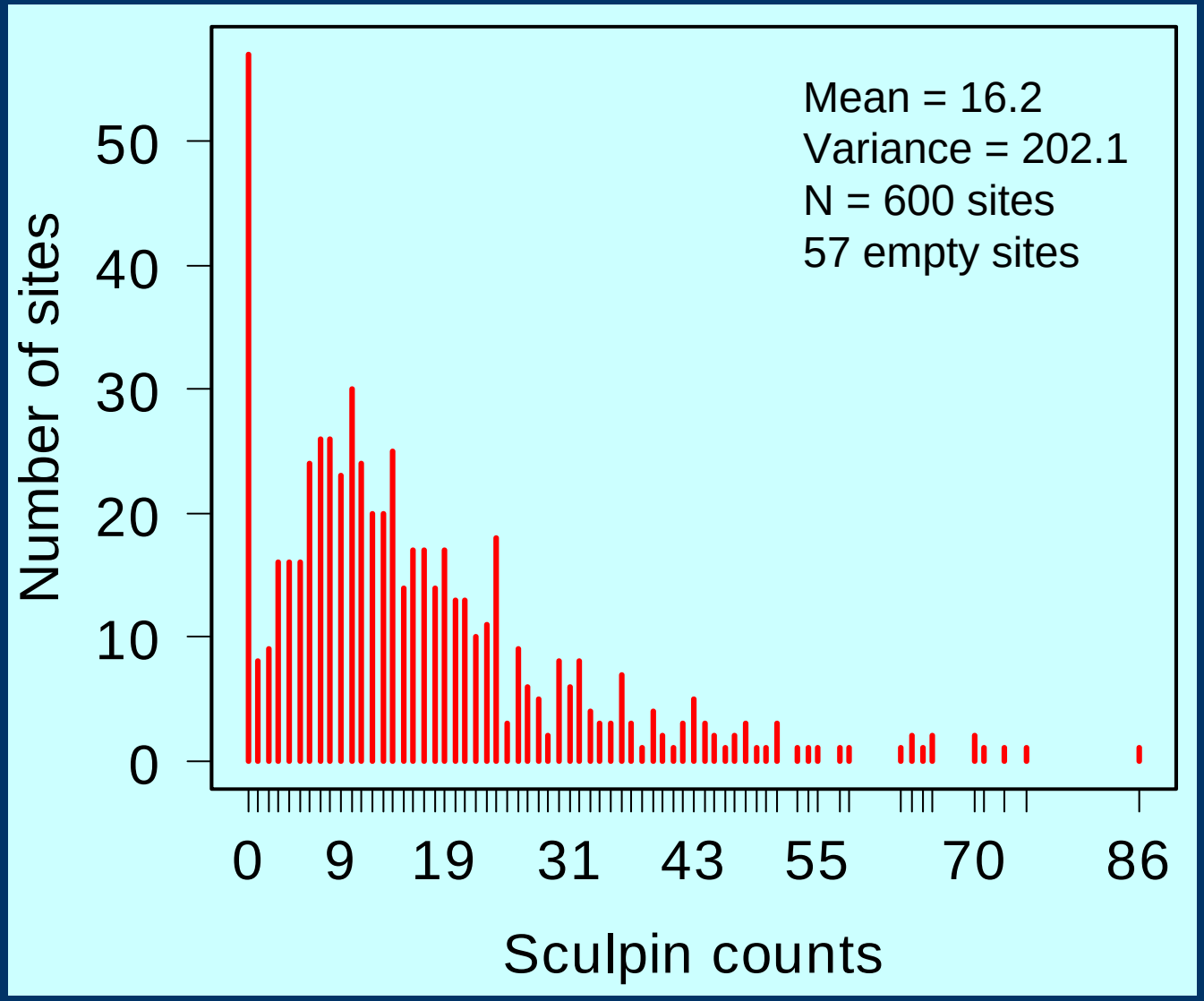
Zero-inflated Regression for Modeling Species Abundance in Relation to Habitat: A Bayesian approach

M.A. Rodríguez, C.G.B. Demétrio, S.S. Zocchi,
R.A. Leandro, J. Deschênes

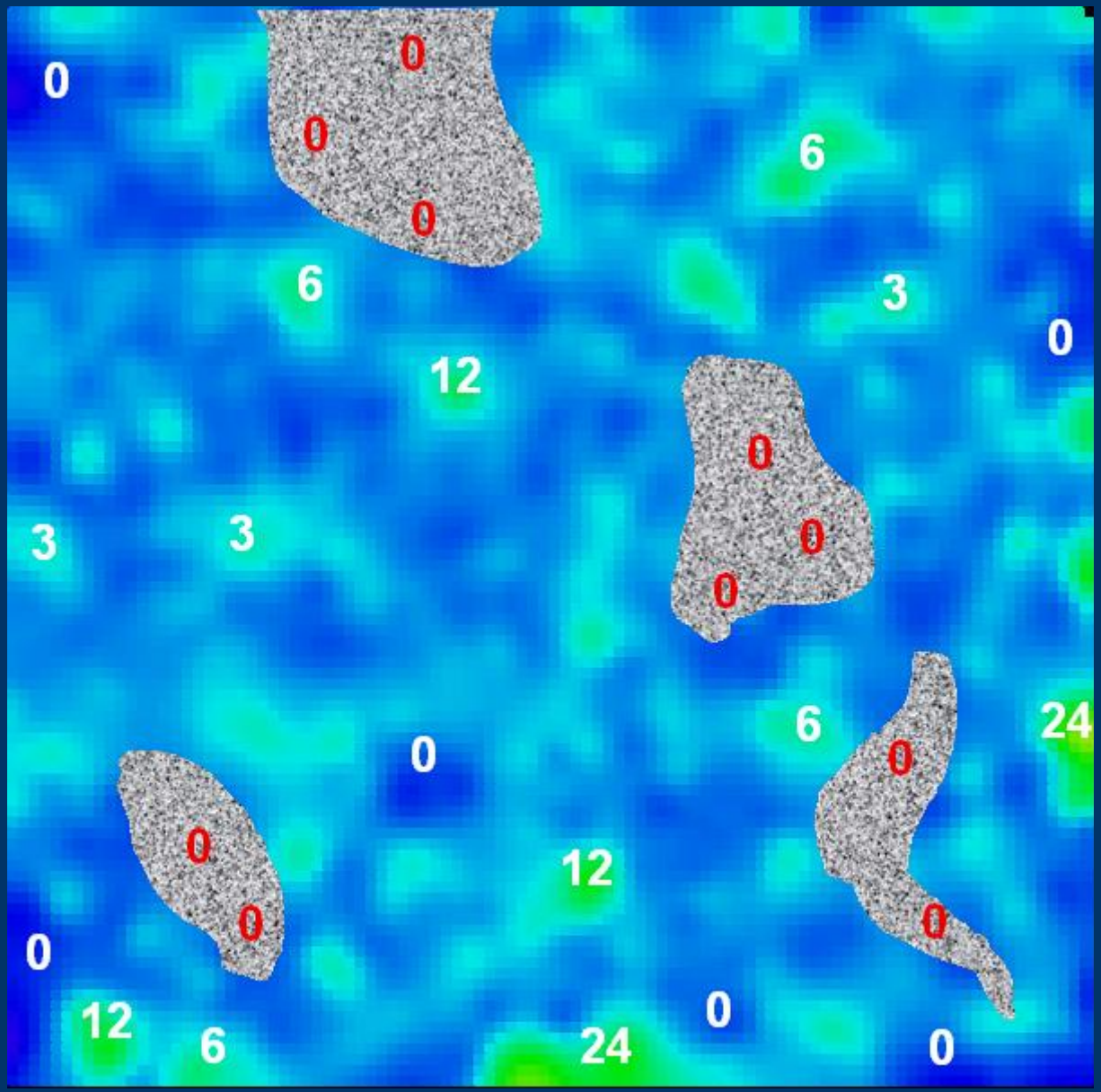
Ecologists have traditionally relied on ordinary least-squares or Poisson regression for linking species abundance to environmental features, but these approaches are often limited by restrictive assumptions

This talk presents a case study illustrating some of the limitations of traditional regression approaches in ecological studies and the use of alternative methods to counter these limitations

When nature does not abhor a vacuum:



Distribution of counts highly over-dispersed with excess zeros; nearly 10% of the sampling units have zero counts



Structural zeros vs. sampling zeros

The zero-inflated negative binomial (ZINB) distribution is a mixture of a Bernoulli distribution and a negative binomial distribution

$$P(Y = 0) = 1 - p + p\, t^{k}, \qquad 0 < p < 1$$

$$P(Y = y) = p\, C(y + k - 1, y)\, t^{k} (1 - t)^{y}, \qquad y = 1, 2, 3, \ldots$$

$t = \dfrac{k}{k + \mu}$, μ = mean of the underlying negative binomial distribution

$$\mathrm{E}(Y) = p\mu$$

$$\mathrm{Var}(Y) = p\mu\,(1 + \mu/k + (1 - p)\,\mu)$$

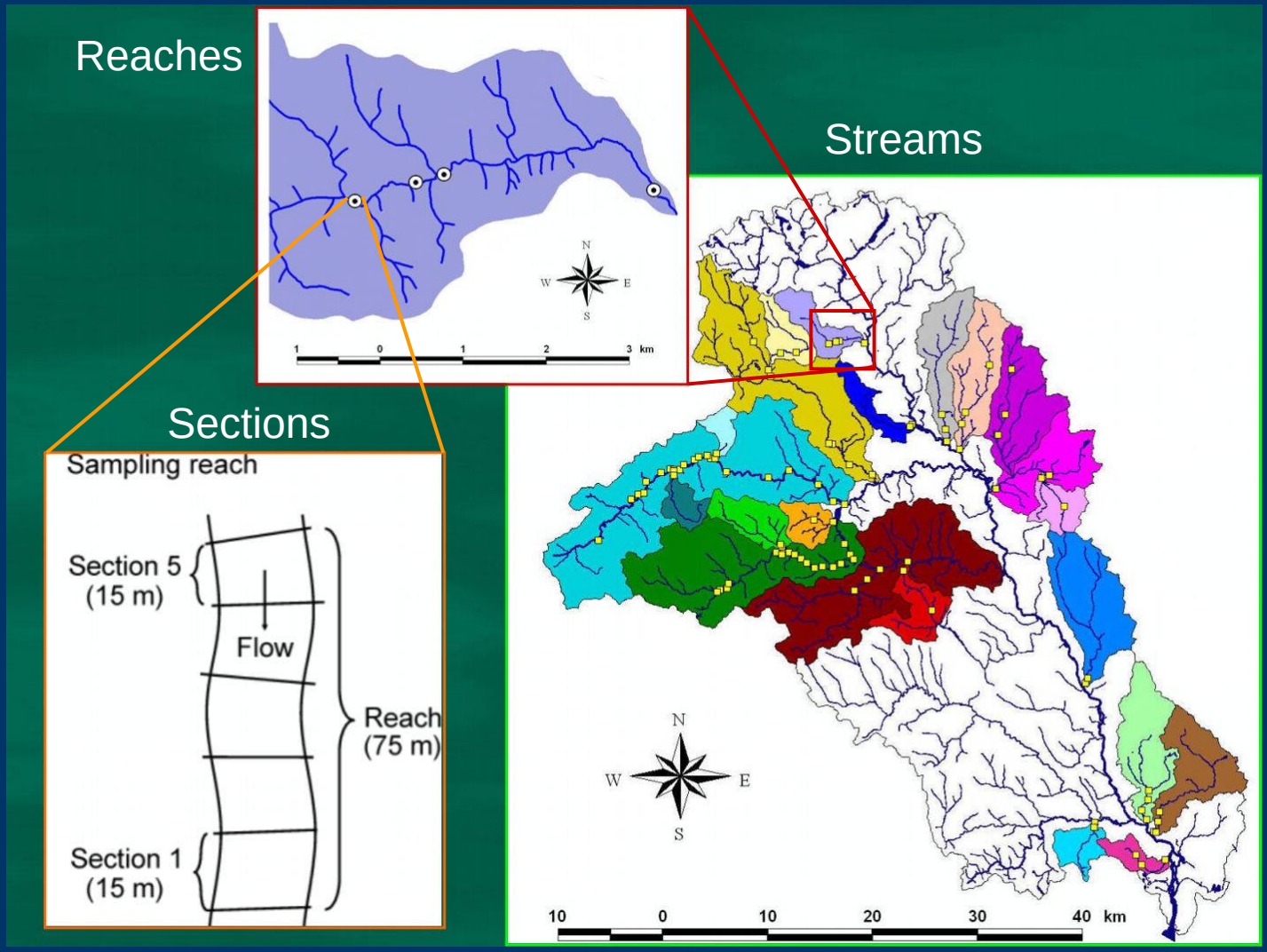
in the ZINB regression model with two levels of random effects:

$$\mathrm{logit}(p) = Z\alpha + u_1 + u_2 \qquad u_i \sim \mathcal{N}(0, \Omega_u)$$

$$\log(\mu) = X\beta + v_1 + v_2 \qquad v_i \sim \mathcal{N}(0, \Omega_v)$$

Case study examining the relationship between abundance of slimy sculpin and stream habitat features

Sampling scheme comprising three hierarchical levels:



600 sections distributed among *120 reaches* and *31 streams* of the Cascapedia River, Québec, Canada

Nine environmental variables considered as potential predictors after preliminary screening:

Spatial scale	Environmental variable
Colonization	Accessibility index
Colonization	Distance to mainstem
Landscape	Height at flood
Landscape	Stream order
Landscape	Sub-basin area
Landscape	Valley width
Local habitat	Cover
Local habitat	Mean depth
Local habitat	Mean wetted width

Bayesian approach:

MCMC – OpenBugs run from R interface

Hierarchical centering for lowest level (“Reaches”)

Two chains; overdispersed initial values

First 60 000 iterations discarded

Further 30 000 iterations monitored with thinning (1 in 10)

Brooks-Gelman-Rubin convergence diagnostics

Comparison of 10 models of differing complexity (DIC)

Evaluation of 10 negative binomial models based on the deviance information criterion (DIC)

Model	Levels	Parameters	Random effects	Deviance	pD	DIC
1	1	$\mu(\boldsymbol{X},\boldsymbol{\beta})$		4420	8.1	4428
2	1	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$		4268	15.8	4284
3	2	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1	4056	10.9	4067
4	2	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	v_1	3615	118.5	3733
5	2	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1, v_1	3572	99.6	**3671**
6	3	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1, u_2	4053	15.8	4069
7	3	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	v_1, v_2	3614	117.0	3731
8	3	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1, v_1, u_2	3569	100.2	3669
9	3	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1, v_1, v_2	3572	99.4	3671
10	3	$\mu(\boldsymbol{X},\boldsymbol{\beta})$, $p(\boldsymbol{Z},\boldsymbol{\alpha})$	u_1, u_2, v_1, v_2	3569	104.0	3673

$\boldsymbol{Z}$: covariates for the logistic component

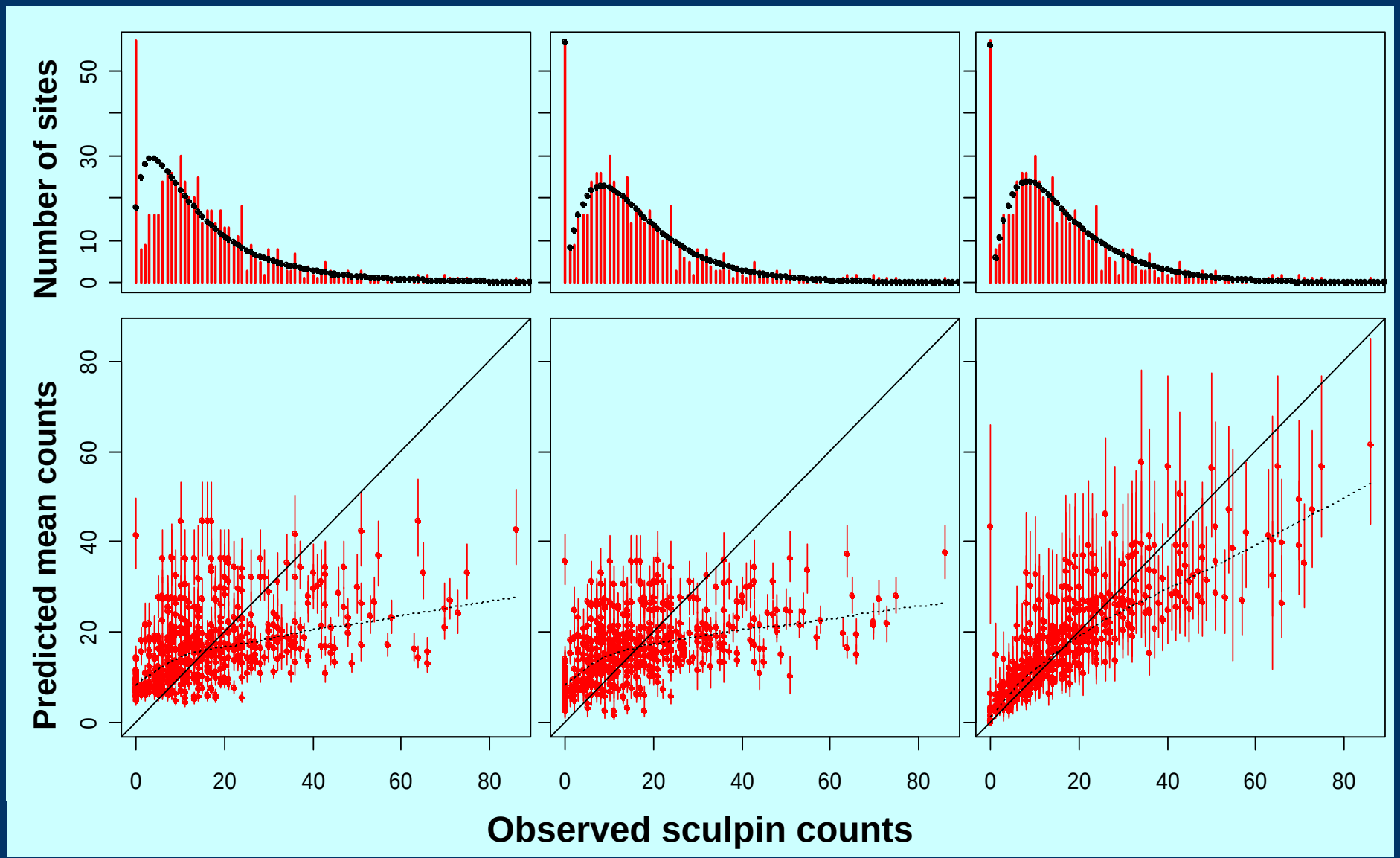
$\boldsymbol{X}$: covariates for the negative binomial component

$\boldsymbol{\alpha}$: regression coefficients for the logistic component

$\boldsymbol{\beta}$: regression coefficients for the negative binomial component

u_i: random effects for the logistic component

v_i: random effects for the negative binomial component


NB
ZINB
ZINB +
random effects
Number of sites
0
10
30
50
Predicted mean counts
0
20
40
60
80
0
20
40
60
80
0
20
40
60
80
Observed sculpin counts

Coefficient estimates for the negative binomial (NB), zero-inflated negative binomial (ZINB), and ZINB with random effects (one level for both the logistic and negative binomial components). Nominally significant effects are in bold characters

	Model								
	NB			ZINB			ZINB + random effects		
	2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Logistic component									
Accessibility				**0.96**	**1.46**	**2.03**	**2.26**	**8.93**	**16.55**
Distance to mainstem				**0.30**	**0.79**	**1.33**	-4.45	2.15	9.01
Mean wetted width				**0.22**	**0.82**	**1.45**	-3.81	1.37	7.38
Stream order				-0.23	0.46	1.19	-5.93	1.43	8.52
Mean depth				**0.01**	**0.44**	**0.90**	-2.61	1.13	4.95
Valley width				-0.30	0.13	0.57	-5.90	0.63	7.50
Cover				**0.01**	**0.45**	**0.93**	-3.58	0.45	4.60
Negative binomial component									
Accessibility	**0.28**	**0.36**	**0.44**	**0.17**	**0.24**	**0.31**	**0.11**	**0.26**	**0.40**
Distance to mainstem	**0.11**	**0.18**	**0.26**	**0.09**	**0.15**	**0.22**	**0.01**	**0.12**	**0.24**
Mean wetted width	**0.22**	**0.34**	**0.47**	**0.08**	**0.19**	**0.29**	**0.12**	**0.27**	**0.41**
Stream order	**0.23**	**0.32**	**0.40**	**0.20**	**0.27**	**0.34**	**0.11**	**0.25**	**0.41**
Sub-basin area	**-0.58**	**-0.44**	**-0.31**	**-0.41**	**-0.30**	**-0.18**	**-0.51**	**-0.31**	**-0.10**
Height at flood	**0.06**	**0.14**	**0.21**	**0.09**	**0.16**	**0.22**	-0.03	0.04	0.13
k	1.41	1.62	1.86	2.23	2.59	2.97	6.59	8.07	9.78
σ_{u1}							10.61	18.91	30.20
σ_{v1}							0.49	0.58	0.69

Colonization

Landscape processes

Local habitat

Conclusions

Incidence was strongly related only to accessibility; *abundance* was influenced both by accessibility and landscape features

Heterogeneity in count data is common in population studies and is probably best viewed as a potentially rich source of information rather than as a nuisance

Zero-inflated regression allows one to detect and interpret ecologically interesting heterogeneity in the data

Accounting for intra-group correlations allows for improved assessment of environmental effects

Acknowledgements