Ecological forecasting with dynamic GAMs

Nicholas Clark

School of Veterinary Science, University of Queensland, Australia

November, 2023

“Because all decision making is based on what will happen in the future, either under the status quo or different decision alternatives, decision making ultimately depends on forecasts”

Dietze et al. 2018

Properties of ecological series

Temporal autocorrelation

Lagged effects

Non-Gaussian data and missing observations

Measurement error

Time-varying effects

Nonlinearities

Multi-series clustering

Properties of ecological series

Temporal autocorrelation

Lagged effects

Non-Gaussian data and missing observations

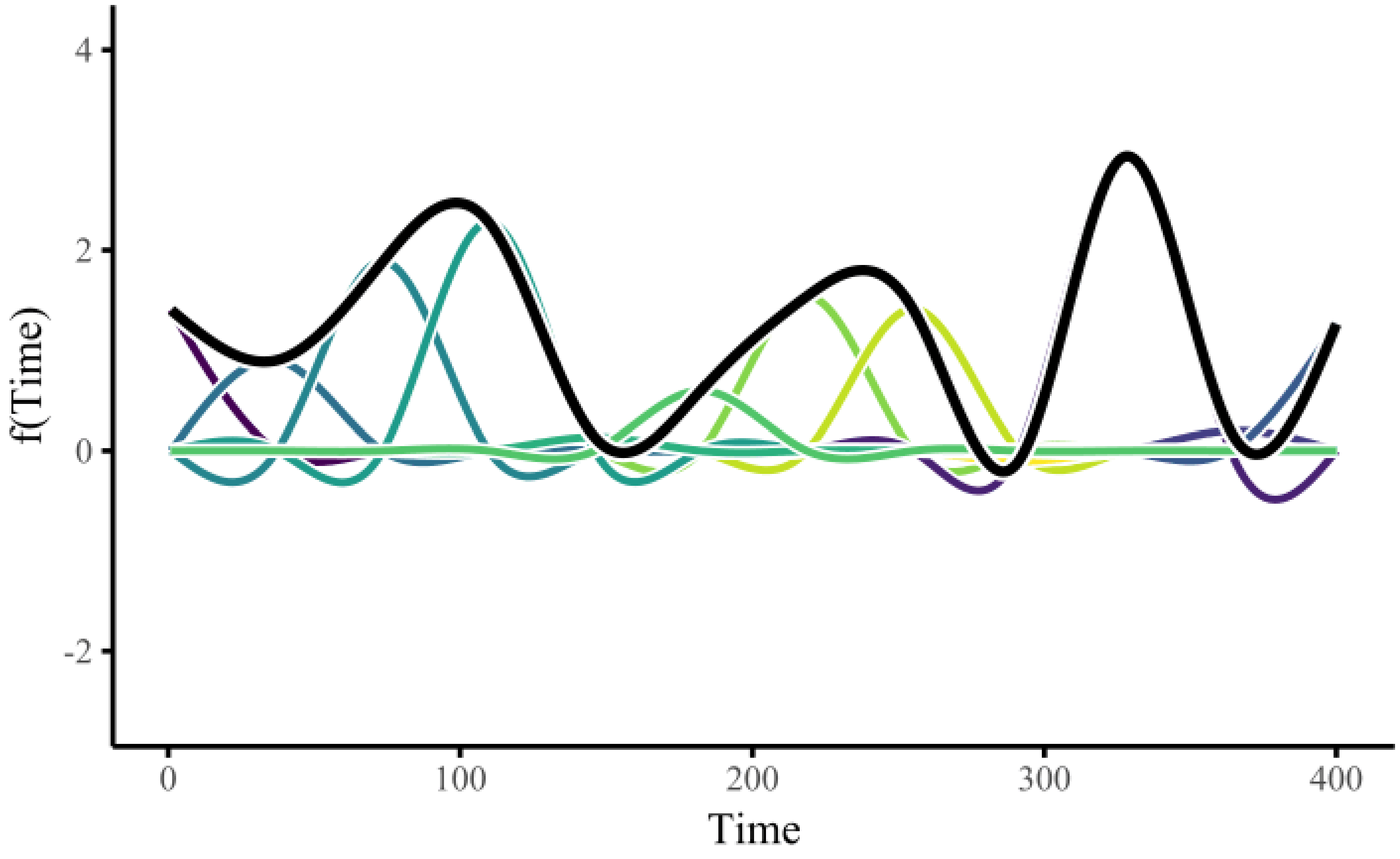
Measurement error

Time-varying effects

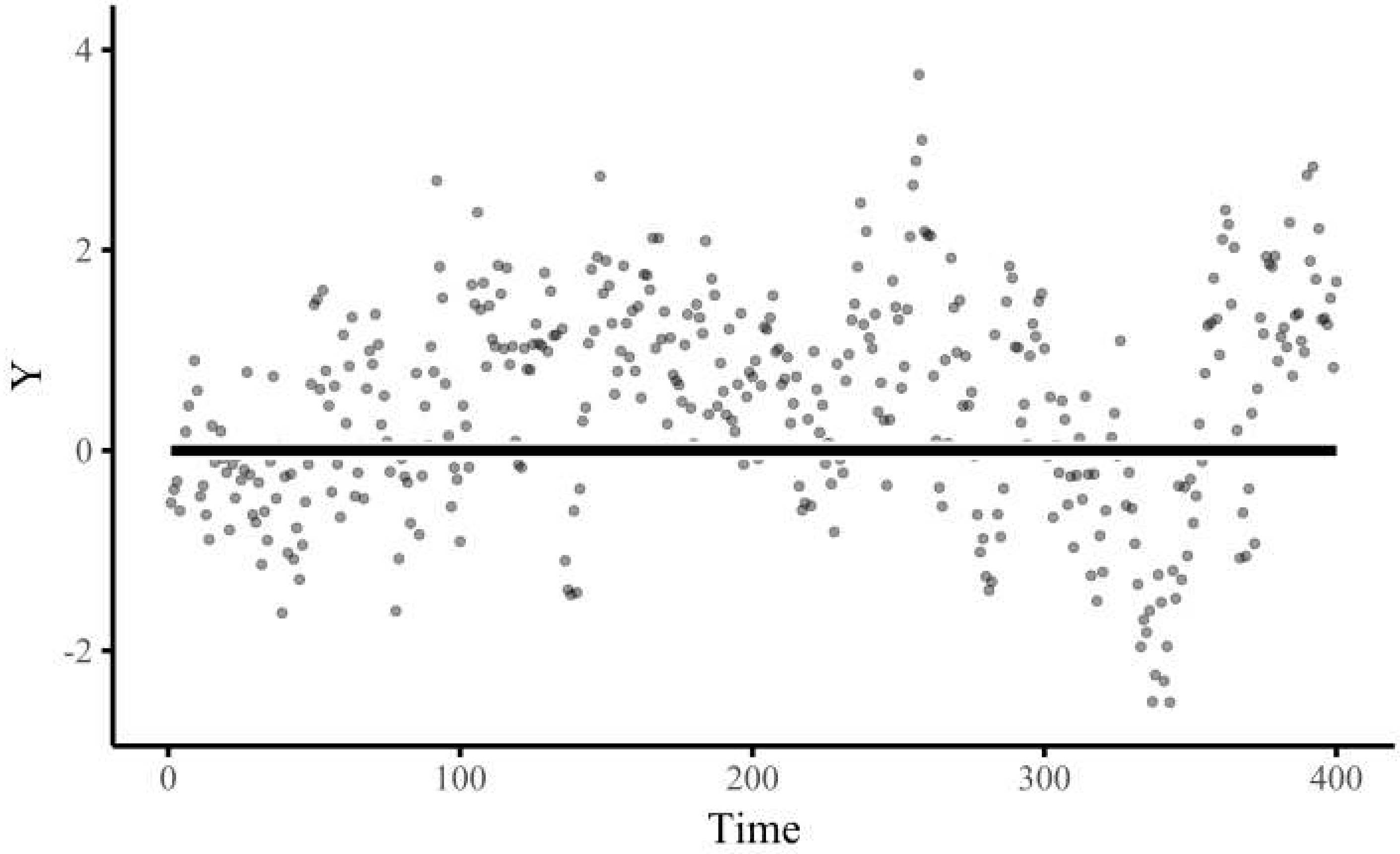
Nonlinearities

Multi-series clustering

GAMs use splines...



...penalized to fit data



Easy to fit in R

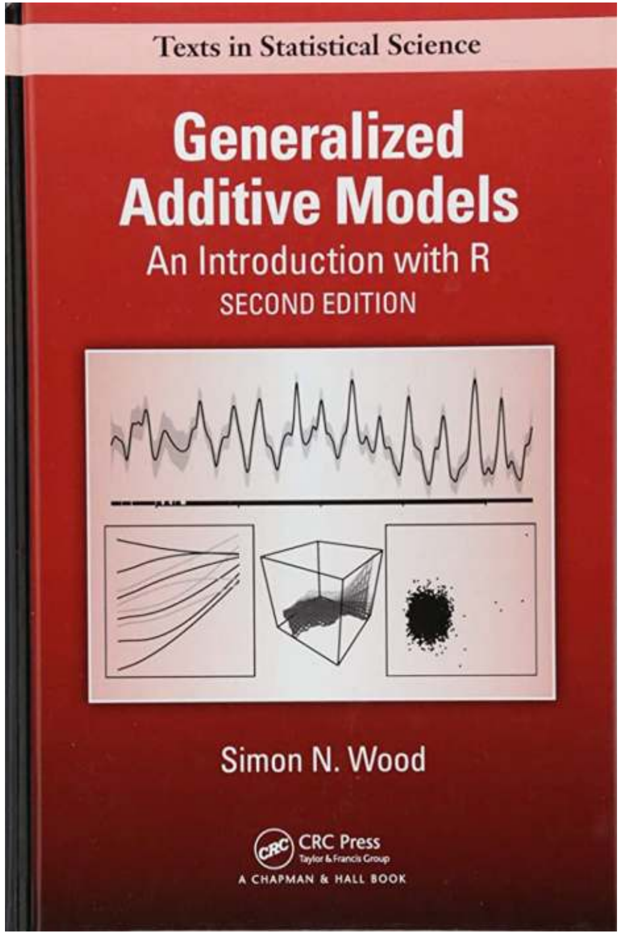
$$\mathbb{E}(\boldsymbol{Y}_t|\boldsymbol{X}_t) = g^{-1}(\alpha + \sum_{j=1}^{J} f(x_{jt}))$$

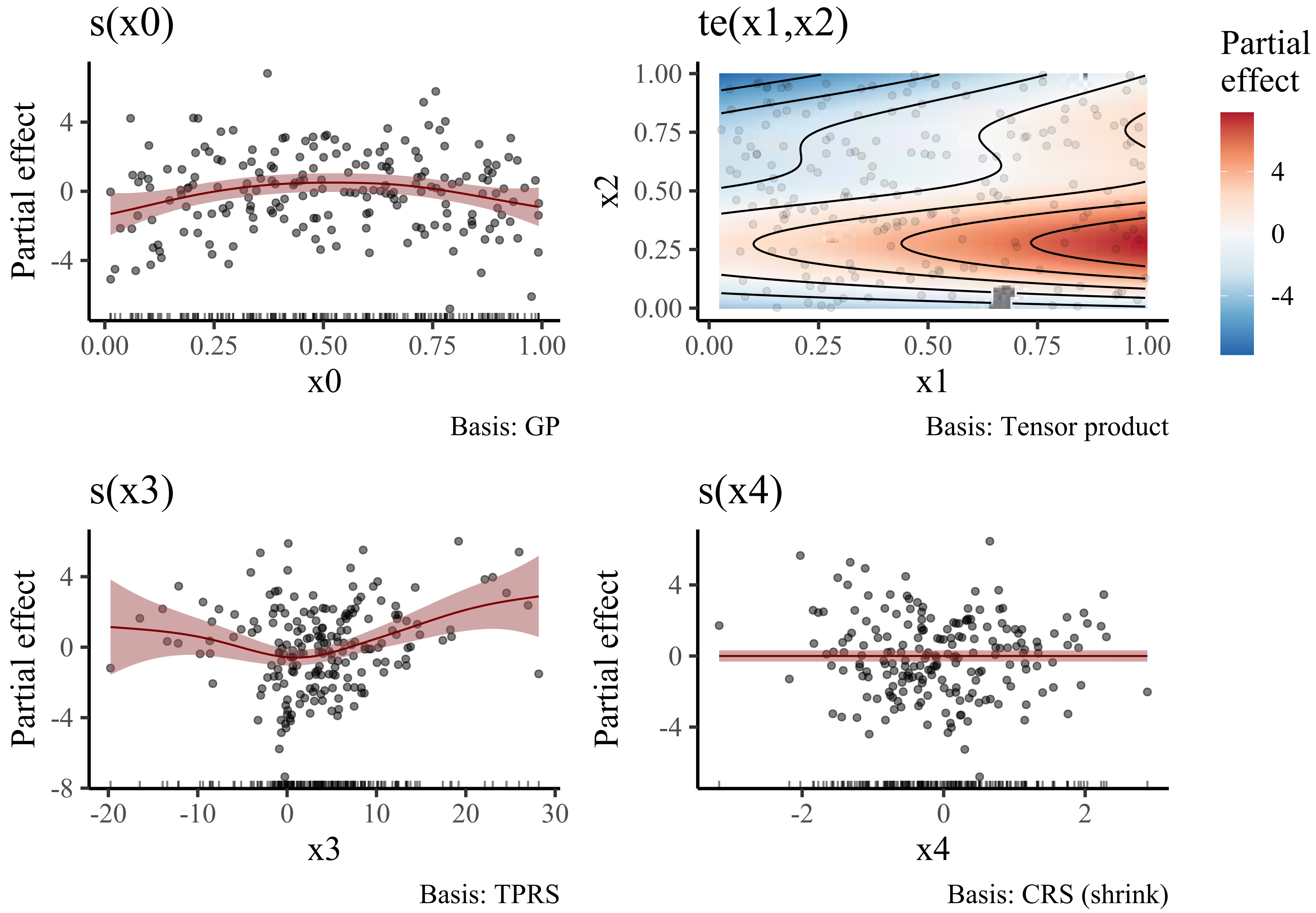
Where:

g^{-1} is the inverse of the link function

α is the intercept

$f(x)$ are potentially nonlinear functions of the J predictors





s(x0)
Partial effect
4
0
-4
0.00
0.25
0.50
0.75
1.00
x0
Basis: GP
te(x1,x2)
x2
1.00
0.75
0.50
0.25
0.00
x1
Basis: Tensor product
Partial
effect
4
0
-4
s(x3)
Partial effect
-8
-20
-10
0
10
20
30
x3
Basis: TPRS
s(x4)
Partial effect
-2
0
2
x4
Basis: CRS (shrink)

Need more on GAMs?

Gavin has you covered ☺

Generalized Additive Models with R and mgcv

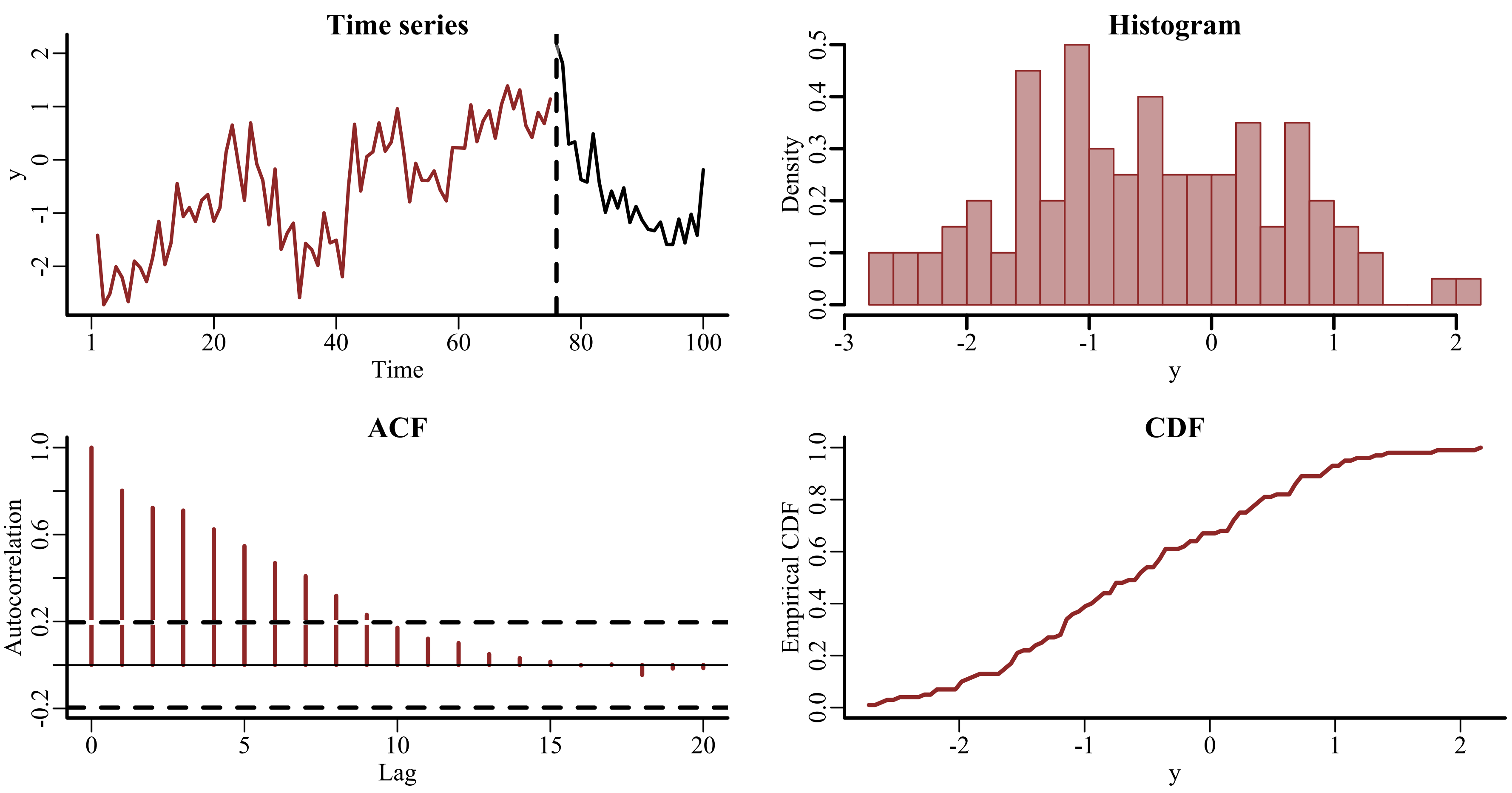
Gavin Simpson

January 3, 2022

GAMs are just fancy GLMs, where some (or all) of the predictor effects are estimated as (possibly nonlinear) smooth functions

But the complexity these smooth functions can handle is *enormous*

What's the catch?

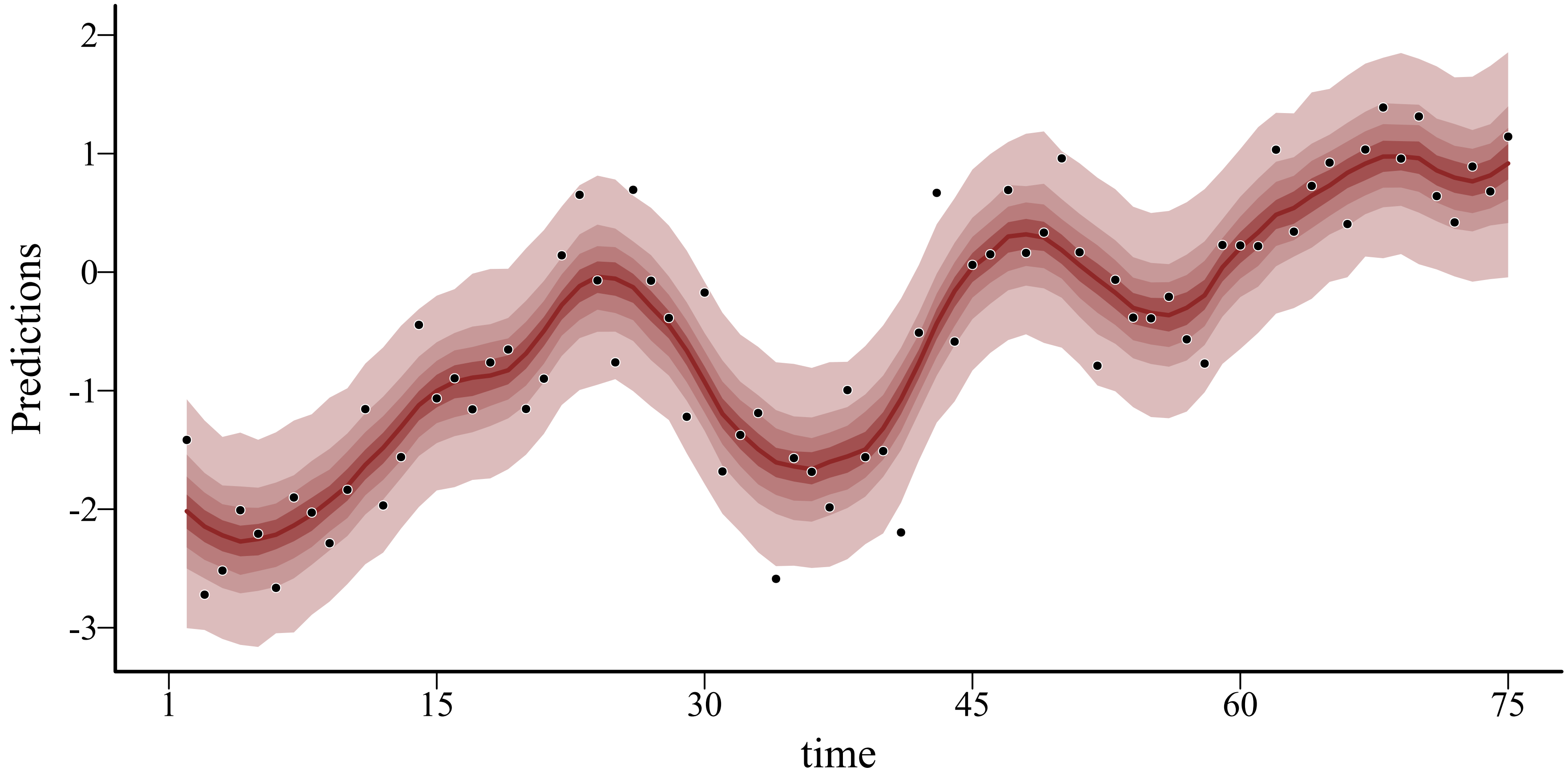


A spline of `time`

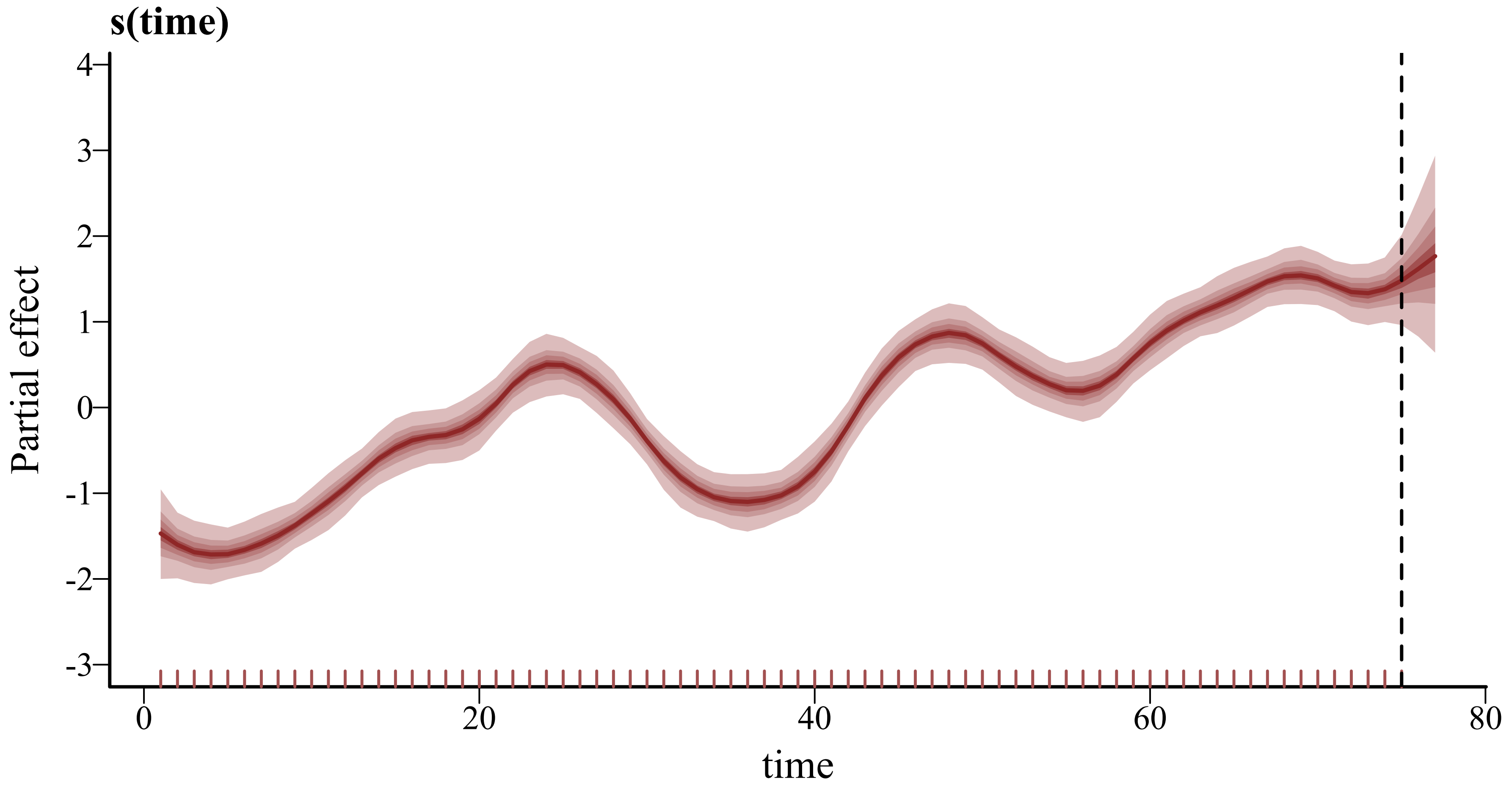
```r
library(mgcv)
model <- gam(y ~ s(time, k = 20, bs = 'bs', m = 2),
             data = data,
             family = gaussian())
```

A B-spline (`bs = 'bs'`) with `m = 2` sets the penalty on the second derivative

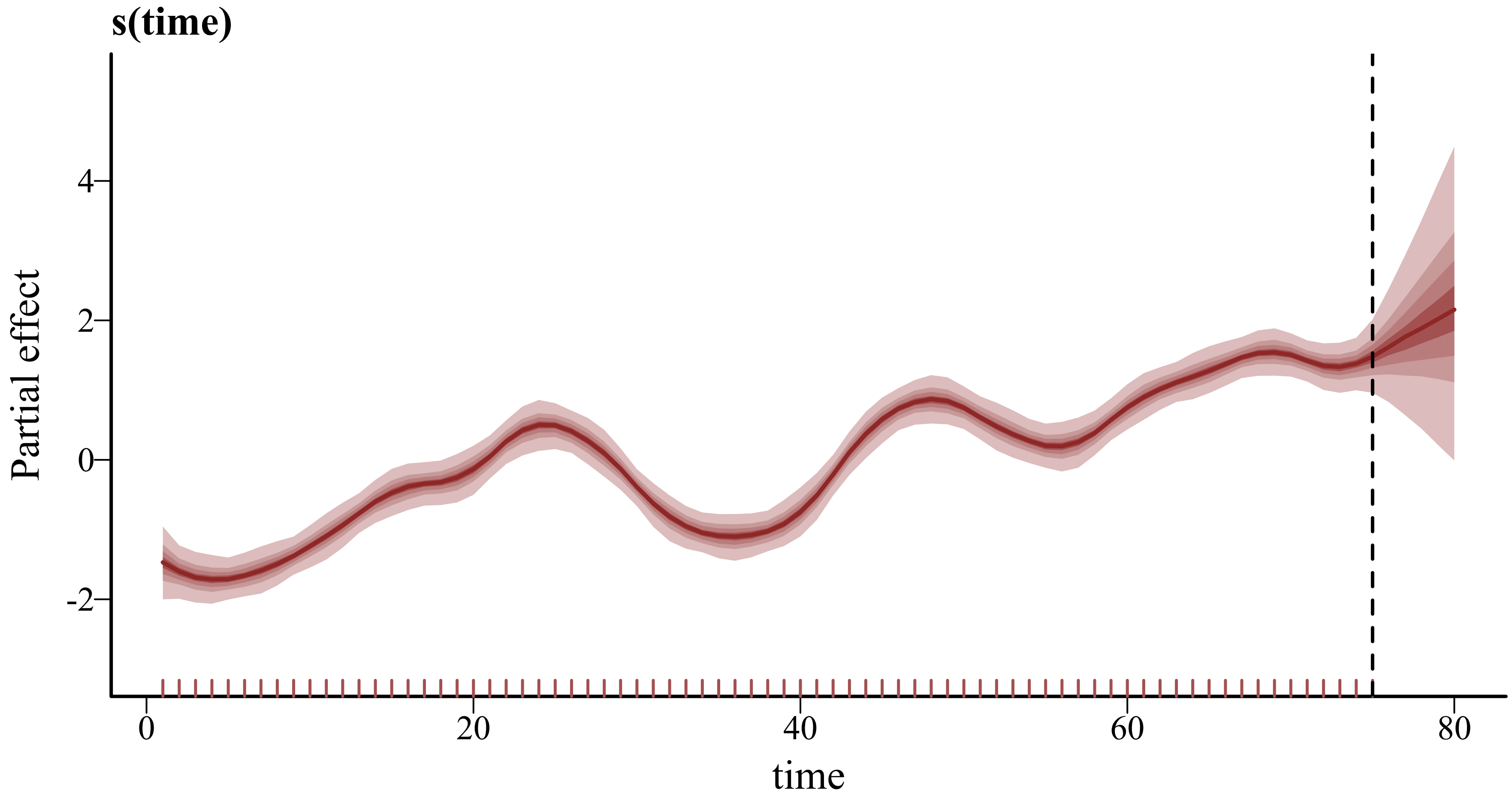
Hindcasts 😊



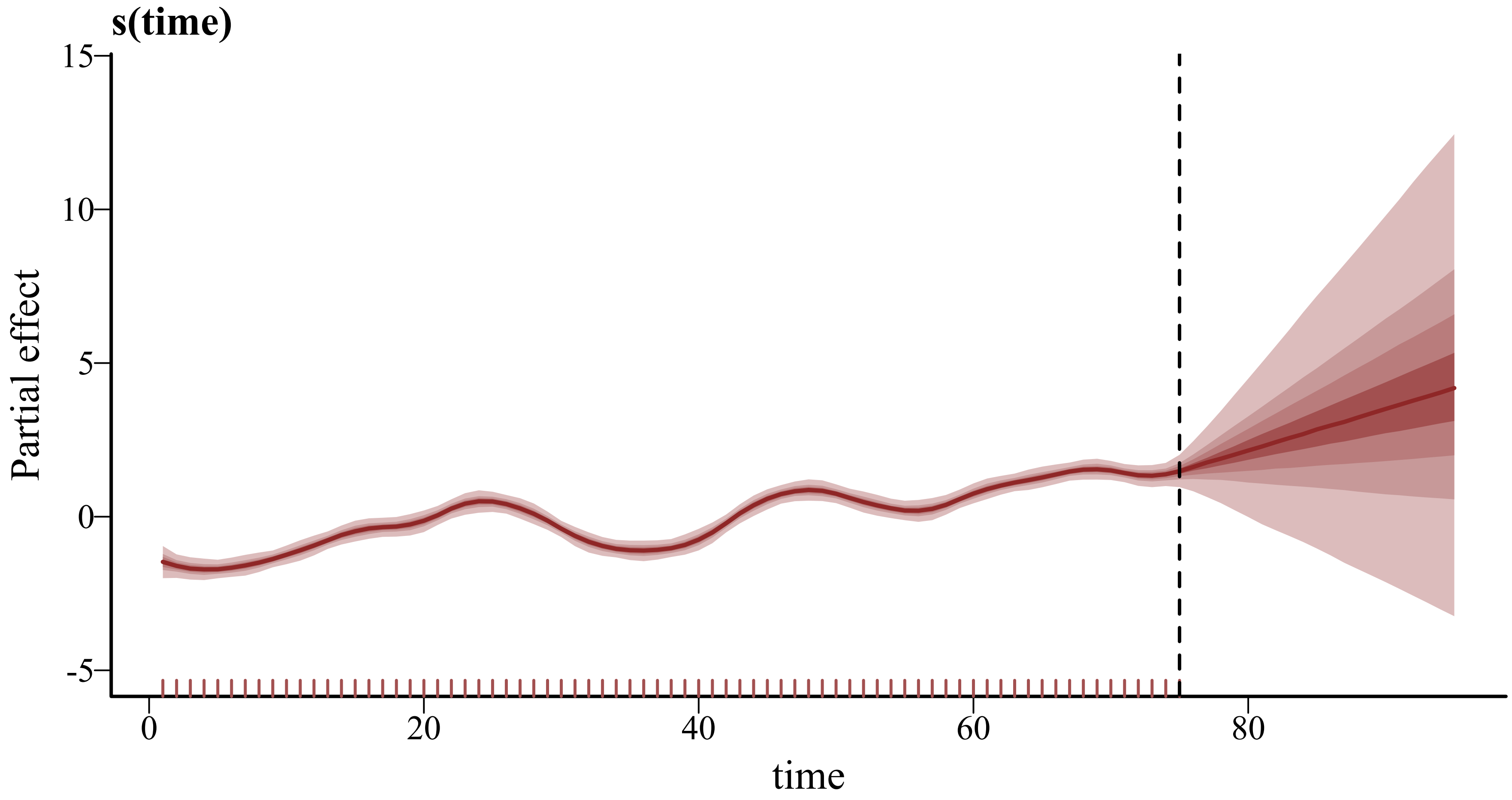
Extrapolate 2-steps ahead ☺



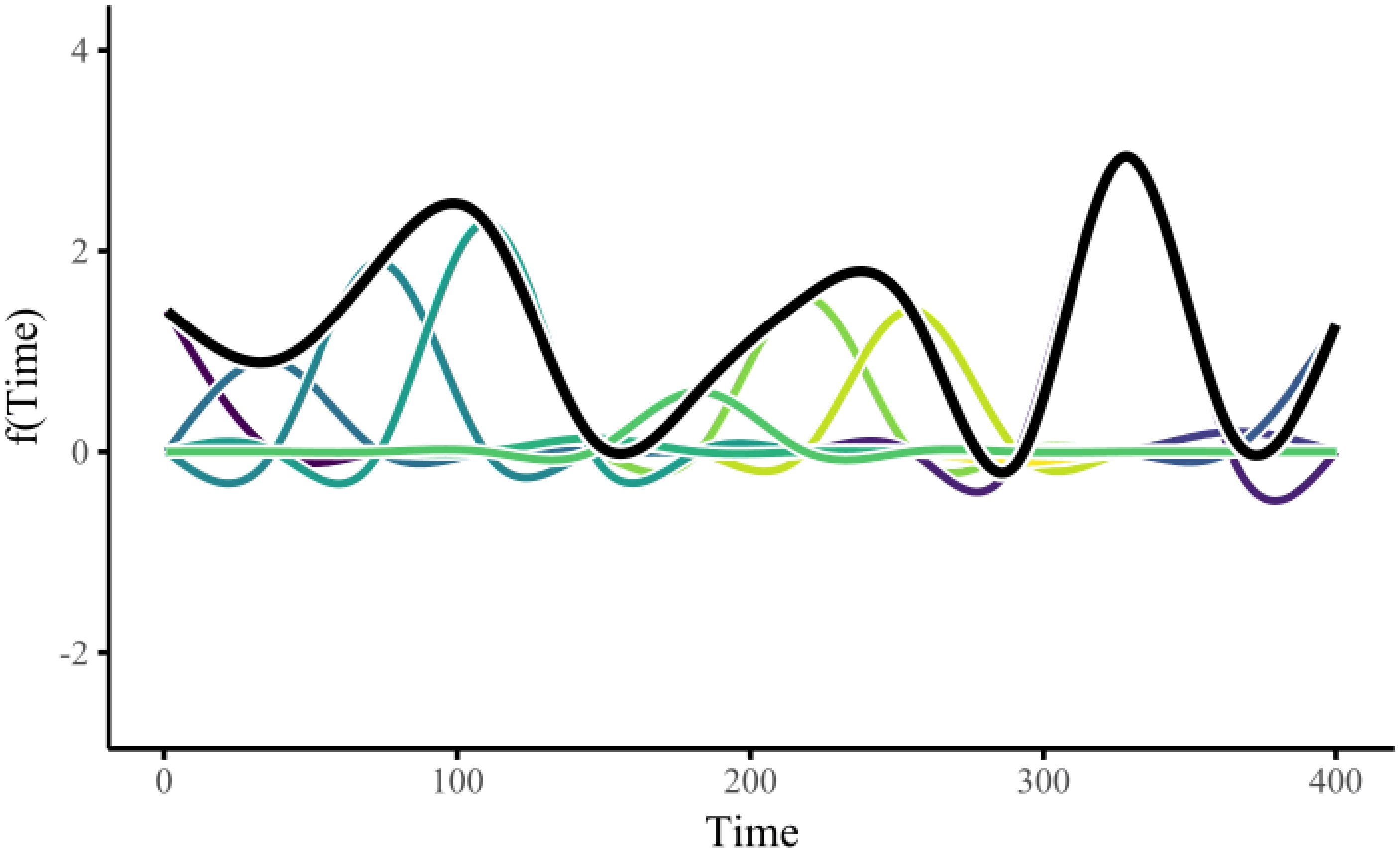
5-steps ahead ☹



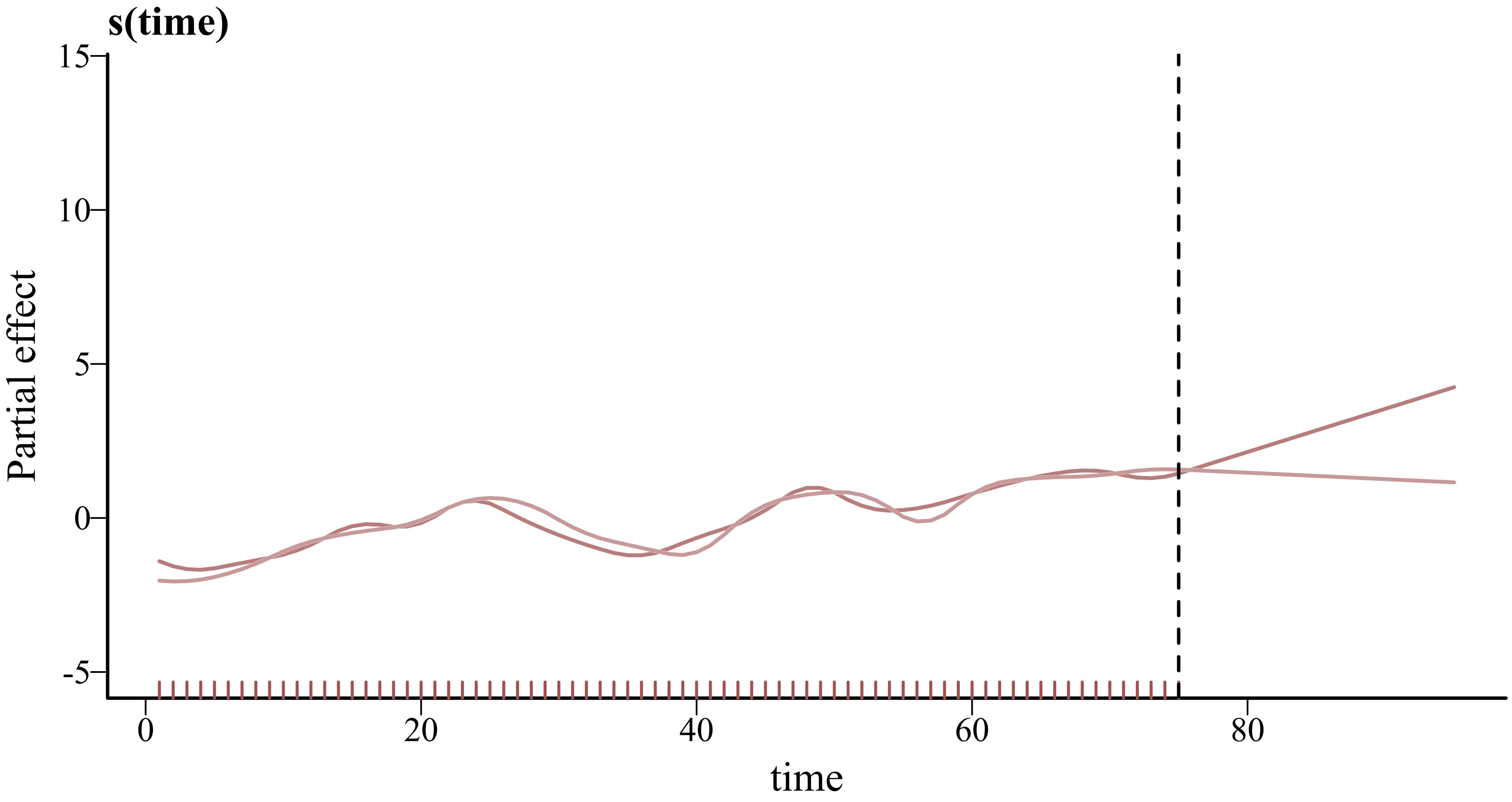
20-steps ahead ☹



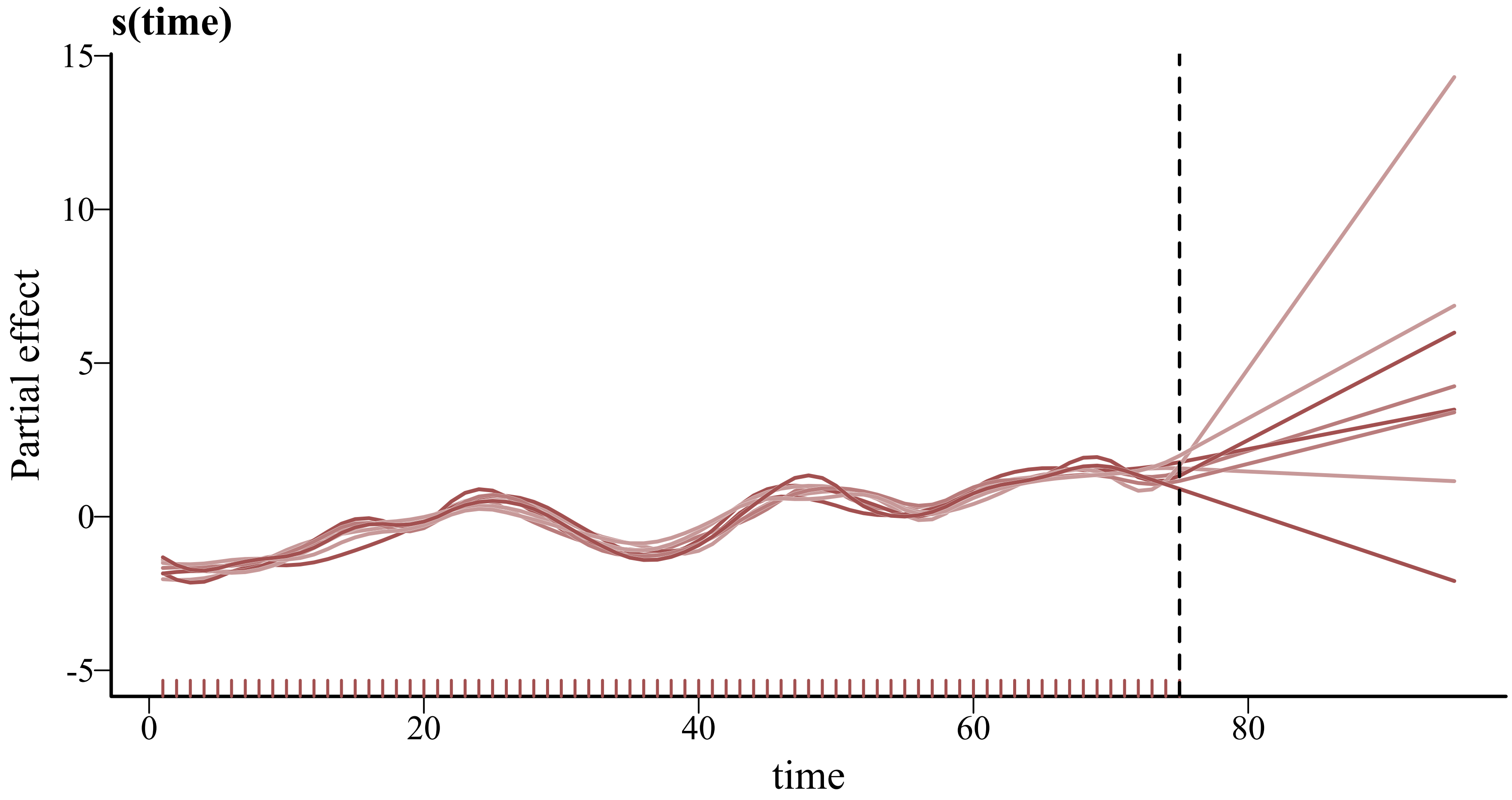
Basis functions ⇨ local knowledge



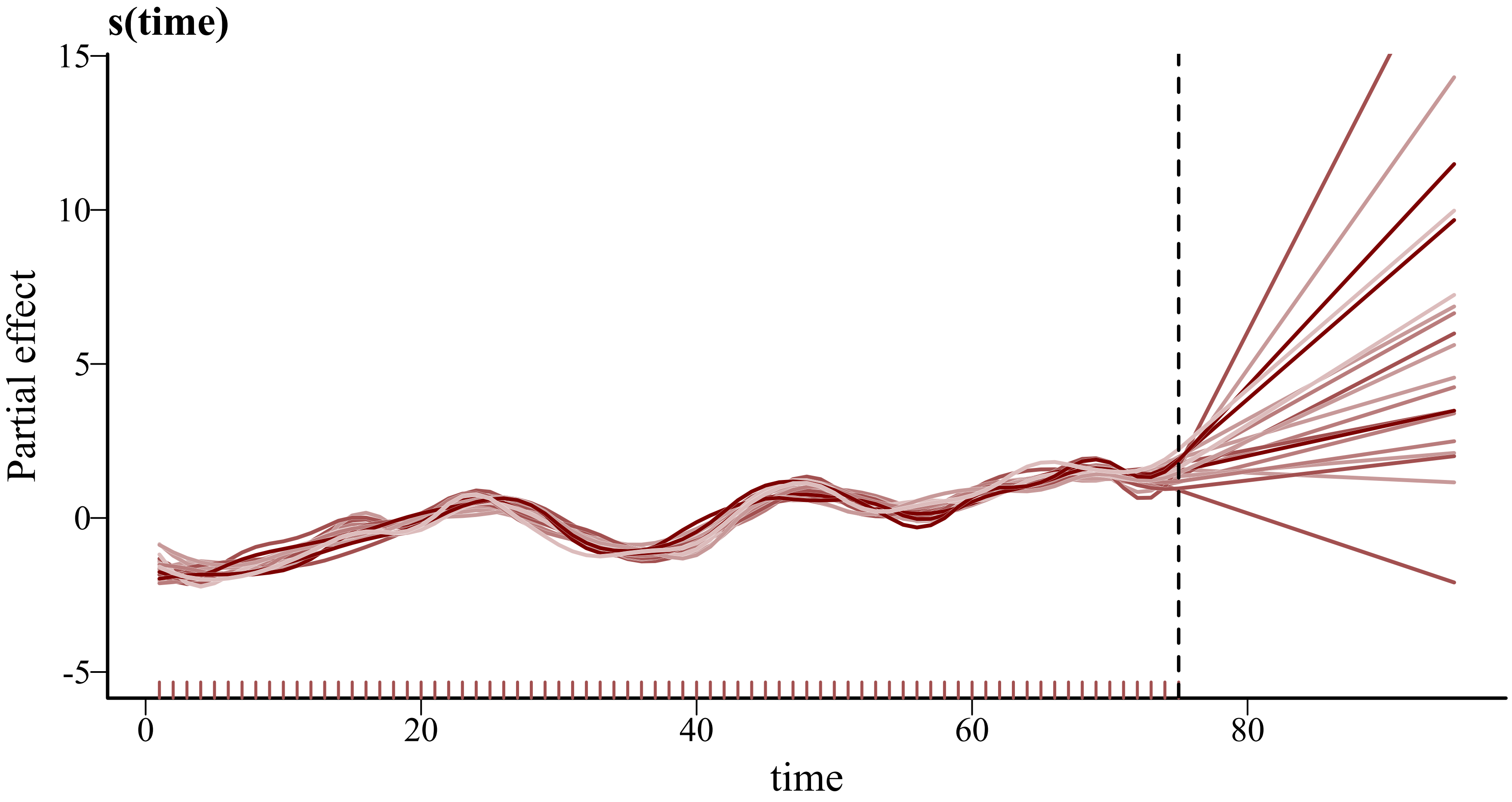
Basis functions ⇨ local knowledge



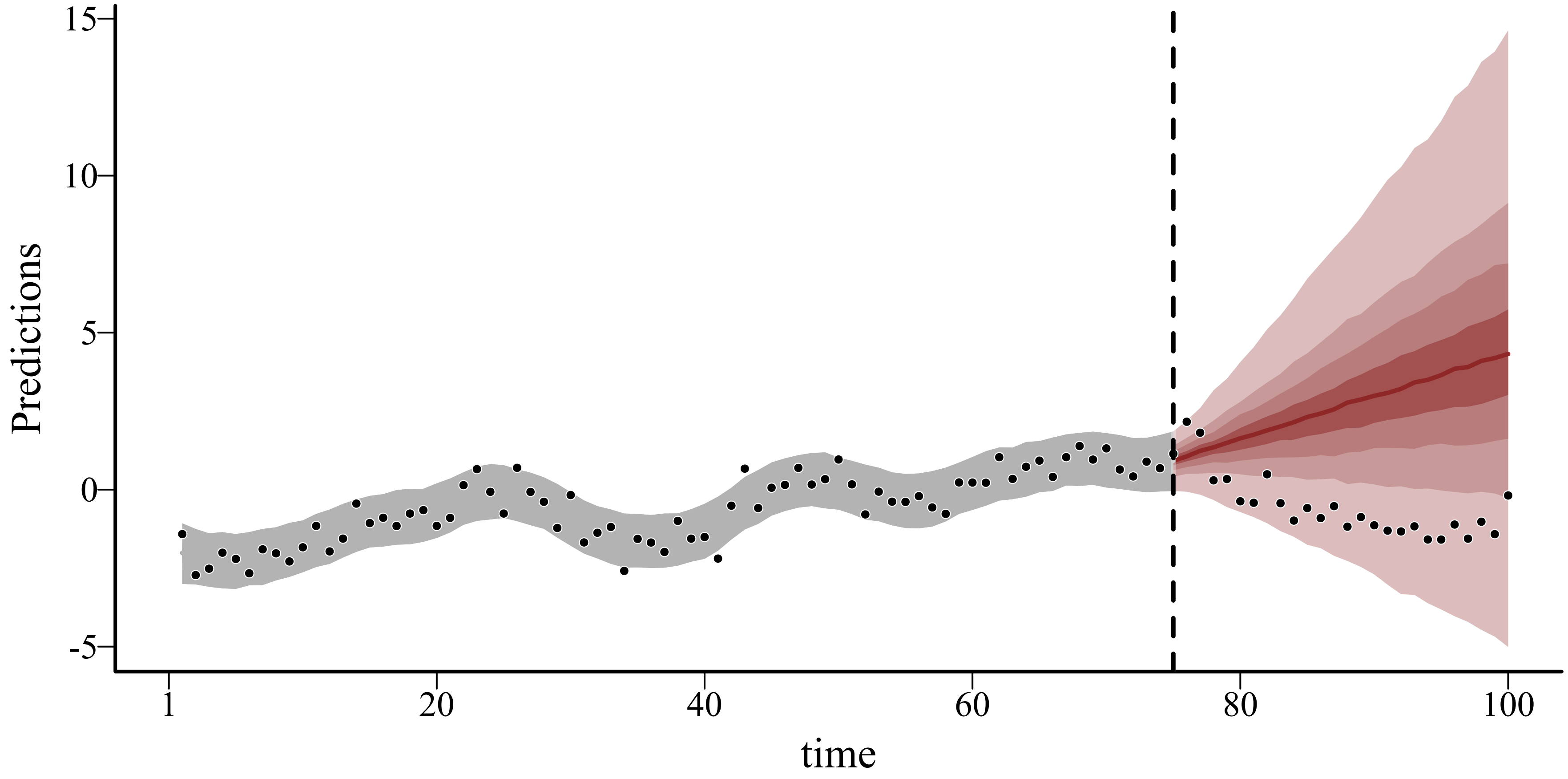
Basis functions ⇨ local knowledge



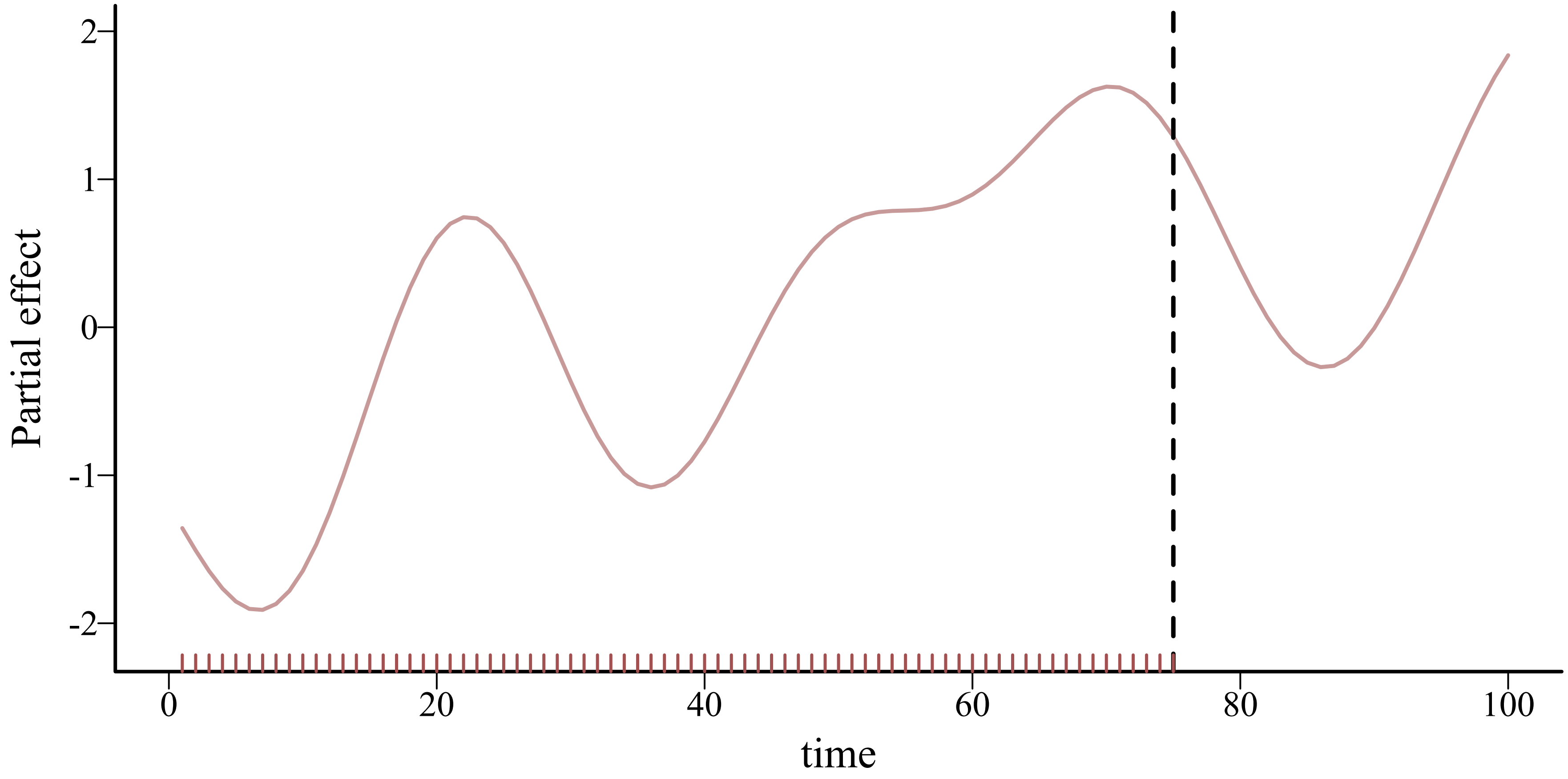
Basis functions ⇨ local knowledge



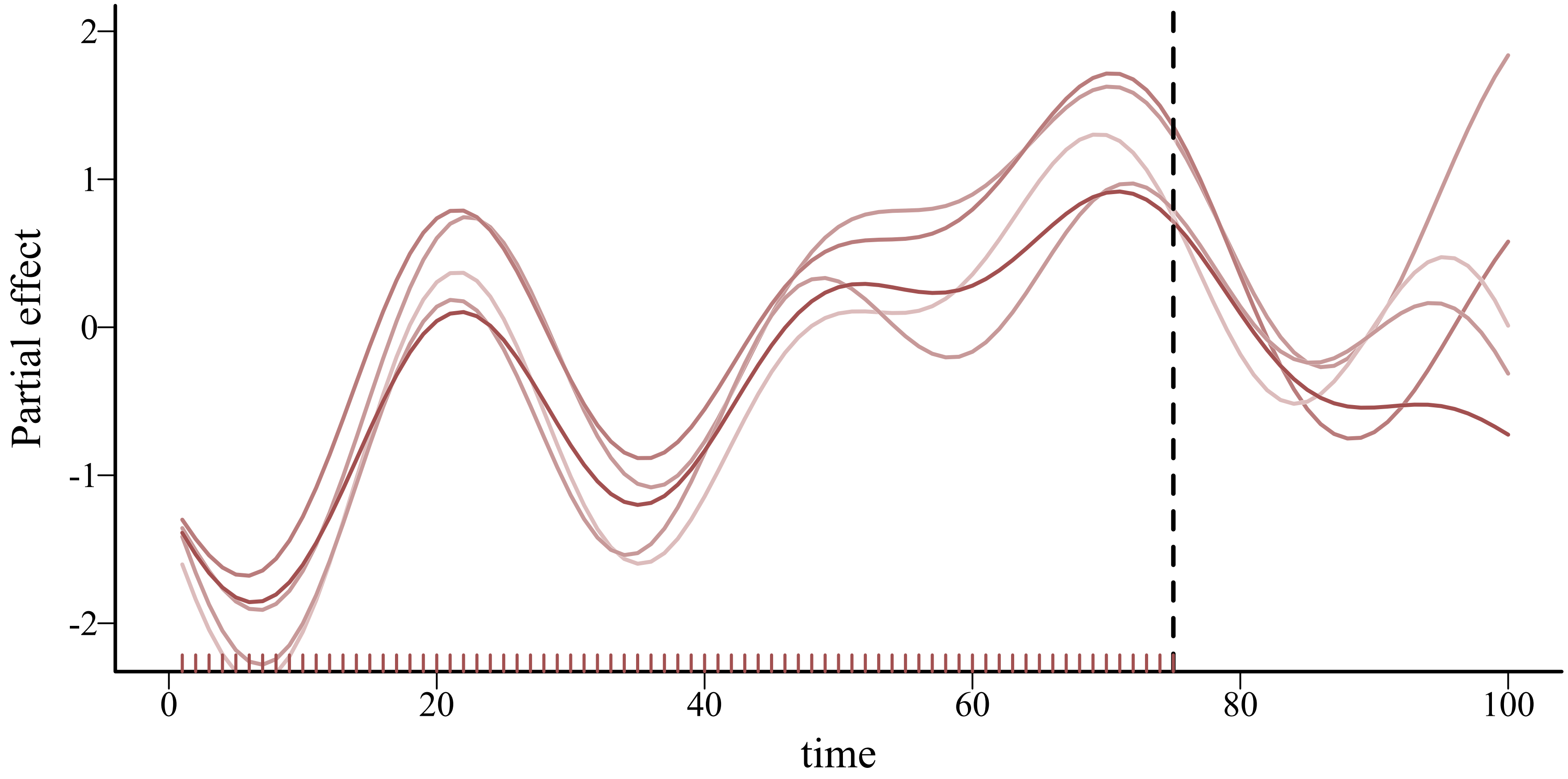
Forecasts 😰



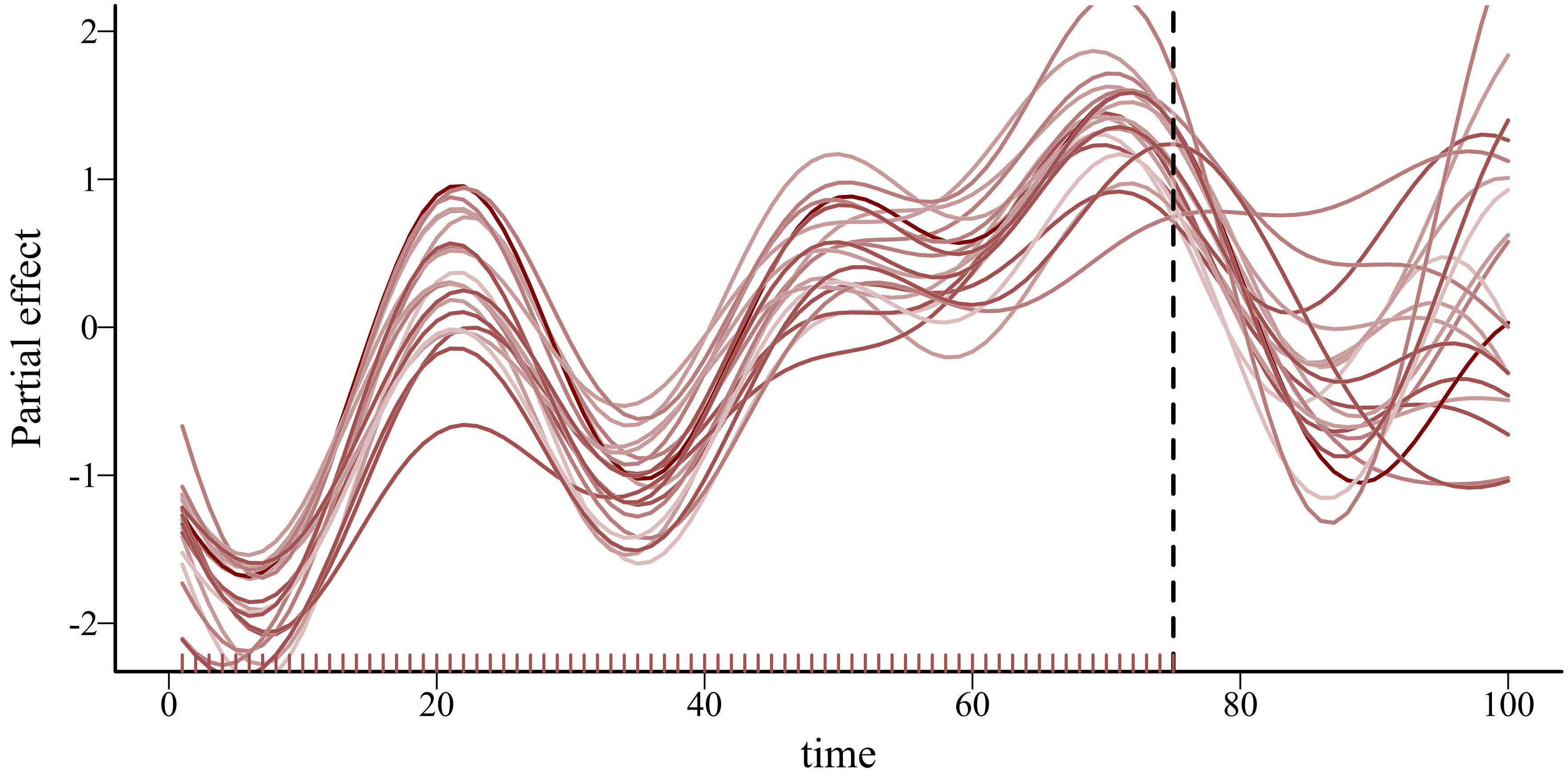
We need *global knowledge*



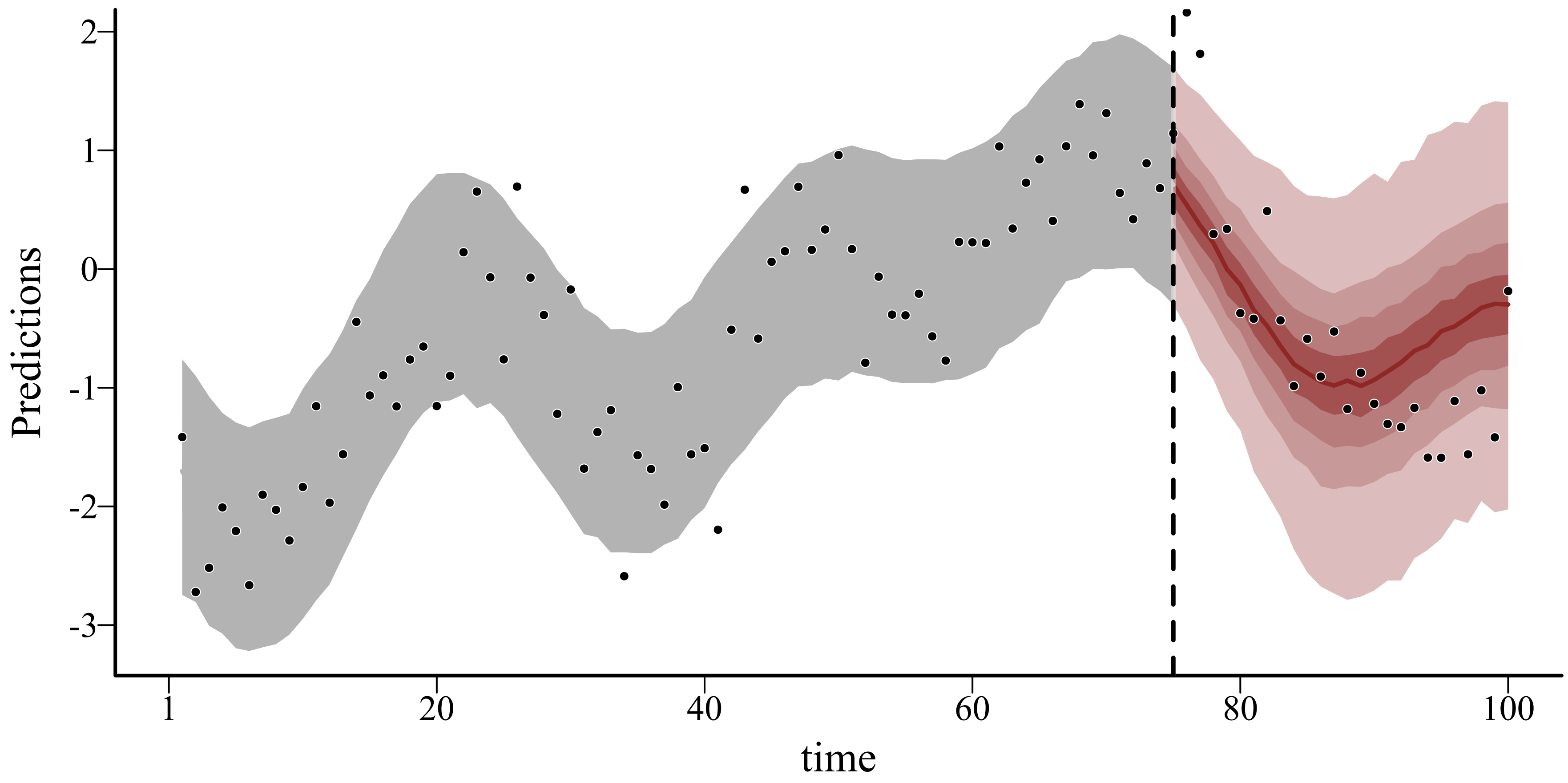
We need *global knowledge*



We need *global knowledge*



We need *global knowledge*



Dynamic GAMs

$$\mathbb{E}(\boldsymbol{Y_t}|\boldsymbol{X_t}) = g^{-1}(\alpha + \sum_{j=1}^{J} f(x_{jt}) + z_t)$$

Where:

g^{-1} is the inverse of the link function

α is the intercept

$f(x)$ are potentially nonlinear functions of the J predictors

z_t is a ***latent dynamic process***

Modelling with the `mvgam` 📦

- Bayesian framework to fit Dynamic GLMs and Dynamic GAMs
  - Hierarchical intercepts, slopes and smooths
  - Latent dynamic processes
  - State-Space models with measurement error
- Built off the `mgcv` 📦 to construct penalized smoothing splines
- Convenient and familiar R formula interface
- Uni- or multivariate series from a range of response distributions
- Uses Stan for efficient Hamiltonian Monte Carlo sampling

Observation families

`gaussian()`, `student-t()` ⇨ real values in $(-\infty, \infty)$

`lognormal()`, `Gamma()` ⇨ positive real values in $[0, \infty)$

`betar()` ⇨ real values (proportional) in $[0, 1]$

`poisson()`, `nb()` ⇨ non-negative integers in $(0, 1, 2, \dots)$

Extended predictor effects

`s()` ⇨ Smoothing spline of one or more covariates

`s(bs = 're')` ⇨ Hierarchical slopes or intercepts

`te(), ti(), t2()` ⇨ Tensor product smoothing spline of two or more covariates

`gp()` ⇨ Gaussian Process function (with squared exponential kernel) of one covariate

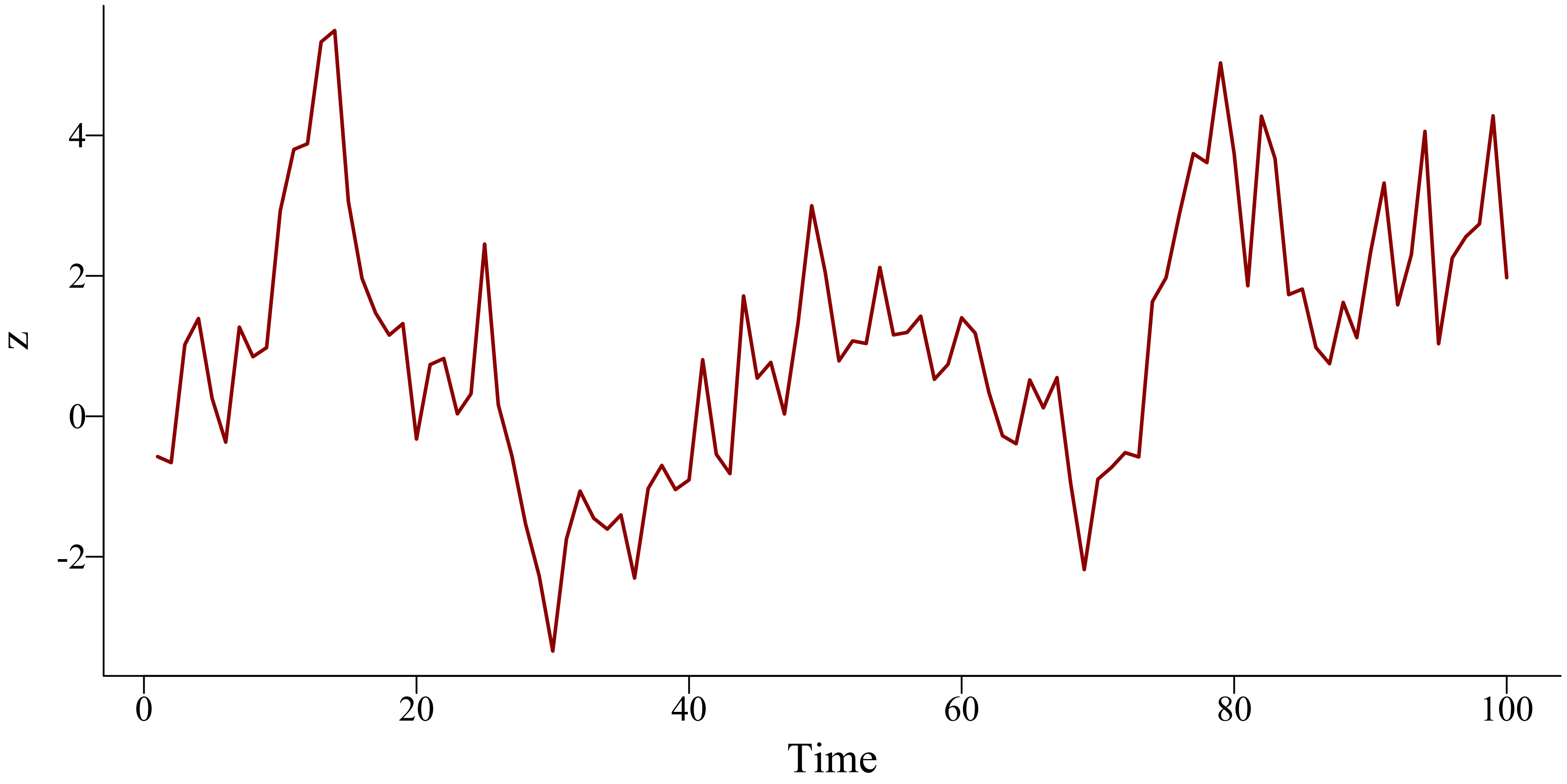
`dynamic()` ⇨ Time-varying effect of one covariate

We can fit models that include random effects, nonlinear effects and complex multidimensional smooth functions. All these effects can operate *on both process and observation models*

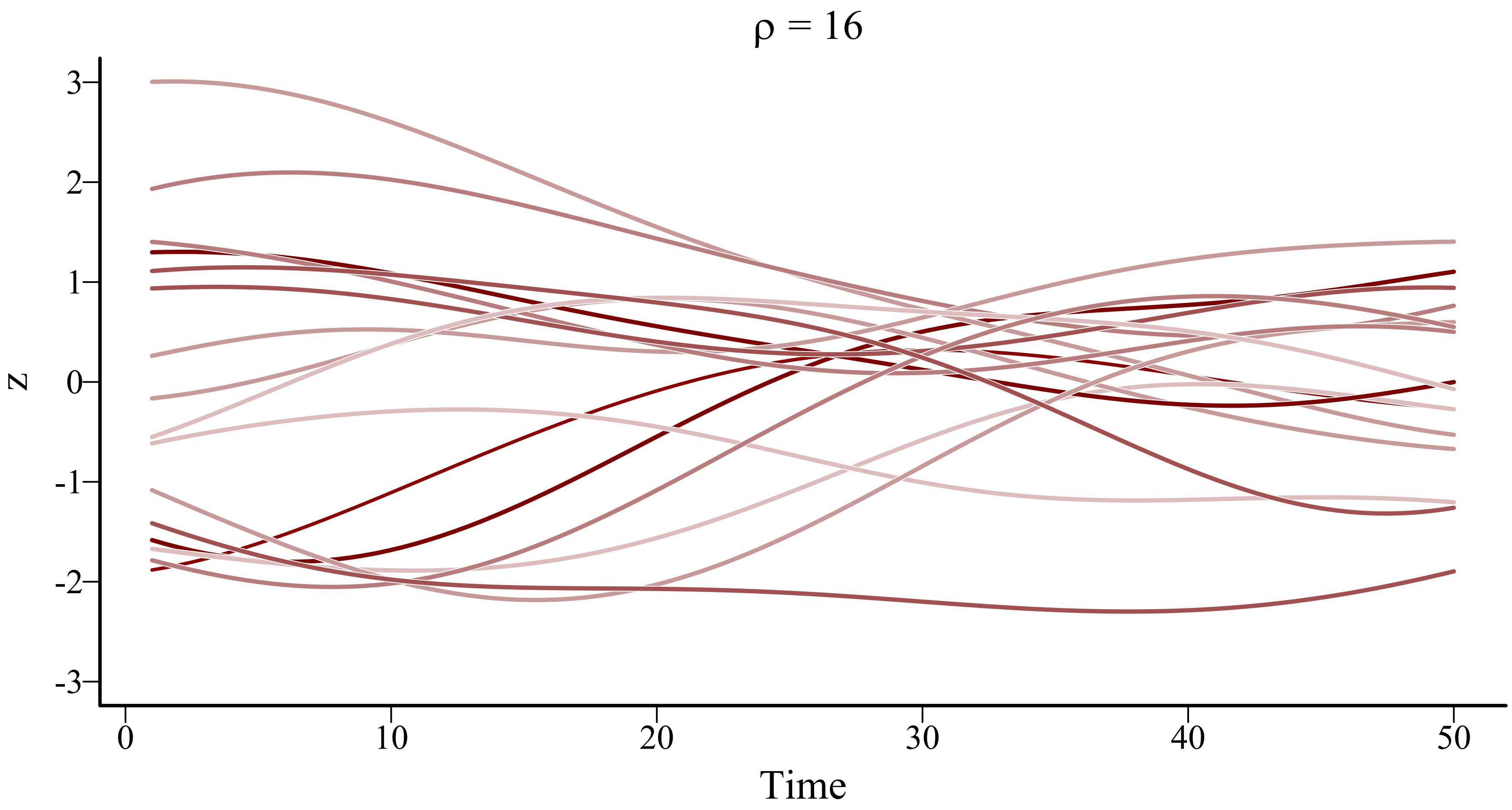
Can incorporate unobserved temporal dynamics; do not need to regress the outcome on its own past values

Very advantageous for ecological time series. But what kinds of dynamic processes are available in the `mvgam` 📦 ?

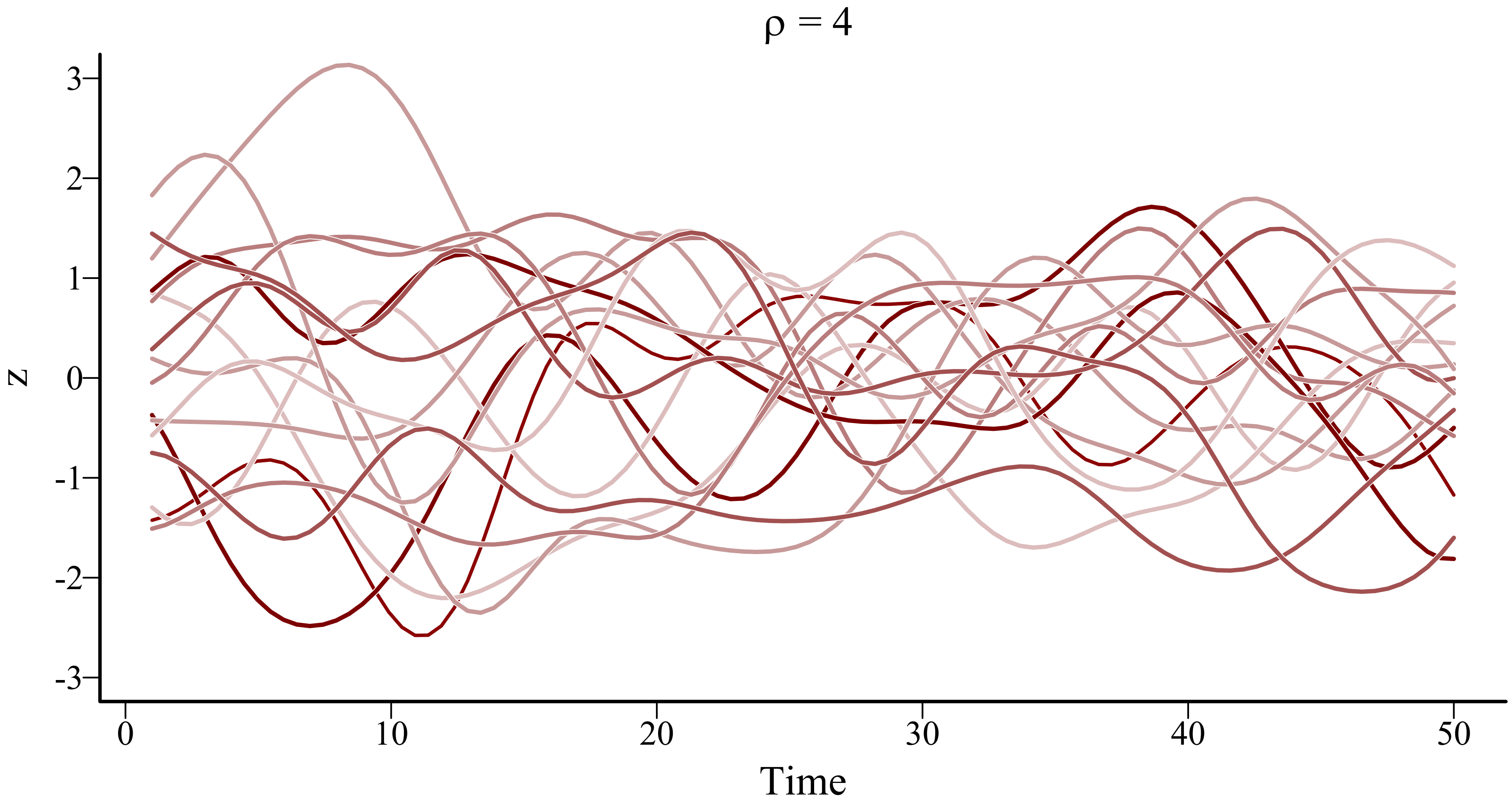
Random Walk or AR(1-3)



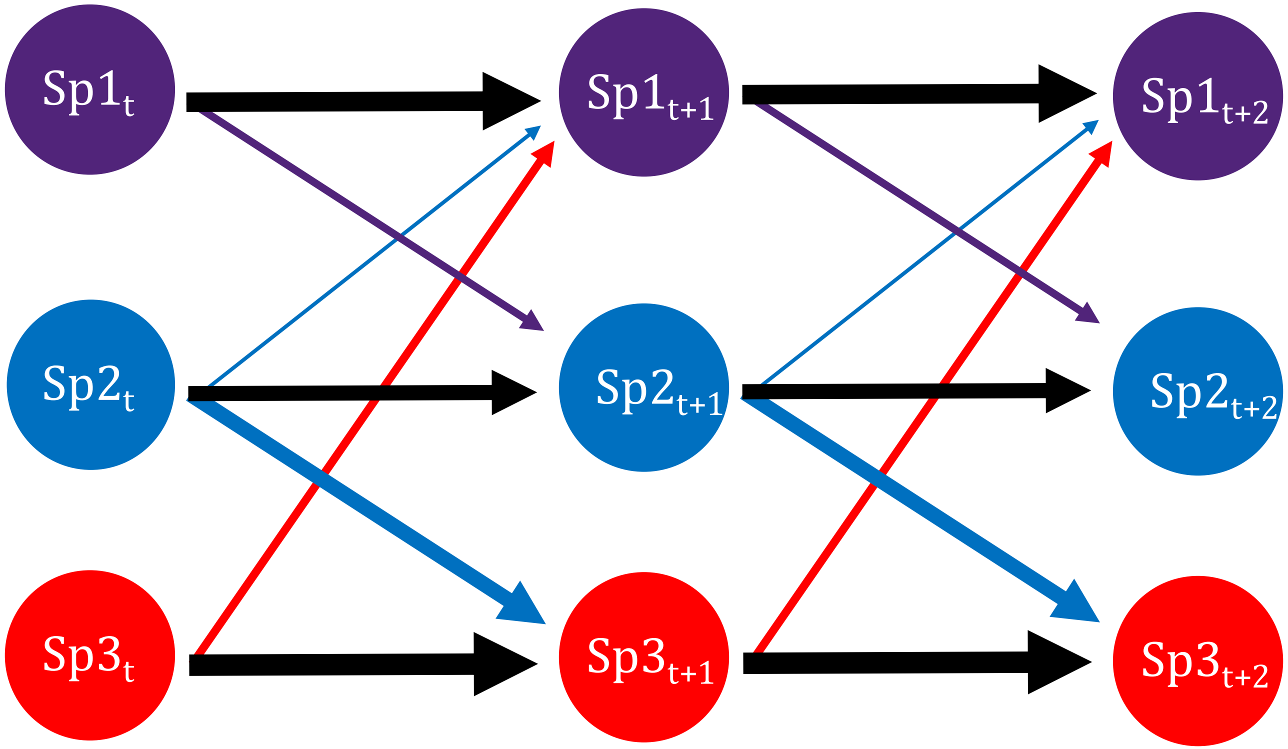
Gaussian Process...



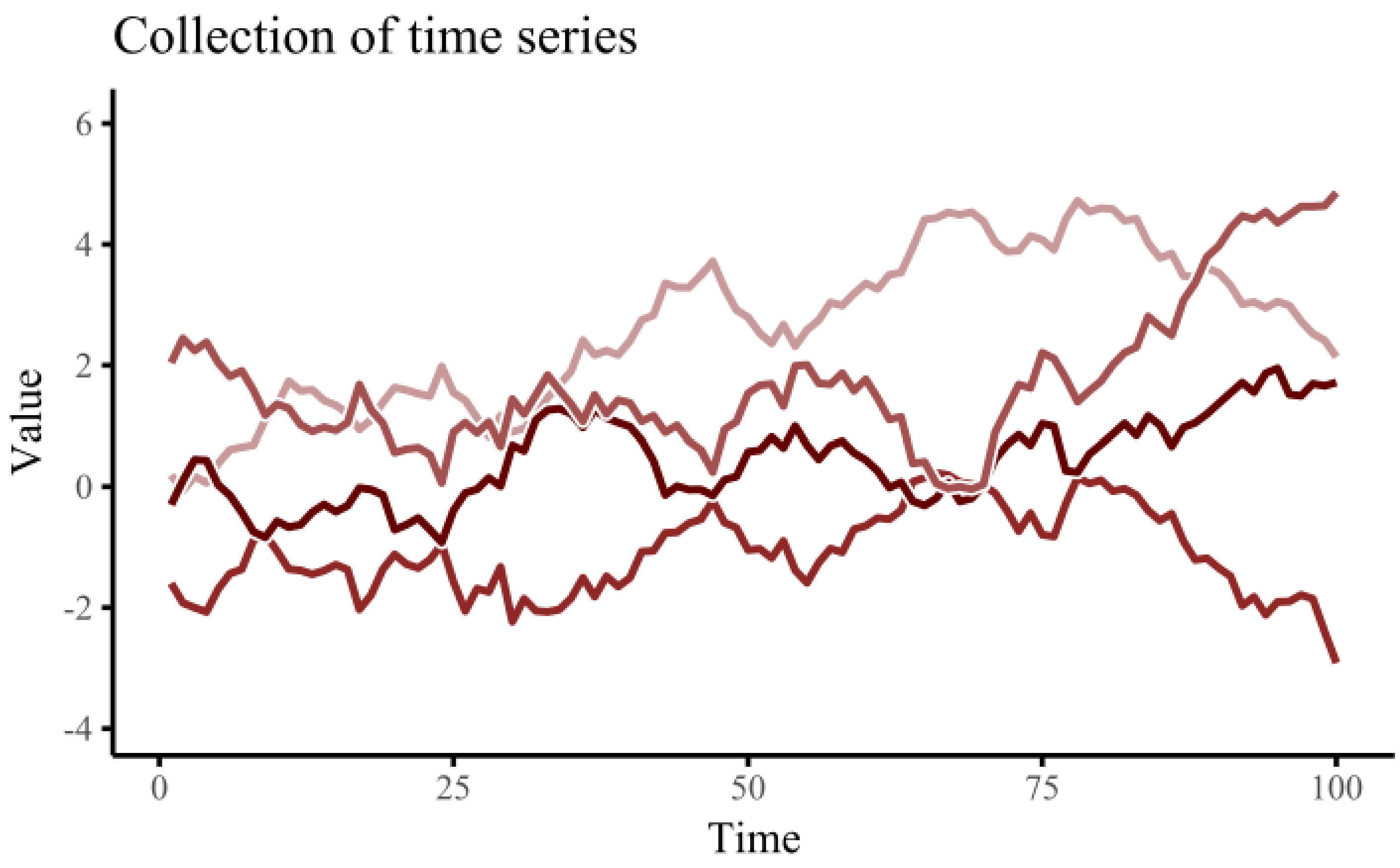
...where length scale ⇨ *memory*



VAR1 ⇨ Granger causality



Factors ⇨ induced correlations



Example of the interface

```r
model <- mvgam(
  formula = y ~
    s(series, bs = 're') +
    s(x0, series, bs = 're') +
    x1 +
    gp() +
    te(x3, x4, bs = c('cr', 'tp')),
  data = data,
  family = poisson(),
  trend_model = 'AR1',
  burnin = 500,
  samples = 500,
  chains = 4,
  parallel = TRUE
  )
```

Example data (long format)

y	series	time
2	species_1	1
0	species_2	1
NA	species_3	1
NA	species_4	1
1	species_1	2
0	species_2	2
3	species_3	2
5	species_4	2

Response (NAs allowed)

y	series	time
2	species_1	1
0	species_2	1
NA	species_3	1
NA	species_4	1
1	species_1	2
0	species_2	2
3	species_3	2
5	species_4	2

Series indicator (as `factor`)

y	series	time
2	species_1	1
0	species_2	1
NA	species_3	1
NA	species_4	1
1	species_1	2
0	species_2	2
3	species_3	2
5	species_4	2

Time indicator

y	series	time
2	species_1	1
0	species_2	1
NA	species_3	1
NA	species_4	1
1	species_1	2
0	species_2	2
3	species_3	2
5	species_4	2

Any other predictors

y	series	time	x0	x1	x2	x3	x4
2	species_1	1	-0.38	A	0.20	1.18	-0.72
0	species_2	1	-0.71	A	-2.67	1.02	0.67
NA	species_3	1	0.05	B	-0.33	0.12	1.50
NA	species_4	1	0.77	B	0.65	0.86	-0.49
1	species_1	2	0.29	A	-0.25	1.18	-0.82
0	species_2	2	0.34	A	-0.15	2.12	0.20
3	species_3	2	-0.38	B	-0.81	1.33	-1.15
5	species_4	2	1.32	B	0.22	-0.72	1.36

Types of mvgam predictions

$$\mathbb{E}(Y_t)$$

type = 'expected'
gives draws from the expected value of the posterior predictive distribution (i.e. the average of type = 'response')

type = 'response'
gives draws from a random Poisson distribution using λ_t

$$Y_t \sim \text{Poisson}(\lambda_t)$$

$$\log(\lambda_t) = \alpha + f(season)_t + z_t$$

In Poisson regression, this is the inverse of λ_t (i.e. $\exp(\lambda_t)$)

type = 'link'
gives posterior draws from the linear predictor on the log scale

modified from Heiss 2022

Workflow in `mvgam` 📦

Fit models that can include nonlinear splines, GPs, and multivariate dynamic processes to ecological time series

Use posterior predictive checks and Randomized Quantile (Dunn-Smyth) residuals to assess model failures

Use `marginaleffects` 📦 to generate interpretable (and reportable) model predictions

Produce probabilistic forecasts

Evaluate forecasts from competing models with proper scoring rules

More resources

Vignette ⇨ Overview of the package

Vignette ⇨ Formatting data for use in `mvgam`

Vignette ⇨ Shared latent process models

Vignette ⇨ Time-varying effects

Vignette ⇨ Multivariate State-Space models

Motivating publication ⇨ Clark & Wells 2023 *Methods in Ecology and Evolution*

Example

Long-term rodent monitoring

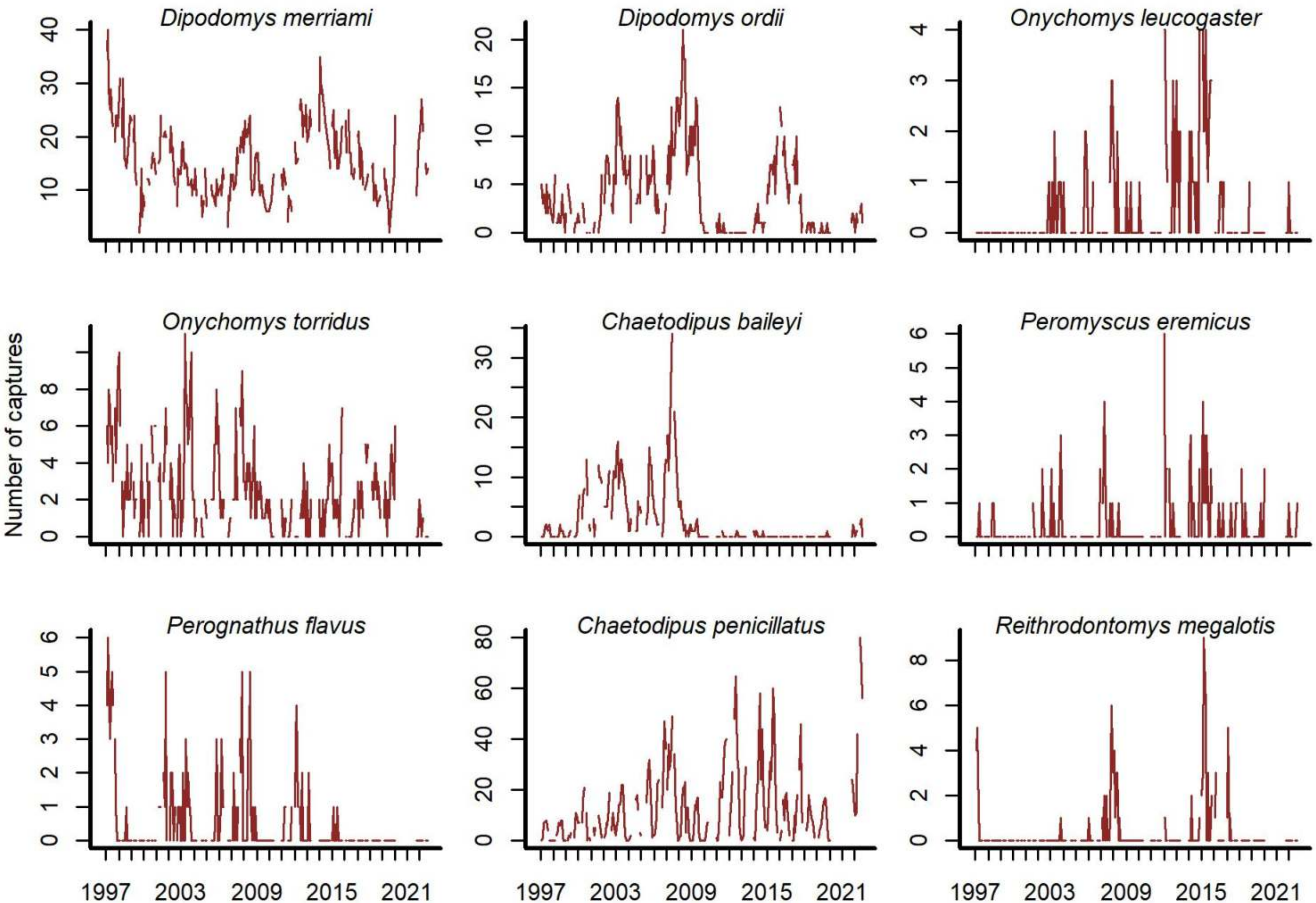
Monthly trapping in Portal, AZ, from 1977

20 rodent species recorded

Experimental evidence for interactions

Both gradual and sudden regime transitions

What drives community dynamics?



Dipodomys merriami
Dipodomys ordii
Onychomys leucogaster
Onychomys torridus
Chaetodipus baileyi
Peromyscus eremicus
Perognathus flavus
Chaetodipus penicillatus
Reithrodontomys megalotis
Number of captures
1997
2003
2009
2015
2021

$$for\ i = 1, \ldots, 9\ species$$
$$for\ t = 1, \ldots, 319\ time\ points$$
$$\mathrm{Y}_{1:9,t} \sim Poisson(\exp(X_{1:9,t}))$$
$$X_{1:9,t} \sim MVNormal\left(\mu_{1:9,t} + A * (X_{1:9,t-1} - \mu_{1:9,t-1}), \sigma * C * \sigma\right)$$
$$\mu_{i,t} = \beta_{NDVI\,[i]} * \mathrm{NDVI}_{\mathrm{MA}\,12\,[t]} + f_{global\ [t]}(\mathrm{lag}, \mathrm{temp}) + f_{species\ [i,t]}(\mathrm{lag}, \mathrm{temp})$$

GAM component

$$\beta_{NDVI} \sim Normal(\mu_{NDVI}, \sigma_{NDVI})$$
$$\mu_{NDVI} \sim Normal(0, 1)$$
$$\sigma_{NDVI} \sim InvGamma(2.37, 0.73)$$

Random slopes for species' responses to NDVI moving average

$$f_{global} = \sum(\beta_{global} * b_{global})$$
$$\beta_{global} \sim MVNormal(0, (\sum(\mathrm{S}_{global} * \lambda_{global})^{-1})$$
$$\lambda_{global} \sim Normal(30, 25)$$

Distributed lag for community response to lagged minimum temperature

$$f_{species\ [i]} = \sum(\beta_{species\ [i]} * b_{species\ [i]})$$
$$\beta_{species\ [i]} \sim MVNormal(0, (\sum(\mathrm{S}_{species\ [i]} * \lambda_{species\ [i]})^{-1})$$
$$\lambda_{species} \sim Normal(30, 25)$$

Distributed lags for species' responses to lagged minimum temperature

Multivariate dynamic component

$$A \in P(\mathbb{R})$$
$$P \sim Normal(0, 0.67)$$

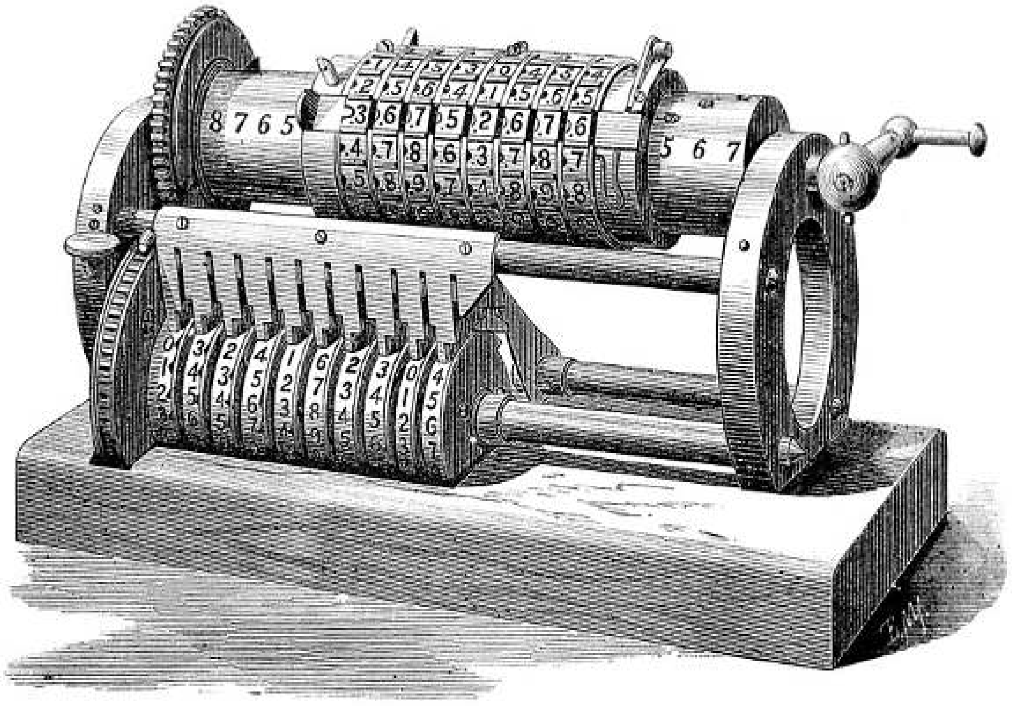
Vector autoregression for lagged multispecies dependence

$$\sigma \sim Beta(10, 10)$$
$$C \sim LKJcorr(2)$$

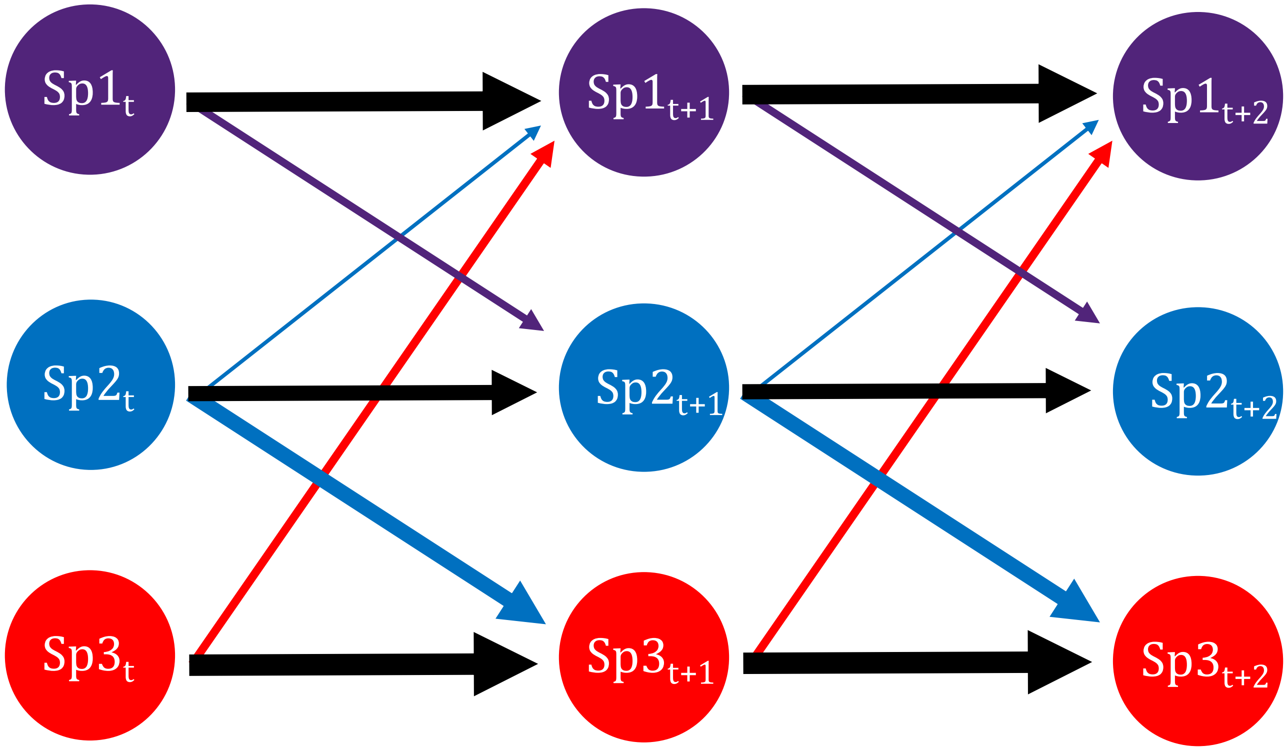
Temporally correlated process errors

Model components

1. Multispecies temporal dynamics



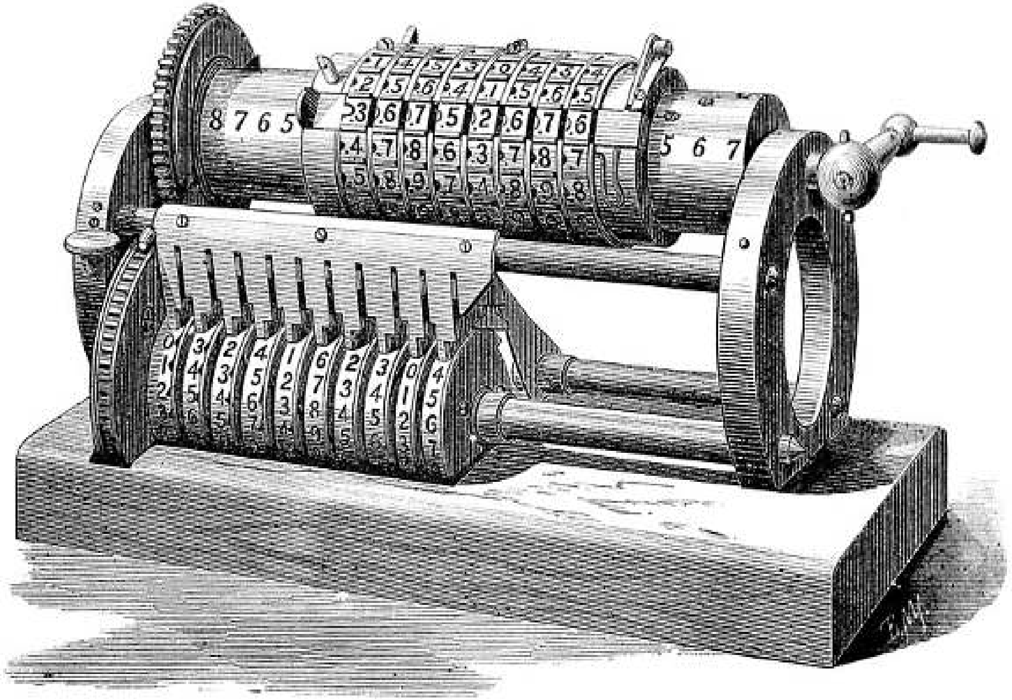
VAR1 ⇨ Granger causality

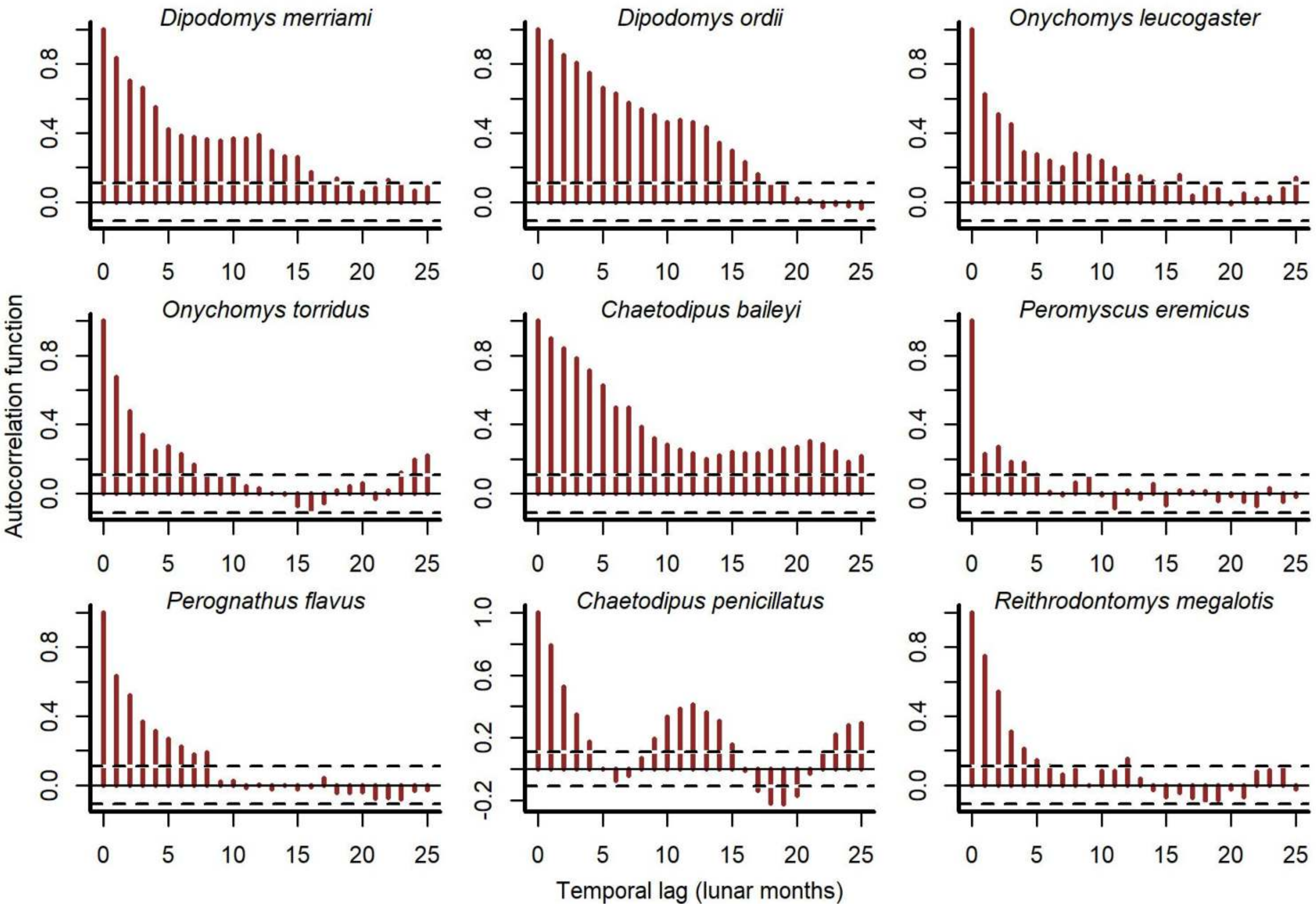


Model components

1. Multispecies temporal dynamics

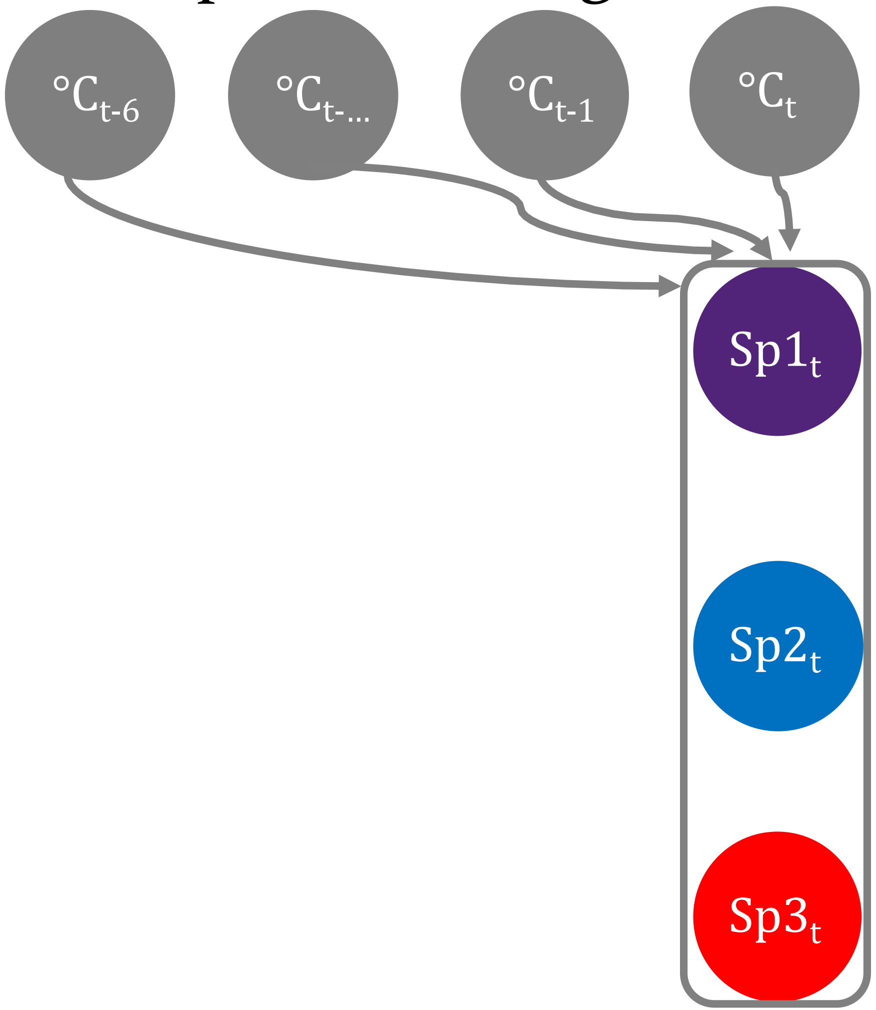
2. Seasonal variation



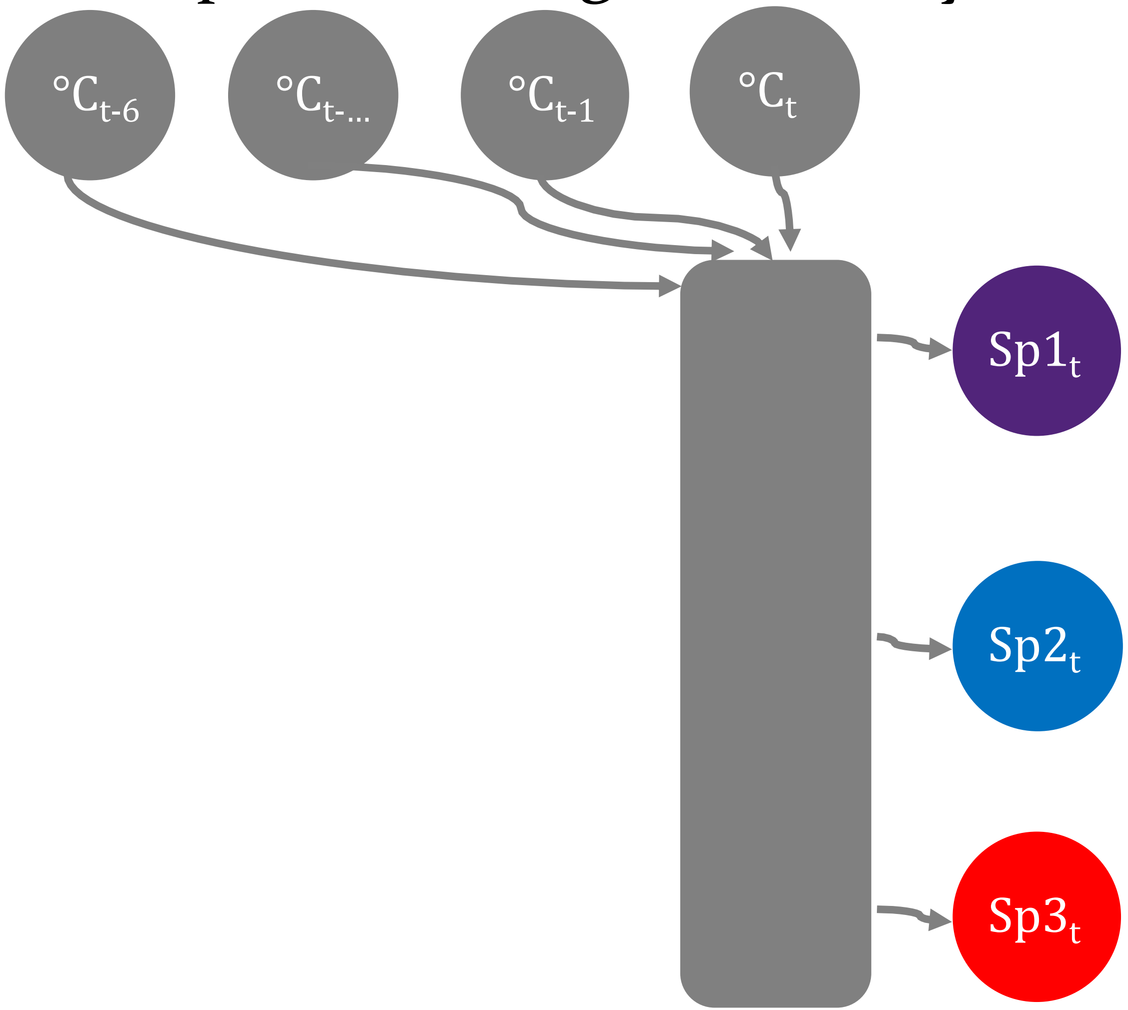


Dipodomys merriami
Dipodomys ordii
Onychomys leucogaster
Onychomys torridus
Chaetodipus baileyi
Peromyscus eremicus
Perognathus flavus
Chaetodipus penicillatus
Reithrodontomys megalotis
Autocorrelation function
Temporal lag (lunar months)
0.0
0.4
0.8
-0.2
0.2
0.6
1.0
0
5
10
15
20
25

Community mintemp effect changes smoothly over lags



Community mintemp effect changes smoothly over lags



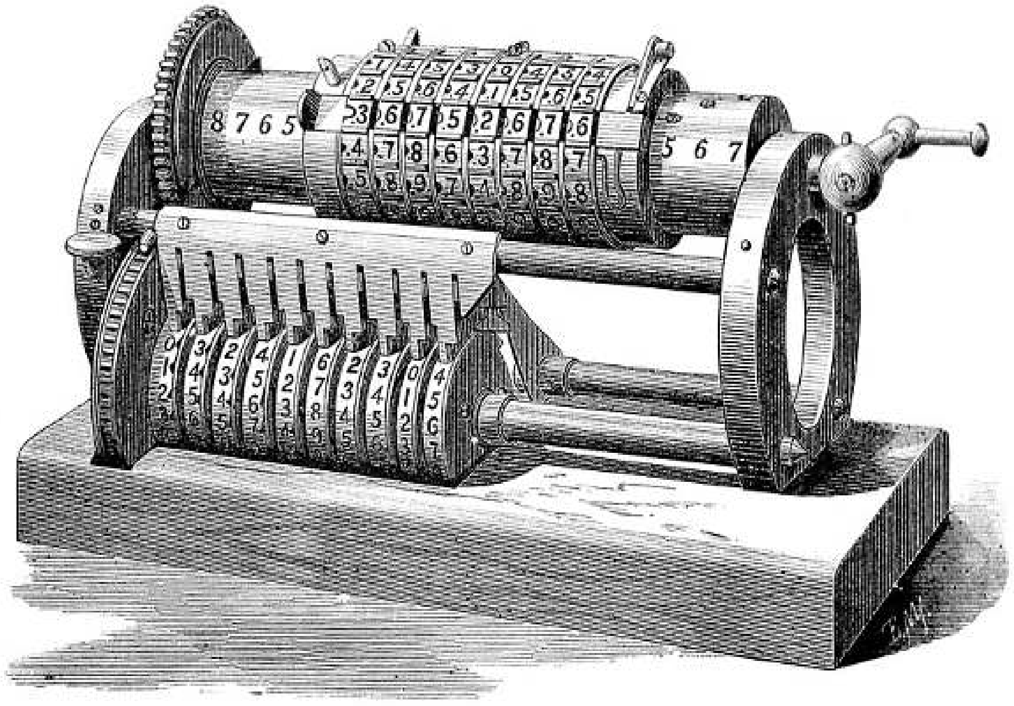
Species mintemp effects deviate smoothly from the community effect

Model components

1. Multispecies temporal dynamics

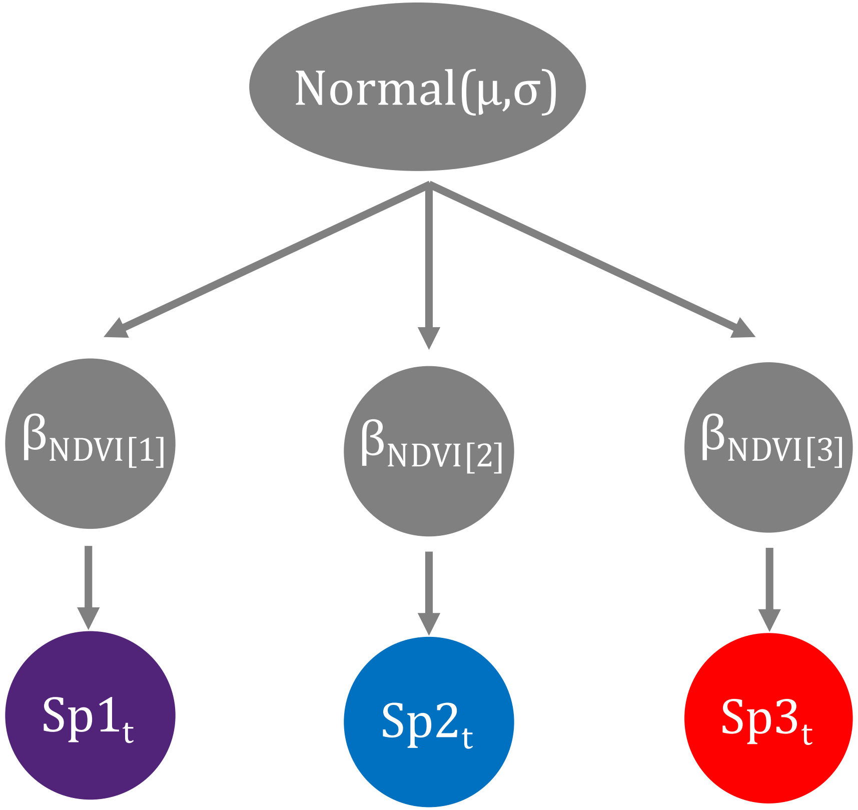
2. Seasonal variation

3. ***Vegetation associations***

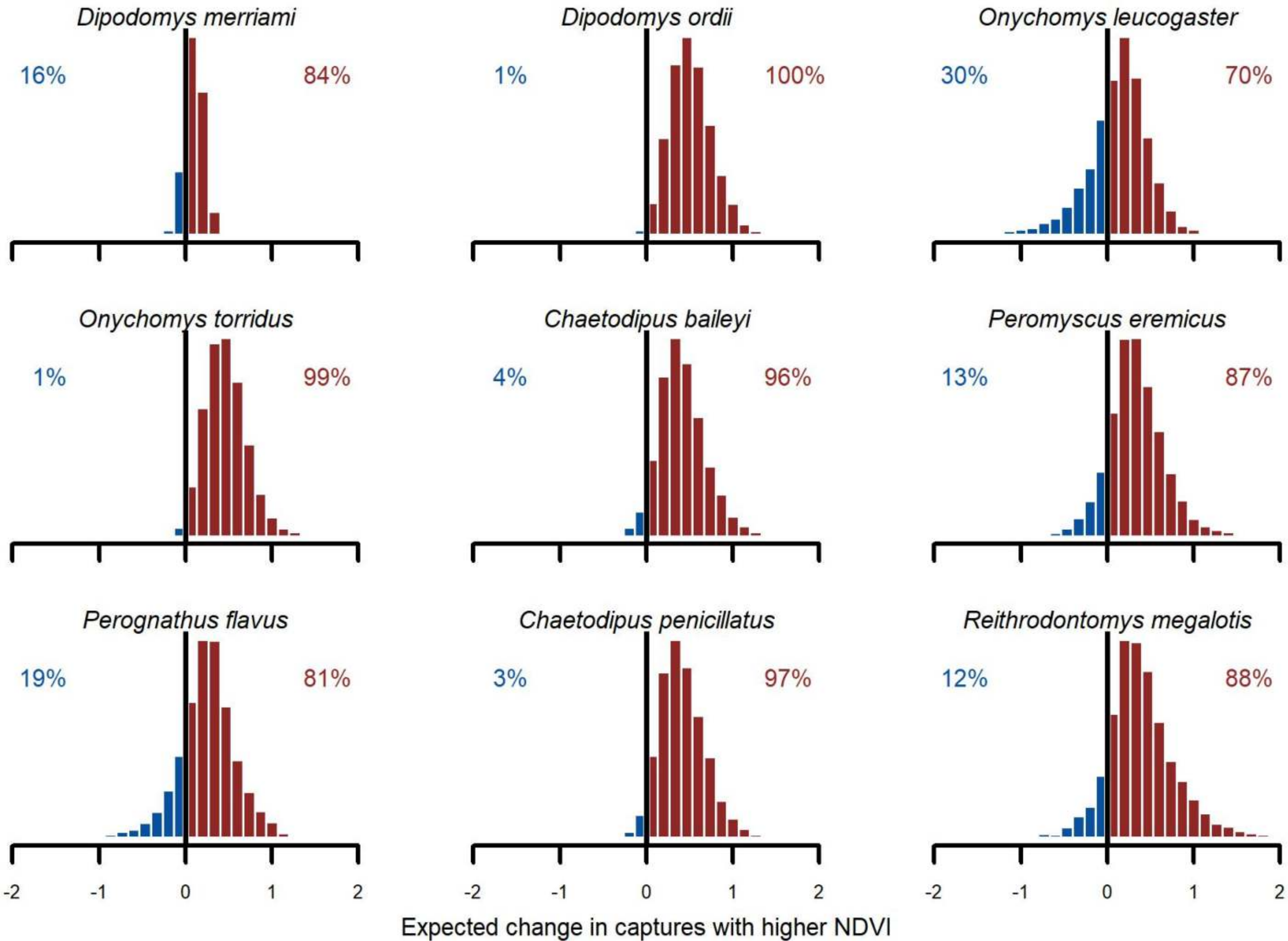



Sunrise in Cave Creek Canyon
Debra Davison

Hierarchical slopes for associations with NDVI moving average

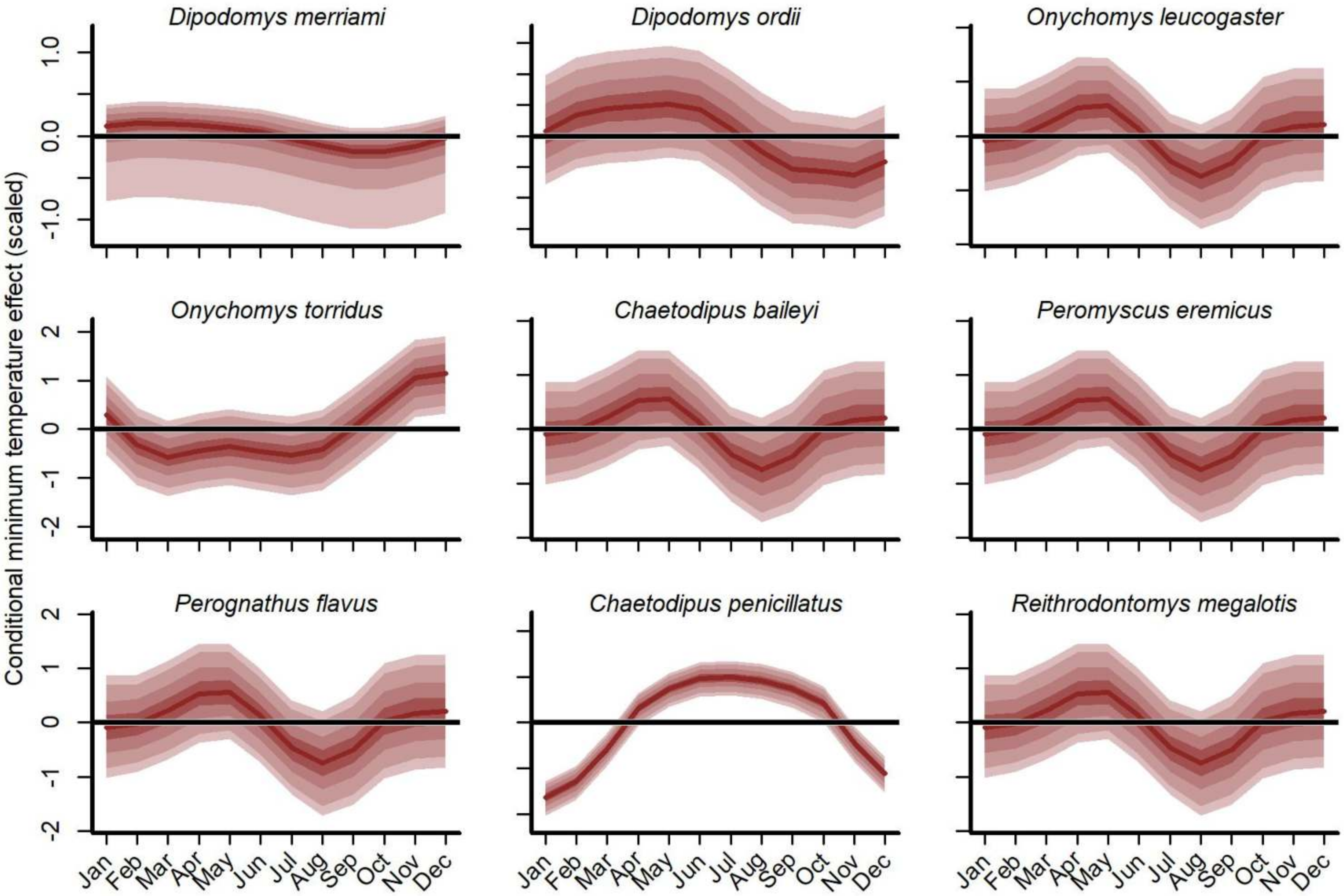


What did we find?


Dipodomys merriami
16%
84%
Dipodomys ordii
1%
100%
Onychomys leucogaster
30%
70%
Onychomys torridus
1%
99%
Chaetodipus baileyi
4%
96%
Peromyscus eremicus
13%
87%
Perognathus flavus
19%
81%
Chaetodipus penicillatus
3%
97%
Reithrodontomys megalotis
12%
88%
-2
-1
0
1
2
Expected change in captures with higher NDVI

What did we find?

1. Higher NDVI = more captures

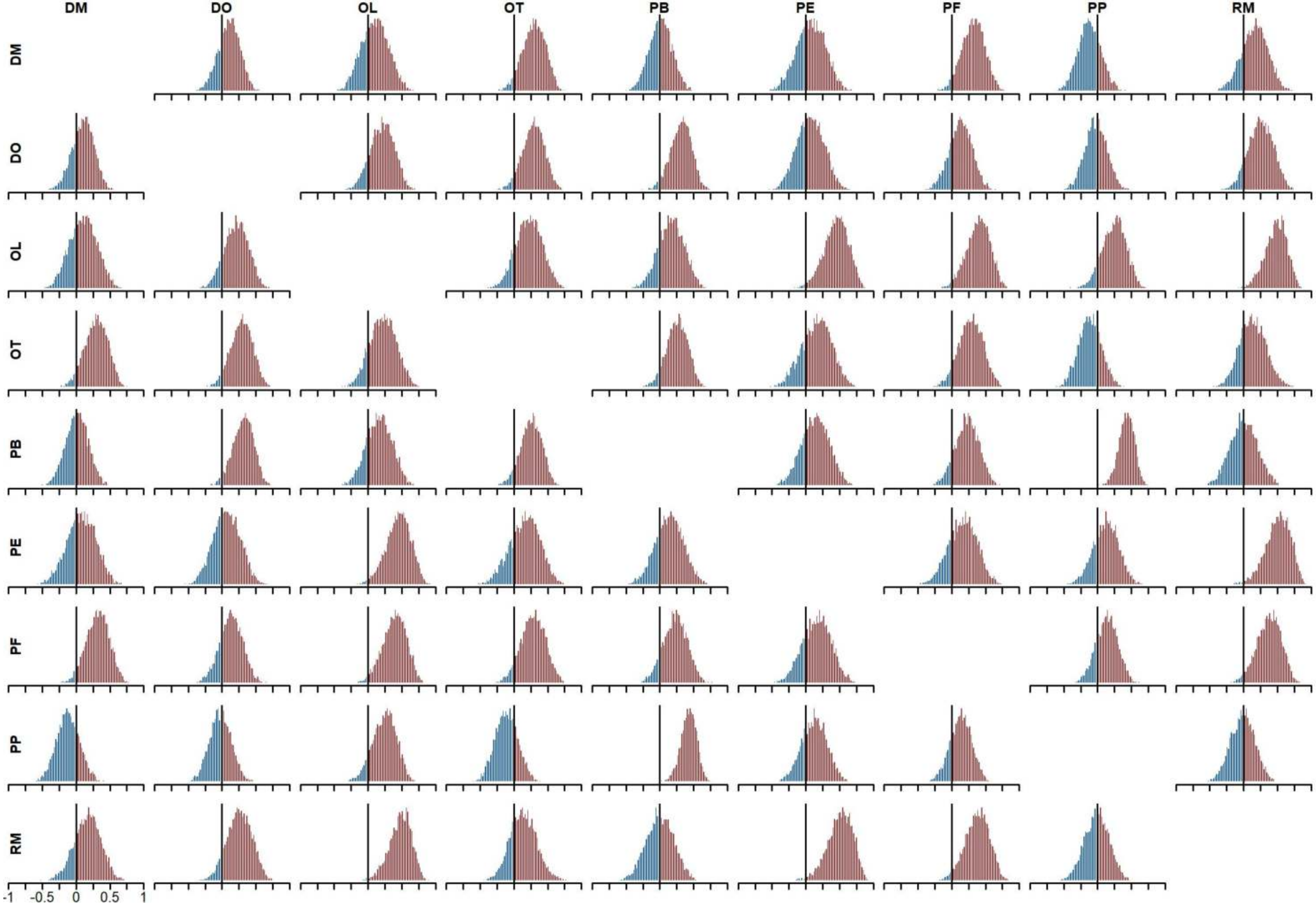


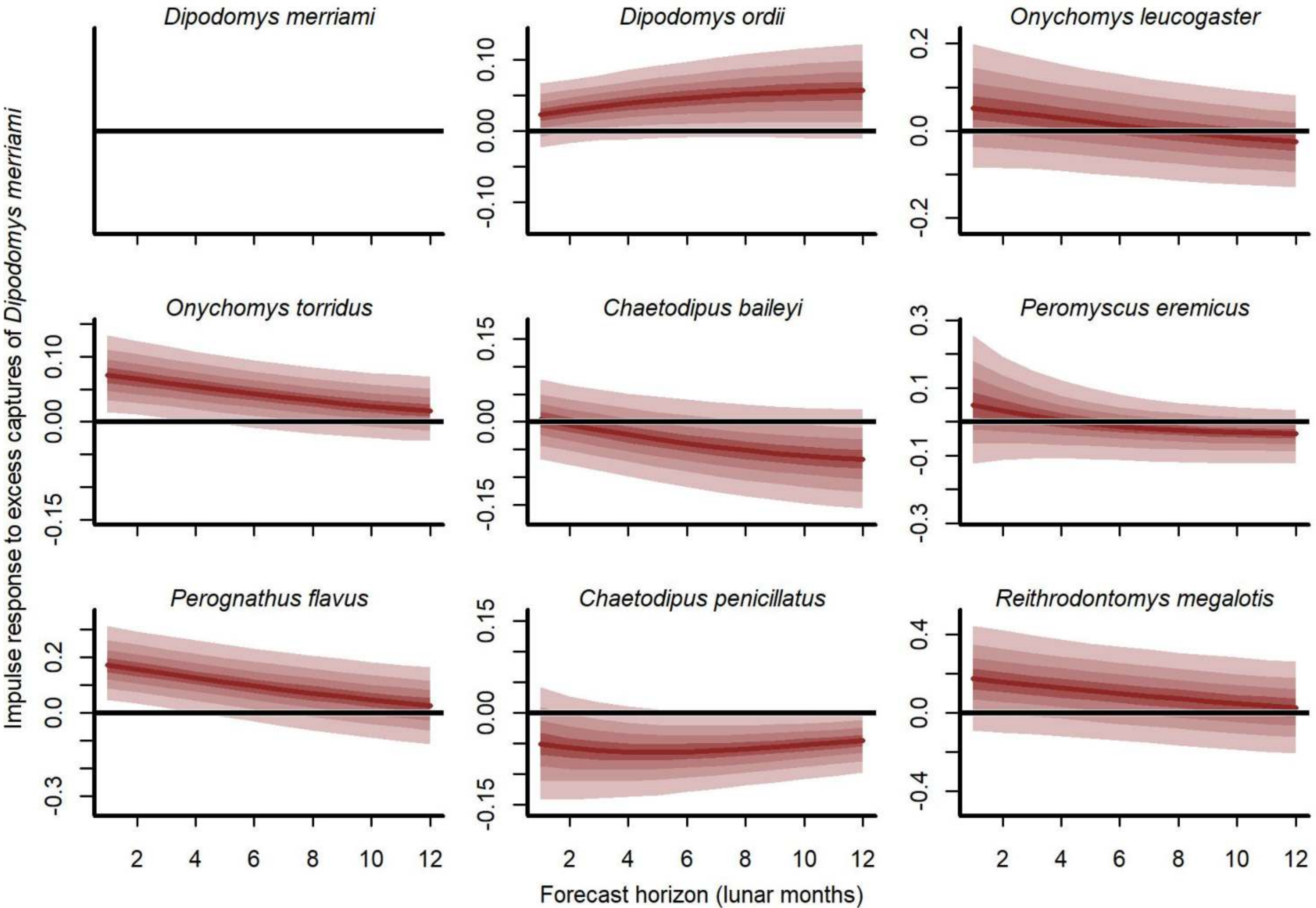
Dipodomys merriami
Dipodomys ordii
Onychomys leucogaster
Onychomys torridus
Chaetodipus baileyi
Peromyscus eremicus
Perognathus flavus
Chaetodipus penicillatus
Reithrodontomys megalotis
Conditional minimum temperature effect (scaled)
1.0
0.0
-1.0
2
1
0
-1
-2
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

What did we find?

1. Higher NDVI = more captures

2. Varying seasonal effects


DM
DO
OL
OT
PB
PE
PF
PP
RM
-1
-0.5
0
0.5
1



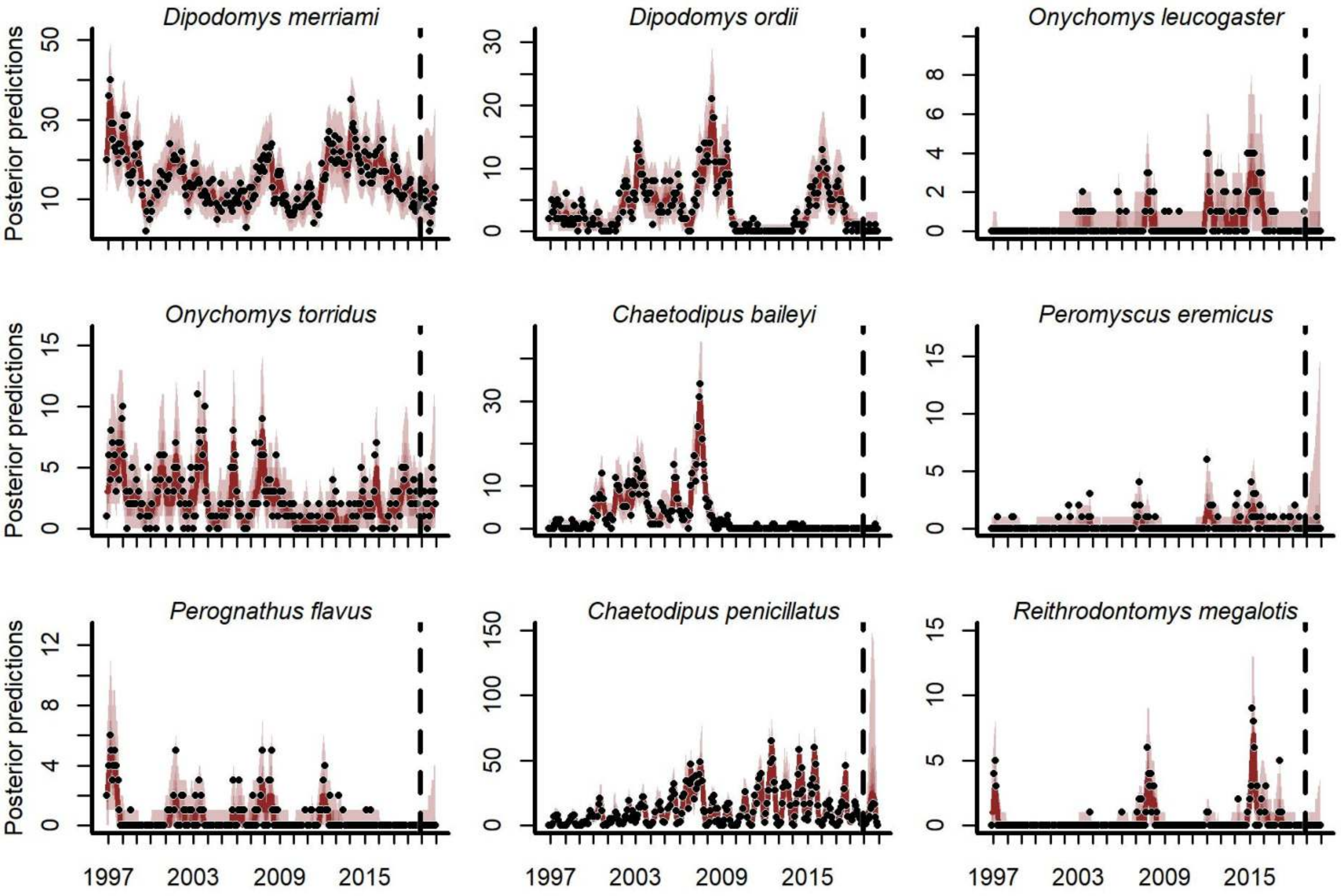
Dipodomys merriami
Dipodomys ordii
Onychomys leucogaster
Onychomys torridus
Chaetodipus baileyi
Peromyscus eremicus
Perognathus flavus
Chaetodipus penicillatus
Reithrodontomys megalotis
Impulse response to excess captures of Dipodomys merriami
Forecast horizon (lunar months)

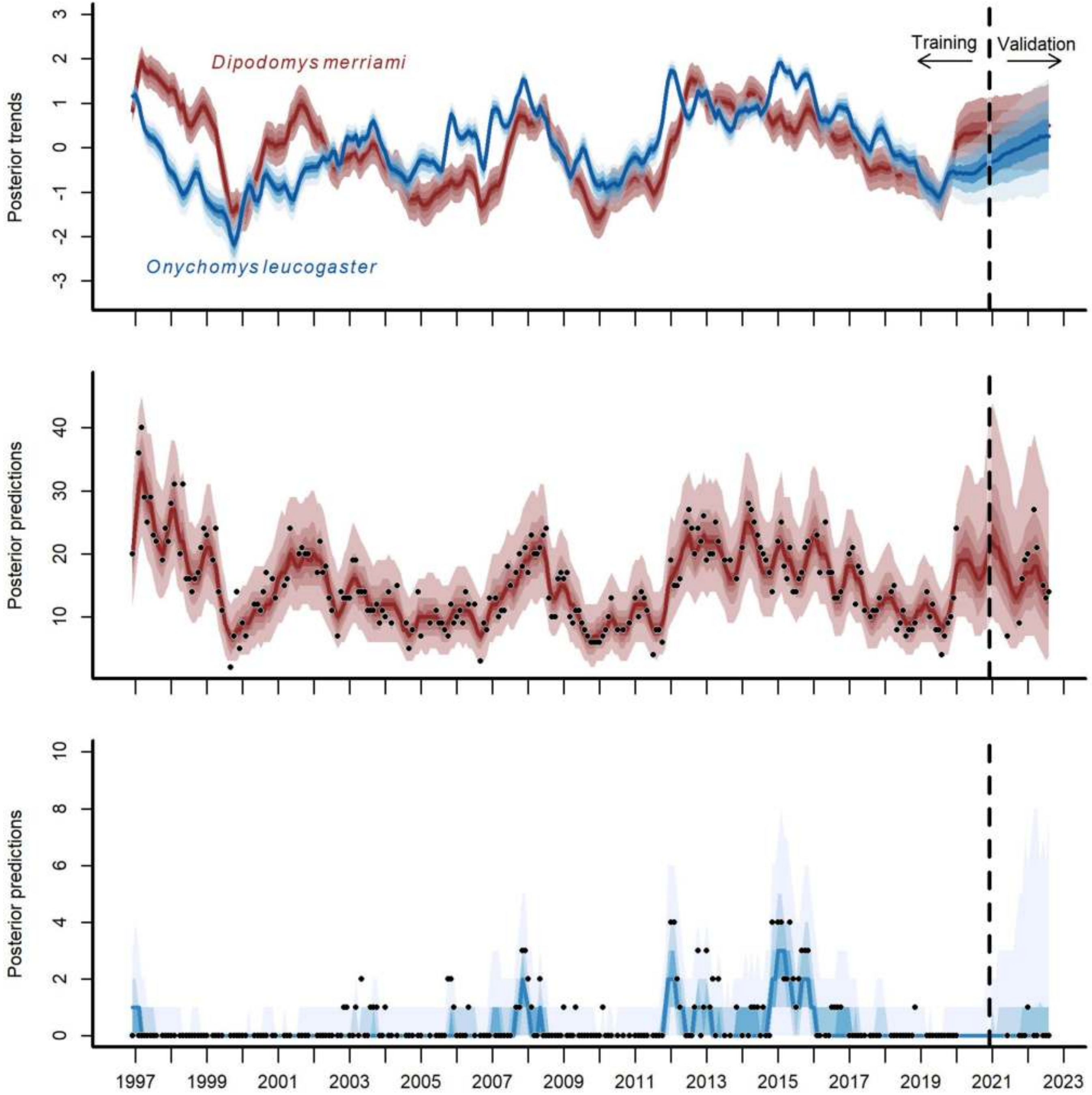
What did we find?

1. Higher NDVI = more captures

2. Varying seasonal effects

3. Complex dependencies


Dipodomys merriami
Dipodomys ordii
Onychomys leucogaster
Onychomys torridus
Chaetodipus baileyi
Peromyscus eremicus
Perognathus flavus
Chaetodipus penicillatus
Reithrodontomys megalotis
Posterior predictions
1997
2003
2009
2015



Dipodomys merriami
Onychomys leucogaster
Training
Validation
Posterior trends
Posterior predictions
Posterior predictions
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021
2023

What did we find?

1. Higher NDVI = more captures

2. Varying seasonal effects

3. Complex dependencies

4. Improved forecasts

Thank you.
Questions?