Excess zero models:

Generalized linear models

+

lots and lots of zeros

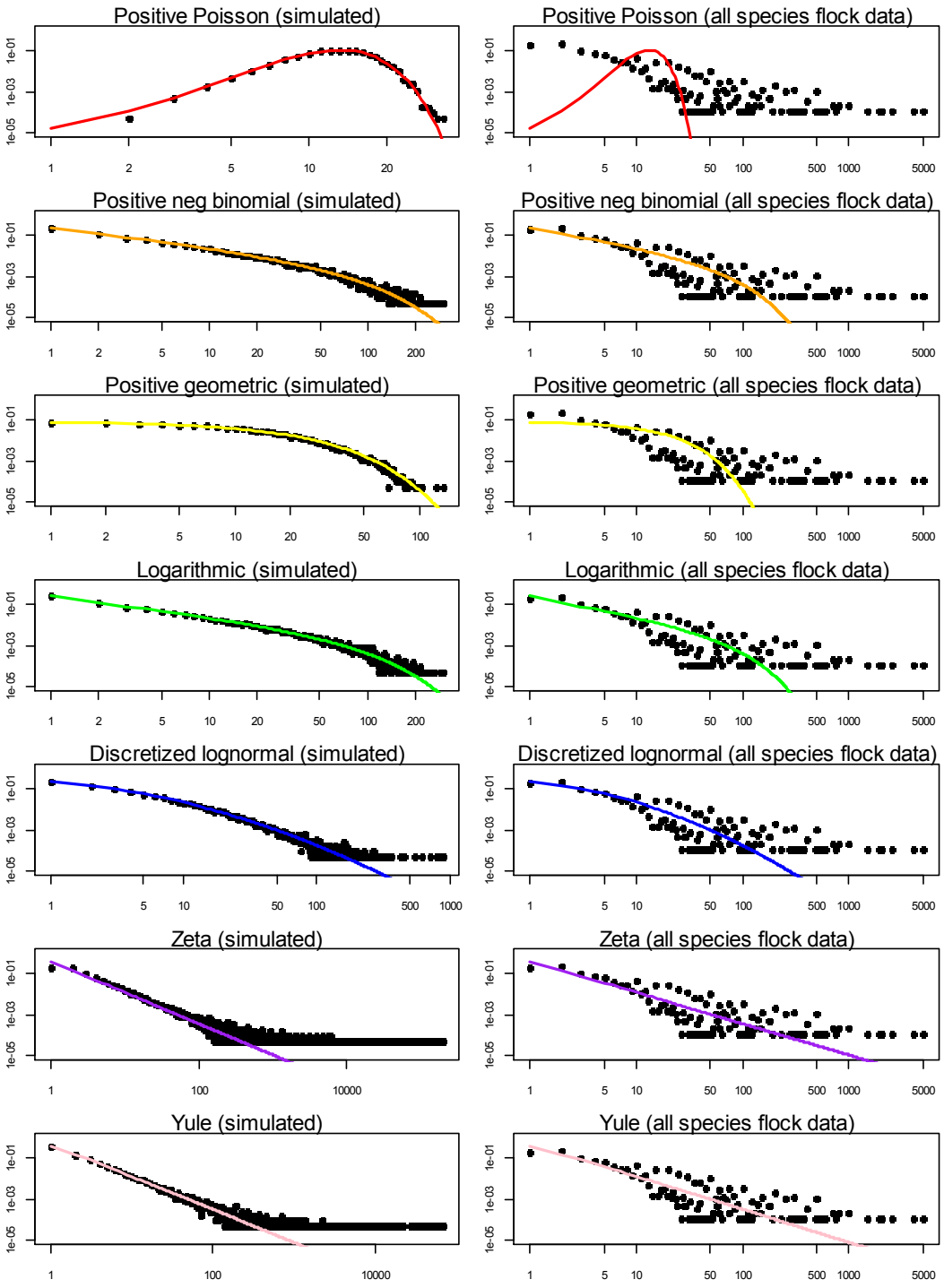
=

zero-inflated or zero-truncated models

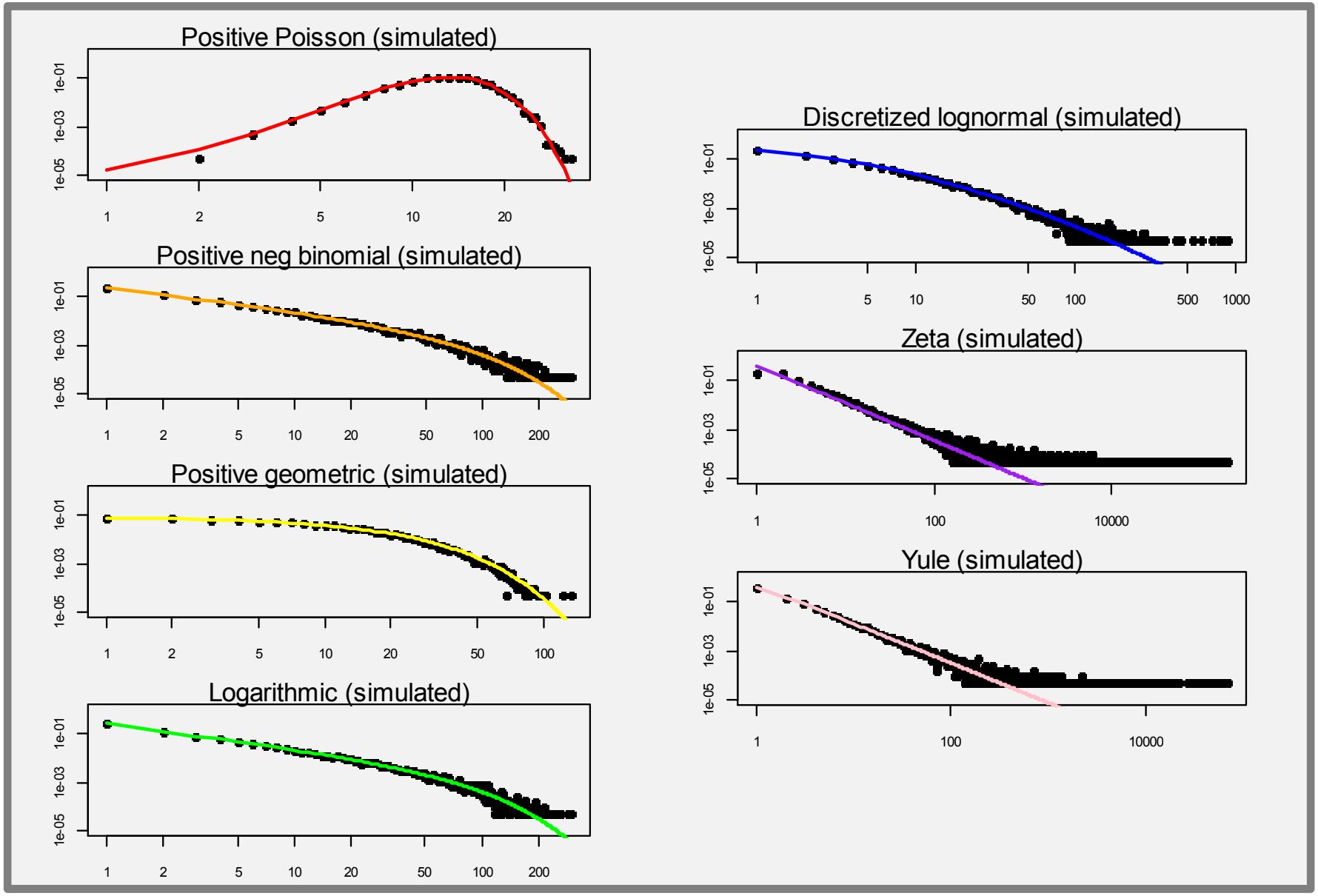
Generalized linear models

Count models:

- Modeling counts is really common
- We've discussed two distributions for doing this
  - Poisson distribution
  - Negative binomial distribution
- But there are many other distributions


Positive Poisson (simulated)
Positive Poisson (all species flock data)
Positive neg binomial (simulated)
Positive neg binomial (all species flock data)
Positive geometric (simulated)
Positive geometric (all species flock data)
Logarithmic (simulated)
Logarithmic (all species flock data)
Discretized lognormal (simulated)
Discretized lognormal (all species flock data)
Zeta (simulated)
Zeta (all species flock data)
Yule (simulated)
Yule (all species flock data)
1e-01
1e-03
1e-05

Examples of distributions



Generalized linear models

Count models

- Often the numbers of zeros in the sample cannot be accommodated properly by a Poisson or negative binomial model.
  - Count distributions, generally, under-predict zeros

- New models are needed to deal with “the excess zeros” observed in many real data sets.

Excess zero models

Count models

- These models (also known as two-part models) allow for two different processes:
  - First level drives whether the value is 0 or positive (binary indicator)
  - Second level drives the value of the count (values greater than or equal to zero)
- Excess zero models are a hierarchical extension of count models!

Excess zero models

Two main types

- Zero inflated models
  - Zeros can be included in both components of the model
- Zero truncated models (Hurdle model)
  - Zeros can only be included in the first component of the model and are not included in models of the count

Zero-inflated model

- Basic idea is that there are two types of zeros:
  - Structural zeros
  - Zeros that are part of the count
- Structural zeros are defined as such because there is no possible way that a value other than zero is possible
- Other zeros just happen as part of the random variation associated with the count distribution (e.g., Poisson, negative binomial) and in other trials could take nonzero values

Zero-inflated model

Consider count data of a particular species at a location *j*, denoted y_j

Define the mean of y_j: $\mu_j = \lambda_j * x_j$

$$x_j \sim \text{Bern}(p_i)$$

$$(y_j \mid x_j = 1) \sim Pois(\lambda_j)$$

The parameter λ_j is the estimated mean count when $x_j = 1$, otherwise the mean is zero

Zero-inflated model

Consider count data of a particular species at a location *j*, denoted y_j

Define the mean of y_j: $\mu_j = \lambda_j * x_j$

$$x_j \sim \text{Bern}(p_i)$$

$$(y_j \mid x_j = 1) \sim Pois(\lambda_j)$$

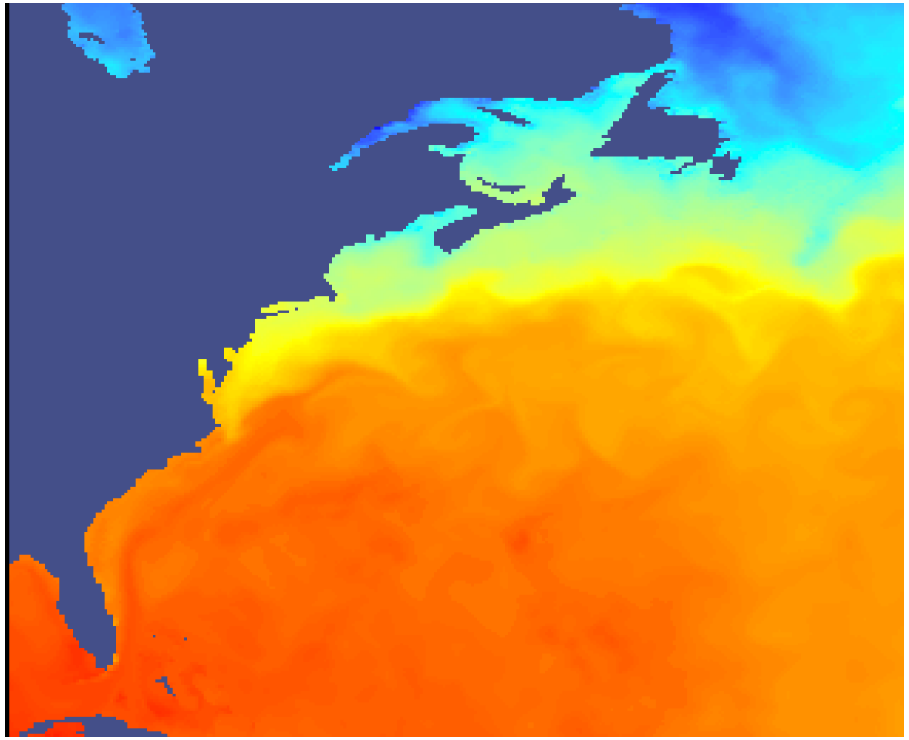
The parameters λ_j and p_j can be estimated relative to covariates:

$$log(\lambda_j) = some\ linear\ model$$
$$logit(p_j) = some\ linear\ model$$

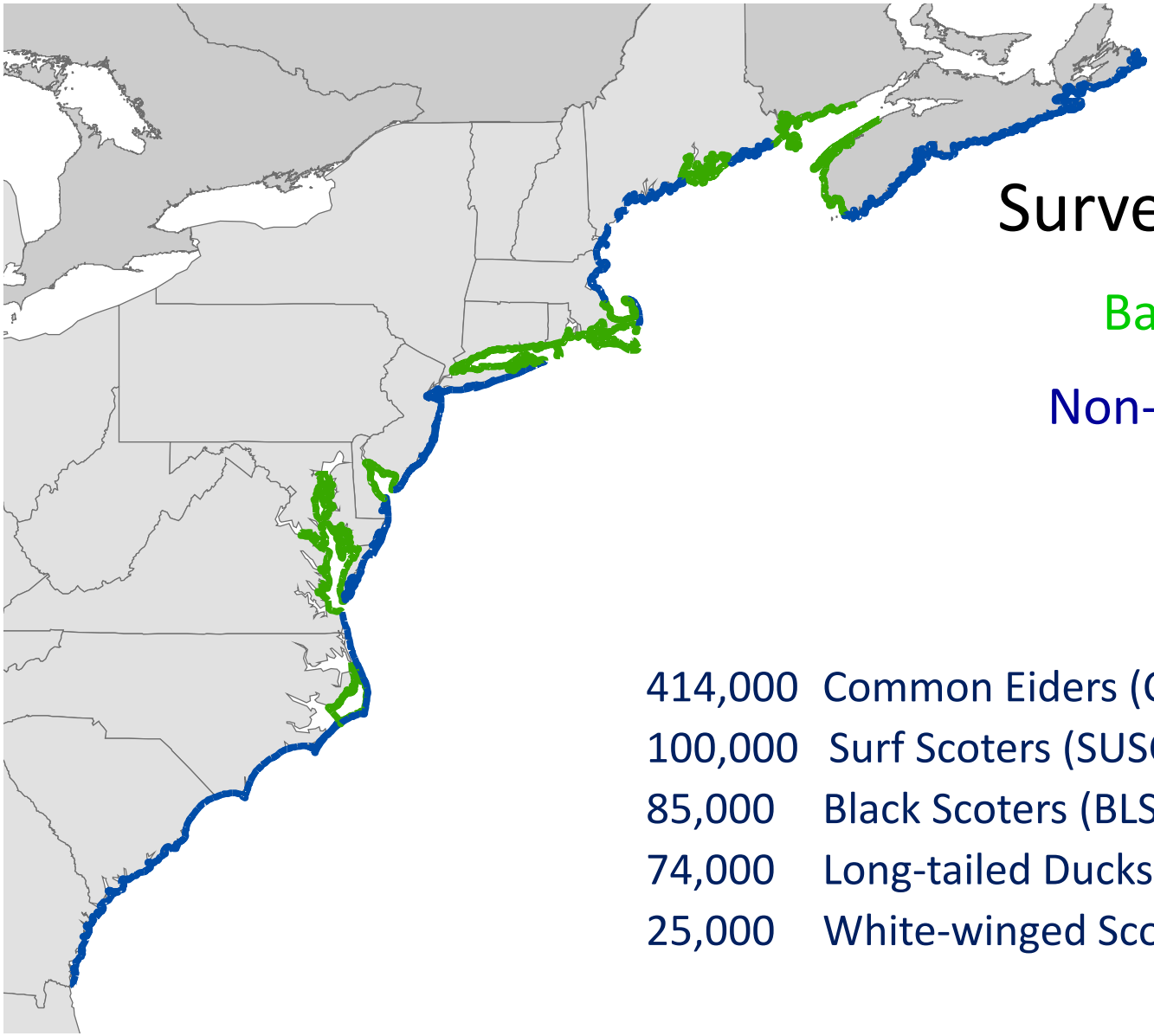
Zero inflated model

Question: what influences the wintering distributions of sea ducks on the Atlantic coast?

- Climate variables
  - The North Atlantic Oscillation (NAO)
  - Localized sea surface temperature
- Local environmental characteristics



Zero-inflated model



Survey transect

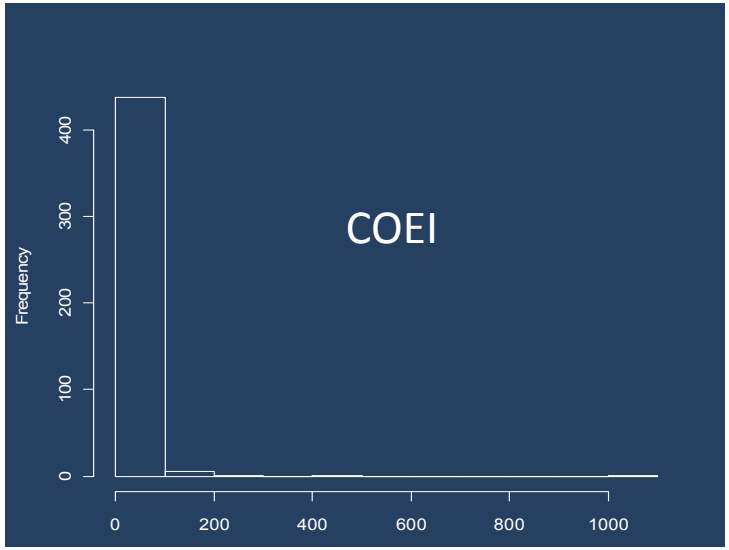
Bay habitat

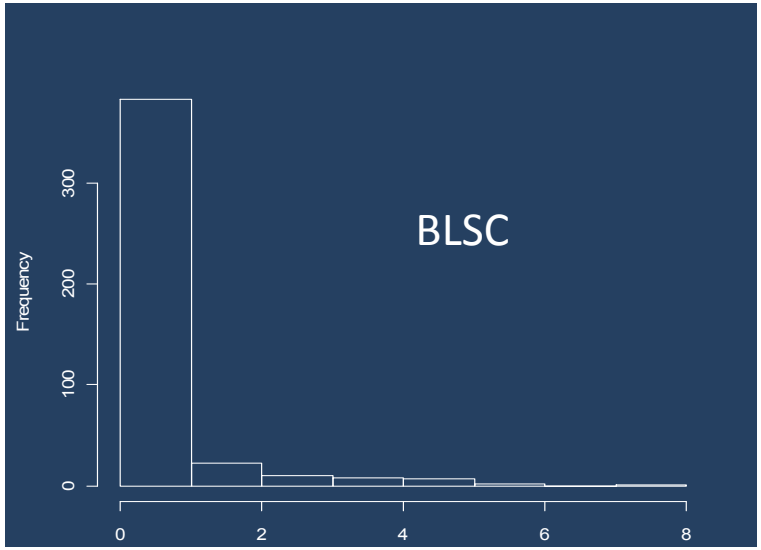
Non-bay habitat

414,000 Common Eiders (COEI)
100,000 Surf Scoters (SUSC)
85,000 Black Scoters (BLSC)
74,000 Long-tailed Ducks (LTDU)
25,000 White-winged Scoters (WWSC)

Zero-inflated model

$y_{i,j,t}$ are spatially referenced counts for each species *i* counted at segment *j* year *t*





- Zero-inflation -> segment inclusion (yes/no)
- Poisson distribution -> count, if segment included

Zero-inflated model

Define the mean of $y_{i,j,t}$:

$$y_{i,j,t} \sim Pois(\mu_{i,j,t})$$

$$\mu_{i,j,t} = \lambda_{i,j,t} * x_{i,j,t}$$

$$x_{i,j,t} \sim \text{Bern}(p_{i,j})$$

$$\text{logit}(p_{i,j}) = Some\ linear\ model$$

$$\log(\lambda_{i,j,t}) = Some\ linear\ model$$

Zero-inflated model

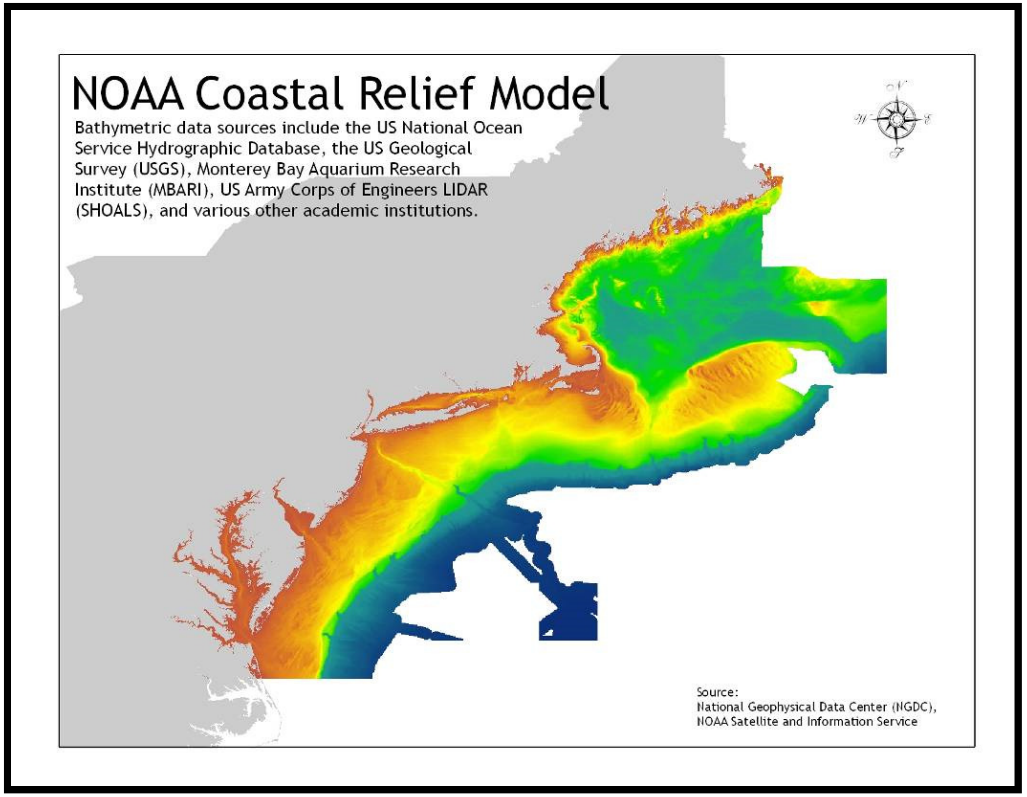
- What affects the inclusion of a segment?
  - Latitude
- To capture the range of the species – some of which did not cover the whole survey range

$$x_{i,j,t} \sim \text{Bern}(p_{i,j})$$

$$\text{logit}(p_{i,j}) = \alpha 0_i + \alpha 1_i * latitude_j$$

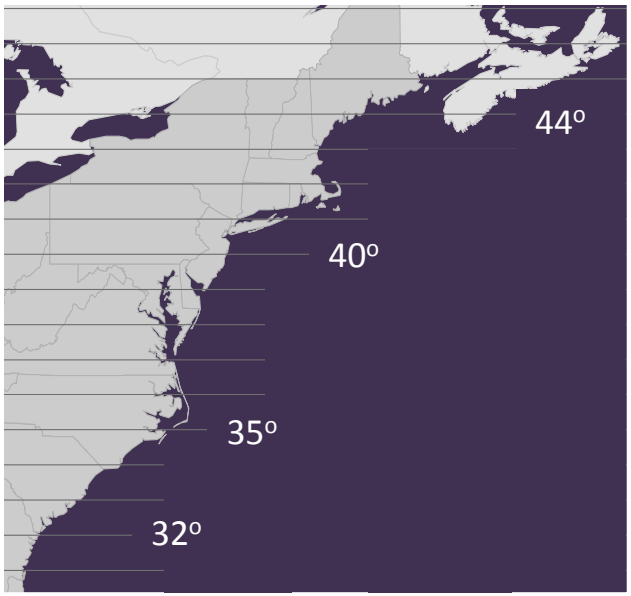
Zero-inflated model

- North Atlantic Oscillation (NAO)
- Sea surface temperature (SST)
- Bays indicator
- NAO·SST interaction

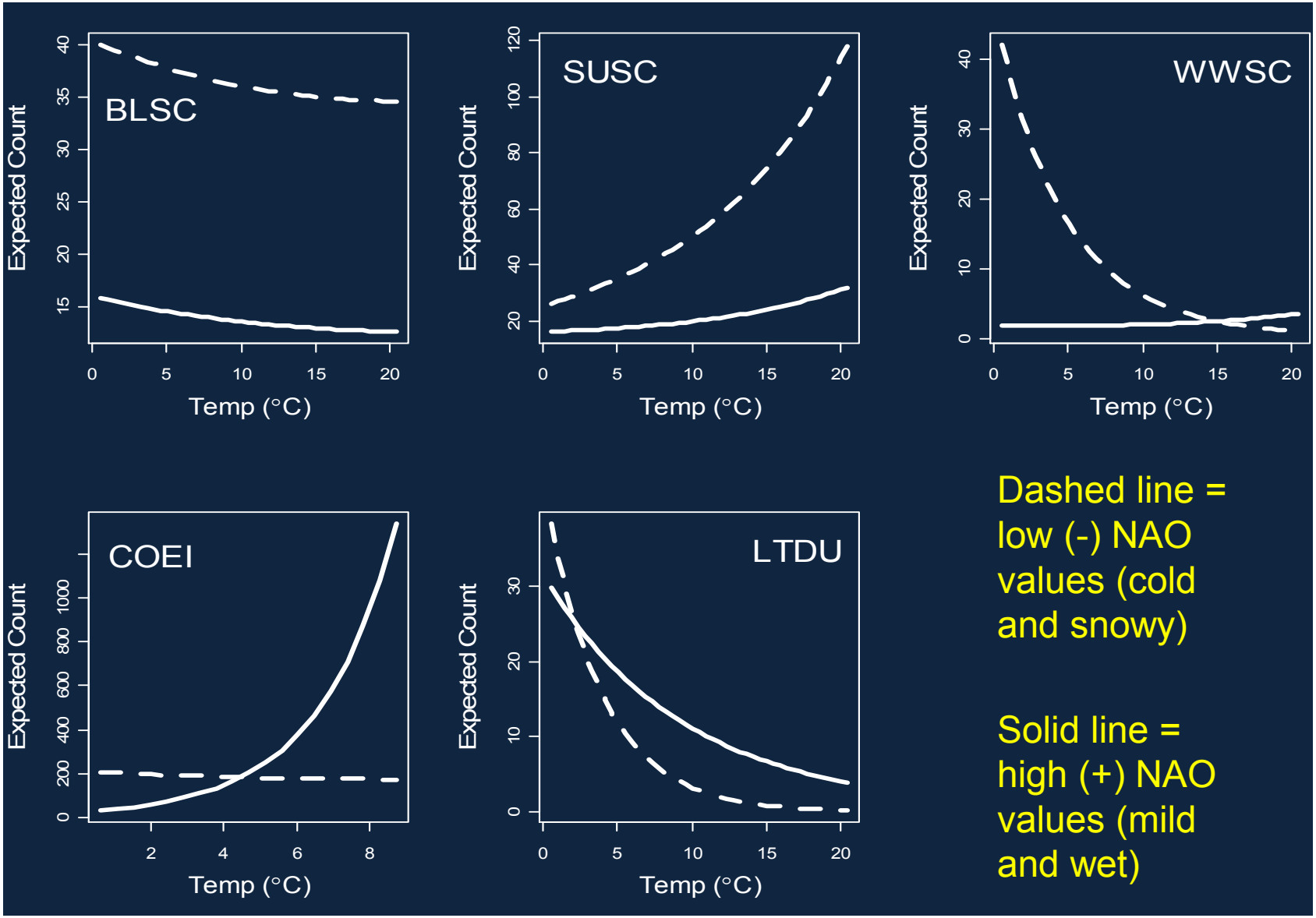


$$\log(\lambda_{i,j,t}) = \beta 0_i + \beta 1_i * NAO_t + \beta 2_i * SST_j + \beta 3_i * SST_j * NAO_t + \beta 4_i * bays_j$$

Zero-inflated model



Zero-inflated model



Zero-inflated model

- Two types of zeros were possible in our sea duck model:
  - Structural zeros: segment was unsuitable due to latitude (too far south)
  - Zeros that are part of the count: segment was suitable but count was zero anyways

- Zero-inflated models are useful in the scenario where zeros can still be a part of the count

Zero-truncated model

- Basic idea is that there is only one kind of zero:
  - Structural zeros
  - Counts can only be positive values: x = (1,2,3…)
- We again using a Bernoulli distribution to describe the zero portion of the models
- But we must use a truncated version of the other distribution to model only the positive counts
- Referred to as "***Hurdle models***"

Zero truncated model

Question: what is the prevalence and flock size of northern gannets in the nearshore Atlantic Ocean?

- Factors affecting prevalence
  - Season
- Factors affecting flock size
  - Food availability
  - Habitat quality

Zero-truncated model

Consider the data first in terms of presence/absence, $y_i = 0$ if northern gannets where absent at location *i*:

$$(y_i = 0) \sim Bern(1 - p_i)$$

$$\mathrm{P}\,(y_i = 0) = 1 - p_i$$

$$\mathrm{logit}(p_{i,j}) = \alpha 0 + \alpha 1 * season_i$$

Define the probability that $y_i = k$, where $k > 0$ (counts at every location *i*):

$$(y_i = k) \sim p_i PosPois(\lambda_i)$$

$$\mathrm{P}(y_i = k) = p_i \frac{{\lambda_i}^k e^{-\lambda_i}}{k!(1-e^{-\lambda_i})}$$

$$\log(\lambda_{i,j,t}) = \beta 0 + \beta 1 * food_i + \beta 2 * habitat_i$$

Excess zero models

Does it ever makes sense to conceive of two distinct mechanisms underlying a realized abundance and distribution in space?

- Excess zeros may simply be due to a failure to include all relevant covariates in the count model
- Some argue that it may not be interesting to try to attribute much biology to what might merely a deficiency of the abundance model

Excess zero models

Of course, there are exceptions!

- Imagine the abundance of some terrestrial species in an archipelago
- Clearly, any abundance greater than zero requires the colonization of an island beforehand
- Colonization is a stochastic process with binary outcome:
  - Colonized or it is not colonized
  - This may have nothing to do with the factors that determine abundance on that island once it is colonized

Excess zero models

- In most cases, excess zero models are simply used for convenience as a modeling trick to make up for our lack of perfect knowledge of covariates governing abundance
- You may want to adopt zero-inflated models to account for poor model fit but be cautious in the biological interpretation of the zero-inflation part
- Probably not a good idea to develop complicated covariate models in the zero-inflation part
- ***Never want to use the same covariates in both the zero-inflation part and in the abundance part of the model*** (resulting model is probably near-unidentifiable; see also Ghosh et al., 2012).

Excess zero models

Homework: Zero-inflated distributions to model counts of fish that were caught in a state park