Estimating the seal pup abundance in the Greenland Sea with Bayesian hierarchical modeling

Martin Jullum (NR)

Joint with Thordis Thorarinsdottir (NR)
and Fabian Bachl (Uni. of Edinburgh)

Oslo, 15.06.17

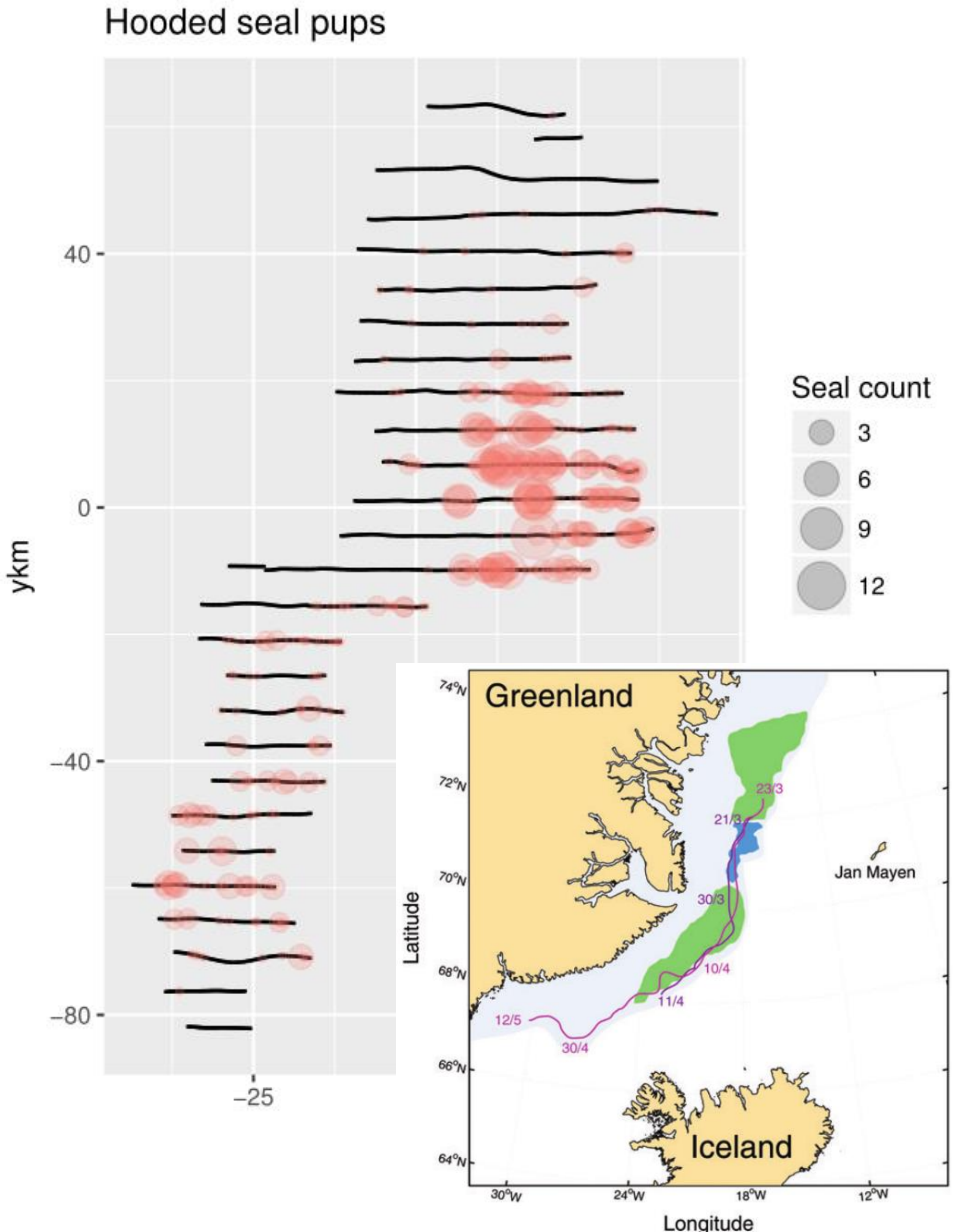
Problem

- Ultimale goal: Monitor seal abundance in the North Atlantic
- Well established dynamic abundance model for seals
  - Key component is **estimate + uncertainty of the number of seal pups**
  - Existsing methods
    - Very basic ad-hoc scaling method
    - Spatial GAM (splines) model

- **Our task: Propose method to estimate the total number of pups with uncertainty + validate it!**

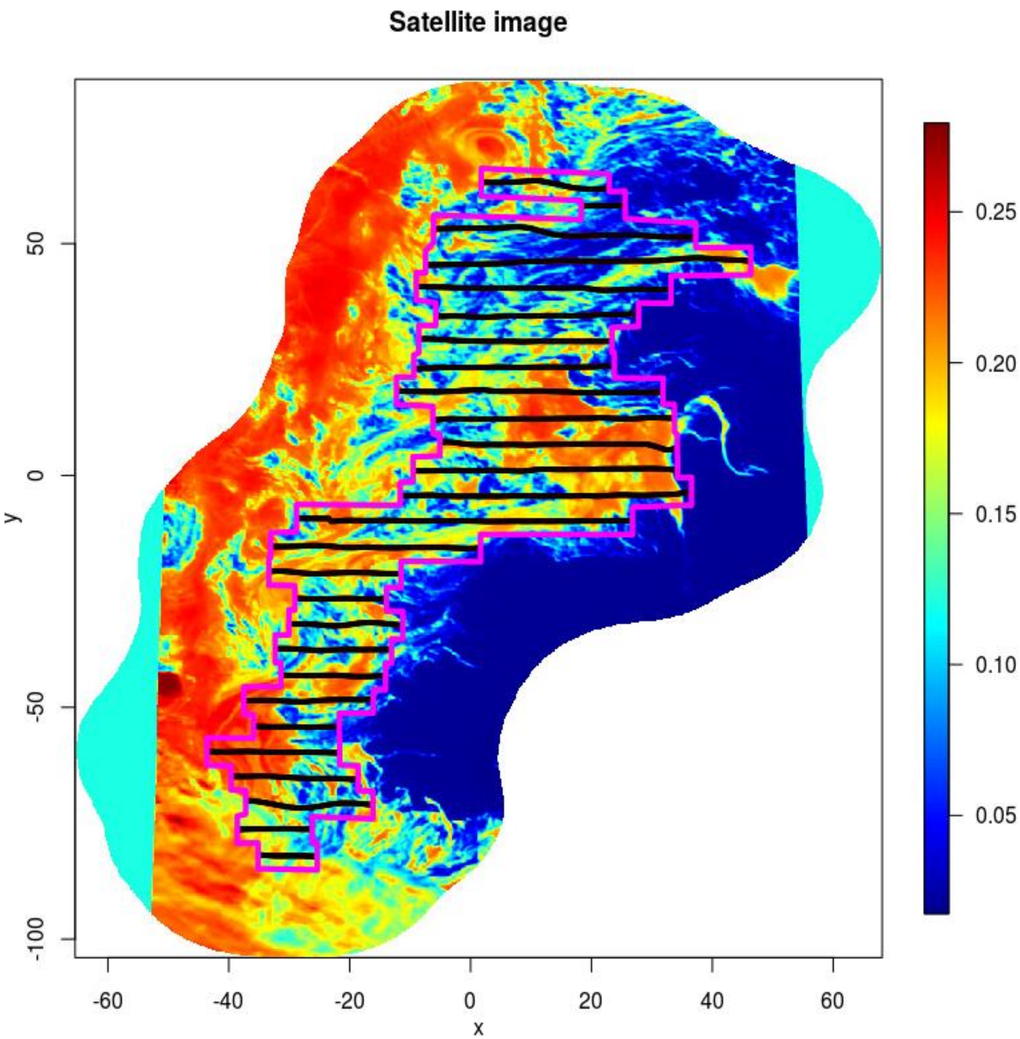
Data

- From an aerial photo survey conducted east of Greenland in 2012
- Number of pups in 2792 photos (A) in 27 transects sparsely covering the seal domain



Data

- From an aerial photo survey conducted east of Greenland in 2012
- Number of pups in 2792 photos (A) in 27 transects sparsely covering the seal domain
- Additional info: Quantified satellite image to indicate ice thickness
- Seal domain Ω shown in pink



A model for the seal pup appearance

- Model the spatial distribution of the seal pups with a Log-Gaussian Cox Process (LGCP)
  - Gaussian latent field Z
  - Point pattern $Y|Z \sim \mathrm{PoissonProcess}(\lambda(s) \; = \exp(Z(s)))$
  - LGCP property: Given Z, counts $N(B)$ in disjoint Borel sets B indep. and distributed as $\mathrm{Poisson}(\lambda \; = \int_B \; \exp\big(Z(s)\big)\, ds)$
  - LGCP Log-likelihood

$$|A| - \int_A \exp(Z(s))\mathrm{d}s + \sum_{i=1}^{n} Z(s_i),$$

Discretizing the LGCP-model

- Data are aggregated counts per photo
- Solution: Discretize the LGCP-model to the set where our data lives
- Let $N_1, \ldots, N_n$ be the counts in the $n = 2792$ photos, covering the space $A_1, \ldots, A_n$
- Discretize the LGCP-model to

  $$p(N_1, \ldots, N_n | Z) = \prod_{i=1}^{n} \mathrm{Poisson}\left(k = N_i, \lambda = \int_{A_i} Z(s)\, \mathrm{d}\, s\right),$$

  with $\mathrm{Poisson}(\mathrm{k}, \lambda) = \lambda^k \exp(-\lambda)/k!$

INLA and SPDE-INLA

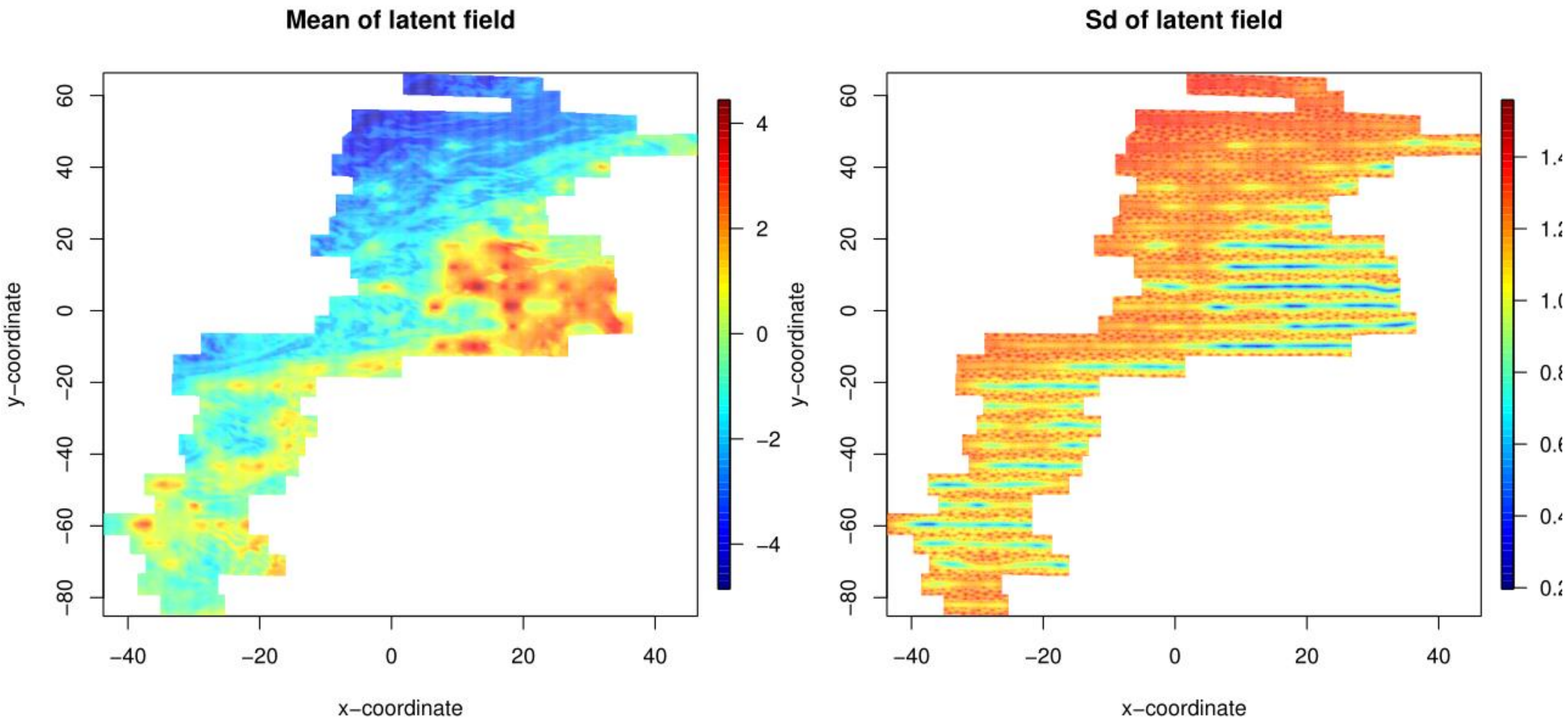
- Integrated Nested Laplace Approximation (INLA)
  - Computational feasible approximate Bayesian inference for Gaussian latent models with discrete latent field (GMRF)
  - Based on extensive Laplace approx. and numerical optimization.

- Spatial Partial Differential Equation (SPDE) approach
  - Makes INLA applicable to Gaussian latent models with continuous fields
  - Triangulates continuous latent field which translates to certain GMRF by formulation through solution to a SPDE

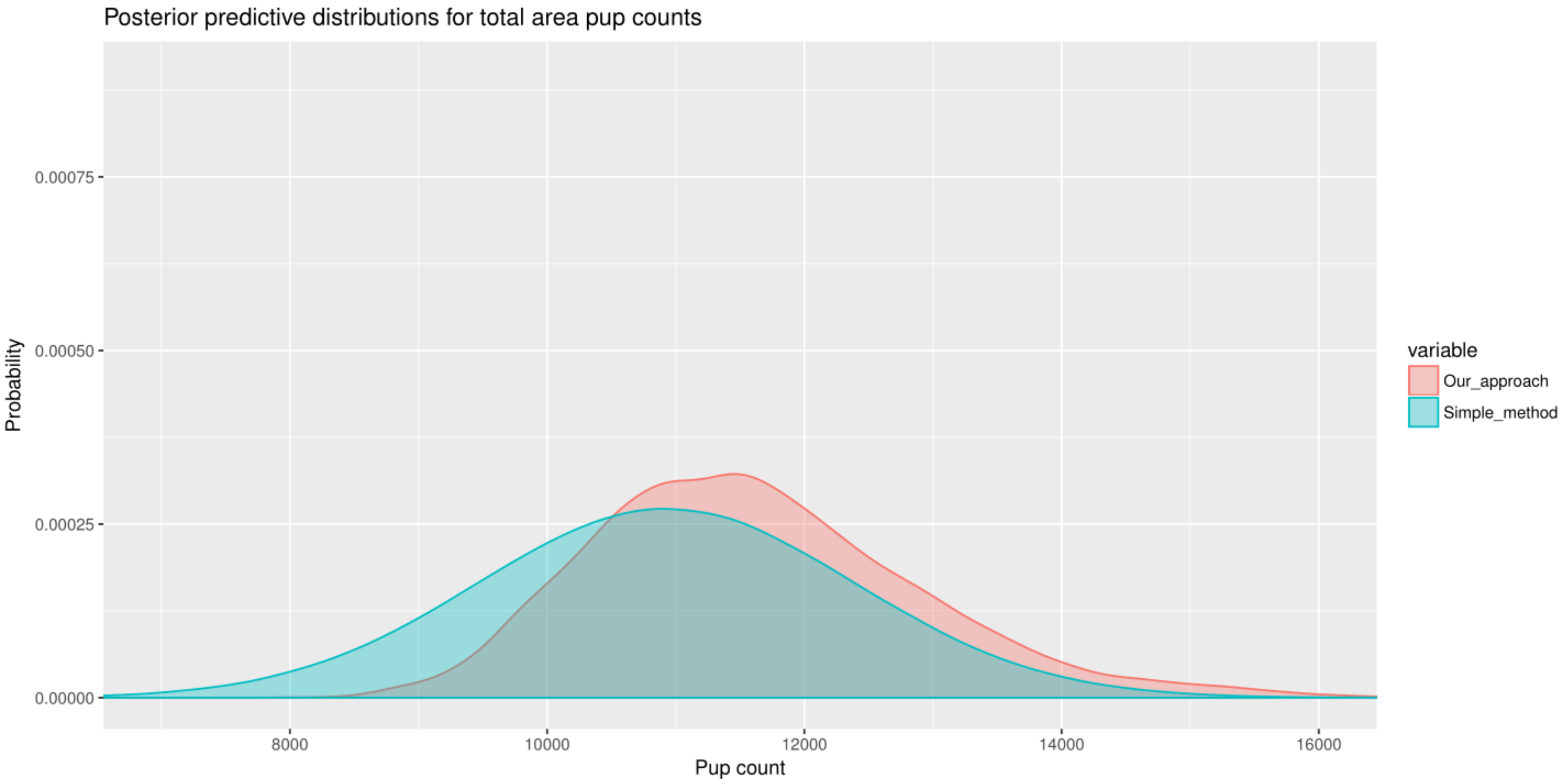
Modeling approach

- Latent field $\mathrm{Z}(s) = \alpha + \beta^t \boldsymbol{x}_s + g(s)$, $\boldsymbol{x}_s$ satellite information, $g(s)$ zero-mean Gaussian field with a Matern covariance structure
- Bayesian approach with vague priors on all parameters
- The Bayesian solution to our problem is the «**posterior predictive distribution**» of seal pup counts in the seal domain $p(N(\Omega)|Y)$
  - Easy to compute with samples from $p(Z|Y)$
- Use SPDE-INLA to fit the model and perform the posterior sampling

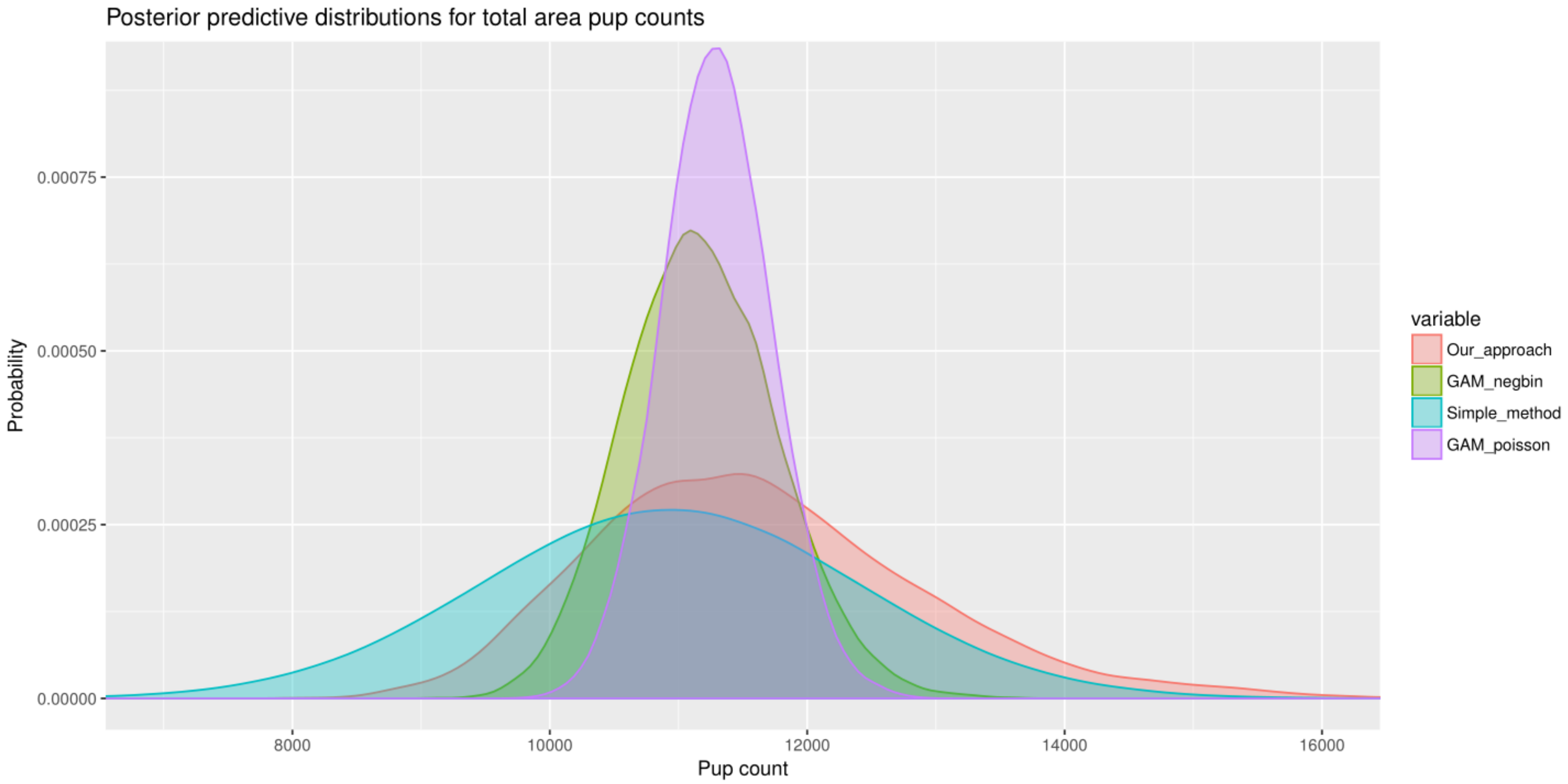
Results our approach



Method comparison

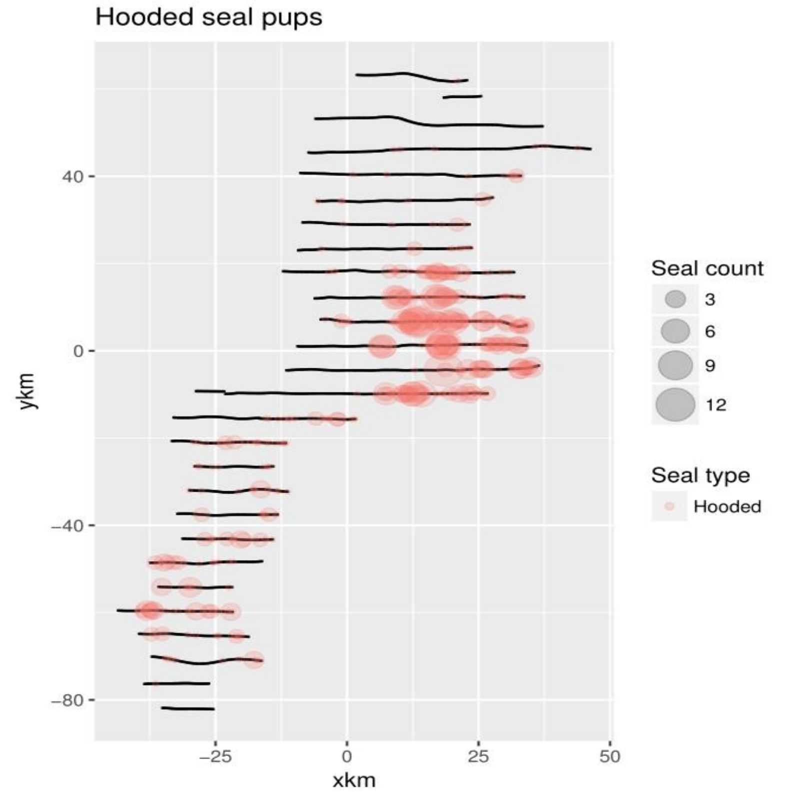


Method comparison



Result validation

- 2 different CV schemes
  - Leave out random photos
  - Leave out all photos on transect
  - Evaluate posterior predictive distribution both per photo and per transect
- Evaluation criterion

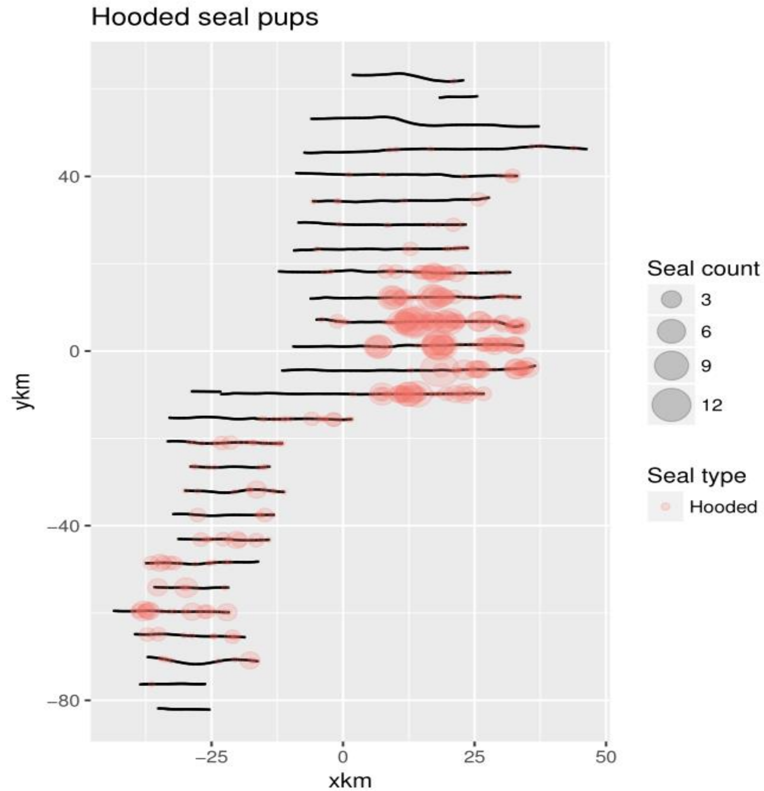


$$CRPS(F,y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}(x-y)\right)^2 \mathrm{d}x$$

$$\mathrm{logscore}(f,y) = \log\left(f(y)\right)$$

Result validation

- 2 different CV schemes
  - Leave out random photos
  - Leave out all photos on transect
  - Evaluate posterior predictive distribution both per photo and per transect



- Evaluation criterion

$$CRPS(F,y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}(x-y)\right)^2 \mathrm{d}x$$

$$\mathrm{logscore}(f,y) = \log\left(f(y)\right)$$

- Long story short
  - Our method is significantly better on photo level
  - We do as well as the GAM approaches (better than the simple method) on transect level

Result validation

PHOTO LEVEL

	Random 10-fold CV CRPS	Leave-out full transect CRPS
Our approach	***0.18 (0.16, 0.19)***	*0.22 (0.20, 0.25)*
GAM_negbin	0.21 (0.19, 0.23)	*0.22 (0.20, 0.24)*
GAM_poisson	0.22 (0.20, 0.24)	0.24 (0.22, 0.26)
Simple_method	0.26 (0.24, 0.28)	0.26 (0.24, 0.29)

AGGREGATE/TRANSECT LEVEL

	Random 10-fold CV CRPS	Leave-out full transect CRPS
Our approach	5.43 (4.04, 6.99)	9.91 (5.99, 14.80)
GAM_negbin	5.93 (4.95, 7.00)	*9.37 (5.66, 13.63)*
GAM_poisson	5.90 (4.49, 7.42)	10.14 (5.86, 15.09)
Simple_method	*4.83 (3.27, 6.66)*	15.57 (11.77, 19.68)

Alternative model

- Arnt-Børre + Tor Arne + others (2009)
  - Let $Z_0(s) = f_{GAM}(s)$, with $f_{GAM}(s)$ a (spatial) smooth spline.
  - $\mu_i = |A_i|\exp(Z_0(s_i^*))$ for s_i^* the mid-point in cell A_i
  - Fit the counts per photo as a negative binomial regression with constant shape κ and $|A_i|$ as offset
  - Frequentist approach
  - Smoothness of $f_{GAM}(s)$ chosen through generalized CV

  - We test this formulation, also with satellite data and Poisson distributions
  - Use sampling to produce predictive distribution of total pup counts for comparison