Ecological Modelling 265 (2013) 74–84

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Review

Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data

J.B. Lecomte[a,b,*], H.P. Benoît[c], M.P. Etienne[a,b], L. Bel[a,b], E. Parent[a,b]

[a] INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France
[b] AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France
[c] Gulf Fisheries Centre, Fisheries and Oceans Canada, Moncton, NB E1C 9B6, Canada

ARTICLE INFO

Article history:
Received 10 April 2013
Received in revised form 7 June 2013
Accepted 8 June 2013
Available online 5 July 2013

Keywords:
Zero-inflated data
Bayesian hierarchical modeling
Habitat associations
Spatial dependencies
Macro-invertebrates

ABSTRACT

Biomass samples from marine scientific surveys are commonly used to investigate spatial and temporal variations in stock abundances. Biomass records are often characterized by a high proportion of zeros on the one hand, and occasional large catches on the other. These features induce a modeling challenge when trying to understand the state of populations and their ecological associations with one another and with habitat. We develop a hierarchical Bayesian model to represent the spatial structure of biomass and analyze the spatial distribution and habitat associations of three species of macro-invertebrates sampled in the southern Gulf of St. Lawrence (Canada). A zero-inflated distribution based on a compound Poisson with Gamma marks is used for the observation layer, and a linear model with spatial correlated errors accounts for the role of habitat variables (temperature, depth and sediment type) in the process layer. Maps of quantities of interest (e.g. probability of presence, quantity of biomass) are produced, taking into account the uncertainty of the estimated parameters and observation errors. This hierarchical Bayesian modeling approach provides a useful tool for spatial management of human activities that may affect living resources that may affect living resources, such as marine protected areas.

Contents

* Corresponding author at: INRA, UMR 518, 16 rue Claude Bernard, 75005 Paris, France. Tel.: +33 144087292.
E-mail address: jean-baptiste.lecomte@agroparistech.fr (J.B. Lecomte).

0304-3800/$ – see front matter.
http://dx.doi.org/10.1016/j.ecolmodel.2013.06.017

1. Introduction

Understanding species spatial distribution and habitat associations are key challenges when managing harvested, endangered or invasive species (Welsh et al., 1996; Engler et al., 2004; Cook et al., 2007). Marine spatial management measures, in which the spatial and temporal distribution of human activities is restricted to achieve ecological, social and economic objectives (e.g., marine protected areas), have been the focus of many studies in the recent decades (Shea, 1998; Hilborn et al., 2004; Hobday and Hartmann, 2006; Hartog et al., 2011). In many applications, these management approaches require knowledge of habitat use by the targeted species to be effective (Perry and Smith, 1994; Williams and Bax, 2001).

Linear or additive models are often developed to infer distributions and habitat use and preferences using survey or other ecological data (as reviewed by Guisan and Thuiller, 2005). Efficient models must be able to address two common characteristics of ecological data: observations can be dominated by a large number of null values combined with skewed positive values, and abundance can be strongly spatially correlated. Failure to address both of these characteristics is well known to impact model parameter estimates and their uncertainty, leading to incorrect statistical inference and therefore, in turn, potentially inappropriate management actions (Zuur et al., 2009; Sileshi et al., 2009). Ideally, the models should also be able to address possible spatial misalignment between the available data for abundances and for habitat characteristics.

High proportions of zeros in survey data stem from three general causes. An observed zero value can be a true zero if the species is not present in the studied area, while a false zero, also called pseudo-absence, results from a low probability of detection even though the species is present. A third class of zeros results from an observer effect, whereby a species normally found in the study area is frightened away by some inappropriate data collection procedure. Numerous approaches exist for such zero-inflated data when dealing with counts, as reviewed in Martin et al. (2005). The two main approaches, Zero-inflated Poisson (ZIP) and Zero-inflated binomial (ZIB), are mixture models and the presence–absence is modeled separately from the number of counts (i.e. individuals). The development of zero-inflated models for continuous abundance data (i.e. densities or biomasses) has also received attention (Stefansson, 1996; Maunder and Punt, 2004; Fletcher et al., 2005; Shono, 2008; Ancelet et al., 2010). The simplest approach consists in adding a positive constant to all the observations, typically followed by a logarithmic transformation, as is often performed in generalized linear modeling (GLM). This approach requires choosing an arbitrary constant that could severely bias model estimates (Maunder and Punt, 2004; Shono, 2008). An alternative is to remove the zero catches from data prior to the analysis. However removing zero values often affects the results and can also bias the analysis (Martin et al., 2005), though this is not necessarily the case (Maunder and Punt, 2004). A common and slightly more complex approach for continuous data, named the delta approach (Stefansson, 1996; Shono, 2008), models separately the presence–absence using a binomial distribution and positive values using a standard probability distribution function such as the log-normal (leading to a delta-lognormal model) or the gamma (delta-gamma). The approach reduces bias since the expected biomass is the product of the probability of presence and the average positive biomass. This family of models treats all absences as true zeros. Furthermore, sampling effort, which can vary between sites for a number of logistical and operational reasons, is mostly addressed by a prior standardization of the data (Stefansson, 1996). However, performing such a standardization may obscure the relationship that exists between expected values (for a given sampling effort) and their associated variance for count probability density functions.

In this paper, we develop a hierarchical Bayesian spatial model for biomass data that overcomes these shortcomings. We apply this approach to describe the distribution and habitat associations of epibenthic invertebrates in the southern Gulf of St. Lawrence (sGSL), Canada. The biomass records come from an annual bottom trawl survey in which invertebrates and fish are collected at randomly chosen locations by sweeping the ocean floor over targeted distances which can vary between sites. We use a model based on two substructures that are linked probabilistically using a hierarchical approach. The first substructure, the observation layer, consists of a compound Poisson model with Gamma marks, which heuristically models the process of observing a Poissonian number of patches of a species, each containing a random biomass given by the Gamma mark. This approach constitutes a generalization of the one proposed by Bernier and Fandeux (1970) and applied in ecology by Ancelet et al. (2010) which used exponential marks. It also allows for explicit accounting for the duration or volume of sampling for individual sampling events. The second model substructure explicitly models habitat associations using a linear model that accounts for spatial autocorrelation using a geostatistical approach. Jointly, these model substructures result in a modeling approach that is very flexible, likely making it a useful tool for spatial analysis and planning.

2. Methods

2.1. Data description

Fisheries and Oceans Canada has conducted an annual bottom-trawl survey in the sGSL each September since 1971 (Chadwick et al., 2007; Benoît et al., 2009). Since its inception, the main objective of this survey has been to quantify the abundance and the distribution of marine fishes and certain commercially important invertebrates. Since 1988, data for epibenthic invertebrates such as urchins, starfish, whelks and anemones have been collected. The domain for the sGSL survey is split into 27 strata defined so as to be homogeneous in terms of depth and geographic location. Every year, since the mid 1980s, 140–200 sites have been chosen according to a stratified random design. The number of sites per stratum is generally proportional to stratum size, making the selection of sites at the survey level approximately randomly balanced. Sites are sampled using a straight-line tow for a target duration of 30 min. at 3.5 knots. All captured organisms are identified to the lowest taxonomic level possible and weighed in kilograms per tow. Habitat information, such as bottom temperature (°C) and depth (m), is also collected at each bottom-trawl site. Moreover the type of sediments is interpolated at each sampling site from an existing map of surficial geology for the Gulf of St. Lawrence (Loring and Nota, 1973). This study focuses on three epibenthic macroinvertebrates sampled during the 1997 survey to illustrate the modeling approach: green sea urchin (*Strongylocentrotus droebachiensis*), starfish (*Asterias sp*), and sea cucumber (*Cucumaria frondosa*). These three taxa were chosen for their differences in density distribution and habitat preferences so as to demonstrate the model's ability to confront different data situations (Figs. 1 and 2). In fact, the majority of epibenthic macroinvertebrates in the sGSL are distributed in patches of localized variable abundance, interspersed by numerous and relatively large areas where the species is absent. Consequently, the dataset contains a very large proportion of sites where the species are not observed.

2.2. The statistical model for zero-inflated continuous positive data

The model description is split into two parts, as is classically done in hierarchical Bayesian modeling. The first section describes

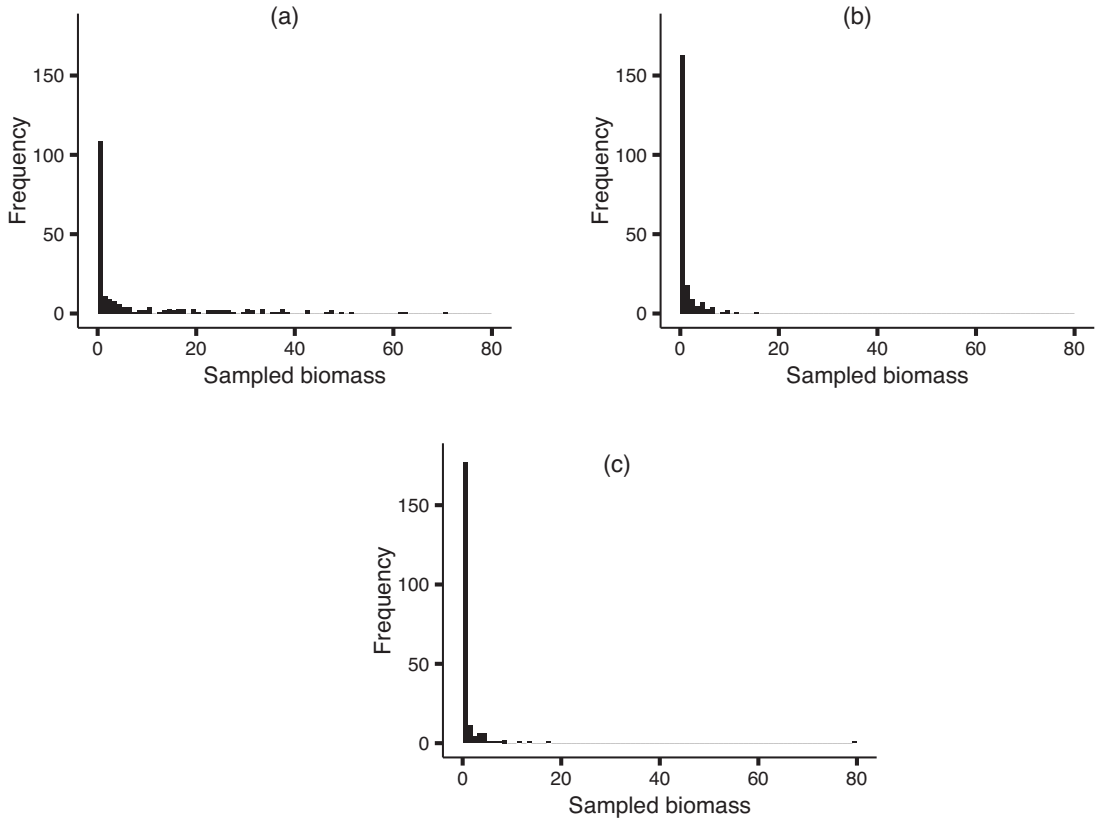

Fig. 1. Histogram of the sampled biomass from individual tows in the 1997 sGSL bottom-trawl survey year (in kg/tow) for the three selected species: (a) urchin, (b) starfish, and (c) cucumber.

how the observation is linked to the actual biomass in a given site. The second section models the spatial distribution of the latent biomass field at the scale of the survey.

2.2.1. Observation layer

The observation layer consists of a compound Poisson process. Let us define N_s, the unknown number of patches of organisms sampled at a site s, which results from a possibly non-homogeneous Poisson process:

$$N_s \sim \text{Poisson}(E_s \mu_s) \qquad \forall \ \ s \in \{1, \ldots, S\} \tag{1}$$

where μ_s is the locally expected number of patches and E_s is the sampling effort at site s. Every sampled patch i is defined as containing an unknown random quantity of biomass $M_{s,i}$. We assume that all the marks $M_{s,i}$ are independent and identically Gamma distributed with scale and rate parameters a and b:

$$M_{s,i} \sim \text{Gamma}(a, b). \tag{2}$$

The average biomass of a patch is a/b. Note that when $a = 1$, $M_{s,i}$ is exponentially distributed, which corresponds to the observation model used by Ancelet et al. (2010).

The quantities N_s and $M_{s,i}$ can be interpreted heuristically in terms familiar to ecologists and allow modeling of observed biomass at a location s, denoted Y_s:

$$Y_s = \begin{cases} \sum_{i=1}^{N_s} M_{s,i} & \text{if } N_s > 0 \\ 0 & \text{if } N_s = 0 \end{cases} \tag{3}$$

If there is at least one patch at the sampling site, the observed biomass Y_s is the random sum of the existing biomass in each patch. Conversely, if there are no patches at the sampled site, then nothing is caught. The main quantities of interest of the model are summarized in Table 1.

2.2.2. Spatial distribution layer

The spatial coherence of biomass distribution relies on the spatial distribution of the covariates and some unexplained latent spatial structure. In the following, we first describe how covariates are introduced to the model, and then how additional spatial structure is included.

2.2.2.1. Covariate effects. Bottom temperature, depth and sediment type are selected as available covariates related to habitat that potentially explain the distribution of the invertebrates. Their effects on μ_s, the average number of patches at site s, are included in the model through a logarithm link function. The full specification of μ_s is then given by:

$$\log(\mu_s) = \alpha_0 + \beta_{Sed_s} + \gamma_{Depth_s} + \zeta_{Temp_s} + \epsilon_s \tag{4}$$

where β_{Sed_s}, γ_{Depth_s} and ζ_{Temp_s} are the site specific effects of sediment type, depth and temperature, respectively, and ϵ_s is a Gaussian noise that accounts for potentially spatially structured process error. Four classes of sediments, based on the granulometry, are distinguished in the model: pelite, fine sand, coarse sand, gravel with occasional sand patches. Depth (in meters) is split into

Table 1
Quantities of interest for the proposed model.

Probability of presence	$1 - \exp^{(-\mu_s)}$
Expected positive biomass	$\left(\frac{\mu_s a}{b}\right)\left(\frac{1}{1-\exp^{(-\mu_s)}}\right)$
Expected biomass	$\frac{(\mu_s a)}{b}$
Variance of the biomass	$\left(\frac{\mu_s a}{b}\right)\left(\frac{a+1}{b}\right)$

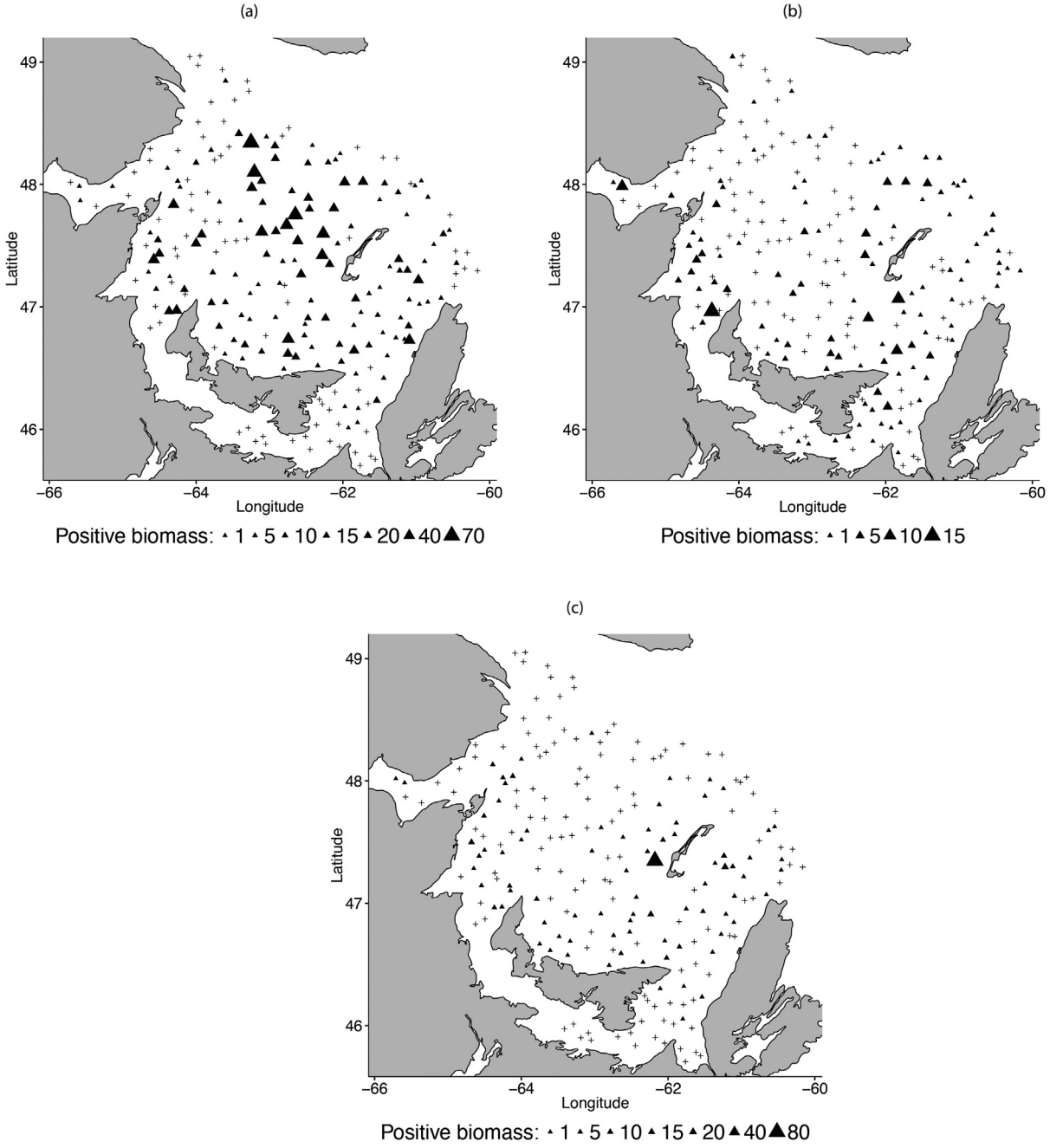

Fig. 2. Spatial distribution of macro-invertebrate biomass from individual tows in the 1997 sGSL bottom-trawl survey. Triangles represent sites where positive biomass was collected. Their widths are proportional to the biomass sampled (in kg/tow). Crosses represents sites where no biomass was caught: (a) urchin, (b) starfish, and (c) sea cucumber.

four classes to account for a possible non-linear response: [0, 50[; [50, 100[; and, [100, 400[. Likewise, bottom temperature (in °C) is split into three classes: [− 1, 1[; [1, 5[; and, [5, 15[. For the purpose of identifiability, some covariate classes have to be chosen as references. The most prevalent class for each covariate is defined as the baseline effect: fine sand for sediment type, [50, 100[for depth and [− 1, 1[for temperature.

2.2.2.2. Spatial effects. Covariates are often able to capture a part of the spatial dimension of a species' distribution, but they are unlikely to fully explain the spatial distribution. Even after accounting for the deterministic effect of environmental covariates, nearby locations are much more likely to resemble each other compared to more distant stations. In order to account for this residual spatial structure, spatially correlated errors are included. This approach has the benefit of improving inferences on the covariates by accounting for correlation and is very useful for creating interpolated maps of biomass in the study area (Lichstein et al., 2002). In practice, the random spatial process is included as a spatial Gaussian noise, $(\epsilon_s)_{s=1,\ldots,S}$, already defined in Eq. (4). The simple exponential covariance function, known to be robust in environmental applications, is used:

$$Cov(\epsilon_s, \epsilon_{s'}) = \sigma^2 \left(\exp\left(-\frac{h}{\Phi} \right) \right) \qquad \text{with} \quad h = d(s, s'). \tag{5}$$

It describes a decreasing exponential neighborhood covariation with increasing distance h between two sites s and s'. Φ is the

parametric range, which controls the rate of correlation decline as a function of distance, and σ^2 is the variance parameter.

Modeling the effect of the covariates and the spatial correlation could also have been done on the expected biomass in a patch (a/b). The inference of expected number of patch μ and the expected biomass in one patch a/b produces highly correlated estimates (Ancelet et al., 2010). Therefore, it is preferable to include covariates in only one of these quantities. Because the expected number of patches controls species presence as well as abundance, we choose to add both covariates and spatial effects to this latent layer.

2.3. Bayesian inference

Hierarchical models such as the one proposed in this paper have been developed under the Bayesian paradigm (Gelman et al., 2004). Bayesian analysis requires setting prior distributions for all the parameters (*i.e.* $a, b, \alpha_0, \zeta, \beta, \gamma, \sigma^2$, and Φ). We choose standard flat priors for the regression parameters ($\alpha_0, \zeta, \beta, \gamma$) because the number of patches and the effect of the three covariates are not known *a priori*. We choose a uniform prior distribution for the standard deviation σ as recommended by Gelman (2006).

The range parameter Φ is known to be difficult to estimate in hierarchical models (Cressie, 1993; Stein, 1999; Zhang, 2004; Zhang and Wang, 2009). We therefore devise a strategy based on data from other years for direct standard estimation. It is assumed that the latent spatial structure of the organisms changes little between neighboring years because of their limited dispersal abilities over large spatial scales and their slow biological turnover rates. For that matter, preliminary analyses involving data from 1996, 1997 and 1998 confirmed that the constant latent spatial structure assumption is reasonable (unpublished results). Therefore inference is conducted on the data sampled in 1996 by first fitting a model including all the covariates but no latent spatial structure and then applying a classical spatial analysis (kriging) to the resulting residuals. The range estimate is then plugged into the hierarchical analysis for 1997.

We choose a sufficiently vague prior for the expected biomass in a patch ($\mathbb{E}(M) = a/b$), with 5%, 50% and 95% quantiles that are respectively: 0.005, 1, and 135 kg. This prior distribution allows the expected biomass of one patch to take realistic values, which helps remove some of the possible confounding between a situation with a high number of small patches each containing a small amount of biomass and a low number of high biomass patches.

Eqs. (4) and 5 describe the full process model, but variants of this full model, with subsets of the explanatory covariates, are also considered as candidate alternatives (described below). Each model run is implemented in OpenBUGS, the open source version of WinBUGS (Ntzoufras, 2011) and its add-on GeoBUGS. The code for the model is presented in Appendix A. Previous tries with a reduced set of simulated data (to make sure that the inference algorithm worked) showed a strong autocorrelation of the MCMC iterations but the Gelman–Rubin convergence test became acceptable after 30,000 iterations of each of the three MCMC chains. Due to computational costs (about 3 days for each model), only two chains are launched for 60,000 iterations with a burn-in period of 30,000 iterations and MCMC convergence checked by visual inspection. A thinning of 100 iterations is performed in order to get rid of within-chain autocorrelation. Prediction and validation steps are computed in R 2.15.0 with the geoR package for the spatial correlation (Diggle and Ribeiro, 2001).

2.4. Validation and model selection

2.4.1. Posterior predictive checking

Posterior predictive checking is used to evaluate the model's ability to fit the observed data as recommended by Gelman et al. (1996). N draws $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)})$ are sampled from the posterior distribution for vector $\theta = (a, b, \alpha_0, \beta, \gamma, \zeta, \sigma^2, \Phi)$. For each draw $\theta^{(i)}$, artificial replicated data $(\hat{Y}_s^{(i)})_{s=1,\ldots,S}$ are generated using the model $Y_s|\theta^{(i)}$. When the model fits well the data, the replicated data $\hat{Y}_s$ are expected to be close to the observed data Y_s. The discrepancy between $\hat{Y}_s$ and Y_s can be calculated using the omnibus statistics denoted $T(Y, \theta)$ as suggested in Gelman et al. (1996). In that case, the chosen T statistic is a Bayesian residual sum of squares, which is asymptotically distributed as a χ^2 distribution Gelman et al. (1996). Model fit is assessed by comparing the observed $T(Y, \theta^{(i)})$ with the $T(Y, \theta^{(i)})$ distributions. A Bayesian p-value is then computed to account for the cases in which the $T(\hat{Y}^{(i)}, \theta^{(i)})$ statistic exceeds the $T(Y, \theta^{(i)})$. Bayesian p-value close to 0.5 indicate a well fitted model.

$$\begin{aligned} T(Y, \theta^{(i)}) &= \sum_{s=1}^{S} \left\{ \frac{(Y_s - \mathbb{E}(Y_s|\theta^{(i)}))^2}{\mathbb{V}(Y_s|\theta^{(i)})} \right\} = \sum_{s=1}^{S} \left\{ \frac{(Y_s - \frac{a^{(i)}\mu_s^{(i)}}{b^{(i)}})^2}{\frac{\mu_s^{(i)}a^{(i)}}{b^{(i)}} \frac{a^{(i)}+1}{b^{(i)}}} \right\} \\ T(\hat{Y}^{(i)}, \theta^{(i)}) &= \sum_{s=1}^{S} \left\{ \frac{(\hat{Y}_s^{(i)} - \mathbb{E}(\hat{Y}_s^{(i)}|\theta^{(i)}))^2}{\mathbb{V}(\hat{Y}_s^{(i)}|\theta^{(i)})} \right\} = \sum_{s=1}^{S} \left\{ \frac{(\hat{Y}_s^{(i)} - \frac{a^{(i)}\mu_s^{(i)}}{b^{(i)}})^2}{\frac{\mu_s^{(i)}a^{(i)}}{b^{(i)}} \frac{a^{(i)}+1}{b^{(i)}}} \right\} \end{aligned} \quad (6)$$

2.4.2. Model comparison

Many competing submodels of Eq. (4) can be defined based on combinations of the covariates and the spatially correlated errors. The full model (M_{s1}) includes all covariates (depth, temperature and sediment type) and the spatial correlation. For each species, this full model is compared with several submodels summarized in Table 2. Two criteria are used to compare the models. The first is the Bayesian Information Criterion (*BIC*) proposed by Schwarz (1978):

$$BIC = -2 \times \log(L) + k \times \log(n) \quad (7)$$

where n is the number of observations, k is the number of parameters to be estimated in the submodel and L is the maximized value of the likelihood function for the estimated submodel. The *BIC* measures how well the model fits the data, with respect to the number of model parameters.

The second comparison criterion, the mean squared prediction error (*MSPE*) is used to assess the accuracy of the model predictions:

$$MSPE_s = \sum_{s=1}^{S} \mathbb{E}((\hat{Y}_s - Y_s)^2) = \sum_{s=1}^{S} \{(\mathbb{E}(\hat{Y}_s) - Y_s)^2 + \mathbb{V}(\hat{Y}_s)\}. \quad (8)$$

This criterion takes into account the bias of the prediction relative to the true value, and includes a term for the predictive variance. Here, we consider the spatial average of this criterion over the prediction dataset.

2.5. Predictions

The main advantage of models such as the ones proposed in this paper is the capability of making predictions, Y_{new}, conditional on the observations, while preserving the inferred spatial structure. The predictive distribution of the biomass quantity is given by:

$$[Y_{new}|Y_{obs}] = \iint [Y_{new}, \theta, \mu|Y_{obs}] d\mu \, d\theta. \quad (9)$$

In practice, for each iteration i of the MCMC chains, we perform a conditional simulation $[\epsilon_{new}^{(i)}|\sigma^{2(i)}, \Phi, \epsilon^{(i)}]$ to obtain by kriging a realization of the latent Gaussian field $\epsilon_{new}^{(i)}$ at the sites where predictions are wanted. Prediction for a new site s_0 requires knowledge of the values of the covariates at this site. These are obtained using a linear interpolation of values at neighboring sites for the

Table 2
Alternate submodels for the spatial hierarchical model and results of the model fit for the three taxa based on the model selection criteria BIC and MSPE, and the Bayesian *p*-value model checking criterion. *s* denotes a model with spatial correlation and a lack of *s* means no spatial correlation. M_{s1} is the more complex model, which includes all covariates: sediment type (*Sed*), depth (*Dep*) and bottom temperature (*Temp*) and a spatial correlation. Model M_1 include all covariates without spatial structure.

	Covariates			*BIC*	B. *p*-value	*MSPE*
Model	*Sed*	*Dep*	*Temp*			
Strongylocentrotus droebachiensis						
M_{s1}	×	×	×	844.41	0.51	38.41
M_{s2}	×	×		888.76	0.46	53.13
M_{s3}	×		×	853.28	0.43	38.69
M_{s4}	×			874.16	0.41	42.45
M_{s5}		×	×	927.55	0.47	74.187
M_{s6}		×		961.34	0.49	91.10
M_{s7}			×	999.78	0.54	122.67
M_{s8}				1078.47	0.53	225.21
M_1	×	×	×	1022.67	0.54	160.91
Asterias sp						
M_1	×	×	×	317.76	0.48	0.69
M_2	×	×		343.33	0.45	0.71
M_3	×		×	338.61	0.45	0.73
M_4	×			365.63	0.41	0.82
M_5		×	×	392.24	0.49	1.54
M_6		×		399.77	0.55	1.36
M_7			×	461.06	0.41	2.76
M_8				412.38	0.53	2.56
Cucumaria frondosa						
M_{s1}	×	×	×	215.38	0.54	0.11
M_{s2}	×	×		270.12	0.55	0.12
M_{s3}	×		×	244.05	0.46	0.15
M_{s4}	×			263.09	0.44	0.17
M_{s5}		×	×	244.80	0.52	0.14
M_{s6}		×		270.12	0.42	0.18
M_{s7}			×	282.09	0.41	0.22
M_{s8}				305.34	0.56	0.26
M_1	×	×	×	482.19	0.57	5.65

temperature and depths, while sediment type is obtained from an interpolated map from the study of Loring and Nota (1973). Then, the latent layer $\mu^{(i)}_{new}$ is generated to account for the effects of the included covariates. The biomass of the studied species in unsampled locations, $Y^{(i)}_{new}$, is then merely drawn from the observation model sub-component. These posterior predictive joint distributions can be summarized by maps showing various statistics of the species distribution (e.g. mean, or median biomass, or the proportion of zeros). Given the assumption of little movement between successive years, we rely on data from the 1998 survey (not used for model fitting) to evaluate the predictive ability of the competing models.

3. Results

The results are presented by species, following a common presentation format.

1. Results of analyses to establish whether there is spatial structure in the residuals, and in the affirmative case, estimating the range parameter using data from the 1996 survey.
2. The model is fitted to the 1997 data under the Bayesian paradigm and its predictive ability is checked using the 1998 data.
3. Results of submodel comparisons and the implied effect of covariates are presented.

3.1. Green sea urchin

Spatial structure is apparent in the residuals of the inference performed on the green sea urchin biomass sampled in 1996 (Fig. 3a). The estimated value for the parametric range is $\hat{\Phi} = 23$ km. This range is considerably smaller than the average inter-station distance in the annual survey of the sGSL, 156 km. *BIC* and *MSPE* scores are considerably smaller for the full spatialized model, M_{s1}, compared to a similar model without spatial correlation, M_1 (Table 2), indicating that adding a spatial structure improves model fit. The evaluation of competing models with different subsets of covariates is therefore limited to models that include spatial correlation. Detailed results of submodel comparisons for the green sea urchin are provided in Table 2.

The validation of the models by posterior predictive checking gave acceptable results for all models, with the *p*-values around 0.5. Based on *BIC*, the best fitting model is the one that includes all three covariates. This model also has the best *MSPE*. The different effects of the covariates included in this model are presented by their posterior distributions in Fig. 4. Sediment type has an important effect on the biomass of the green sea urchin, with pelite having a negative effect and both coarse sand

Table 3
Proportion of predictions of the best model for the three species (urchin: M_{s1}, starfish: M_1, cucumber: M_{s1}) for the biomass sampled in 1998.

	Predictions			
	True zero	False positive	True positive	False zero
Strongylocentrotus droebachiensis	0.71	0.29	0.76	0.24
Asterias sp	0.55	0.45	0.51	0.49
Cucumaria frondosa	0.79	0.21	0.52	0.48

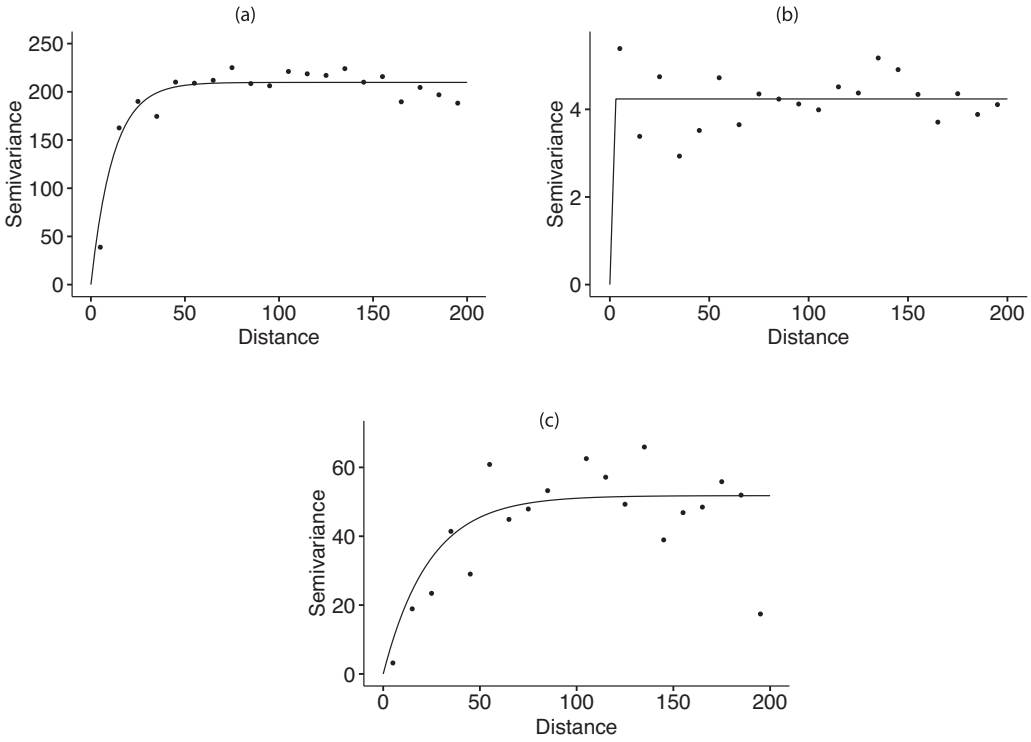

Fig. 3. Variogram for the biomass sampled in 1996 obtained from the residuals of the model including all covariates without spatial correlation for the three species: (a) green sea urchin, (b) starfish, and (c) sea cucumber. Distance in kilometers.

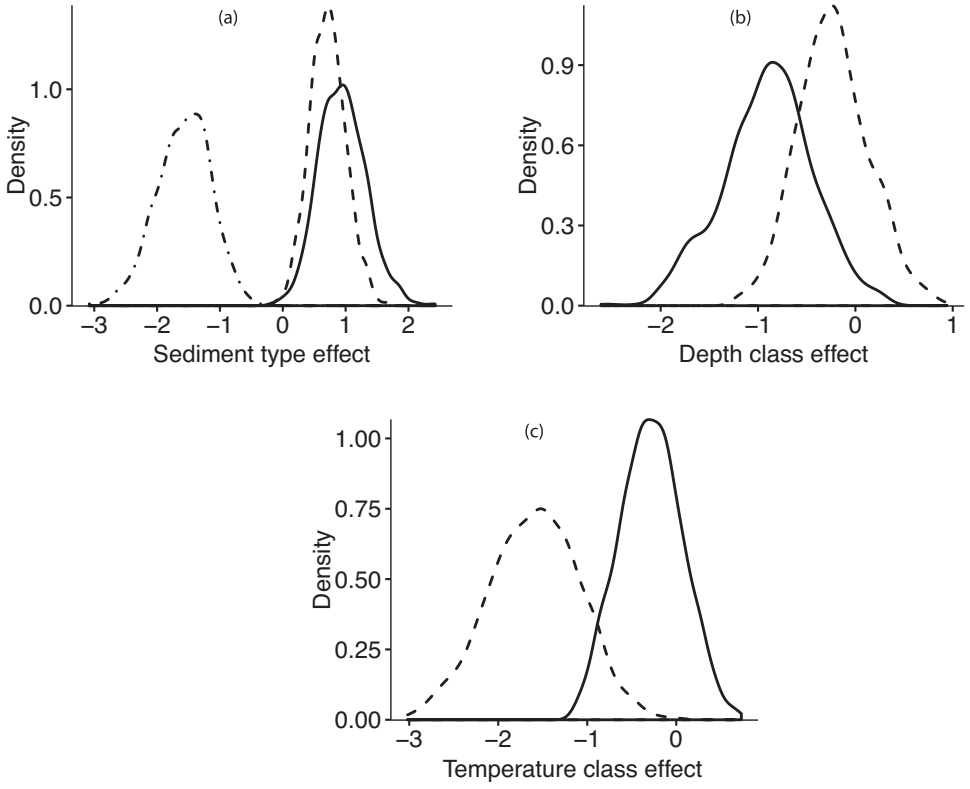

Fig. 4. Posterior distributions of the parameters included in the model M_{s1} for green sea urchin (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

and gravel having a positive effect relative to fine sand (Fig. 4a). Depth also appears to be important with large depths having a negative effect on the urchin biomass (Fig. 4b). The effect of bottom temperature is negative for both classes, [1, 5[and [5, 15[, relative to the effect of the lower temperature class. The negative effect is strongest for the warmest temperature class (Fig. 4c).

The majority of the 1998 biomass records are predicted well, though there are some misclassifications with the positive biomass (Table 3). Qualitatively, the interpolated map of the median biomass quantities predicted with Model M_{s1} matches well the survey observations (Fig. 5). This interpolation allows a good identification of areas with a high quantity of biomass as well as areas without biomass. Note that urchins are widely distributed in the sGSL.

3.2. Starfish

No spatial structure is detected in the residuals of the model fitted to the starfish data in 1996 (Fig. 3b). The favored model, based on the two model selection criteria includes the effects of the sediment type, depth and temperature classes (Table 2; Fig. 6).

Pelite type sediment has a negative effect and the gravel type has a positive effect on starfish biomass, relative to fine sand (Fig. 6a). The effect of depth is manifested by a small positive effect of the shallow depths relative to intermediate depths (Fig. 6b). Temperature classes [1, 5[and [5, 15[have a positive effect on the starfish biomass relative to temperatures between −1 and 1 °C (Fig. 6c). The model is not able to provide good predictions of starfish biomass sampled in 1998 (Table 3).

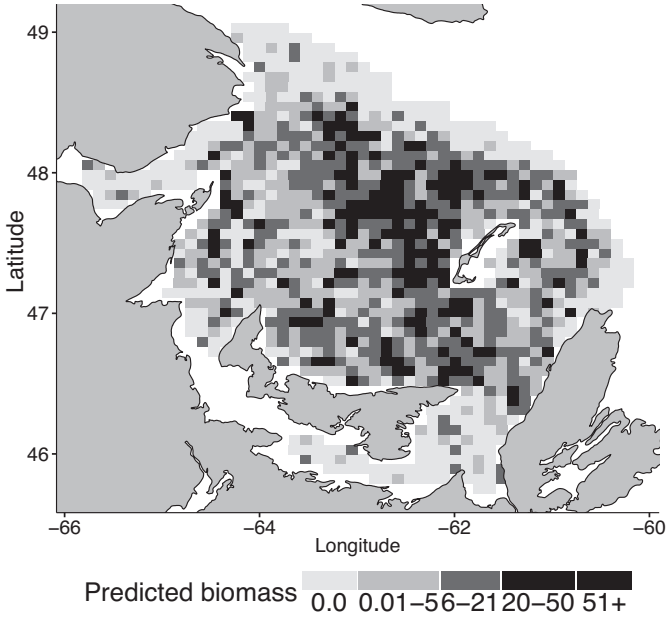

Fig. 5. Prediction of the median quantity of green sea urchin biomass (in kg per standard tow) on a grid in the sGSL.

3.3. Sea cucumber

Some spatial structure is detected in the 1996 distribution of sea cucumber biomass (Fig. 3c). As for the urchins, the estimated parametric range $\hat{\Phi} = 22$ km, is considerably smaller than the average interstation range in the survey. *BIC* and *MSPE* scores confirm that

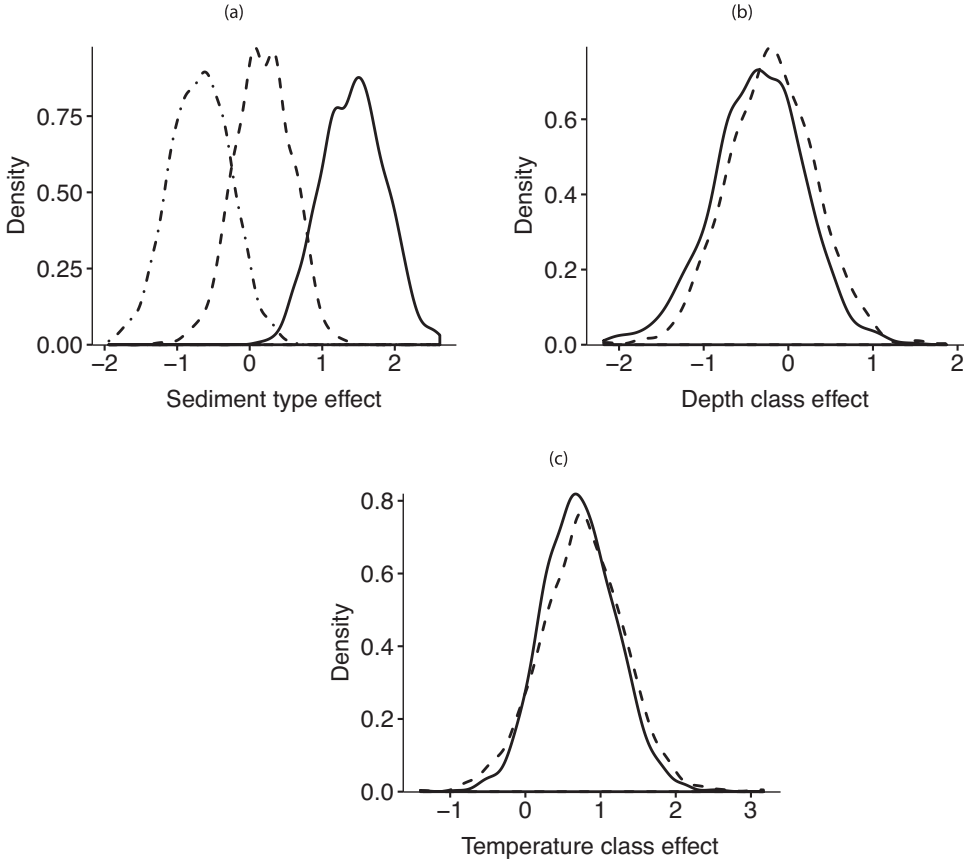

Fig. 6. Posterior distributions of the parameters included in the model M_1 for starfish (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

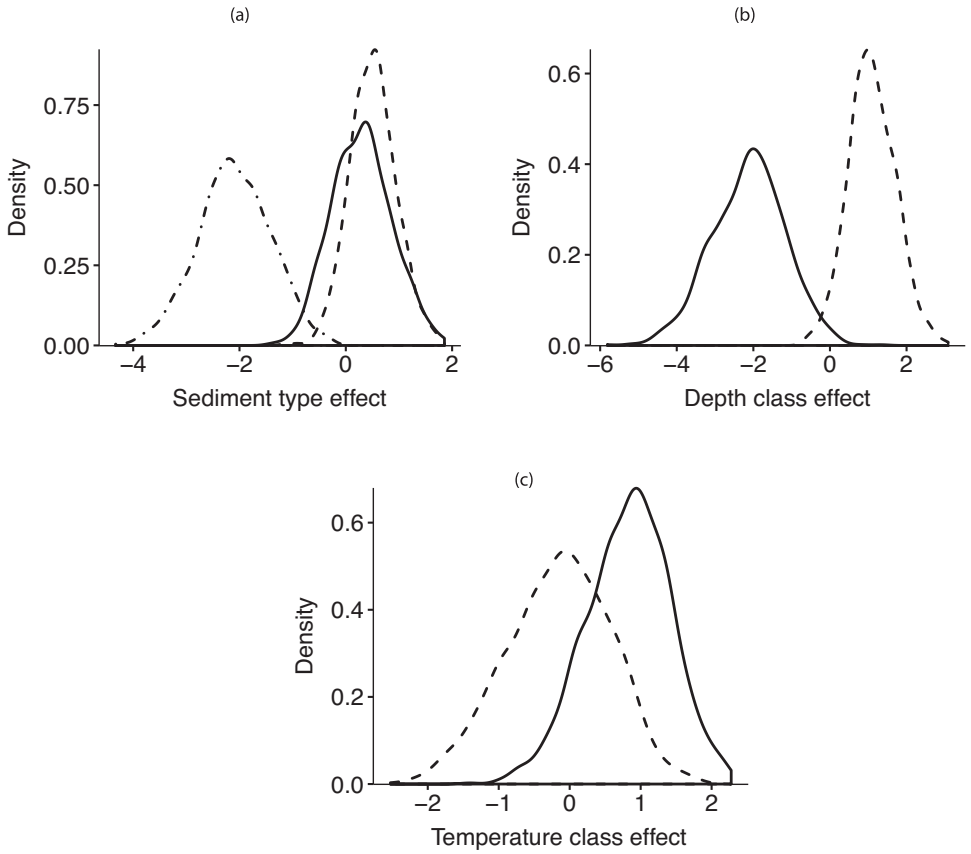

Fig. 7. Posterior distributions of the parameters included in the model M_{s1} for sea cucumber (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

a model including spatially correlated error is preferred (Table 3). The analysis of competing models reveals that a model including all three covariates is favored (Table 3). Relative to fine sand, pelite sediment has a negative effect on sea cucumber biomass, while coarse sand and gravel has a positive effect (Fig. 7a). Sea cucumber biomass is predicted to be greater in shallow depths, and lower in deeper areas, relative to intermediate depths (Fig. 7b). The effect of temperatures is positive for the [1, 5[class and null for the [5, 15[class, compared to the coldest temperature class. (Fig. 7c).

The sites without sea cucumber in 1998 are well predicted by the model (Table 3). However, the model predicts a high number of false positive values. Fig. 8 represents the median quantity of cucumber biomass predicted by Model M_{s1}. This map allows the identification of a small area of high biomass in the center of the area near the Magdalen Islands archipelago and points out large areas without sea cucumbers in the northern portion of the survey area where depth declines rapidly into a large channel, the Laurentian channel.

4. Discussion

In this work, we propose a Bayesian hierarchical model to handle spatialized continuous zero-inflated data. We expand the model proposed by Ancelet et al. (2010) by considering Gamma marks instead of Exponential ones, which allows more flexibility in the modeling approach. We also add locally recorded covariates to the model in order to describe the habitat associations as well as complex spatial dependence structures.

The distributions of the three species appear to be affected by the sediment type and depth, which is expected given that both are key variables affecting the distribution of epibenthic invertebrates in other marine ecosystems (Freeman and Rogers, 2003; Hily et al., 2008). The species are also affected by temperature, a variable known to rule the distribution of marine biota. Noticeable residual spatial variation remains for urchins and sea cucumbers, but not for starfish. Model performance for starfish was weaker than for the other two species. A probable explanation is that contrary to urchins and sea cucumbers, the starfish data represent catches for an assemblage of species, each with potentially unique habitat

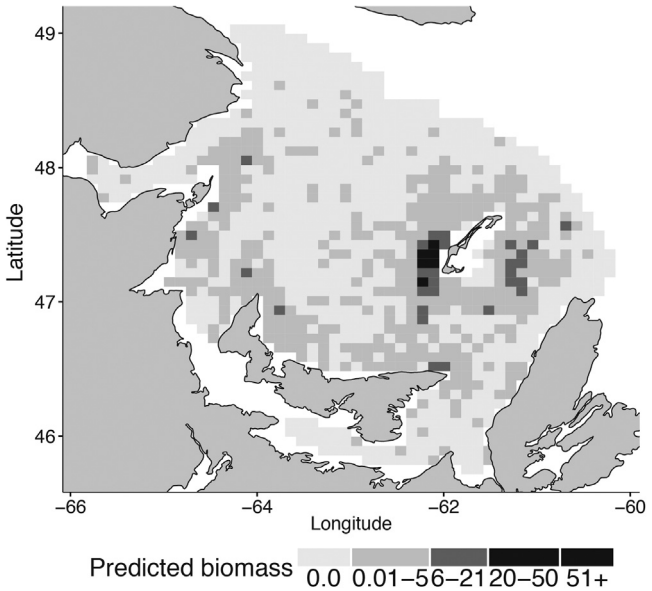

Fig. 8. Prediction of the median quantity of sea cucumber biomass (in *kg* per standard tow) on a grid in the sGSL.

associations. Modeling the assemblage as a unit effectively means modeling a weighted average habitat association, which may not be strong and which will change as the relative abundance of the component species changes. This is likely to lead to a weakened ability to predict distribution patterns over time based on habitat covariates. Furthermore, the performance of this model will be influenced by sampling stochasticity given that the data from any one year of the survey represent a single sampling realization that will only on average reflect the true patterns in distribution and habitat associations even if the model is correct. Such an effect may explain some of the deficiencies in model performance observed for all three invertebrate taxa.

The small estimated values for the range parameters are consistent with the limited dispersal abilities of the organisms considered here. For example, the maximum displacement of green sea urchin is estimated at 20 cm per day *i.e.* approximately 17 km per year (Lauzon-Guay and Scheibling, 2007). Possible explanations for this small scale effect are small scale habitat features, other habitat features not included in the model, or a result of demographic or behavioral characteristics of the organisms. This small correlation distance helps to capture a spatial structure at small scales, which is not otherwise related to the deterministic effect of covariates that vary over much broader scales. Taking into account this spatial correlation greatly improves the fitting and the predictive capacity of the model. However, the model still misclassifies some predictions of the biomass sampled in 1998. These errors could be the result of changes in the hydrodynamic and physico-chemical properties of the water column, which impacts macroinvertebrates over short temporal scales relative to depth or sediment types (Warwick and Uncles, 1980; Ysebaert and Herman, 2002; Freeman and Rogers, 2003; Bolam et al., 2008). Another possible explanation is that the range parameter ϕ is not well estimated because it is based on a single realization of the survey and because the survey sampling density is considerably coarser than the apparent range (Zimmerman, 2006; Irvine et al., 2007). Additional surveys with finer grain sampling might help in better defining the range parameter.

An attractive feature of the approach is its flexibility. For example, instead of a Poisson distribution for the number of patches, it is possible to use the Negative Binomial distribution, a Gamma mixture of Poisson. This alternative extension would allow for over-dispersion of the number of patches in the observation submodel, resulting in an over-dispersed quantity of biomass. A GLM with a logarithm link function is used here to model the intensity of the Poisson distribution but it would also be possible to use a generalized additive model in more complex cases (Guisan et al., 2002; Zuur et al., 2009).

The two latent quantities (number of patches and biomass quantity in each patch) must not be over-interpreted. These two conceptual quantities reflect a heuristic construction of a zero-inflated model but their direct ecological interpretation cannot be confirmed by observation. In addition, the two hidden quantities μ and a/b are highly correlated (Ancelet et al., 2010). While this would hamper the interpretation of their individual values, it is not an issue in the application presented here because we are interested in their joint effect, *i.e.* the predicted biomass. By linking μ instead of a/b to the covariates, covariates play a role in both the probability of presence and the quantity of biomass, also partially controlled by the number of patches.

Interpolation relying on covariates and spatial structure exploits a large amount of information from the data collected during marine surveys. Maps predicting ecological properties such as the mean biomass, areas of high density, or the presence–absence of organisms can be easily produced, together with their uncertainties. Such maps are commonly used for ecological analyses and data-based approaches (Shea, 1998; Hilborn et al., 2004; Hobday and Hartmann, 2006; Hartog et al., 2011; Dutertre et al., 2012). Furthermore, the approach provides preliminary answers with probabilistic predictions for the quantitative effect of long-term changes in important habitat variables, for example as might be expected under global warming.

Appendix A. Model code

```
model{
### Latent Layers
for (s in 1:nsample){
  log(mu[s])        <        - alpha0 + beta2[Tp[s]] +
beta1[Sed[s]] + beta3[Pr[s]] + w[s]
  w[s] <- v[s] - mean(v[1:nsample])
  mupres[s] <- mu[s]*(dtow[s]/dstandard)
}
v[1:nsample]      ~      spatial.exp(mu_v[],x[],y[],
tau_v,phi_v,kappa_v)
### Observation Model
### Strictly positive biomass data
for(k in 1:npres){
  ### Number of patches
  ngis[k] ~ dpois(mupres[pres[k]])C(1,)
  ### Observed positive biomass
  Y_a[k] <-a*ngis[k]
  Y[pres[k]] ~ dgamma(Y_a[k],b)
}
### Evaluation of probabilities of zeros
for(j in 1:nabs){
  ### Probability of presence at site j
  proba[j] <- 1 - exp(-mupres[abs[j]])
  Y[abs[j]] ~ dbern(proba[j])
}
### Priors
alpha0 ~ dnorm(0,0.01)
beta1[2] <- 0
beta1[1] ~ dnorm(0,0.01)
beta1[3] ~ dnorm(0,0.01)
beta1[4] ~ dnorm(0,0.01)
beta3[2] <- 0
beta3[1] ~ dnorm(0,0.01)
beta3[3] ~ dnorm(0,0.01)
beta2[1] <- 0
beta2[2] ~ dnorm(0,0.01)
beta2[3] ~ dnorm(0,0.01)
a ~ dgamma(2,5)
b ~ dgamma(2,5)
sigma ~ dunif(0,100)
sigma2 <-sigma*sigma
tau_v <-1/(sigma2)
    }
```

References

Ancelet, S., Etienne, M.-P., Benoît, H.P., Parent, E., 2010. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process? Environmental and Ecological Statistics 17 (3), 347–376.

Benoît, H.P., Swain, D.P., Chouinard, G.A., 2009. Using the long-term bottom-trawl survey of the southern Gulf of St. Lawrence to understand marine fish populations and community change. AZMP Bulletin 8, 19L 27.

Bernier, J., Fandeux, D., 1970. Théorie du renouvellement application à l'étude statistique des précipitations mensuelles? Revue de Statistiques Appliquées 18 (2), 75–87.

Bolam, S., Eggleton, J., Smith, R., Mason, C., Vanstaen, K., Rees, H., 2008. Spatial distribution of macrofaunal assemblages along the English Channel. Journal of the Marine Biological Association of the UK 88 (04), 675–687.

Chadwick, E., Brodie, W., Colbourne, E., Clark, D., Gascon, D., Hurlbut, T., 2007. History of annual multi-species trawl surveys on the Atlantic coast of Canada. Atlantic Zonal Monitoring Program Bulletin 6, 25L 42.

Cook, A., Marion, G., Butler, A., Gibson, G., 2007. Bayesian inference for the spatio-temporal invasion of alien species? Bulletin of Mathematical Biology 69 (6), 2005–2025.
Cressie, N., 1993. Statistics for Spatial Data, revised edition. New York, Wiley.
Diggle, P., Ribeiro, P., 2001. geoR: a package for geostatistical analysis? R news 1 (2), 14–18.
Dutertre, M., Hamon, D., Chevalier, C., Ehrhold, A., 2012. The use of the relationships between environmental factors and benthic macrofaunal distribution in the establishment of a baseline for coastal management? ICES Journal of Marine Science 70 (2), 294–308.
Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data? Journal of Applied Ecology 41 (2), 263–274.
Fletcher, D., MacKenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression? Environmental and Ecological Statistics 12 (1), 45–54.
Freeman, S., Rogers, S., 2003. A new analytical approach to the characterisation of macro-epibenthic habitats: linking species to the environment. Estuarine. Coastal and Shelf Science 56 (3-4), 749–764.
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). Bayesian Analysis 1, 515L 534.
Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian Data Analysis. CRC press.
Gelman, A., Meng, X.-l., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733L 807.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene? Ecological Modelling 157 (2-3), 89–100.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models? Ecology Letters 8 (9), 993–1009.
Hartog, J.R., Hobday, A.J., Matear, R., Feng, M., 2011. Habitat overlap between southern bluefin tuna and yellowfin tuna in the east coast longline fishery-implications for present and future spatial management? Deep Sea Research Part II: Topical Studies in Oceanography 58 (5), 746–752.
Hilborn, R., Stokes, K., Maguire, J.-J., Smith, T., Botsford, L.W., Mangel, M., Orensanz, J., Parma, A., Rice, J., Bell, J., Cochrane, K.L., Garcia, S., Hall, S.J., Kirkwood, G., Sainsbury, K., Stefansson, G., Walters, C., 2004. When can marine reserves improve fisheries management? Ocean & Coastal Management 47 (3–4), 197–205.
Hily, C., Le Loc'h, F., Grall, J., Glémarec, M., 2008. Soft bottom macrobenthic communities of North Biscay revisited: long-term evolution under fisheries-climate forcing. Estuarine, Coastal and Shelf Science 78 (2), 413–425.
Hobday, A.J., Hartmann, K., 2006. Near real-time spatial management based on habitat predictions for a longline bycatch species? Fisheries Management and Ecology 13 (6), 365–380.
Irvine, K.M., Gitelman, A.I., Hoeting, J.A., 2007. Spatial designs and properties of spatial correlation: effects on covariance estimation. Journal of Agricultural, Biological, and Environmental Statistics 12 (4), 450–469.
Lauzon-Guay, J.-S., Scheibling, R.E., 2007. Seasonal variation in movement, aggregation and destructive grazing of the green sea urchin (*Strongylocentrotus droebachiensis*) in relation to wave action and sea temperature. Marine Biology 151 (6), 2109–2118.
Lichstein, J.W., Simons, T.R., Shriner, S.A., Franzreb, K.E., 2002. Spatial autocorrelation and autoregressive models in ecology? Ecological Monographs 72 (3), 445–463.
Loring, D., Nota, D., 1973. Morphology and sediments of the Gulf of St. Lawrence. Fisheries and Marine Service, Ottawa.
Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A., Possingham, H., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations? Ecology Letters 8 (11), 1235–1246.
Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches? Fisheries Research 70 (2–3), 141–159.
Ntzoufras, I., 2011. Bayesian Modeling using WinBUGS, vol. 698. Wiley, Hoboken, NJ.
Perry, R.I., Smith, S.J., 1994. Identifying habitat associations of marine fishes using survey data: an application to the Northwest Atlantic? Canadian Journal of Fisheries and Aquatic Sciences 51 (3), 589–602.
Schwarz, G., 1978. Estimating the dimension of a model? The Annals of Statistics 6 (2), 461–464.
Shea, K., 1998. Management of populations in conservation, harvesting and control. Trends in Ecology & Evolution 13 (9), 371–375.
Shono, H., 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis? Fisheries Research 93 (1-2), 154–162.
Sileshi, G., Hailu, G., Nyadzi, G.I., 2009. Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data? Ecological Modelling 220 (15), 1764–1775.
Stefansson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches? ICES Journal of Marine Science 53 (3), 577–588.
Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.
Warwick, R., Uncles, R., 1980. Distribution of benthic macrofauna associations in the Bristol channel in relation to tidal stress. Marine Ecology Progress Series 3, 97L 103.
Welsh, A., Cunningham, R., Donnelly, C., Lindenmayer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros? Ecological Modelling 88 (1-3), 297–308.
Williams, A., Bax, N., 2001. Delineating fish-habitat associations for spatially based management: an example from the south-eastern Australian continental shelf. Marine and Freshwater Research 52 (4), 513–536.
Ysebaert, T., Herman, P., 2002. Spatial and temporal variation in benthic macrofauna and relationships with environmental variables in an estuarine, intertidal soft-sediment environment. Marine Ecology Progress Series 244, 105L 124.
Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics? Journal of the American Statistical Association 99 (465), 250–261.
Zhang, H., Wang, Y., 2009. Kriging and cross-validation for massive spatial data? Environmetrics 21 (3-4), 290–304.
Zimmerman, D.L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17 (6), 635–652.
Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Statistics for Biology and Health. Springer, New York.