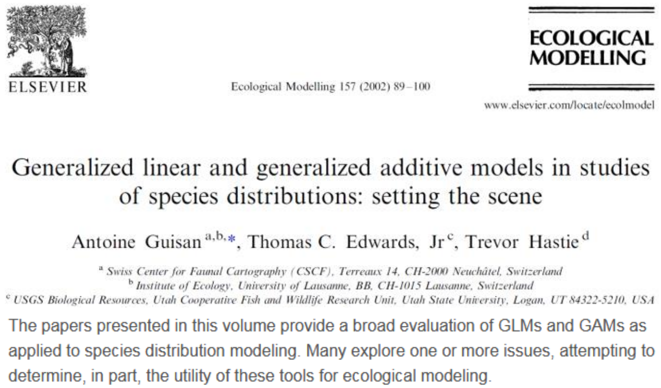


Global Change Biology (2013) **19**, 2965–2975, doi: 10.1111/gcb.12286

Soil moisture's underestimated role in climate change impact modelling in low-energy systems

PETER CHRISTIAAN LE ROUX, JUHA AALTO and MISKA LUOTO
Department of Geosciences and Geography, University of Helsinki, Helsinki FI-00014, Finland

Fig. 4 Relationship between species richness and the five predictor variables. Fitted lines represent best generalized additive models, with the associated explained deviance (D^2) reported for each relationship (degrees of smoothness optimized during calculation; ranging between 2.5 and 3). See Fig. S1 for colour version distinguishing between northern and southern plots.



ELSEVIER

Ecological Modelling 157 (2002) 89–100

ECOLOGICAL MODELLING

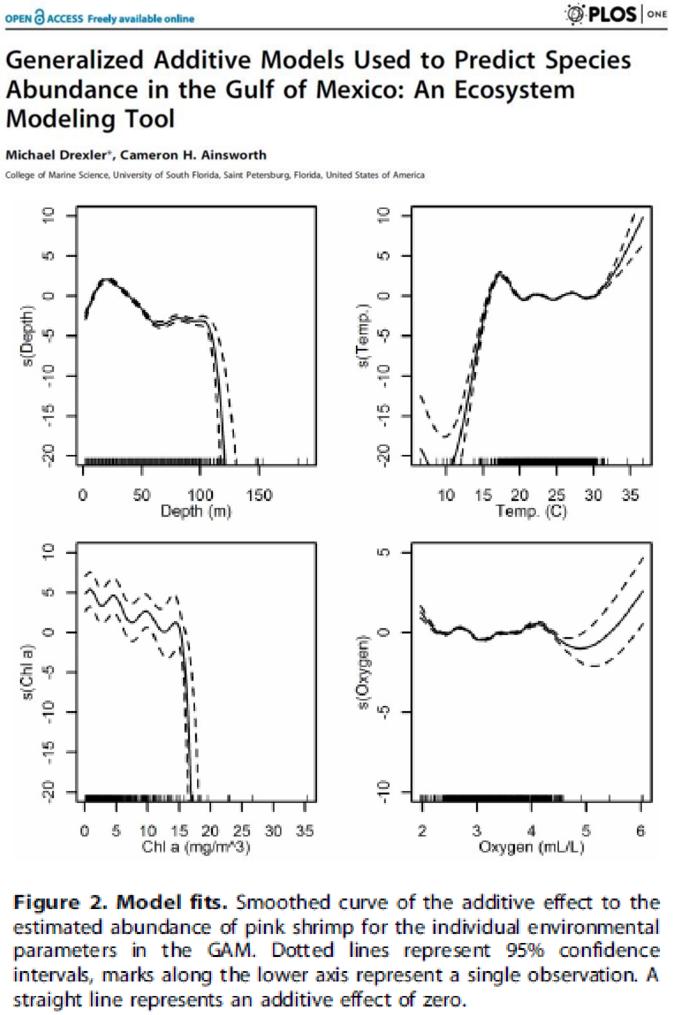
www.elsevier.com/locate/ecolmodel

Generalized linear and generalized additive models in studies of species distributions: setting the scene

Antoine Guisan[a,b,*], Thomas C. Edwards, Jr[c], Trevor Hastie[d]

[a] Swiss Center for Faunal Cartography (CSCF), Terreaux 14, CH-2000 Neuchâtel, Switzerland
[b] Institute of Ecology, University of Lausanne, BB, CH-1015 Lausanne, Switzerland
[c] USGS Biological Resources, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, UT 84322-5210, USA

The papers presented in this volume provide a broad evaluation of GLMs and GAMs as applied to species distribution modeling. Many explore one or more issues, attempting to determine, in part, the utility of these tools for ecological modeling.



OPEN ACCESS Freely available online

PLOS | ONE

Generalized Additive Models Used to Predict Species Abundance in the Gulf of Mexico: An Ecosystem Modeling Tool

Michael Drexler*, Cameron H. Ainsworth
College of Marine Science, University of South Florida, Saint Petersburg, Florida, United States of America

Figure 2. Model fits. Smoothed curve of the additive effect to the estimated abundance of pink shrimp for the individual environmental parameters in the GAM. Dotted lines represent 95% confidence intervals, marks along the lower axis represent a single observation. A straight line represents an additive effect of zero.

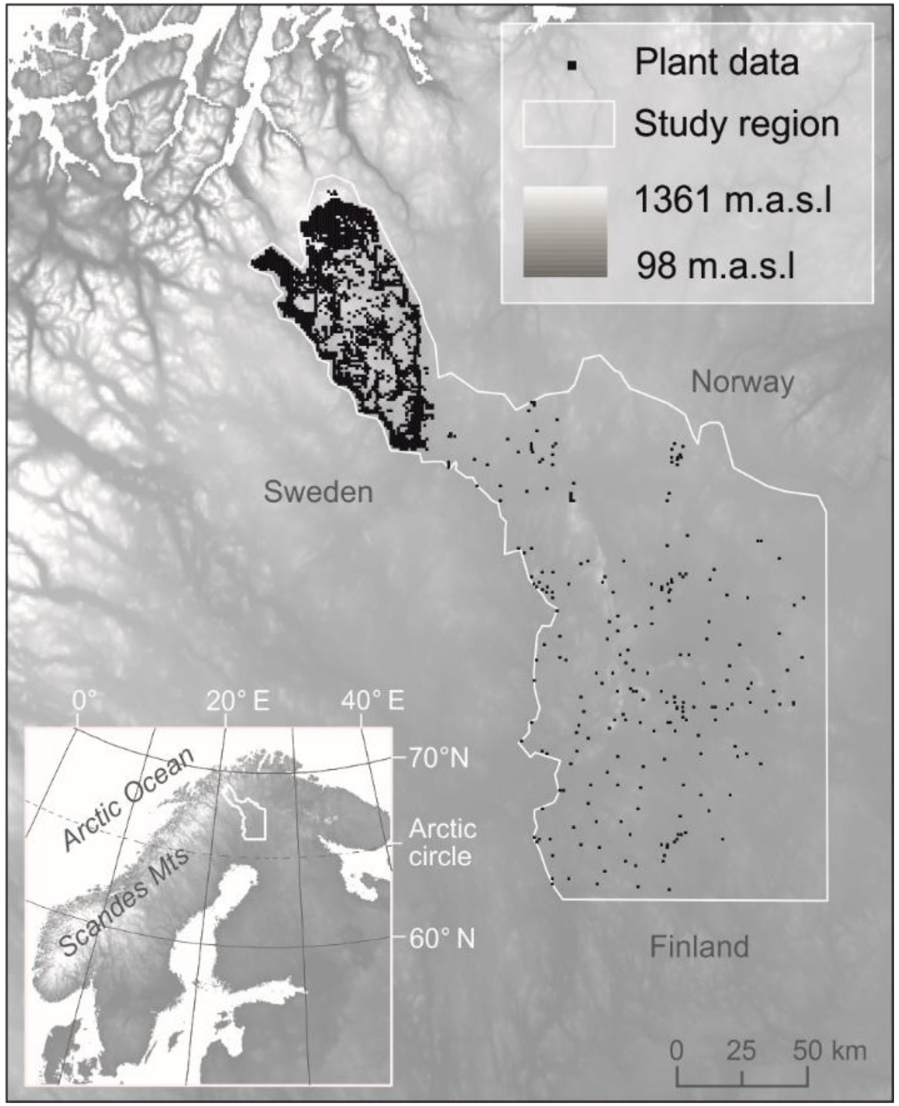
Practical – modelling species richness and distributions in Northwestern Lapland with GAM

R-script: *”GAM_practical.R”*

Data: *”NW_Lapland_data.csv”*

- totalspr = total vascular plant species richness
- rarespr = rare vascular plant species richness
- fdd = freezing degree days; overwintering temp conditions
- gdd = growing degree days; growing season temp conditions
- wab = water balance; moisture conditions
- calc = calcareousness of soil ; indicates soil pH
- relalt = relative altitude; topographic roughness
- altitude = mean altitude
- *betnan*
- *dryoct*
- *empher*
- *gersyl*
- *linbor*
- *phycae*
- *rangla*
- *vacmyr*
- *vacvit*

distributions of vascular plants (presence/absence) within a given cell



GAM - Practical

Using GAMs to explore the effect of environmental variables on:

Total plant species richness in a given 1km x 1km area

Empetrum hermaphroditum

Dryas octopetala