Navigating Challenges in Spatio-temporal Modelling of Antarctic Krill Abundance: Addressing Zero-inflated Data and Misaligned Covariates

André Victor Ribeiro Amaral[†,*], Adam M. Sykulski[†], Emma Cavan[†], Sophie Fielding[‡]

[†]Imperial College London. London, UK.

[‡]British Antarctic Survey (BAS). Cambridge, UK.

[*]Corresponding author. E-mail: a.ribeiro-amaral@imperial.ac.uk

arXiv:2412.01399v1 [stat.AP] 2 Dec 2024

Abstract

Antarctic krill (*Euphausia superba*) are among the most abundant species on our planet and serve as a vital food source for many marine predators in the Southern Ocean. In this paper, we utilise statistical spatio-temporal methods to combine data from various sources and resolutions, aiming to accurately model krill abundance. Our focus lies in fitting the model to a dataset comprising acoustic measurements of krill biomass. To achieve this, we integrate climate covariates obtained from satellite imagery and from drifting surface buoys (also known as drifters). Additionally, we use sparsely collected krill biomass data obtained from net fishing efforts (KRILLBASE) for validation. However, integrating these multiple heterogeneous data sources presents significant modelling challenges, including spatio-temporal misalignment and inflated zeros in the observed data. To address these challenges, we fit a Hurdle-Gamma model to jointly describe the occurrence of zeros and the krill biomass for the non-zero observations, while also accounting for misaligned and heterogeneous data sources, including drifters. Therefore, our work presents a comprehensive framework for analysing and predicting krill abundance in the Southern Ocean, leveraging information from various sources and formats. This is crucial due to the impact of krill fishing, as understanding their distribution is essential for informed management decisions and fishing regulations aimed at protecting the species.

Keywords: Antarctic krill; Hurdle model; Misaligned data; Spatial statistics; Zero-inflated.

1 Introduction

In environmental statistics, modelling complex ecological systems often involves substantial methodological challenges, many of which are widely encountered across various applications. A common challenge is misaligned data, where variables collected at different spatio-temporal resolutions must be integrated into a unified model. For instance, remotely sensed data, such as satellite imagery, commonly provide information on environmental phenomena at various gridded resolutions—which is fundamentally different from, e.g., data collected along transects or continuous trajectory data. Additionally, ecological datasets are often zero-inflated, containing an excess of zero observations due to the natural absence of a species or resource in certain areas. These challenges highlight the need for more sophisticated modelling frameworks that can handle such complexities, producing accurate and interpretable results while remaining computationally feasible for inference. Such aspects are central to our approach to modelling the abundance of krill in the Southern Ocean.

Antarctic krill (*Euphausia superba*), hereafter referred to as "krill," are one of the largest species of crustacean that lives in the water column (Cavan et al., 2019) and have one of the highest biomasses of any species on Earth (Atkinson et al., 2009; Bar-On et al., 2018; Yang et al., 2022). Growing up to 6 cm in size and occupying a low level in the food chain, krill efficiently transfer energy by feeding on phytoplankton and serving as prey for numerous predators, including whales, seals, and penguins (Ruck et al., 2014). Their keystone role highlights their importance to the structure and functioning of the Southern Ocean ecosystem (McCormack et al., 2021). In addition, krill are the target of the largest fishery in the region (Nicol et al., 2012). Over the past two to three decades, research on krill abundance has primarily aimed to protect krill and their predators from the impacts of fishing (Nicol et al., 2012). More recently, their role in biogeochemical cycling, particularly the carbon cycle (Cavan et al., 2019), has provided another compelling reason for conservation. Antarctic krill contribute significantly to carbon sequestration by producing long strings of carbon-rich faecal pellets that sink hundreds of metres per day, reaching deep ocean layers where the carbon can remain stored for over a century. For instance, using a combination of krill abundance data (KRILLBASE) (Atkinson et al., 2017) and outputs from a physical ocean circulation model, Cavan et al. (2024) demonstrated that krill can sequester approximately 20 MtC (megatonnes of carbon) annually in the ocean interior.

At the simplest level, protecting krill from overfishing through spatial conservation policies requires knowledge of their abundance and spatio-temporal distribution across the Southern Ocean. Although often classified as plankton, there is ongoing debate about whether krill should instead be considered "nekton," as they are capable of swimming and forming massive swarms that can move against currents. As a result, while they inhabit all regions of the Southern Ocean, their distribution is highly patchy at any given time. Thus, to achieve dynamic conservation measures that adapt to the changing locations of krill, we must be able to understand their patterns in space and time. Currently, the best estimates of spatial krill biomass or abundance come from historic net haul data (KRILLBASE) and acoustic surveys, which are limited to discrete observations from research vessels (Fielding et al., 2014; Atkinson et al., 2017). The Southern Ocean's remoteness and harsh conditions restrict access to research

vessels to just half the year when weather permits, making sampling both challenging and expensive. This highlights the critical need for a comprehensive modelling framework to enhance the spatial (and temporal) coverage of krill monitoring.

Integrating remotely sensed data and *in situ* measurements may provide a robust approach to addressing many challenges in modelling krill abundance. Satellite imagery offers large-scale, high-resolution information on key environmental variables (e.g., sea surface temperature, chlorophyll concentration, salinity, etc.), while *in situ* data provides precise, location-specific observations that capture dynamic oceanographic processes with finer detail. In this paper, the *in situ* covariates come from drifters, which track near-surface ocean currents and provide valuable insights into localised water movements. These drifter trajectories allow the derivation of additional environmental covariates, such as speed and mass flux, as we shall detail. When combined with satellite imagery, these datasets further enhance our ability to model the physical and biological factors influencing krill distribution. To estimate krill abundance, we use acoustic data (see Section 2.1.1), while the KRILLBASE dataset is employed for validation when extrapolating beyond the observed area.

The remainder of this paper is structured as follows. In Section 2, we introduce the krill abundance data and additional datasets used to construct covariates. Section 3 outlines the spatio(-temporal) hurdle model applied to krill abundance in the South Georgia region and detail the mathematical framework for deriving spatial products from drifter trajectories. In Section 4, we present and interpret the model estimates for krill biomass. Finally, in Section 5, we provide an overall discussion of our modelling approach and findings, highlight key limitations, and suggest potential extensions for future work.

2 Materials

In this section, we present the datasets used in the analysis, including krill biomass measurements from acoustic and net haul data (KRILLBASE), along with remotely sensed data (e.g., satellite imagery) and *in situ* measurements from drifters, which are used as covariates in our model.

2.1 Study Area and Sampling Approach

Throughout this paper, we focus our analysis on subregions within the Southern Ocean, specifically around South Georgia, located in Subarea 48.3. This subarea, as defined by the CCAMLR (Commission for the Conservation of Antarctic Marine Living Resources, 2015), is a key ecological and management region due to its critical importance as both a krill habitat and a significant fishing area. The availability of both acoustic data and net haul data (KRILLBASE) for some parts of this region provides the necessary information to model krill abundance and distribution, making it a suitable focus for our study. Figure 1 illustrates the study area and shows the sampling locations for both acoustic (from 2016) and net haul data (spanning 1926 to 2016), and already highlights the inherent challenges of heterogeneity in the datasets, here exhibited through the highly irregular spatial sampling locations in both cases.

Note that, although both datasets provide information on krill biomass, they do not measure it

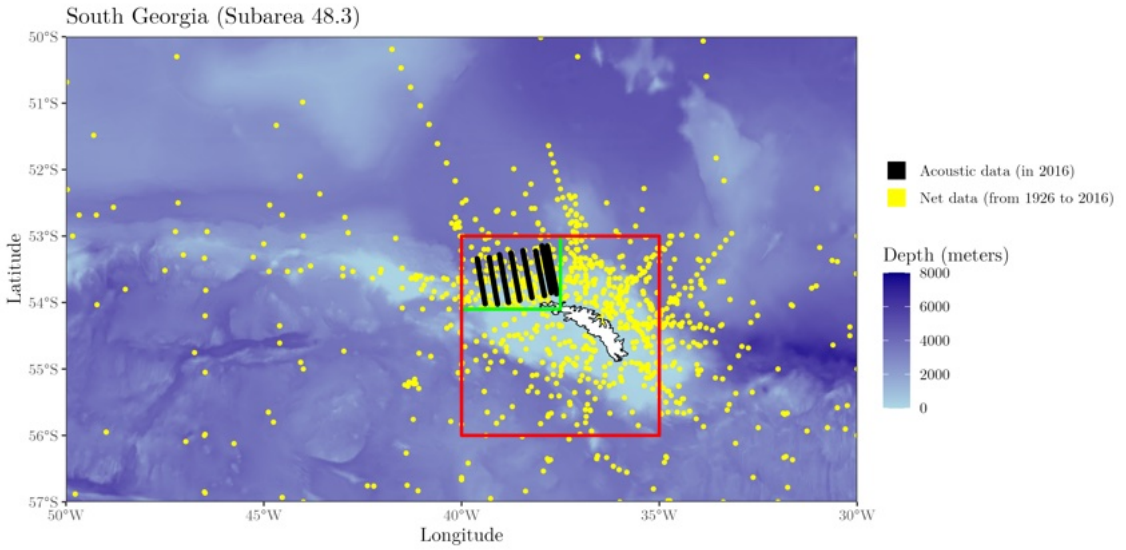

Figure 1: Study area (South Georgia, Subarea 48.3), showing acoustic data collected in 2016 and net haul data from KRILLBASE collected between 1926 and 2016. The green box indicates the region where acoustic data were collected, while the red box marks the area with sufficient net haul data to serve as a validation set for model predictions (see Section 4).

in the same manner. The net data serves only as a proxy for true krill abundance at a given location and time—since it is possible for a net to miss krill swarms even when deployed in krill-populated areas. In contrast, acoustic data enables high-resolution sampling of krill density along a vessel's path, offering a more precise measurement of krill biomass and serving as the primary data source for our analysis. In the following sections, we provide further details on these two datasets.

2.1.1 Acoustic Data

Acoustic surveys in South Georgia (Polar Ocean Ecosystem Time Series, Western Core Box) were conducted annually from 1997 to 2020, excluding the years 2002 and 2008, in intervals of 3 to 8 days within the December to February period (with December data considered as observations for the following year) (Fielding et al., 2014). Figure 4 (left column) presents the raw data for the first and last years, while plots for the remaining years are available in Figures SF2–SF21 (Supplementary Material). The surveys typically cover 8 transects, each 40 nautical miles in length, with a minimum separation of 10 nautical miles and a resolution of 500 metres.

The analysis of this dataset involves several challenges. Firstly, as shown in the Figure 4 (left column) and Figures SF2–SF21 (Supplementary Material), many observations are zero, indicating an absence of detected krill. Secondly, krill biomass can vary substantially even over short distances; in some instances, neighbouring observations span from zero to hundreds of g/m^2, highlighting the spatial heterogeneity of krill biomass distribution. Lastly, while we we would like to make predictions across the entire region shown in Figure 1, our data is limited to a much smaller area (green box, Figure 1). This limitation constrains our ability to generate reliable predictions for regions distant from the sampled

locations. In Section 4, we present results from two analyses covering the regions outlined by the green and red boxes (Figure 1).

2.1.2 KRILLBASE

KRILLBASE is a large-scale dataset documenting the krill biomass (g/m^2) based on net sampling conducted throughout the Southern Ocean from 1926 to 2016 (Atkinson et al., 2017). This dataset offers valuable, long-term insights into krill abundance, which will be useful when validating our model predictions. In this paper, we pre-processed this dataset following the same procedure described in Cavan et al. (2024), adjusting observations to estimate the expected krill density as of January each year (aligning with the acoustic data collection season), based on the collection date of each sample. Figure 1 shows the KRILLBASE sampling locations in South Georgia over the entire study period, with the corresponding spatio-temporally aggregated krill biomass, grouped into 0.2° longitude by 0.125° latitude cells, shown in the right-most plot of Figure 6.

2.2 Covariates

To effectively model krill abundance, we need to incorporate relevant covariates that capture environmental conditions influencing krill distribution. To obtain this information, we rely on multiple data sources, specifically satellite imagery and data products derived from the drifters. Satellite imagery offers large-scale, high-resolution coverage of environmental variables, while drifters provide valuable *in situ* measurements of ocean currents and other local conditions.

2.2.1 Satellite Imagery

We utilise ocean-related products from the Copernicus Marine Service (Copernicus Programme of the European Union, 2024), which provide high-resolution information on key ocean features within the study region—all of which can impact krill distribution patterns (Whitehouse et al., 2009; Warwick-Evans et al., 2022). By incorporating these satellite-derived covariates into our model, we account for large-scale environmental conditions that may drive changes in krill abundance across the South Georgia region. However, it is important to note that these datasets are not direct observations; rather, they are derived products created from satellite measurements and data processing techniques, and thus have associated uncertainty and loss of resolution due to instrumentation noise, and change of support and smoothing during processing.

Table 1 details all covariates used in our analysis, including products derived from satellite imagery, drifter trajectories (see Sections 2.2.2 and 3.2) and other environmental factors in the form of bathymetry (depth) and slope.

2.2.2 Drifter Data

The second data source, and the most challenging to incorporate, is the drifters. Part of NOAA's (National Oceanic and Atmospheric Administration) "Global Drifter Program," this dataset comprises

Table 1: Potential covariates for describing the spatial (and spatio-temporal) distribution of krill abundance. † indicates covariates obtained from satellite imagery, and ‡ indicates covariates derived as products from drifter trajectories. § denotes covariates observed only during the months of December, January, and February (to match the acoustic data time window). ⋆ indicates interpolation as described in Section SS1.1 (Supplementary Material).

Covariate	Spatial resolution (°)	Temporal resolution	Source	Label
Bathymetry (depth)†	0.01×0.01	NA	NOAA (10.25921/fd45-gt74)	depth
Slope	0.01×0.01	NA	Computed based on bathymetry	slope
Chlorophyll†	0.25×0.25	Yearly§	Copernicus Marine Service (10.48670/moi-00019)	chlor
Potential temperature†	0.083×0.083	Yearly§	Copernicus Marine Service (10.48670/moi-00021)	pot_temp
Salinity†	0.083×0.083	Yearly§	Copernicus Marine Service (10.48670/moi-00021)	salinity
Speed (satellite)†	0.083×0.083	Yearly§	Copernicus Marine Service (10.48670/moi-00021)	speed_sat
Surface temperature†	0.05×0.05	Yearly§	Copernicus Marine Service (10.48670/mds-00329)	surf_temp
Speed (drifters)‡	0.01×0.01—after interpolation⋆	1997–2020§	Computed based on drifter trajectories	speed_drif
Expected frequency‡	0.01×0.01—after interpolation⋆	1997–2020	Computed based on drifter trajectories	expect_freq
Residence time‡	0.01×0.01—after interpolation⋆	1997–2020	Computed based on drifter trajectories	res_time
Mass flux‡	0.01×0.01—after interpolation⋆	1997–2020	Computed based on drifter trajectories	mass_flux
Density of drifters‡	0.25×0.25	1997–2020	Computed based on drifter trajectories	density_drif

thousands of floating buoys known as *drifters* deployed in the ocean, whose positions are tracked over time by satellites, most typically using GPS. Figure 2 (left) shows the trajectories of all drifters that were observed in the South Georgia region during the analysed time period, with a zoomed-in view of the area where the acoustic data were collected (right). These data provide valuable *in situ* information about the study region which might inform krill abundance. Drifter data has previously been used to inform abundance and dynamics of a broad range of ocean-borne species and objects (O'Malley et al., 2021), including plankton (Laso-Jadart et al., 2023). While krill are not like plankton and can swim against weak currents, the impacts of ocean dynamics and currents on krill abundance and krill flux is nonetheless well documented (Murphy et al., 2004), therefore, there is reasonable scientific rationale for drifter data being informative in predicting krill abundance.

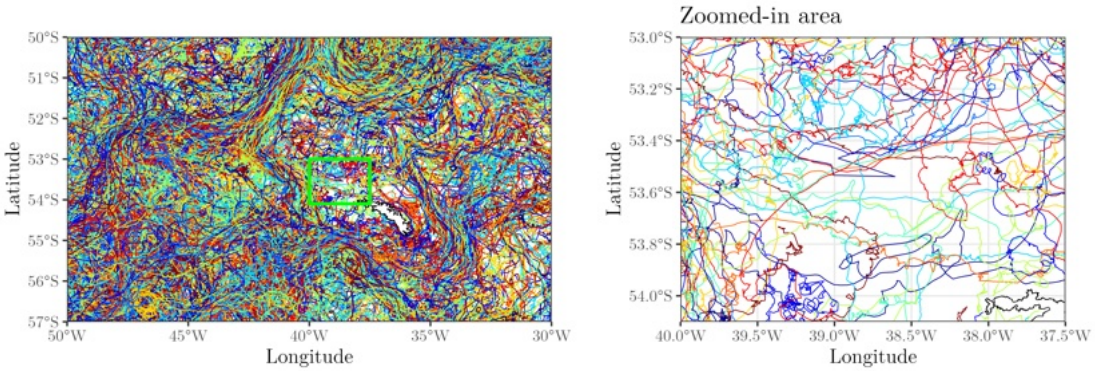

Figure 2: Left: drifter trajectories observed in South Georgia (Subarea 48.3) from 1997 to 2020, with different colours representing distinct trajectories. The green box indicates the region where acoustic data were collected, and the colours of the different trajectories are only used to ease the visualisation. Right: zoomed-in view of the green box.

In this paper, we consider all the trajectories presented in Figure 2 (left). Specifically, the positions of the buoys are recorded on an hourly basis (Elipot et al., 2016), with a total of 1,294 trajectories observed from 1997 to 2020. These trajectories vary in length from 122 to 8,797 points, adding up to

1,475,178 unique observations (or approximately 168.4 years' worth of data).

However, the drifter trajectories are not yet ready-to-use covariates as they are stored in the form of timestamped trajectories (as in Figure 2) rather than gridded spatio(-temporal) products as would be typical from e.g., satellite imagery. To proceed, we therefore aim to transform the drifters into spatially gridded data by extracting specific features of interest from them, which we shall describe in detail in Section 3.2. As we will show, various spatial data products can be derived from drifter trajectories, providing potentially orthogonal information to satellite imagery and enhancing our model. While drifter data have previously been applied in krill abundance modelling (Siegel et al., 2013), some of the products introduced in this study represent a novel use of this dataset as environmental covariates to describe krill distribution.

3 Methods

In this section, we outline the modelling framework and inference approach used to model krill abundance in the South Georgia region, as well as the methods for deriving valuable products from drifter trajectories.

3.1 Spatio-temporal Modelling

Throughout this paper, we use a Hurdle-Gamma model (Cragg, 1971; Min and Agresti, 2002) to address the challenges in modelling krill abundance, where the data consist of a non-negative continuous outcome with excess zeros. The Hurdle-Gamma model is particularly useful in this context, as it jointly models the probability of krill absence (i.e., presence-absence) and the distribution of non-zero abundance values.

Let $\mathcal{X} \subset \mathbb{R}^2$ denote the continuous spatial domain, with observed locations $(s_1, \cdots, s_n) \subset \mathcal{X}$. Similarly, let $\mathcal{T}$ denote the temporal domain, where $t \in \mathcal{T}$. Following the notation in Krainski et al. (2018), let

$$z_{it} = \begin{cases} 1, & \text{if the krill biomass is non-zero at location } s_i \text{ at time } t \\ 0, & \text{otherwise} \end{cases}$$

and y_{it} denotes the krill biomass at location s_i at time t, given that the biomass is non-zero. Specifically, we model the presence-absence component as $z_{it} \sim \text{Bernoulli}(\pi_{it})$ and the positive biomass as $y_{it} \sim \text{Gamma}(a_{it}, b_{it})$. The Gamma distribution is parametrized such that, $\mathbb{E}(y_{it}) = \mu_{it} = a_{it}/b_{it}$ and $\text{Var}(y_{it}) = a_{it}/b_{it}^2$.

The linear predictors for the presence-absence indicator z_{it} and the positive biomass y_{it} are specified as follows

$$\text{logit}(\pi_{it}) = \beta_0^z + \beta_1^z \texttt{cov}_1^z + \cdots + \beta_{\ell_1}^z \texttt{cov}_{\ell_1}^z + \psi_{it}, \tag{1}$$

and

$$\log(\mu_{it}) = \beta_0^y + \beta_1^y \texttt{cov}_1^y + \cdots + \beta_{\ell_2}^y \texttt{cov}_{\ell_2}^y + \gamma \cdot \psi_{it} + \xi_{it}, \tag{2}$$

where ℓ_1 and ℓ_2 denote the number of covariates in each model, which may overlap, and ψ_{it} and ξ_{it} are spatio-temporal random effects. Here, γ serves as a "copy" factor to scale the shared random effect ψ_{it} in the biomass model, allowing for dependencies between the presence-absence and biomass components.

For the random effects, we define ψ_{it} (similarly, ξ_{it}) as an autoregressive process to capture temporal correlation while allowing for spatial dependency at each time point (Moraga, 2019). In particular, we set

$$\psi_{it} = \alpha\psi_{i,(t-1)} + \phi_{it}, \tag{3}$$

where $|\alpha| < 1$, $\psi_{i1} \sim \text{Normal}(0, \sigma_\phi^2/(1-\alpha^2))$, and ϕ_{it} is a temporally independent but spatially dependent Gaussian Process (GP) at each year with covariance given by a Matérn kernel, i.e.,

$$\text{Cov}(\phi_{it}, \phi_{jt}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa \cdot h)^\nu K_\nu(\kappa \cdot h), \tag{4}$$

where $h = ||s_{it} - s_{jt}||$ is the Euclidean distance between the locations s_{it} and s_{jt}, and σ^2 denotes the marginal variance. $\Gamma(\cdot)$ is the Gamma function, and $K_\nu(\cdot)$ is a modified Bessel function of the second kind, such that $\nu > 0$ determines the mean square differentiability of the corresponding process. Lastly, $\kappa > 0$ is related to the range ρ, such that $\rho = \sqrt{8\nu}/\kappa$.

Finally, while we assume Gaussianity for the spatial field, this may not fully capture high spatial heterogeneity in the data, potentially leading to oversmoothing of distinct features such as sharp valleys or peaks. To address this, we perform a sensitivity analysis to assess the robustness of our results under this assumption before drawing any conclusions (see discussion in Section 4). This approach balances interpretability and computational feasibility, ensuring that the model remains practical to fit without imposing an excessive computational burden (Section 3.1.2).

3.1.1 Spatial Modelling

In Section 4.2, we will perform an aggregated spatial analysis and, in this instance, drop the temporal component from Equations (1) and (2). In this setting, the corresponding spatial Hurdle-Gamma model will be defined as before; however, the random effects ψ_i and ν_i will be modelled as Gaussian processes in space only, using a Matérn kernel similar to that in Equation (4).

3.1.2 Inference

Inference is conducted within a Bayesian framework using Integrated Nested Laplace Approximations (INLA) (Rue et al., 2009) to efficiently approximate posterior distributions in latent Gaussian models, which is particularly advantageous for complex spatio-temporal structures (as in Section 3.1). Model fitting also relies on the Stochastic Partial Differential Equation (SPDE) approach, where the Gaussian

field with a Matérn covariance structure is expressed as the solution of a SPDE (Whittle, 1963) and then approximated by a Gaussian Markov Random Field (GMRF) on a triangulated mesh (Lindgren et al., 2011), enabling a scalable representation of spatial dependence. Finally, we use Penalised Complexity (PC) priors (Simpson et al., 2017) for the parameters in the random effects, following the recommendations of Krainski et al. (2018). In practice, we implement our models using `R-INLA` (Lindgren and Rue, 2015), and the corresponding code is available at `https://github.com/avramaral/krill_abundance`.

3.2 Deriving Products from Drifter Trajectories

In this section, we use the drifter trajectory data introduced in Section 2.2.2 to derive spatial products for use as covariates in our krill abundance model. We note in passing that these products may also be valuable for predicting other ocean phenomena, such as the spread of oil spills, plankton, and plastic pollution.

We begin by establishing some notation. The observed position of drifter i in a spatial region of interest $\mathcal{X}$ at time t will be denoted by $q_i(t) \in \mathcal{X}$, representing its latitude-longitude coordinates. The collection of consecutive positions for each drifter i observed in region $\mathcal{X}$ will be denoted by $\{q_i(t)\}$ and is known as the *trajectory* of drifter i. Note that if the drifter leaves the spatial region of interest $\mathcal{X}$, but then re-enters, then multiple trajectories may be collected from the same drifter, and for simplicity we will denote each such trajectory with its own drifter index value i.

A primary use of drifter trajectory data is to track the velocity of the drifter along its path—often referred to in fluid dynamics as the *Lagrangian* velocity, named so because the drifter is deliberately designed to mimic a buoyant particle as it moves through time and space and thus has a Lagrangian perspective of the horizontal fluid flow near the surface. There are many works focussed on deriving statistics from Lagrangian velocities, see e.g., LaCasce (2008); Sykulski et al. (2016), where we employ similar notation and modelling principles here. As is typical in ocean flow analysis, the Lagrangian velocity of drifter i at time t will be modelled in the complex plane by $z_i(t) = u_i(t) + \mathrm{i}v_i(t)$ where $u_i(t)$ and $v_i(t)$ correspond to the zonal (eastward) and meridional (northward) velocities respectively, and are obtained in practice from $\{q_i(t)\}$ by some form of differencing or gradient modelling over time for each drifter i (Elipot et al., 2016). Representing two-dimensional time series in the complex plane is common in signal processing applications, especially when the two dimensions are measuring the same quantity (in this case, velocities) in orthogonal directions, and offers computational and modelling advantages over vector or bivariate representations, as reviewed in Sykulski et al. (2017), and as we shall take advantage of here.

Therefore we have at our disposal a collection of trajectories $\{q_i(t)\}$ and corresponding velocities $\{z_i(t)\}$ for drifter i inside region $\mathcal{X}$, where in Section 4 the region $\mathcal{X}$ will be the entire Subarea 48.3 shown earlier in Figure 1. We now seek to derive or "engineer" spatial covariates from $\{q_i(t), z_i(t)\}$ that can be utilised in our Hurdle-Gamma model of Equations (1) and (2). The key opportunity in deriving such covariates is to capture the local information content inherent in drifter trajectories (and their velocity gradients) that cannot be captured from satellite imagery. We now propose five such covariates, as shown in Table 1, which each capture different characteristics of the drifter data.

- **Speed (drifters):** we compute the speed of all drifter observations given by $|z_i(t)|$ and then map these to their corresponding locations $q_i(t)$. After which we create a spatially gridded product at the desired resolution by interpolating using Gaussian processes with a Matérn kernel, as detailed in Section SS1.1 (Supplementary Material). Note, importantly, that this covariate is expected to be different from the speed (satellite) covariate in Table 1, as the satellite data we use provides estimates of the *geostrophic* velocity computed from sea surface height (SSH) gradients, whereas drifter speeds are expected to be a mix of geostrophic and ageostrophic velocities (caused for example by surface winds). This difference is explained in detail by O'Malley et al. (2023). Figure 3 shows maps of the speed estimates in Subarea 48.3 from satellite data and drifter observations for comparison.
- **Expected frequency:** the speed from drifters is potentially informative, but ignores the information contained in the *shape* of the drifter trajectories $\{q_i(t)\}$. As detailed in Section 2.7 of LaCasce (2008), and Section 2.2 of Lilly et al. (2017), one of the best ways to understand the shape of drifter trajectories is via the *Lagrangian frequency spectrum* defined by

$$S_z(\omega) = \int_{-\infty}^{\infty} s_z(\tau) e^{-\mathrm{i}\omega\tau} d\tau, \quad \omega \in \mathbb{R},$$

where $s_z(\tau)$ is the autocovariance of the complex-valued velocity process $z(t)$ given by

$$s_z(\tau) = \mathbb{E}(z(t)z^*(t+\tau)) - \mathbb{E}(z(t))\mathbb{E}(z^*(t)), \quad \tau \in \mathbb{R},$$

where $z(t)$ is a second-order stationary stochastic process such that $s_z(\tau)$ is invariant over time t, and $z^*(t)$ denotes the complex conjugate of $z(t)$. The Lagrangian frequency spectrum can therefore be interpreted as the power spectral density of the velocity process, as it decomposes the second-order variability, or *power*, of the velocity process by frequency. Drifters that have a tendency to oscillate or jitter will have more power at high frequencies, and drifters that have a tendency to move in straighter lines will have more power at low frequencies. An informative covariate that summarises this content is the *expected frequency* of the velocity process given by

$$\mathrm{EF}_z = \int_{-\infty}^{\infty} \frac{|\omega| S_z(\omega)}{\int_{-\infty}^{\infty} S_z(\omega) d\omega} d\omega, \tag{5}$$

where the density $S_z(\omega)$ above has been normalised to integrate to 1 (such that it can be interpreted as a probability density over ω in some sense) thus explaining the term "expected frequency." In practice, we have at our disposal sampled velocity time series $z_i(t) = \{z_i(t_1), \ldots, z_i(t_{n_i})\}$ for each drifter i (of length n_i). Here, we can approximate the Lagrangian frequency spectrum via a tapered spectral estimate as follows

$$\hat{S}_i(\omega) = \frac{\Delta}{n_i} \left| \sum_{j=1}^{n_i} h_j z_i(t_j) e^{-\mathrm{i}j\omega\Delta} \right|^2, \tag{6}$$

where Δ is the temporal sampling interval and is assumed constant (which it is with the drifter data used in this paper; see Section 2.2.2). The sequence $\{h_j\}$ in Equation (6) is known as a data taper that satisfies $\sum_{j=1}^{n_i} h_j^2 = 1$, where we select $\{h_j\}$ to be a DPSS (discrete prolate spheroidal sequence) of order 1 (with bandwidth parameter set to 4), and is used to remove bias in the estimate of the spectrum, see (Percival and Walden, 1993, Chapter 6) for more details. We can then approximate the expected frequency in Equation (5) by

$$\widehat{\mathrm{EF}}_i = \frac{1}{\kappa} \sum_{k=1}^{n_i} |\omega_k| \hat{S}_i(\omega_k),$$

where

$$(\omega_1, \ldots, \omega_{n_i}) = \frac{2\pi}{n_i \Delta}(-\lceil n_i/2 \rceil + 1, \ldots, -1, 0, 1, \ldots, \lfloor n_i/2 \rfloor),$$

are the observed Fourier frequencies and $\kappa = \sum_{k=1}^{n_i} \hat{S}_i(\omega_k)$. The calculation of expected frequency requires the velocity time series to be approximately stationary, which will not generally be the case for an entire drifter trajectory in our region of interest. Therefore, we compute the expected frequency for each drifter trajectory over temporal windows (with 50% overlap) of length 5 days, which is considered to be a good approximation of the "decorrelation timescale" (i.e., the timescale at which a drifter "forgets" its history of movement), and is a standard choice in ocean drifter analysis (see O'Malley et al. (2021) and references therein). Finally, we derive a spatial gridded map by taking the set of computed expected frequencies and mapping them onto the midpoint location of each trajectory segment and then spatially smoothing onto a grid using Gaussian processes, as detailed in Section SS1.1 (Supplementary Material).

- **Residence time:** the expected frequency summarises the non-zero frequency content of a drifter trajectory. On the other hand, the zero frequency of the Lagrangian frequency spectrum yields a quantity known as the *diffusivity* which from Section 2.3 of Lilly et al. (2017) can be related via several quantities such that

$$\kappa_z = \frac{1}{4} S_z(0) = \frac{1}{4} \int_{-\infty}^{\infty} s_z(\tau) d\tau = \lim_{t \to \infty} \frac{1}{4} \frac{d}{dt} \mathbb{E}\left\{ |q(t)|^2 \right\}, \tag{7}$$

where $q(t) = \int_0^t z(\tau) d\tau$ is the (complex-valued) displacement of the drifter at time t where $z(t)$ is a zero-mean velocity process. The diffusivity can therefore also be understood as the integral of the autocovariance sequence over all lags, or as the expected rate of change over time of the squared displacement of the drifter after its mean is removed (i.e., the rate of diffusion)—thus linking Equation (7) to the physical notion and definition of diffusivity. Therefore, we propose a covariate from the drifters which can capture spectral information missing in the expected frequency, namely the diffusivity. However, diffusivity is difficult to estimate individually from single drifter trajectories, as the spatio-temporally varying local mean velocity (also known as the mean flow) must be removed and separated (Oscroft et al., 2020), but in the Global Drifter Program

the local mean flow is in general unknown due to drifter sparsity. We therefore instead estimate a quantity known as the *residence time*, commonly used in fluid dynamics and chemistry (Nauman, 2008), which estimates how long a fluid particle spends within a control volume of fixed size, thus incorporating both diffusivity and mean flow features. Specifically, in our case, the residence time is estimated by dividing the spatial region into overlapping circular windows of constant radius. Then, within each circle, we compute the median length of time a drifter trajectory consecutively remains inside the circle as our estimate of the residence time. We then map onto a spatial grid as with speed (drifters) and expected frequency. Further implementation details specific for the krill analysis and Subarea 48.3 can be found in Section SS1 (Supplementary Material).

- **Mass flux:** the residence time computes the *average time* a drifter continuously spends in a fixed spatial region. A natural orthogonal covariate to also include is the *number* of drifters that pass through this region over time. This can be interpreted as the *mass flux* of drifters as it measures the rate at which drifters move across a unit area per unit of time. The motivation to include this covariate also comes from Murphy et al. (2004), who find associations between water volume flux (which the drifters are mimicking near the surface) and krill flux. For our analysis, mass flux is computed in exactly the same way as residence time, see Section SS1 (Supplementary Material) for details.
- **Density of drifters:** lastly, as the drifters are freely floating then they are not uniformly sampling the ocean and instead are likely to be preferentially sampling the ocean due to the impact of, for example, convergent or divergent zones (Middleton and Garrett, 1986). Although the density of drifters and krill will not necessarily aggregate in the same way, it nonetheless could be informative as a covariate. We therefore include a basic estimate of the density of drifters which corresponds to the total number of hours spent by drifters in each pixel of the spatial image, as detailed in Section SS1 (Supplementary Material).

The five proposed drifter products are plotted in Figure 3 for Subarea 48.3. Additionally, we include a spatial plot of the "speed (satellite)" covariate for comparison, which, as expected, shows related drifter speeds but also reveals some differing structures. While the five drifter products are clearly not entirely orthogonal (e.g., the mass flux is higher in regions of increased speed, as expected), none of them appear to be collinear. Thus, considering the rich information content of the drifter data, with approximately 1.5 million unique observations, we incorporate all these products into our spatio(-temporal) analysis of krill abundance in the next section. However, we emphasise that in other applications, it may be more appropriate to include only a subset of these products. For instance, in Section 4.2, we applied stepwise forward selection and added only residence time and mass flux from the drifters, in addition to other satellite-based covariates. Lastly, it is worth noting that drifters are attached to a drogue (also known as a sea anchor) and measure near-surface currents at approximately 15 meters below the water's surface, whereas krill swarms can occur at greater depths. Consequently, while the relationship between some of the maps in Figure 3 and krill abundance may be significant, it might not be as strong across all locations.

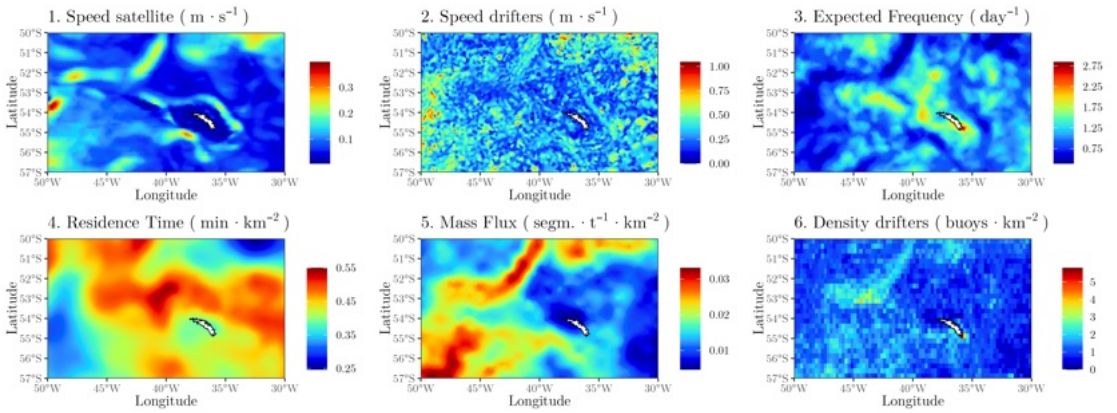

Figure 3: Covariates in South Georgia (Subarea 48.3), as described in Table 1. All drifter-derived products were computed based on the trajectories collected from 1997 to 2020. 1: Speed ($\mathrm{m} \cdot \mathrm{s}^{-1}$) from the satellite, averaged over 1997–2020. 2: Speed ($\mathrm{m} \cdot \mathrm{s}^{-1}$) from the drifters. 3: Expected frequency (day^{-1}). 4: Residence time ($\mathrm{min} \cdot \mathrm{km}^{-2}$). 5: Mass flux ($\mathrm{segments} \cdot \mathrm{t}^{-1} \cdot \mathrm{km}^{-2}$), where "time" refers to the entire observational period, i.e., 24 years. 6: Density of drifters ($\mathrm{buoys} \cdot \mathrm{km}^{-2}$).

4 Results

Following the modelling framework described in Section 3.1 and incorporating covariates from satellite imagery and drifter products (Table 1), we fit a Hurdle-Gamma model for krill biomass from acoustic data (Section 2.1.1) under two settings. First, we apply the model to the disaggregated data at its original spatio-temporal resolutions and focus on the region where we observed the data (green box, Figure 1). In this setting, our primary interest lies in the interpretability of some model parameters. Second, to enhance predictive capability outside the observed window (red box, Figure 1) and in line with the approach of Warwick-Evans et al. (2022), we fit a spatial-only version of our model to the acoustic data aggregated across space and time. This setup also enables variable selection at a feasible computational cost.

4.1 Disaggregated Spatio-temporal Modelling

First, we fit the spatio-temporal Hurdle-Gamma model introduced in Section 3.1, including all covariates listed in Table 1 in the linear predictors for both the presence-absence and positive biomass components—i.e., Equations (1) and (2), respectively. Full details on this model are provided in Section SS2.1 (Supplementary Material). This approach was chosen to avoid the need for multiple model re-fits, as the associated computational cost was prohibitively high, despite the optimised inference specifications detailed in Section 3.1.2.

Table ST1 (Supplementary Material) shows the estimated coefficients, and Figure 4 shows the predicted values for both presence-absence and positive biomass components in 1997 and 2020. The corresponding results for the remaining years are shown in Figures SF2–SF21 (Supplementary Material). In the right-most plots of Figure 4, we masked out predicted values at locations with high uncertainty—specifically, where the standard deviation is greater than 3 (on the log scale, or approximately $20\mathrm{g/m}^2$).

As previously noted, the variability in the disaggregated data makes any spatio-temporal extrapolation beyond the observed locations extremely challenging using the modelling framework from Section 3.1. This also explains our focus on the region delineated by the green box (Figure 1). To address such a limitation and improve our ability to make predictions in non-observed areas, we shift to an aggregated analysis in Section 4.2.

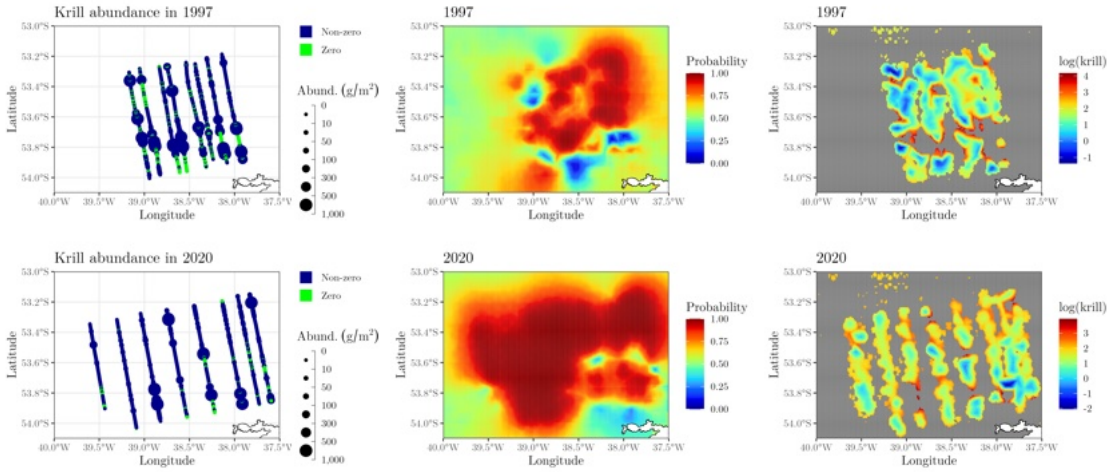

Figure 4: Left column: observed acoustic krill biomass data. Middle column: estimated probability of non-zeros. Right column: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

In addition to the predicted processes shown in Figure 4, we may also be interested in interpreting certain model hyperparameters, particularly those related to the estimated random effects. Figure 5 presents the posterior distributions of key parameters, including the range ρ, as in Equation (4), for the random effects both in the presence-absence linear predictor and in the positive krill biomass linear predictor, i.e., Equations (1) and (2), respectively.

Using the mode of the corresponding posterior distribution as a point estimate for the range, we find that, in the presence-absence component, it is approximately 20 nautical miles (approx. 37 km), whereas in the positive krill biomass component, it is approximately 3.7 nautical miles (approx. 6.8 km). As noted in Section 2.1.1, the transects are positioned at least 10 nautical miles apart to ensure independent samples across different transects. In this context, our estimates could further refine sampling routes for future surveys, as $\rho = \sqrt{8\nu}/\kappa$ indicates the distance at which spatial correlation is close to 0.1 (Cameletti et al., 2013).

However, before using these estimates to guide adjustments in data collection strategies, it is essential to assess their robustness under the Gaussianity assumption for the latent field. In Section SS2.1.1 (Supplementary Material), we conducted a sensitivity analysis by re-estimating the hyperparameters for observations generated from a latent non-Gaussian model. The results suggest that, while a potentially misspecified model may introduce a small bias in the range parameter (in particular, in our experiment, we noticed an upward bias of 5-10% for the parameter κ), the overall conclusions regarding acoustic survey sampling remain unaffected, as these differences are not substantial enough to meaningfully impact

interpretation.

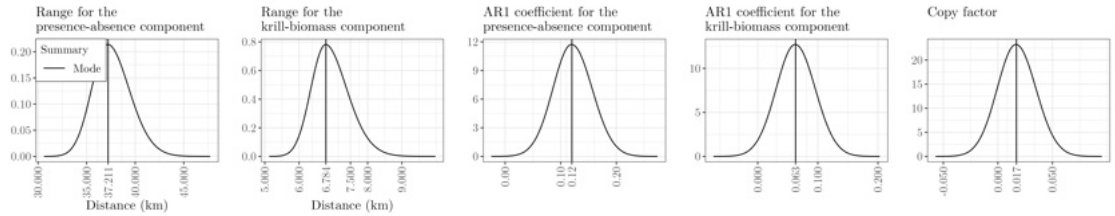

Figure 5: Posterior distributions for the range, AR1 coefficients from Equation (3), and "copy" factor γ from Equation (2).

4.2 Aggregated Spatial Modelling

Following the approach of Warwick-Evans et al. (2022), who estimated krill abundance in the northern Antarctic Peninsula region, we fit a spatial Hurdle-Gamma model (as stated in Section 3.1.1) to an aggregated version of the original acoustic krill data. Specifically, we will aggregate the acoustic data both temporally and spatially into cells measuring 0.067° longitude by 0.036° latitude, corresponding to approximately 4×4 km (as shown in the left column of Figure 6). Modelling this aggregated version of the data reduces issues of high spatial variability over short distances, making the Gaussianity assumption more reasonable and decreasing uncertainty in predictions beyond the observed area. While this approach sacrifices spatio-temporal resolution, it may offer valuable insights into the spatial distribution of krill biomass. In this section, we focus on making predictions within the red box (Figure 1), where there are more observations from KRILLBASE (net haul data), enabling us to compare and evaluate the accuracy of our extrapolated predictions—although, as discussed in Section 2.1, the net haul data do not measure krill biomass in the same way as the acoustic surveys and thus serve only as a proxy for the true spatial distribution.

In this scenario, since the model is computationally much cheaper to fit, we can perform variable selection. Specifically, we will perform stepwise forward selection based on the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2013; Gelman et al., 2014). Additionally, we tested alternative models with simpler random effect structures (also using stepwise forward variable selection). However, the original model, i.e., the spatial Hurdle-Gamma model with linear predictors as in Equations (1) and (2), consistently outperformed these alternatives (Section SS2.2.1, Supplementary Material), reinforcing our choice to use it. Full details on the selected model are provided in Section SS2.2, where Table ST3 (Supplementary Material) shows the estimated coefficients. The covariates included in the linear predictors, in addition to the intercept, were as follows: for the presence-absence component, as in Equation (1), we considered chlorophyll, potential temperature, speed (satellite), surface temperature, mass flux, and residence time. For the positive krill biomass predictor, as in Equation (2), we considered depth, salinity, and surface temperature. Notably, some drifter-derived products improved the model's performance, as indicated by the WAIC—suggesting that *in situ* and remotely sensed data may provide complementary information to the model.

Figure 6 (middle column) shows the predicted krill biomass (log scale) based on the posterior mean. Model estimates were derived from the aggregated data (left column), with predictions based on covariate data from 2020—i.e., the upper limit of the acoustic krill biomass observations. The corresponding prediction uncertainty, represented by the 2.5^{th} and 97.5^{th} quantiles, is shown in Figure SF23 (Supplementary Material). The same figure shows the estimated probabilities (with uncertainty) of observing non-zero krill biomass.

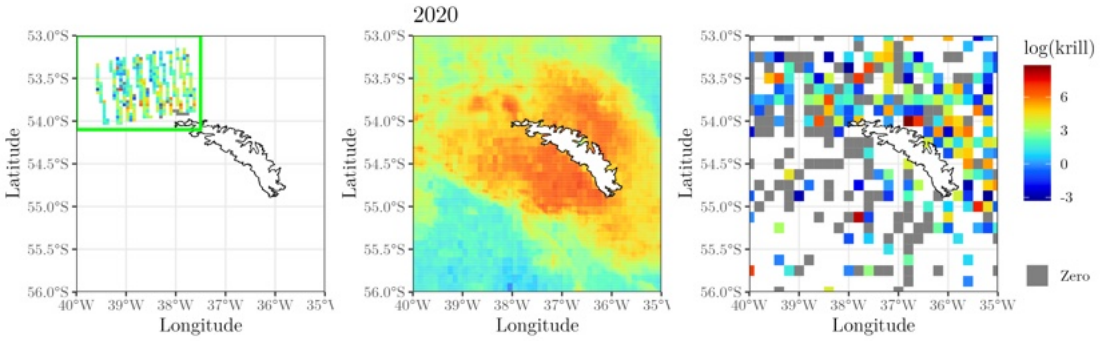

Figure 6: Left column: aggregated acoustic krill biomass data. Middle column: predicted krill biomass in 2020. Right column: KRILLBASE (net haul data), aggregated temporally and spatially. Krill biomass is in g/m^2.

Finally, Figure 6 also allows us to visually compare the spatial distribution of krill biomass extrapolated from the acoustic data (middle column) with the corresponding distribution observed in the net haul data (KRILLBASE, right column). Despite the sparse KRILLBASE coverage, we can still identify hotspot areas, particularly in the north-east region around South Georgia Island, which align with findings in the literature (Schmidt et al., 2016) and are well captured by our model predictions. However, predictions for the south-west region are more challenging to compare with the net data, meaning that the observed symmetry between the south-west and north-east portions of the map in our predictions (likely driven by environmental factors) should be interpreted with caution. In fact, Brierley et al. (1999) demonstrated substantial differences between the eastern and western parts of the South Georgia shelf, indicating possible fine-scale variability. Taken together, these analyses suggest that our model captures several key spatial patterns observed in the data and may, to a certain extent, reasonably extend these patterns into unsampled regions. Although, as seen in Figure 6, the predicted krill biomass lacks the patchy nature observed in the acoustic and net haul data.

5 Discussion

In this paper, we presented a statistical framework for modelling krill abundance in the Southern Ocean, with a specific focus on the South Georgia region. By integrating heterogeneous data sources collected at various spatio(-temporal) resolutions and of different types, we tackled key challenges commonly encountered in ecological modelling, such as misaligned datasets and zero-inflated observations.

These data sources included acoustic observations of krill biomass, net haul data used for validation, remotely sensed satellite imagery, and drifter-derived covariates. Our approach demonstrated the benefits of combining multiple data sources to enhance both interpretability—evidenced by insights gained from hyperparameter estimation to inform sampling strategies (see Section 4.1)—and predictive accuracy, particularly in the aggregated spatial analysis (see Section 4.2).

More broadly, integrating remotely sensed data, such as satellite imagery, with timestamped trajectories from drifters offers a powerful approach to modelling marine ecosystems. While satellite imagery provides an overview of environmental conditions, drifter-derived products offer complementary insights into the physical and biological factors influencing the target distribution. These data sources can deliver potentially orthogonal information, even when describing the same phenomenon. In such cases, drifter-derived products can also function as a calibration data source for remotely sensed observations (Villejo et al., 2024).

The findings of this work may contribute to the development of more effective conservation and management strategies for krill in the Southern Ocean. As highlighted by Warwick-Evans et al. (2022), identifying regions with high krill density enables the determination of areas where krill fishing would have the least ecological impact. Moreover, an important consideration when making decisions based on model estimates is the need to account for uncertainty. Misinterpreting or neglecting uncertainty can lead to overconfidence in predictions and potentially harmful management outcomes. By providing credible intervals for key estimates, our framework enables managers to make informed decisions supported by the data while accounting for the inherent variability of ecological systems and the limitations of the modelling process.

Finally, while our work demonstrates the potential of integrating multiple data sources, there are notable limitations. First, relying on acoustic surveys from a smaller region within Subarea 48.3 (green box in Figure 1) makes it challenging to extrapolate predictions beyond the observed window with reasonable uncertainty, particularly for the non-aggregated analysis (Section 4.1). Second, for the non-aggregated analysis, the assumption of Gaussianity for the underlying random effects in the Hurdle-Gamma model may lead to oversmoothed prediction maps. While this approach reduces computational burden during inference, exploring latent non-Gaussian models (e.g., Cabral et al. (2024)) in future work could offer greater flexibility and more effectively capture abrupt variations in the target ecological variable, even over short distances. Third, although we used net haul data (KRILLBASE) as a validation source for our extrapolated predictions, its spatial and temporal sparsity limits the reliability of conclusions about krill abundance at fine scales. Consequently, comparisons such as those presented in Figure 6 should be interpreted with caution, as noted in Section 4.2. Overall, additional extensions could further improve our model. An alternative approach is to explicitly account for spatially varying preferential sampling in the acoustic observations, combining this with data fusion techniques (Amaral et al., 2024; Zhong et al., 2024), as sampling locations are often influenced by ecological factors, potentially introducing bias into the analysis. Additionally, incorporating other *in situ* data sources, such as profiling floats (Roemmich et al., 2009), could provide complementary oceanographic measurements and refine the modelling of krill abundance.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank the Turner-Kirk Trust for supporting this research. We also thank Jeffrey Early for helpful advice in forming drifter-derived spatial products. E. C. was supported by a Natural Environment Research Council (NERC) grant NE/Y004515/1 and a WWF research grant (GB085708). S. F. and data collection for the Western Core Box acoustic survey were funded by the Natural Environment Research Council (NERC).

References

Amaral, A. V. R., Krainski, E. T., Zhong, R. and Moraga, P. (2024). Model-based geostatistics under spatially varying preferential sampling. Journal of Agricultural, Biological and Environmental Statistics *29*, 766–792.

Atkinson, A., Hill, S. L., Pakhomov, E. A., Siegel, V., Anadon, R., Chiba, S., Daly, K. L., Downie, R., Fielding, S., Fretwell, P., Gerrish, L., Hosie, G. W., Jessopp, M. J., Kawaguchi, S., Krafft, B. A., Loeb, V., Nishikawa, J., Peat, H. J., Reiss, C. S., Ross, R. M., Quetin, L. B., Schmidt, K., Steinberg, D. K., Subramaniam, R. C., Tarling, G. A. and Ward, P. (2017). KRILLBASE: a circumpolar database of Antarctic krill and salp numerical densities, 1926–2016. Earth System Science Data *9*, 193–210.

Atkinson, A., Siegel, V., Pakhomov, E. A., Jessopp, M. J. and Loeb, V. (2009). A re-appraisal of the total biomass and annual production of Antarctic krill. Deep Sea Research Part I: Oceanographic Research Papers *56*, 727–740.

Bar-On, Y. M., Phillips, R. and Milo, R. (2018). The biomass distribution on Earth. Proceedings of the National Academy of Sciences *115*, 6506–6511.

Brierley, A. S., Watkins, J. L., Goss, C., Wilkinson, M. T. and Everson, I. (1999). Acoustic estimates of krill density at South Georgia, 1981 to 1998. CCAMLR Science *6*, 47–57.

Cabral, R., Bolin, D. and Rue, H. (2024). Fitting latent non-Gaussian models using variational Bayes and Laplace approximations. Journal of the American Statistical Association.

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. AStA Advances in Statistical Analysis *97*, 109–131.

Cavan, E. L., Belcher, A., Atkinson, A., Hill, S. L., Kawaguchi, S., McCormack, S., Meyer, B., Nicol, S., Ratnarajah, L., Schmidt, K., Steinberg, D. K., Tarling, G. A. and Boyd, P. W. (2019). The importance of Antarctic krill in biogeochemical cycles. Nature communications *10*, 4742.

Cavan, E. L., Mackay, N., Hill, S. L., Atkinson, A., Belcher, A. and Visser, A. (2024). Antarctic krill sequester similar amounts of carbon to key coastal blue carbon habitats. Nature Communications *15*, 7842.

Commission for the Conservation of Antarctic Marine Living Resources (2015). Subarea 48.3. https://www.ccamlr.org/en/node/74757.

Copernicus Programme of the European Union (2024). Copernicus Marine Service. https://data.marine.copernicus.eu/products.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica: journal of the Econometric Society *39*, 829–844.

Elipot, S., Lumpkin, R., Perez, R. C., Lilly, J. M., Early, J. J. and Sykulski, A. M. (2016). A global surface drifter data set at hourly resolution. Journal of Geophysical Research: Oceans *121*, 2937–2966.

Fielding, S., Watkins, J. L., Trathan, P. N., Enderlein, P., Waluda, C. M., Stowasser, G., Tarling, G. A. and Murphy, E. J. (2014). Interannual variability in Antarctic krill (Euphausia superba) density at South Georgia, Southern Ocean: 1997–2013. ICES Journal of Marine Science *71*, 2578–2588.

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. Statistics and computing *24*, 997–1016.

Krainski, E., Gómez-Rubio, B., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. (2018). Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC.

LaCasce, J. H. (2008). Statistics from Lagrangian observations. Progress in Oceanography *77*, 1–29.

Laso-Jadart, R., O'Malley, M., Sykulski, A. M., Ambroise, C. and Madoui, M. A. (2023). Holistic view of the seascape dynamics and environment impact on macro-scale genetic connectivity of marine plankton populations. BMC Ecology and Evolution *23*, 46.

Lilly, J. M., Sykulski, A. M., Early, J. J. and Olhede, S. C. (2017). Fractional Brownian motion, the Matérn process, and stochastic modeling of turbulent dispersion. Nonlinear Processes in Geophysics *24*, 481–514.

Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with `R-INLA`. Journal of Statistical Software *63*, 1–25.

Lindgren, F., Rue, H. and Lindström, H. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *73*, 423–498.

McCormack, S. A., Melbourne-Thomas, J., Trebilco, R., Blanchard, J. L., Raymond, B. and Constable, A. (2021). Decades of dietary data demonstrate regional food web structures in the Southern Ocean. Ecology and Evolution *11*, 227–241.

Middleton, J. F. and Garrett, C. (1986). A kinematic analysis of polarized eddy fields using drifter data. Journal of Geophysical Research: Oceans *91*, 5094–5102.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. Journal of the Iranian Statistical Society *1*, 7–33.

Moraga, P. (2019). Geospatial health data: Modeling and visualization with R-INLA and shiny. Chapman and Hall/CRC.

Murphy, E. J., Watkins, J. L., Meredith, M. P., Ward, P., Trathan, P. N. and Thorpe, S. E. (2004).

Southern Antarctic Circumpolar Current Front to the northeast of South Georgia: horizontal advection of krill and its role in the ecosystem. Journal of Geophysical Research: Oceans *109*.

Nauman, E. B. (2008). Residence time theory. Industrial & Engineering Chemistry Research *47*, 3752–3766.

Nicol, S., Foster, J. and Kawaguchi, S. (2012). The fishery for Antarctic krill–recent developments. Fish and Fisheries *13*, 30–40.

Oscroft, S., Sykulski, A. M. and Early, J. J. (2020). Separating mesoscale and submesoscale flows from clustered drifter trajectories. Fluids *6*, 14.

O'Malley, M., Sykulski, A. M., Laso-Jadart, R. and Madoui, M. A. (2021). Estimating the travel time and the most likely path from Lagrangian drifters. Journal of Atmospheric and Oceanic Technology *38*, 1059–1073.

O'Malley, M., Sykulski, A. M., Lumpkin, R. and Schuler, A. (2023). Probabilistic Prediction of Oceanographic Velocities with Multivariate Gaussian Natural Gradient Boosting. Environmental Data Science *2*, e10.

Percival, D. B. and Walden, A. T. (1993). Spectral analysis for physical applications. Cambridge University Press.

Roemmich, D., Johnson, G. C., Riser, S., Davis, R., Gilson, J., Owens, W. B., Garzoli, S. L., Schmid, C. and Ignaszewski, M. (2009). The Argo Program: Observing the global ocean with profiling floats. Oceanography *22*, 34–43.

Ruck, K. E., Steinberg, D. K. and Canuel, E. A. (2014). Regional differences in quality of krill and fish as prey along the Western Antarctic Peninsula. Marine Ecology Progress Series *509*, 39–55.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) *71*, 319–392.

Schmidt, K., Schlosser, C., Atkinson, A., Fielding, S., Venables, H. J., Waluda, C. M. and Achterberg, R. P. (2016). Zooplankton gut passage mobilizes lithogenic iron for ocean productivity. Current Biology *26*, 2667–2673.

Siegel, V., Reiss, C. S., Dietrich, K. S., Haraldsson, M. and Rohardt, G. (2013). Distribution and abundance of Antarctic krill (Euphausia superba) along the Antarctic Peninsula. Deep sea research part I: oceanographic research papers *77*, 63–74.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science *32*, 1–28.

Sykulski, A. M., Olhede, S. C., Lilly, J. M. and Danioux, E. (2016). Lagrangian time series models for ocean surface drifter trajectories. Journal of the Royal Statistical Society Series C: Applied Statistics *65*, 29–50.

Sykulski, A. M., Olhede, S. C., Lilly, J. M. and Early, J. J. (2017). Frequency-domain stochastic modeling of stationary bivariate or complex-valued signals. IEEE Transactions on Signal Processing *65*, 3136–3151.

Villejo, S. J., Martino, S., Lindgren, F. and Illian, J. (2024). A Data Fusion Model for Meteorological Data using the INLA-SPDE method. arXiv:2404.08533.

Warwick-Evans, V., Fielding, S., Reiss, C. S., Watters, G. M. and Trathan, P. N. (2022). Estimating the average distribution of Antarctic krill Euphausia superba at the northern Antarctic Peninsula during austral summer and winter. Polar Biology *45*, 857–871.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. The Journal of Machine Learning Research *14*, 867–897.

Whitehouse, M. J., Atkinson, A., Ward, P., Korb, R. E., Rothery, P. and Fielding, S. (2009). Role of krill versus bottom-up factors in controlling phytoplankton biomass in the northern Antarctic waters of South Georgia. Marine Ecology Progress Series *393*, 69–82.

Whittle, P. (1963). Stochastic-processes in several dimensions. Bulletin of the International Statistical Institute *40*, 974–994.

Yang, G., Atkinson, A., Pakhomov, E. A., Hill, S. L. and Racault, M. F. (2022). Massive circumpolar biomass of Southern Ocean zooplankton: Implications for food web structure, carbon export, and marine spatial planning. Limnology and Oceanography *67*, 2516–2530.

Zhong, R., Ribeiro Amaral, A. V. and Moraga, P. (2024). Spatial data fusion adjusting for preferential sampling using integrated nested Laplace approximation and stochastic partial differential equation. Journal of the Royal Statistical Society Series A: Statistics in Society *0*, 1–18.

Supplementary Material for "Navigating Challenges in Spatio-temporal Modelling of Antarctic Krill Abundance: Addressing Zero-inflated Data and Misaligned Covariates"

André Victor Ribeiro Amaral[†,*], Adam M. Sykulski[†], Emma Cavan[†], Sophie Fielding[‡]

[†]Imperial College London. London, UK.

[‡]British Antarctic Survey (BAS). Cambridge, UK.

[*]Corresponding author. E-mail: a.ribeiro-amaral@imperial.ac.uk

arXiv:2412.01399v1 [stat.AP] 2 Dec 2024

SS1 Products from Drifter Trajectories

In this section, we specify the details for computing the drifter-derived products—as described in Section Section 3.2.

- "Speed (drifters):" as stated in Table 1, this quantity was computed based on drifter trajectories observed during the months of December, January, and February, so that it is comparable with the speed estimates obtained from the satellite imagery.
- "Expected frequency:" it was computed over rolling temporal windows of 121 hours in length, with a 60-hour overlap between consecutive segments. The location of a segment is defined by its mid-point.
- "Residence time" and "Mass flux:" in both cases, the spatial windows were defined by circles of radius 50 km, with centres at 20×25 points distributed equidistantly over the study area. The location of a circle is defined by its centre point. Figure SF1 shows these circles. To avoid border effects, the estimates in the circles at the edges were corrected by a multiplicative factor given by the ratio between the area of a full circle and the area of the clipped circle.
- "Density of drifters:" the computation is straightforward and corresponds to the total number of hours spent by drifters in a certain spatial area, specifically a 0.25° longitude by 0.25° latitude cell, divided by the area of this cell.

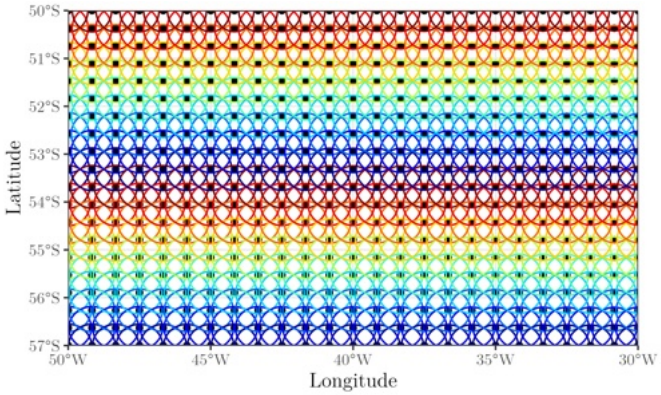

Figure SF1: Spatial windows (circles of radius 50 km) used to compute "residence time" and "mass flux." The colours are only used to ease the visualisation.

SS1.1 Interpolation

After computing all drifter statistics (except for the "density of drifters") as described in Section 3.2, we interpolate these newly generated data points across the entire study area shown in Figure 1 (South Georgia, Subarea 48.3) using the following approach.

Let $y(s) = (y(s_1), \cdots, y(s_n))$ represent the drifter product observed at locations $(s_1, \cdots, s_n) \subset \mathcal{X} \subset \mathbb{R}^2$, where $\mathcal{X}$ is the spatial domain. To spatially interpolate these observations across $\mathcal{X}$, we model

$y(s)$ as follows

$$
\begin{aligned}
y(s) &= \beta_0 + \phi(s) + \epsilon(s), \text{ s.t. } \epsilon(s) \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma_\epsilon^2) \\
\phi(s) &\sim \text{Gaussian Process}(0, r_\phi(h;\theta)) \\
(\beta_0, \sigma_\epsilon^2, \theta) &\sim \text{priors},
\end{aligned}
$$

where $r_\phi(h;\theta)$ is the Matérn kernel, such that $\theta = (\nu, \kappa, \sigma_\phi)$. This model is fitted using `R-INLA` (Lindgren and Rue, 2015) (see Section 3.1.2).

SS2 Additional Results

In this section, we present additional analyses to complement the results from Section 4, including details on both disaggregated spatio-temporal modelling and aggregated spatial modelling.

SS2.1 Disaggregated Spatio-temporal Modelling

In Section 4.1, we implemented a Hurdle-Gamma model (as stated in Section 3.1) with linear predictors specified as follows

$$
\begin{aligned}
\text{logit}(\pi_{it}) = \beta_0^z + \beta_1^z \texttt{month} + \beta_2^z \texttt{depth} + \beta_3^z \texttt{slope} + \\
\beta_4^z \texttt{chlor} + \beta_5^z \texttt{pot_temp} + \beta_6^z \texttt{salinity} + \beta_7^z \texttt{speed_sat} + \beta_8^z \texttt{surf_temp} + \\
\beta_9^z \texttt{speed_drif} + \beta_{10}^z \texttt{expect_freq} + \beta_{11}^z \texttt{res_time} + \beta_{12}^z \texttt{mass_flux} + \beta_{13}^z \texttt{density_drif} + \\
\psi_{it},
\end{aligned} \tag{1}
$$

and

$$
\begin{aligned}
\log(\mu_{it}) = \beta_0^y + \beta_1^z \texttt{month} + \beta_2^y \texttt{depth} + \beta_3^y \texttt{slope} + \\
\beta_4^y \texttt{chlor} + \beta_5^y \texttt{pot_temp} + \beta_6^y \texttt{salinity} + \beta_7^y \texttt{speed_sat} + \beta_8^y \texttt{surf_temp} + \\
\beta_9^y \texttt{speed_drif} + \beta_{10}^y \texttt{expect_freq} + \beta_{11}^y \texttt{res_time} + \beta_{12}^y \texttt{mass_flux} + \beta_{13}^y \texttt{density_drif} + \\
\gamma \cdot \psi_{it} + \xi_{it},
\end{aligned} \tag{2}
$$

where γ is a "copy" factor, and ψ_{it} (similarly, ξ_{it}) is spatio-temporal random effect, such that

$$
\psi_{it} = \alpha_\psi \psi_{i,(t-1)} + \phi_{it},
$$

where $|\alpha| < 1$, $\psi_{i1} \sim \text{Normal}(0, \sigma_{\phi,\psi}^2/(1-\alpha^2))$, and ϕ_{it} is a temporally independent but spatially dependent Gaussian process at each time point, with a Matérn covariance structure with range $\rho_{\phi,\psi}$ and marginal variance $\sigma_{\phi,\psi}^2$.

In Equations (1) and (2), all covariates are defined as in Table 1, and `month` refers to the month in which the krill acoustic data were collected each year, with possible values being December, January,

and February (where December data are considered observations for the following year). For numerical stability, all covariates were scaled to have a mean of zero and a variance of one when fitting the model. The estimated coefficients are shown in Table ST1.

Table ST1: Estimated parameters (with standard deviation and a 95% equal-tail credible interval) for the spatio-temporal model fitted for the disaggregated data.

Parameter	Mean	SD	95% equal-tail CI	Parameter	Mean	SD	95% equal-tail CI	Parameter	Mean	SD	95% equal-tail CI
β_0^z	0.928	0.284	(0.372; 1.485)	β_0^y	2.397	0.063	(2.273; 2.521)	$\rho_{\phi,\psi}$	37.663	1.965	(34.115; 41.843)
β_1^z	−0.104	0.325	(−0.741; 0.534)	β_1^y	−0.287	0.082	(−0.448; −0.126)	$\sigma_{\phi,\psi}$	2.974	0.106	(2.774; 3.193)
β_2^z	0.176	0.170	(−0.156; 0.508)	β_2^y	0.142	0.081	(−0.016; −0.301)	α_ψ	0.122	0.035	(0.054; 0.192)
β_3^z	0.009	0.033	(−0.056; 0.074)	β_3^y	0.035	0.019	(−0.001; 0.072)	$\rho_{\phi,\xi}$	6.954	0.547	(6.001; 8.151)
β_4^z	−0.070	0.098	(−0.263; 0.123)	β_4^y	0.009	0.057	(−0.102; 0.120)	$\sigma_{\phi,\xi}$	2.295	0.086	(2.122; 2.462)
β_5^z	0.209	0.162	(−0.108; 0.527)	β_5^y	−0.308	0.061	(−0.429; −0.188)	α_ξ	0.064	0.032	(0.000; 0.128)
β_6^z	−0.037	0.158	(−0.346; 0.273)	β_6^y	−0.279	0.076	(−0.428; −0.130)	γ	0.017	0.018	(−0.017; 0.053)
β_7^z	0.172	0.115	(−0.052; 0.397)	β_7^y	0.021	0.052	(−0.081; 0.123)				
β_8^z	−0.131	0.096	(−0.319; 0.056)	β_8^y	0.068	0.050	(−0.030; 0.165)				
β_9^z	−0.019	0.064	(−0.145; 0.107)	β_9^y	−0.059	0.042	(−0.142; 0.024)				
β_{10}^z	0.280	0.185	(−0.082; 0.642)	β_{10}^y	−0.020	0.080	(−0.177; 0.138)				
β_{11}^z	0.152	0.223	(−0.284; 0.589)	β_{11}^y	0.102	0.094	(−0.082; 0.286)				
β_{12}^z	−0.106	0.211	(−0.520; 0.307)	β_{12}^y	−0.050	0.083	(−0.213; 0.112)				
β_{13}^z	0.000	0.064	(−0.125; 0.126)	β_{13}^y	−0.055	0.033	(−0.121; 0.010)				

Lastly, Figures SF2–SF21 display the acoustic surveys and predicted values for both the presence-absence and positive biomass components for the years not shown in Section 4.1.

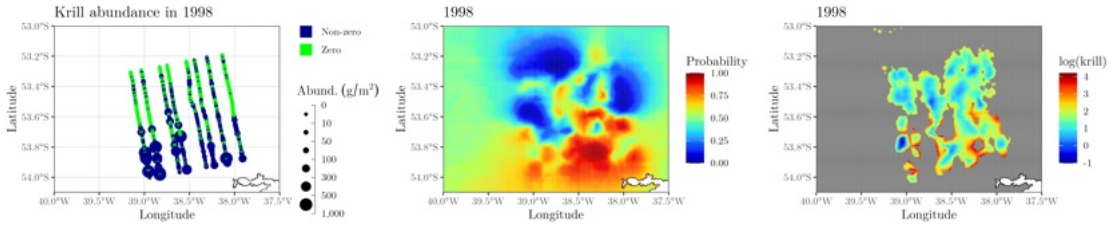

Figure SF2: Year 1998. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

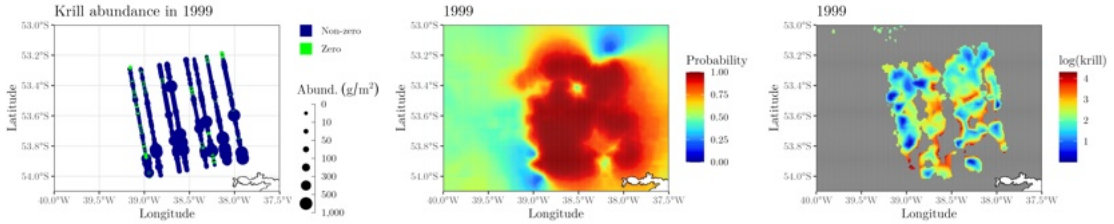

Figure SF3: Year 1999. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

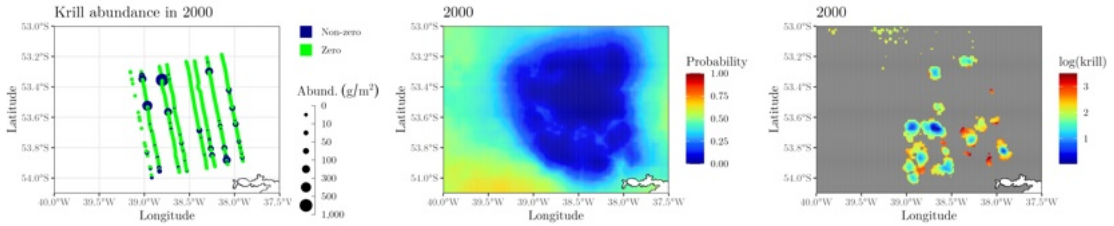

Figure SF4: Year 2000. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

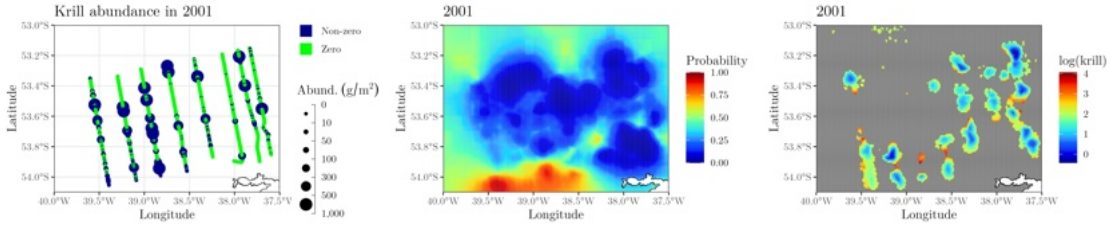

Figure SF5: Year 2001. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

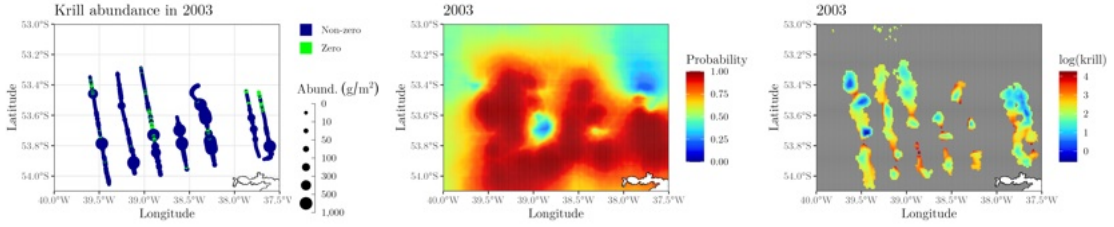

Figure SF6: Year 2003. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

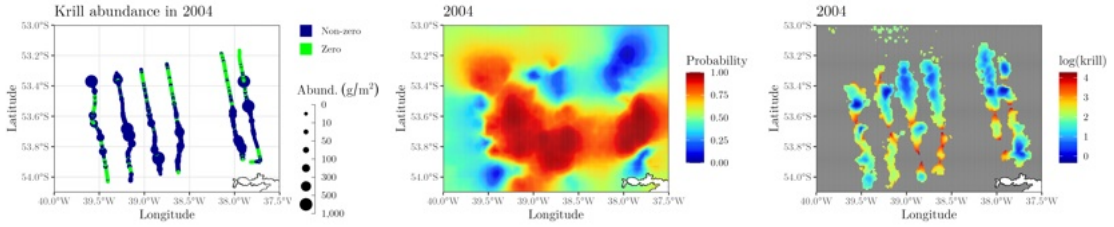

Figure SF7: Year 2004. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

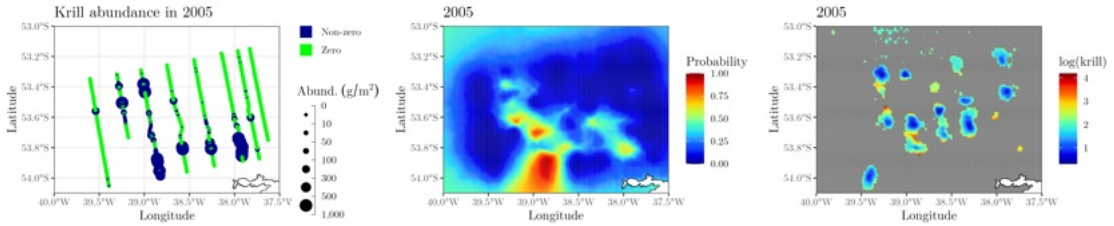

Figure SF8: Year 2005. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

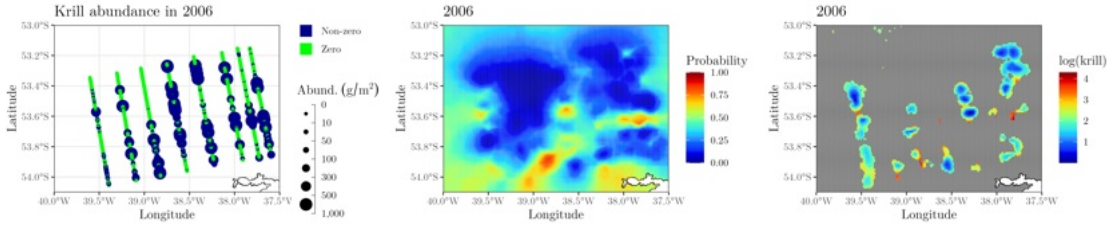

Figure SF9: Year 2006. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

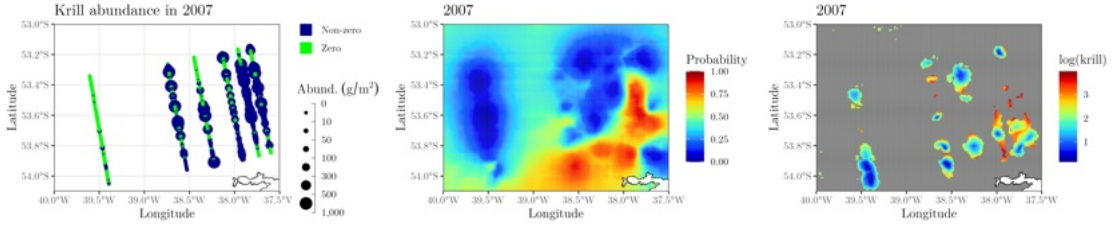

Figure SF10: Year 2007. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

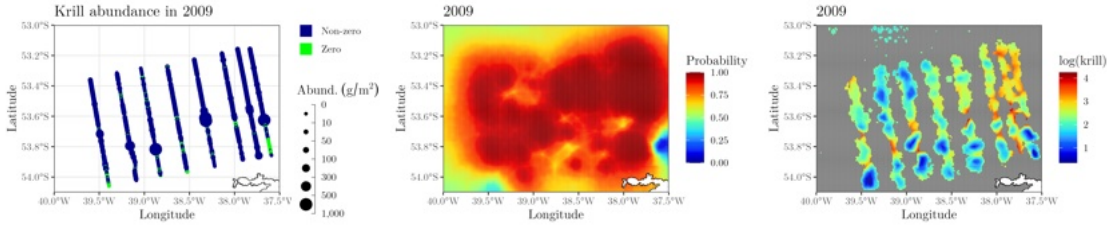

Figure SF11: Year 2009. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

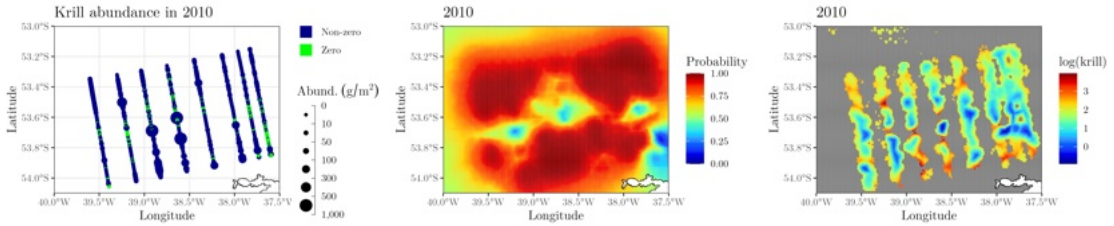

Figure SF12: Year 2010. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

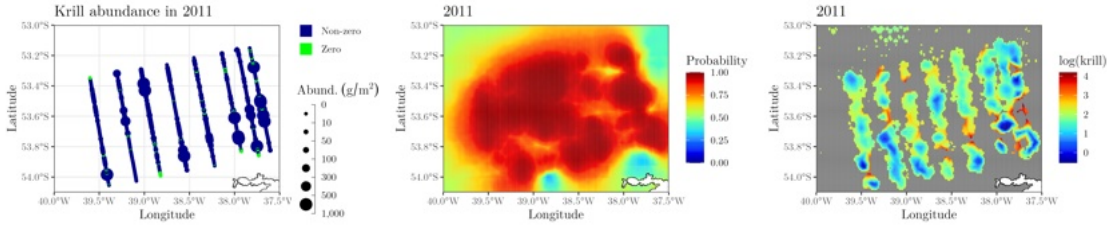

Figure SF13: Year 2011. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

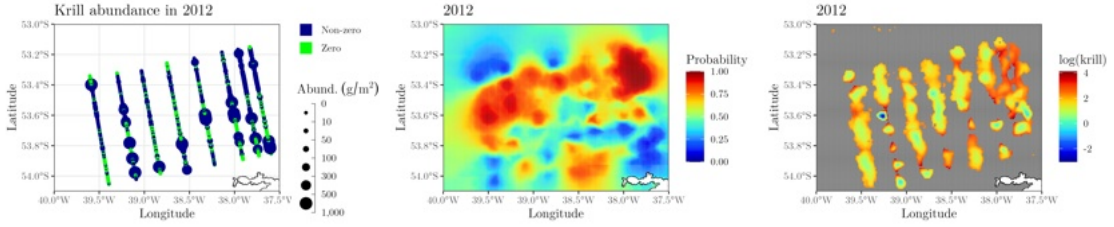

Figure SF14: Year 2012. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

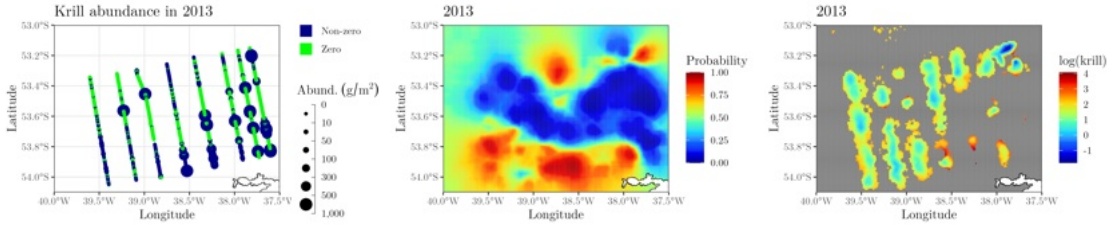

Figure SF15: Year 2013. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

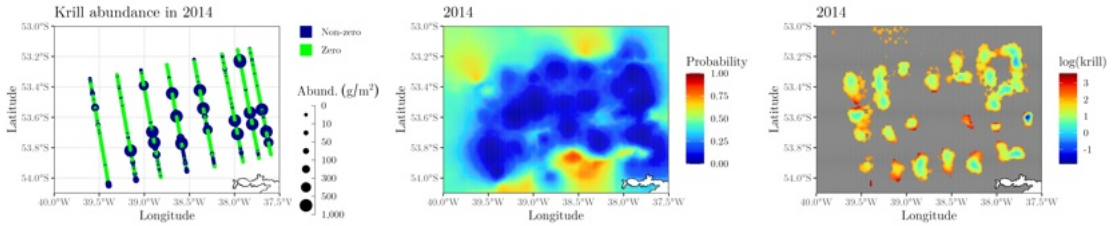

Figure SF16: Year 2014. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

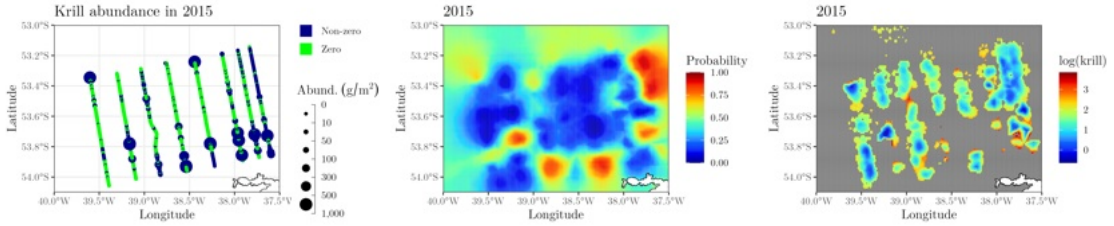

Figure SF17: Year 2015. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

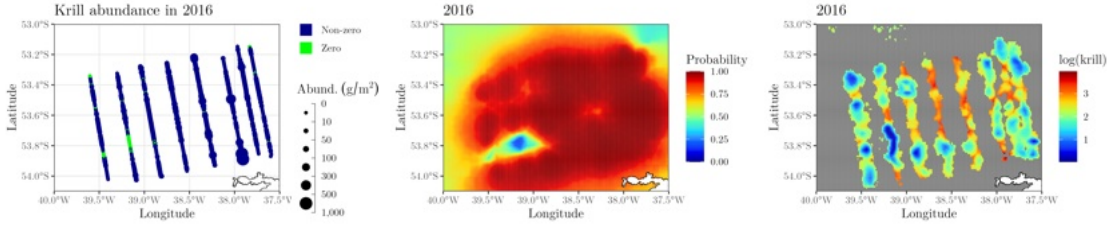

Figure SF18: Year 2016. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

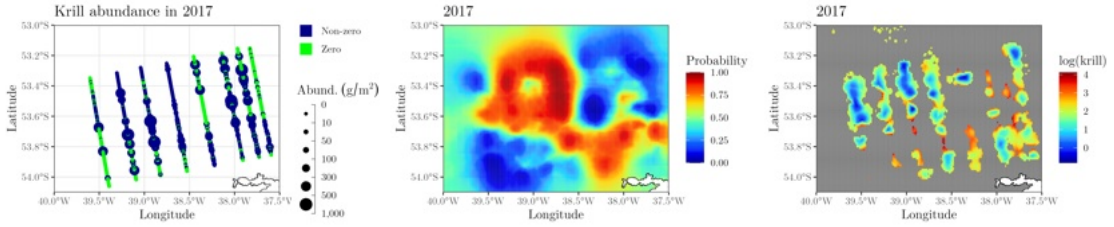

Figure SF19: Year 2017. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

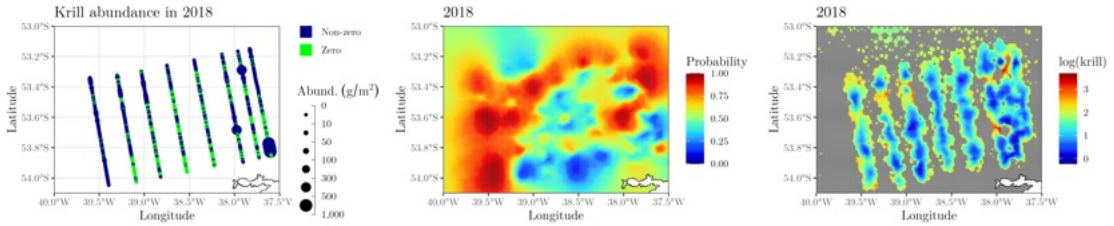

Figure SF20: Year 2018. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

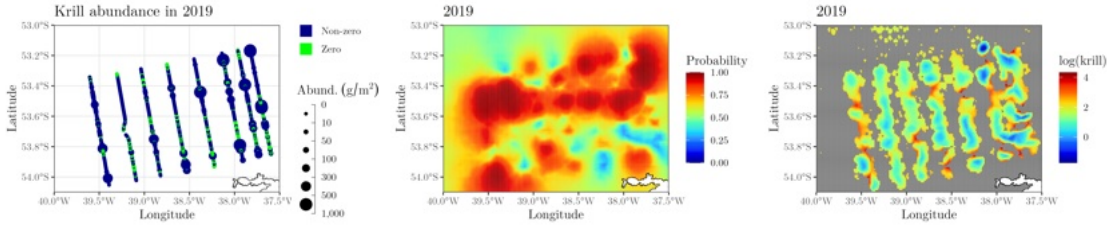

Figure SF21: Year 2019. Left: observed acoustic krill biomass data. Middle: estimated probability of non-zeros. Right: predicted krill biomass (g/m^2), with predictions having a standard deviation greater than 3 (on the log scale) being masked out.

SS2.1.1 Sensitivity analysis

To evaluate the robustness of our hyperparameter estimates under a misspecified model, we will simulate data from a Matérn model with normal inverse Gaussian (NIG) noise and attempt to retrieve key quantities, particularly $\kappa = \sqrt{8\nu}/\rho$, where ρ denotes the range.

The NIG distribution is a continuous probability distribution belonging to the generalized hyperbolic family (Barndorff-Nielsen, 1978). It is particularly useful due to its flexibility in modelling asymmetry and heavy tails. For a random variable X following an NIG distribution with parameters $\nu,\ \sigma > 0$ and $\delta,\ \mu \in \mathbb{R}$, the probability density function is given by

$$f(x;\delta,\mu,\sigma,\nu) = \frac{e^{\nu+\mu(x-\delta)/\sigma^2}}{\pi\sqrt{\nu\sigma^2+(x-\delta)^2}}\sqrt{\nu\mu^2/\sigma^2+\nu^2}K_1\left(\sqrt{(\nu\sigma^2+(x-\delta)^2)\,(\mu^2/\sigma^4+\nu/\sigma^2)}\right), \quad (3)$$

where $K_p(\cdot)$ is modified Bessel function of the second kind of order p. This parametrisation is the same as the one used in the `ngme2` package (Bolin et al., 2024), where δ, μ, σ, and ν represent the parameters for location, skewness, scaling, and shape, respectively.

In this section, we simulate data from a Matérn model with NIG noise in the unit square, i.e., $\mathcal{X} = [0,1] \times [0,1]$, with m replicates. Specifically, we consider the following model. Let $y(s) =$

$(y(s_1), \cdots, y(s_n))$ represent the data set observed at locations $s = (s_1, \cdots, s_n) \subset \mathcal{X}$, then

$$
\begin{aligned}
y(s) &= \beta_0 + \phi(s) + \epsilon(s), \text{ s.t. } \epsilon(s) \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma_\epsilon^2) \\
\phi(s)|\Lambda(s) &= \delta + \mu\Lambda(s) + \sigma\sqrt{\Lambda(s)}\xi(s) \\
\Lambda(s) &\overset{\text{i.i.d.}}{\sim} \text{Inverse Gaussian}(\nu, \nu),
\end{aligned} \tag{4}
$$

where $\xi(s)$ is Gaussian Process with mean 0 and a Matérn kernel given by $r(h;\theta)$, such that $\theta = (\nu_\phi, \kappa_\phi)$.

In particular, we set $\beta_0 = 0$, $\nu_\phi = 1$, $\kappa_\phi = 10$, $\sigma = 1$, $\delta = 0$, $\mu = 0$, $\nu = 10$, and $\sigma_\epsilon = 0.01$ (i.e., for simplicity, we assume the observational noise to be very small), with $m = 10$ replicates. Figure SF22 shows all the simulated surfaces $\phi(s)$.

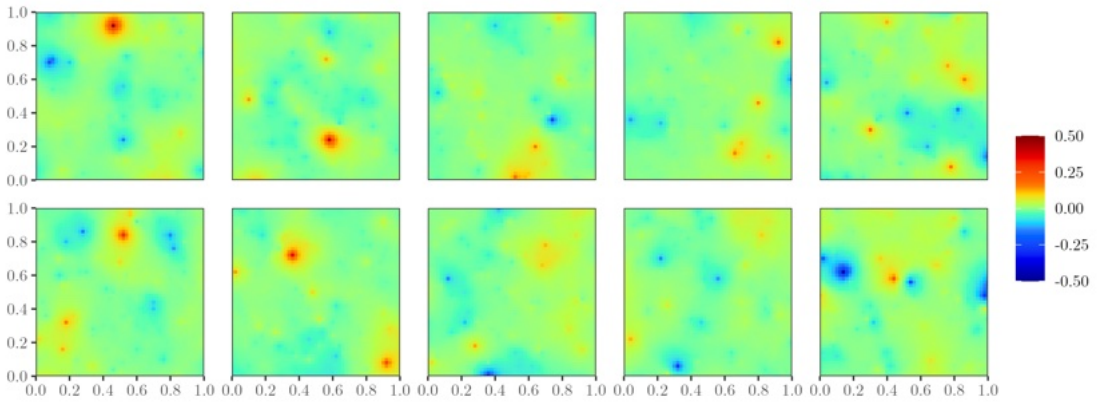

Figure SF22: 10 simulations from a process defined in $[0,1] \times [0,1]$ by a Matérn kernel and NIG noise.

Based on the realizations shown in Figure SF22, we will randomly sample the observed area at $n = 50$, 100, 500, and 1,000 locations and then estimate the Matérn hyperparameters using the correctly specified model (i.e., NIG noise) and a misspecified model with Gaussian noise. Table ST2 presents the obtained estimates for all combinations of fitted models and sample sizes. Inference was made using the `ngme2` package (Bolin et al., 2024).

Table ST2: Estimated hyperparameters for the Matérn structure based on data generated as described in Equation (4) for a correctly specified model (NIG) and a misspecified model (Gaussian). The parameter estimates are based on the mean and standard deviation of the final 2,000 samples, obtained after 10,000 iterations with a burn-in of 8,000. Note that ν_ϕ is not estimated but fixed.

Model	Sample size	Parameters			Model	Sample size	Parameters		
			True	Estimated				True	Estimated
NIG	50	σ	1	2.15 (0.96)	Gaussian	50	σ	1	1.28 (0.20)
		κ_ϕ	10	10.58 (2.70)			κ_ϕ	10	12.36 (1.93)
	100	σ	1	1.03 (0.34)		100	σ	1	1.04 (0.17)
		κ_ϕ	10	9.56 (1.77)			κ_ϕ	10	11.02 (1.78)
	500	σ	1	1.01 (0.17)		500	σ	1	1.07 (0.08)
		κ_ϕ	10	10.21 (0.78)			κ_ϕ	10	11.08 (1.03)
	1,000	σ	1	1.09 (0.20)		1,000	σ	1	1.00 (0.05)
		κ_ϕ	10	9.71 (0.66)			κ_ϕ	10	10.30 (0.78)

From Table ST2, we observe that in the NIG model, σ is more challenging to estimate, particularly for smaller sample sizes (e.g., $n = 50$), which is a known issue due to the parametrisation in Equation (3). On the other hand, κ_ϕ appears to be reasonably well-estimated. In contrast, when analysing the Gaussian model, although the scaling parameter is well-recovered, κ_σ is slightly overestimated (by 5–10%, when $n \geq 100$). Therefore, in a misspecified setting where the observed process exhibits peaks and valleys, the range parameter $\rho_\phi = \sqrt{8\nu_\phi}/\kappa_\phi$ may be underestimated—although the bias is expected to remain modest if the distribution tails of the underlying data are not excessively heavy.

SS2.2 Aggregated Spatial Modelling

In Section 4.2, we implemented a spatial Hurdle-Gamma model (as stated in Section 3.1.1) with linear predictors specified as follows

$$\text{logit}(\pi_i) = \beta_0^z + \beta_1^z \texttt{chlor} + \beta_2^z \texttt{pot_temp} + \beta_3^z \texttt{speed_sat} + \beta_4^z \texttt{surf_temp} + \beta_5^z \texttt{res_time} + \beta_6^z \texttt{mass_flux} + \psi_i, \tag{5}$$

and

$$\log(\mu_i) = \beta_0^y + \beta_1^y \texttt{depth} + \beta_2^y \texttt{salinity} + \beta_3^y \texttt{surf_temp} + \gamma \cdot \psi_i + \xi_i, \tag{6}$$

where γ is a "copy" factor, and ψ_i (similarly, ξ_i) is a spatial random effect modelled as a Gaussian process with a Matérn covariance structure, characterized by the range parameter ρ_ψ and marginal variance σ_ψ^2.

As before, for numerical stability, all covariates were re-scaled. The estimated coefficients are shows in Table ST3. Figure SF23 shows the predicted krill biomass for 2020 along with the associated prediction uncertainty, represented by the 2.5th and 97.5th quantiles. The same figure also shows the probabilities of observing non-zero krill biomass.

Table ST3: Estimated parameters (with standard deviation and a 95% equal-tail credible interval) for the spatial model fitted for the aggregated data.

Parameter	Mean	SD	95% equal-tail CI	Parameter	Mean	SD	95% equal-tail CI	Parameter	Mean	SD	95% equal-tail CI
β_0^z	6.035	1.673	(2.755; 9.315)	β_0^y	3.265	0.249	(2.777; 3.754)	ρ_ψ	50.043	28.115	(13.313; 120.340)
β_1^z	−1.323	0.800	(−2.891; 0.246)	β_1^y	1.502	0.336	(0.845; 2.160)	σ_ψ	3.788	0.943	(2.148; 5.821)
β_2^z	−1.887	0.532	(−2.929; −0.846)	β_2^y	0.681	0.350	(−0.005; 1.367)	ρ_ξ	12.786	5.832	(5.398; 27.757)
β_3^z	−2.114	0.838	(−3.756; −0.472)	β_3^y	0.590	0.094	(0.405; 0.774)	σ_ξ	1.045	0.815	(0.163; 3.177)
β_4^z	1.138	0.458	(0.240, 2.036)					γ	−0.144	0.059	(−0.264; −0.031)
β_5^z	0.430	1.071	(−1.668, 2.529)								
β_6^z	−0.653	1.114	(−2.836, 1.530)								

SS2.2.1 Alternative Random Effects

Alternatively, we tested different random effect structures, performing stepwise forward variable selection for each class of models based on the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2013; Gelman et al., 2014). In particular, we considered three additional models, which are described as follows

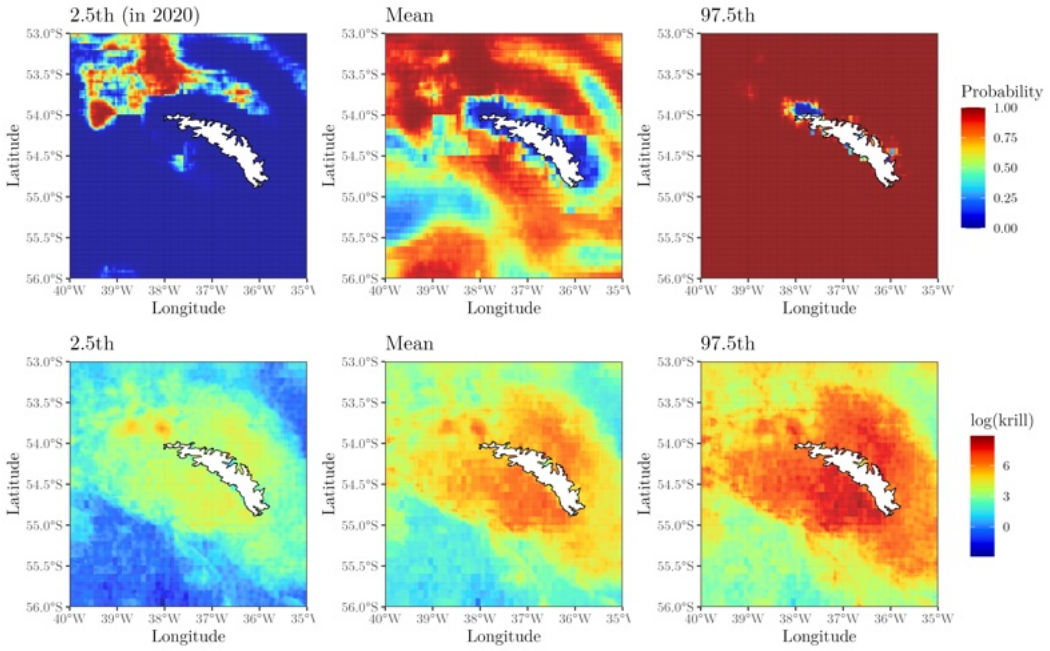

Figure SF23: Year 2020. Top row: predicted probability of non-zero krill biomass (posterior mean) along with associated uncertainty. Bottom row: predicted krill biomass (posterior mean) along with associated uncertainty. Krill biomass is in g/m^2.

1. No random effects

$$\text{logit}(\pi_i) = \beta_0^z + \beta_1^z \texttt{chlor} \tag{7}$$

$$\log(\mu_i) = \beta_0^y + \beta_1^y \texttt{depth} + \beta_2^y \texttt{speed} + \beta_3^y \texttt{surf_temp} + \beta_3^y \texttt{mass_flux} + \beta_4^y \texttt{density_drif} \tag{8}$$

2. Independent random effects

$$\text{logit}(\pi_i) = \beta_0^z + \beta_1^z \texttt{salinity} + \psi_i \tag{9}$$

$$\log(\mu_i) = \beta_0^y + \beta_1^y \texttt{depth} + \beta_2^y \texttt{surf_temp} + \beta_3^y \texttt{expect_freq} + \xi_i \tag{10}$$

3. Shared random effects

$$\text{logit}(\pi_i) = \beta_0^z + \beta_1^z \texttt{expect_freq} + \beta_2^z \texttt{res_time} + \psi_i \tag{11}$$

$$\log(\mu_i) = \beta_0^y + \beta_1^y \texttt{depth} + \beta_2^y \texttt{chlor} + \beta_3^y \texttt{pot_temp} + \beta_4^y \texttt{surf_temp} + \beta_5^y \texttt{speed_drif} + \gamma \cdot \psi_i \tag{12}$$

However, as shown in Table ST4, the model specified by Equations (5) and (6) was selected based on the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) and the WAIC.

Table ST4: Computed DIC and WAIC for alternative spatial models fitted based on the aggregated acoustic krill biomass data. All criterion are negatively oriented, meaning that smaller values are better.

Model	DIC	WAIC
Equations (5) and (6)	2911.852	2937.433
Equations (7) and (8)	2927.787	2956.351
Equations (9) and (10)	2921.207	2948.681
Equations (11) and (12)	3046.863	3058.399

References

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. Scandinavian Journal of statistics *5*, 151–157.

Bolin, D., Jin, X., Simas, A. and Wallin, J. (2024). ngme2: Latent Mixed Effects Models with Flexible Distributions. R package version 0.6.0.

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. Statistics and computing *24*, 997–1016.

Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with `R-INLA`. Journal of Statistical Software *63*, 1–25.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 583–639.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. The Journal of Machine Learning Research *14*, 867–897.