Code: UKCAL-CWF-CON-EIA-RPT-00007-7B29

Volume 7B Proposed Development (Offshore) Appendices

Appendix 6-1 Offshore Ornithology Baseline Characterisation Report Annex 15 MRSea Method Statement

Caledonia Offshore Wind Farm Ltd

5th Floor Atria One, 144 Morrison Street, Edinburgh, EH38EX

OCEAN WINDS

Volume 7B Appendix 6-1 Annex 15 MRSea Method Statement

Code	UKCAL-CWF-CON-EIA-RPT-00007-7B29
Revision	Issued
Date	18 October 2024

This document contains the following report: 'Methods statement – MRSea modelling at Caledonia offshore wind farm' as prepared by Black Bawks Data Science Ltd in August 2023. For the purpose of Consent Application, the document has been retitled to: 'Volume 7B, Appendix 6-1, Annex 15: MRSea Method Statement', alongside the addition of a new front cover.

Black Bawks Data Science Ltd
SC557537
Document: BB0021 – 001
Date: Aug 16, 2023

OCEAN WINDS

Methods statement – MRSea modelling at Caledonia offshore wind farm

Document history

Document ID	Date	Personnel	Approved by
BB0021_001	27 July, 2023	[redacted]	[redacted]
BB0021_001 - 1	3 Aug, 2023	[redacted]	[redacted]
BB0021_001 - final	16 Aug, 2023	[redacted]	-

Table of Contents

Introduction

Between May 2021 and April 2023, 24 digital aerial surveys were flown monthly at the proposed Caledonia Offshore Wind Farm (OWF) wherein transects included the site plus a 4 km buffer. NatureScot recommends that the MRSea modelling framework (Mackenzie et al., 2013) is used for estimating densities of birds for baseline characterisation (NatureScot, 2023).

This methods statement lays out the proposed methodology for applying MRSea to the Caledonia OWF as well as supplementary analyses to support the application including:

- An exploratory analysis to justify which species are able to be modelled
- Adjustments for density estimates based on availability bias and animals unable to be ID'd to species level
- A hotspot analysis based on modelled species
- A displacement analysis based on a distance to turbine parameter
- Scenarios of displacement using a Random Forests model

Exploratory Analysis and Data Preparation

To ensure adherence to the most recent NatureScot recommendations (NatureScot, 2023), we will first conduct an exploratory analysis to identify which species have enough data for model generation. For example, a tight cluster of 30 birds in one location would not be appropriate for modeling, but it's possible that several small clusters of 1–2 birds adding up to 30 individuals may be appropriate. Such an analysis would involve looking at both the quantity of observations available as well as the distribution. We will generate a table of counts for all seabird species recorded by survey and receptor behaviour (i.e., flying, sitting, and all birds). Using the raw counts, any species/survey/behaviour combinations with < 20 observations will be immediately discounted from modelling. Any species/survey/behaviour combinations with 21 – 30 observations will be mapped, and a decision will be made on the appropriateness of inclusion for modelling. This decision will likely have levels of subjectiveness associated with it based on past experience with similar data and this decision-making process will be laid out clearly during the reporting phase.

This analysis will also help identify the most feasible temporal scales for modelling (i.e., by survey, or NatureScot and/or Furness (2015) seasons). This will be done by examining the list of species/survey/behaviour combinations that are able to be modelled. Species that are absent or nearly absent from the site during (for example) the non-breeding season would not be able to be modelled at that temporal scale. To further clarify, if, for example, a species' non-breeding season was from Sept - Apr, and we were only able to model Sept and Oct, then we would present those as survey-level models and not as a full non-breeding season model. We anticipate at this stage; we would also identify which environmental covariates from Table 1 are suitable for modeling. For example, the key environmental parameters that can be used in this analysis might be expected to be different between flying and sitting birds because the covariates that would impact their density and distributions may be different. For example, we might expect wind speed and direction to impact flying birds, but not necessarily birds sitting on the water, and we will explore if a source of µ and v wind components can be acquired at the appropriate spatio-temporal scale for modelling flying birds. This exploration of other environmental covariates could improve density estimates for flying birds which is important because of the need for accurate density estimates of flying birds for collision risk modelling.

Effort and observation data from 24 monthly digital video aerial surveys conducted by APEM between May 2021 and April 2023 will be used for spatial modelling of species distribution and abundance. The Marine Renewables Strategic Environmental Assessment (MRSea) package in R (Scott-Hayward et al., 2013) will be used for density surface modeling/abundance estimation for up to 13 key species within the Caledonia OWF study area. The likely candidate species for consideration of inclusion are guillemot, kittiwake, fulmar, puffin, razorbill, gannet, great black-backed gull, herring gull, great skua, Manx shearwater, common gull, Arctic tern and common tern.

No-ID Apportioning

Birds not identified to species-level will be apportioned using a proportional technique based on species groupings provided by the digital aerial survey provider. For example, if there is a species grouping called "large auks" which is made up of possible Guillemots or Razorbill, then the proportional species composition will be applied within spatial bins that will be used for modelling. In other words, if 80 Guillemot and 20 Razorbill were identified, and a spatial bin/cluster had 10 unidentified large auks, then 8 Guillemot and 2 Razorbills would be apportioned into that bin. In the case that a proportional approach leads to fractional values (for example, if the split was 80/20 for Guillemot to Razorbill, and there were 7 unidentified large auks in a cluster, this would equate to 5.6 Guillemot and 1.4 Razorbill. This would be rounded to 6 Guillemot and 1 Razorbill). The apportioned data would be used in the modelling process.

Proportional no-id apportioning will be attempted first at a transect level (which will ensure a more accurate proportional representation of non-identified species), and if that is unable to be performed, then it will be done at the survey level. In cases where no or low numbers of observations identified to species level versus the species grouping, we will use the average of the proportions from the months within the associated NatureScot season. If the within year seasonal average is not possible to be used, we will use the proportional average from the associated season across both years of surveys, and if that is not possible, we will use the proportional average across all surveys.

Availability Bias

Diving birds, such as guillemots and razorbills, spend time foraging beneath the water surface. As a result, given the snapshot nature of airborne survey methodologies, a significant number of birds might go undiscovered. Thus, an "availability bias" adjustment must be made.

Based on the correction factors suggested by Thaxter et al. (2010) for guillemot, razorbill, and by Spencer (2012) for puffin, the adjustments will be applied to each relevant auk species. The total number of 'unavailable' birds are therefore calculated (23.75% of guillemots, 17.39% of razorbills, and 16.50% of puffins) and added to the monthly sitting bird totals.

Data preparation for modelling

We will use the Complex Regional Spatial Smoother (CReSS) spatial modelling method with Spatially Adaptive Local Smoothing Algorithm (SALSA) based model selection (Scott-Hayward et al., 2013). The models will fit the relationship between the observations (count response variable) and the environment (covariates) at each location, allowing for estimation and prediction of the animal density across the study region.

For each survey, counts of individuals of each species will be assigned to the midpoint of the respective aerial image footprint to produce the input data for each species-specific model. This generates a count variable (i.e., the dependent variable) for each footprint, and the footprint area thus becomes the offset for the model to ensure predicted outputs represent density. Covariate values will be assigned to the midpoint of each segment such that the resulting model input data frame will include survey-specific species counts and covariate values for each transect segment.

Model Inference

Count data from the aerial transect surveys will be correlated to consecutive measurements through space and time. Furthermore, due to environmental and prey conditions, the number of animals in any given area is more likely to be similar for points closer temporally, than those more distant in time. Models fitted to (relative) abundance data attempt to explain animal abundance at any location, but the information (covariate data) that describes why animals are found in high/low numbers at specific locations is frequently missing from the model, leaving patterns in the model's noise component (model residuals). These patterns are also expected to be similar along the track lines. This (positive) correlation in model residuals along the track lines violates a critical assumption for standard statistical models that require an independent set of residuals (such as Generalised Linear Models (GLMs) / Generalised Additive Models (GAMs)). Ignoring this violation can invalidate all model-based precision estimates (e.g., standard errors, confidence intervals, and p-values), resulting in overly complex models that can suggest irrelevant environmental covariates are statistically significant.

Transect data are frequently prone to such spatiotemporal autocorrelation, which violates a core assumption of GLMs/GAMs. Transect ID will thus be included as a blocking factor in the analysis to control for autocorrelation in the model. This informs the model that correlation within a transect is accepted and that transect independence is assumed.

A one-way Analysis of Variance (ANOVA) will be performed to determine the statistical significance of covariates in the predictive model. Partial dependence plots will be used to investigate covariates that have significant relationships with the data in the model. Further model inference can be made by examining the cumulative residual plots output by the models.

Selection of Model Covariates

A full model with all appropriate terms (e.g., as identified from Table 1) will first be fitted for each species without a smooth term for the spatial component. This allows the potential relationships between covariates and species observations to be initially unhindered by spatial information. We will then use Variance Inflation Factors to select terms from the initial model fitting process that should be removed due to collinearity based on a threshold value of 2.

The flexibility of the smoother-related term for each model term will then be chosen, followed by the model selection for the two-dimensional smoother term for the spatial component. Segment area will then be incorporated into the model as an offset term because the transects' division may have resulted in slightly different dimensions. Each model will be permitted to retain the covariates as a smooth or linear term (or omitted completely). SALSA will be used to fit a smooth function for each covariate. Model selection for both the covariates and spatially based smoothers will be conducted by an objective fit measure like a Bayesian Information Criterion (BIC) for quasi-likelihood (QL) models. Models that permit over-dispersion for Poisson-style counts are QL based, necessitating QL-based fit scores.

Knot Placement and Basis Function Details

The number of "knots" used for the model and the effective range of each knot (the spatial extent to which each knot influences the fitted surface) are both key factors in determining the model flexibility for the spatial surfaces in this setting. The candidate models will be chosen from a range of models that vary in the number of knots provided and the effective range (R-value) of each knot because the optimal choices for both these values are always unknown.

The starting knot positions on the spatial surface are chosen to maximize coverage across the spatial area (via a space filling algorithm; John et al., 1995), and these positions are allowed to move according to the SALSA (Walker et al., 2011) model selection technique. The local exponential basis function ((exp(-d/r2)) with d=Euclidean distance) will be used, allowing for varied R-values over the surface.

A variable number of knots (2-40 depending on data sparsity; the number is denoted by the degrees of freedom in the model) will be used for the candidate models, and an objective fit criterion will be employed to select the best model(s). In effect, the position of the knot placement, and to a lesser extent the number of knots, reflect the complexities of the spatial relationship between bird abundance and the covariates chosen for the study.

Knot locations will be identified separately for each survey to accommodate for differences in survey effort and bird distributions across surveys.

Table 1. Candidate covariates to be tested in the MRSea analyses.

Model covariate	Definition	Source
Survey ID	Unique ID for each survey	APEM Aerial Surveying
Seabed sediment type (factor)	Marine habitat classification of seabed substrate for Britain and Ireland	JNCC UK SeaMap 2018 Version 2 (https://hub.jncc.gov.uk/assets/202874e5 -0446-4ba7-8323-24462077561e)
Bathymetry	Depth below sea surface (m)	GEBCO Gridded Bathymetry Data 2019
Bathymetric slope	Change in bathymetry between pixels	GEBCO Gridded Bathymetry Data 2019
Bathymetric aspect	Direction bathymetric slope faces	GEBCO Gridded Bathymetry Data 2019
SST	Interpolated sea surface temperature on hourly 0.01 degree grid	PODAAC (https://podaac.jpl.nasa.gov/dataset/MUR-JPL-L4-GLOB-v4.1)
SST gradient	Change in SST between pixels/ slope of SST	PODAAC (https://podaac.jpl.nasa.gov/dataset/MUR-JPL-L4-GLOB-v4.1)
Sandeel predicted density	Probability of presence of buried sandeel in the North Sea study region.	Marine Scotland (https://spatialdata.gov.scot/geonetwork/srv/eng/catalog.search#/metadata/Marine_Scotland_FishDAC_12377)
Sandeel probability of presence	Predicted density of buried sandeel in the North Sea study region (number per m^2)	Marine Scotland (https://spatialdata.gov.scot/geonetwork/srv/eng/catalog.search#/metadata/Marine_Scotland_FishDAC_12377)
Distance to coast	Distance to coast (m)	NA
Distance to colony	Distance to colony (m)	JNCC Seabird Monitoring database
Segment area	Area of each segment within a transect (m^2)	APEM Aerial Surveying
Spatial component	Northing and Easting	GIS (UTMs)
Distance to wind turbine	Distance to the nearest operational turbine (m)	Moray East Wind Farm, Beatrice Wind Farm
Wind Vectors	μ and v components	NCEP global forecasting system https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast

Geo-Referenced Results

The species-specific fitted surfaces will be generated by making predictions to a grid using the final model at a 1km x 1km resolution. The grid is a series of regular points spaced at 1km resolution across the surface of the area of interest. These regular points are associated with the same environmental covariates as those used in the modelling process. This allows the trained MRSea model to make predictions of animal density on each of those points. Those data can then be visualized or interpolated to create surfaces. These grids will be projected as the Universal Transverse Mercator (Zone 30) projection.

To measure uncertainty spatially and in the population estimates, the model will be bootstrapped 500 times (wherein random subsets of the modelled coefficients are drawn from a multivariate normal distribution and predictions are made for each grid cell, 500 times). From this we will calculate the mean predicted density, the upper and lower 95% confidence limits, and the coefficient of variation (CV; as defined by the ratio of the standard deviation to the mean). These measures of uncertainty will be visualized and presented in the final report.

Abundance Estimates from MRSea Density Surfaces

Abundance estimates will be calculated by summing the grid cells across the prediction surface at the temporal scale specified by the exploratory analysis. To calculate abundance estimates within the survey area, we will sum grid cells that fall within the boundary or touch the edges, ensuring that grid cells at the boundary are clipped to the boundary footprint. The upper and lower confidence limits of the population estimate will be calculated by determining the 95% confidence limits of the 500 bootstrapped surfaces. We note that in previous work in the offshore wind industry in the UK, the term "coefficient of variation" is often applied to population estimates derived from bootstrap exercises. However, this is a misnomer because the standard deviation of a bootstrapped mean is the standard error. The calculation of the ratio of the standard error to the mean is called the relational standard error (RSE), and not the CV as has been previously used. For the sake of accuracy in the vocabulary used and to distinguish this measure of uncertainty from the spatial CV, we will present the uncertainty as the RSE, but note that this is equivalent to what has been referred to as the CV in past work.

A hotspot analysis (task 1b) will be performed upon completion of all species models. The analysis will be informed by the "all birds" survey-level MRSea models. Model outputs from all species and survey-level outputs are normalized on a scale of 0 –1 and then averaged. We also compute the cell-by-cell coefficient of variation by dividing the standard deviation by the mean. This will give us a single magnitude (I.e., mean normalized prediction layer) and persistence (I.e., variability as defined by the CV) layer. The upper and lower 95th percentiles of all values in the magnitude and persistence layers will be computed to use as thresholds for determining if a grid cell is persistent hot or cold spot. This is categorized as per table 2.

Table 2. Classifications of hot spots as defined by percentiles from magnitude and persistence layers

Persistence	Magnitude	Classification
CV > 95th percentile	Mean > 95th percentile	Persistent hot spot
CV > 95th percentile	Mean < 5th percentile	Persistent cold spot
CV < 5th percentile	Mean > 95th percentile	Volatile hot spot
CV < 5th percentile	Mean < 5th percentile	Volatile cold spot
CV > 5th percentile, < 95th percentile	Mean > 95th percentile	Transient hot spot

CV > 5th percentile, < 95th percentile	Mean < 5th percentile	Transient cold spot
CV > 5th percentile, < 95th percentile	Mean > 5th percentile, < 95th percentile	Transient ambient spot
CV > 95th percentile	Mean > 5th percentile, < 95th percentile	Persistent ambient spot
CV < 5th percentile	> 5th percentile, < 95th percentile	Volatile ambient spot

Displacement Modelling

We would make use of the precise Moray East offshore wind farm operational turbine locations and timing of installation to build a series of "distance to turbine" predictor layers that would be used in the modelling process. We will construct monthly "distance to turbine" layers that will be representative of the active turbines in the water at the time of each survey (akin to monthly SST layers). We will examine the partial dependence plots of the distance to turbine layers to make assessments of displacement as inflection points in those figures will indicate if displacement is occurring, and at what spatial scales. This information will be incorporated into the final report as opposed to being a separate analysis, as it will be informed by the partial dependence plots from the MRSea modelling exercise.

Predictive Displacement Scenarios

Subject to engagement with stakeholders, an analysis is proposed which would perform spatial displacement scenarios based on proposed turbine locations. The distance to turbine layers from the operational Moray East offshore wind farm will be used to train a spatial model that uses a machine learning algorithm called random forests. The random forest algorithm is a supervised machine learning algorithm used widely for regression and classification problems in machine learning (Breiman 2001). A random forest is a classifier that includes many decision trees on various subsets of a given dataset. The classifier takes the average decision of that subset to improve the predictive performance. It is built on the idea of ensemble learning, where multiple classifiers are integrated to solve complex problems and improve model performance. In the same way that a forest with many trees is more robust, a random forest algorithm with more decision trees will have greater accuracy and higher predictive ability.

In a spatial context, this is applied in an almost identical fashion to MRSea, where observations or counts are associated with environmental covariates which allows for a model to be trained. This trained model is then applied to a grid of regular points to generate predictions in space. We will use the same environmental covariates as MRSea, however, the spatial component will be replaced by spatial autocorrelation terms, which will help capture the nature of flocking behaviour by marine birds. To generate the predicted scenarios of displacement, we will work with Ocean Winds to create three plausible wind turbine configurations and use each of those scenarios to create distance to wind turbine layers, which will be used to predict displacement.

We propose to use this algorithm in this case as opposed to MRSea because MRSea requires the spatial knots as calculated using the raw observations which have been captured prior to any turbines being installed. We would like to create predictive scenarios of turbines being installed in various spatial configurations without being constrained by the locations of birds as they existed prior to the installation. Working with Caledonia as part of the initial methodology proposal, we would design three wind farm configuration scenarios and create these three predictive models for Guillemot, Razorbill, Kittiwake, Fulmar, Puffin and Gannet.

References:

Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.

Furness, R., 2015. Non-breeding season populations of seabirds in UK waters: Population sizes for Biologically Defined Minimum Population Scales (BDMPS). Nat. Engl. Comm. Rep. 164.

Mackenzie, M.L., Scot-Hayward, L.A., Oedekoven, C.S., Skov, H., Humphreys, E., Rexstad, E., 2013. Statistical Modelling of Seabird and Cetacean data: Guidance Document (No. University of St. Andrews contract for Marine Scotland; SB9 (CR/2012/05)). CREEM, St. Andrews.

NatureScot, 2023. Guidance Note 2: Guidance to support Offshore Wind Applications: Advice for Marine Ornithology Baseline Characterisation Surveys and Reporting.

NatureScot, 2020. Seasonal Periods for Birds in the Scottish Marine Environment: Short Guidance Note Version 2.

Data has the power to change how we interact with the natural world; we can explore its complexities and nuances then make changes for the better

Caledonia Offshore Wind Farm
5th Floor, Atria One
144 Morrison Street
Edinburgh
EH3 8EX

www.caledoniaoffshorewind.com

OCEAN WINDS