Statistical Modelling of Seabird and Cetacean data: Guidance Document

M.L. Mackenzie, L.A. Scott-Hayward, C.S. Oedekoven, H. Skov, E. Humphreys, E. Rexstad

This report constitutes work carried out at the Centre for Research Into Ecological and Environmental Modelling (CREEM) at the University of St. Andrews, performed under contract for Marine Scotland.

Please reference this document as: Mackenzie, M.L, Scott-Hayward, L.A.S., Oedekoven, C.S., Skov, H., Humphreys, E., and Rexstad E. (2013). Statistical Modelling of Seabird and Cetacean data: Guidance Document. University of St. Andrews contract for Marine Scotland; SB9 (CR/2012/05).

Statistical Modelling of Seabird and Cetacean data: Guidance Document

October 17, 2013

Contents

Executive Summary

CREEM were commissioned to review statistical modelling methods currently used in the marine renewables industry. We also compared the performance of these methods and appropriate alternatives not presently in use.

An extensive review of the available literature was undertaken and three modelling methods (GAMs, GAMMs and CReSS) were identified for the methodological comparison. This comparison was carried out using simulated scenarios based on off-shore and near-shore data collected from existing marine renewable developments. In particular, off-shore and near-shore data were each generated with: no-change post-impact, 30% post-impact decrease and post-impact redistribution scenarios.

The ability of CReSS, GAMs and GAMMs to recover genuine impact-related changes was examined. In addition to evaluation based around pre/post-impact changes, the relative performance of the methods at returning both accurate predictions (either pre or post impact) and realistic measures of precision about these predictions was also measured. Further, the ability of each method to correctly identify post-impact effects (if any) was also quantified (e.g. no-change post-impact, post-impact decreases and post-impact redistribution).

CReSS performed better than GAMs and GAMMs at successfully locating spatially explicit impact-related change. CReSS is also recommended for site characterisation because spatial predictions from this method showed fidelity to the simulated animal distributions. Uncertainty in the predicted spatial distribution of animals was close to its nominal (95% coverage) level.

This document contains a discussion about the issues involved with the data collection process and in particular, the differences in survey methods across platforms (e.g. boat, plane, vantage point). Related platform-based issues about the observation process (and associated imperfect detection) for the data collection, and the associated need to correct observed counts prior to input for analysis are also outlined. A description of the methods comparison process and the associated results then follow, along with recommendations based on the results contained therein. Two worked examples (based on the recommended approach) provide not only a suggested approach to analysis, but also offer signposts that consumers of analyses can use to assess reliability of the results.

In addition to this guidance document we have produced a literature review related to statistical modelling of animal distribution in the UK marine renewable industry. Based on the evaluation presented in the guidance document, we have also produced software for the assessment of animal distribution.

1 *Introduction*

This section is the introduction to the entire guidance document. In addition to this guidance document, it is supplemented with a literature review that can be found at

http://creem2.st-andrews.ac.uk/software/Literature_Review.pdf.

This guidance document focuses upon statistical issues related to improving wildlife surveys in the measurement of distribution of animals in areas of near-shore and off-shore renewable energy development. Previous assessment of renewable energy development impact has focused upon measuring differences in animal abundance prior to and following development. This approach suffered from the disadvantages of a) attributing any potential change to development as the causal agent, b) failing to acknowledge other forces that influence animal abundance and distribution and c) insensitivity to more subtle changes in animal populations, e.g. shifts in animal distribution to areas of habitat quality different than prior to renewable development.

The statistical issues to be addressed in assessing animal population distribution and potential changes to those distributions are subtle and complex. If methods for addressing such questions were straightforward, then methods would be universally in use. However, such methods are at the leading edge of statistical development. Herein we describe these statistical methods, data needed to apply the methods, comparison of competing analysis methods based on simulation, diagnostic measures to apply and examine for use of such methods, and guidelines for the proper use and interpretation of spatial modelling methods. Finally we present an appendix containing worked examples demonstrating the steps in analysis and interpretation of impact assessment of off-shore and near-shore renewables upon animal distribution.

1.1 Common features of environmental monitoring and impact data

Defensible baseline characterisation and impact assessment relies on sound statistical models which accommodate crucial features of the input data. While all data sets have their peculiarities, some features tend to persist across studies which make some modelling methods more appropriate than others.

For example, most data sets collected for monitoring and assessment purposes have an imperfect detection process, nonlinear relationships with environmental covariates, complex spatial distributions and spatio-temporal correlation in the input data which is unable to be explained by the model.

1.1.1 Imperfect detection

In the marine environment, there are often two reasons why the animals counted by an observer are fewer than the animals present at that location: either they are at the surface, and thus available to be seen, but are overlooked because they are at distance from the line (perception bias), or they are underwater and therefore unavailable to be counted (availability bias).

Perception bias is likely to be an issue for both birds and marine mammals and tends to be worse for animals further from the observers since animals are harder to see at distance. Failure to account for this imperfect detection process (which returns counts which are systematically too low) leaves the user to model relative numbers of animals across time, at best, but may invalidate the impact assessment process altogether. For instance, if the ability to see animals at the surface differs substantially after impact (for reasons unrelated to the impact) and fewer animals are seen, then the user may be left to conclude an impact-related effect exists. For this reason, some effort must be made to understand the detection process and use this information to inflate the observed counts for the animals which are missed.

For many platforms, availability bias is likely to be predominantly a marine mammal issue and will not affect the bird observation process, however with faster survey speeds (e.g. aerial digital surveys) this may also be an issue for some species of diving seabirds since birds may be underwater when fast moving aircraft passes over.

1.1.2 Nonlinear covariate relationships

Statistical models contain 'covariates' and in this setting these often contain information collected directly from the local environment (environmental covariates) or can include more abstract information (e.g. spatial co-ordinates) thought to act as proxies for unavailable biological/environmental information. The covariates that enter statistical models are either categorical (classifier type) or continu-

ous (numerical) in nature, the latter of which can be related to the response in a variety of ways.

For example, the continuous covariates available for modelling tend to have curved (i.e. nonlinear) rather than straight-line (i.e. linear) relationships with the response/input data[1] and these relationships need to be accommodated by the model if the associated model predictions are to be useful. These nonlinearities make some modelling methods a poor choice (e.g. Generalized Linear Models; GLMs (McCullagh and Nelder, 1989)) and nonlinear methods (e.g. Generalized Additive Models; GAMs (Hastie and Tibshirani, 1990; Wood, 2006)) a better option.

[1] on the 'link' scale

1.1.3 Complex spatial distributions

Seabird and marine mammal distributions are potentially very complex and uneven in the marine environment. For example, the spatial coverage of the survey effort might be patchy (and vary across time) and animal numbers may be highly uneven across the site – there may be some areas which are consistently popular with the animals and some areas where animals are rarely (or never) seen.

This unevenness in animal numbers across the survey area can be difficult to approximate with more traditional smoother-based statistical models which have a single flexibility setting which applies across the whole surface (e.g. GAMs). It might be of value to employ 'spatially adaptive' models that permit the surface flexibility to be targeted into areas of greatest need (e.g. the Complex Region Spatial Smoother; CReSS (Scott-Hayward et al., 2013)).

1.1.4 Spatio-temporal correlation

Baseline monitoring and assessment data are collected across space and time (e.g. along transects from a boat/plane or sampled from a grid from a vantage point) and observations close together in time/space (e.g. consecutive points along transects) tend to be more similar than observations distant in time. Further, the reasons for this similarity are often absent from the model. For example prey density may be a crucial driver of seabird distribution, but since dynamic prey information is rarely available for input to the model the 'hotspots' in some areas are likely to be unexplained by the model. These left-over unexplained patterns violate an assumption of some traditional modelling methods (residual independence) which makes some modelling methods (e.g. GAMs) inappropriate for data of this sort.

While the data are collected close together in space (e.g. along transects) either near-shore or off-shore, the correlation in animal numbers is likely to be predominantly temporal (rather than spatial) in the marine environment. For example, high seabird numbers in response to an abundance of prey in a particular transect on a particular day are unlikely to persist across time since prey tend

to move. In other circumstances, "landscape" level features such as sand banks may constitute more predictable foraging sites and hence induce spatial correlation.

This correlation along transects/within grid-cells needs to be treated appropriately if model outputs are used for decision making. If this (positive) correlation is ignored, the model outputs are likely to 'identify' effects which aren't genuine (e.g. lead us to conclude an impact-related effect is present, when it is not). For this reason, modelling methods which permit correlation of this sort are often more appropriate (e.g. Generalized Estimating Equations; GEEs (Hardin and Hilbe, 2002) and Generalized Additive Mixed Models; GAMMs (Brown and Prescott, 1999)) for these data.

1.2 Pantheon of guidance for offshore renewable impact assessment

Much of the guidance used in site characterisation for offshore renewable developments can be traced to the document published almost 10 years ago by Camphuysen et al. (2004). Recently there has been a proliferation of additional documents prepared for assessing an increasing number of off-shore renewable sites around the UK. Funded by the Centre for Environment Fisheries and Aquaculture Science (CEFAS) and the Marine Mammal Organisation (MMO), a document is scheduled to be completed in Autumn 2013 which reviews post-consent monitoring information from off-shore wind farm developments in relation to licence conditions.

Another document in draft form (Jackson and Whitfield, 2011) attempts to synthesise a wide breadth of information regarding data sources, study designs and survey methods used to assess marine renewables impacts. None of these documents emphasise analysis methods associated with measuring impacts. The document we have written concentrates on the statistical modelling approaches used to detect impacts of marine renewables.

2 Background

Estimating animal abundance via distance sampling (Buckland et al., 2001) constitutes a two-stage estimation process. If sampling is conducted by a method such that not all animals are seen inside the truncation distance (the furthest perpendicular distance from the transect line included in the analysis), then a model is used (via fitting of a detection function) to infer the number of animals in the covered region (the area within the truncation distance from the transect).

Even for data collection methods where detection methods are not susceptible to missing animals away from the transect, uncertainty in animal abundance exists because of extrapolation from the portion of the study area where sampling effort is exerted to portions of the study area where sampling effort does not take place.

Design-based inference relies on the assumption that sampled

portion of study area is similar to the unsampled portion of study area. This is implemented through randomisation; coordinates of the study area are drawn via some stochastic process such that investigator intervention (*judgement sampling*) does not take place.

Design-based sampling designs require a property called "even coverage" which is defined as every point in the study area is as likely to be included in the sample as every other point. This is accomplished by employing random sampling, or preferably systematic sampling with a random start. Systematic sampling ensures that samples are not by chance clustered in some portion of the study area.

Transect spacing relates to between-transect distances employed in systematic sampling. The idea of making transects run parallel to gradients in animal density (or habitat features that may be proxies to animal density) is to cause variability in rates of animal encounter within transects to be large relative to variability in encounter rates between transects. This is because variability in encounter rate between transects is a measure of uniformity of animal density within the sampled region. If there is little variability in encounter rates between transects, then unsampled areas can be predicted to have approximately the same rates of encounter as the sampled transects. Variability in encounter rates between transects contributes to uncertainty in estimated density throughout the study area. Design-based inference acknowledges the existence of variability in encounter rates between transects but does not attempt to explain the patterns in variability.

Distinctions between design-based and model-based modes of inference in studies of animal abundance are drawn by Borchers et al. (2002). Model-based inference presumes the relationship between the response variable and measured covariates in the sampled region (e.g. animal density and water depth) holds in portions of the study area not sampled. Spatial variability in animal densities is modelled through the use of spatially-explicit covariates.

There are three by-products of the process of modelling spatial variability in densities. First, modelling may explain some of the variability and consequently reduce the amount of unexplained (residual) variation in densities. This may produce more precise estimates of animal abundance. The second by-product is a mechanism whereby extrapolation from the sampled area to areas not sampled can be made. The third by-product is insight into the potential drivers that influence animal distribution patterns. It is this third by-product that makes model-based inference the preferred method for investigating consequences of renewable energy development upon seabird and marine mammal populations. Distinctions between design-based and model-based modes of inference in studies of animal abundance are further discussed in Borchers et al. (2002).

Examples of model-based methods of animal estimation include Hedley and Buckland (2004) using an example of minke whales

from the Southern Ocean, Royle et al. (2004) (terrestrial birds) and more recently Johnson et al. (2010) (terrestrial plants).

While valuable, there are challenges associated with using model-based inference procedures. These include unresolved questions regarding the implementation of a sampling design. While there is a potential liberation in that sampling effort does not strictly need to be randomly allocated, the question remains how should sampling effort be allocated when using model-based inference? What criteria should be applied to best use available sampling effort? This document discusses these issues in Section 9.1.

An important challenge associated with employing model-based inference is the potential for the results to be wrong. Theory underpinning design-based inference shows that the answers produced from design-based estimators are, on average and in most circumstances, unbiased. Unfortunately, there is no such guarantee for model-based estimators. An axiom of statistics attributed to G.E.P. Box is that "all models are wrong but some are useful." and since the validity of model-based inference depends on a "correct" model, model-based inference is subject to bias if the models used are incorrectly specified. For this reason, the details of the modelling approach adopted are crucial.

3 *Data acquisition*

Four different survey approaches have been developed by the marine renewable industry to record the at-sea numbers and distributions of seabirds: boat surveys, visual aerial surveys, digital aerial surveys and land-based visual surveys or vantage point surveys.

This section provides a broad summary of the existing guidance and associated reviews relating to these forms of data collection. Also highlighted here are aspects of data collection that are of particular relevance to the data analysis methods described later in this guidance document (section 4).

The protocols relating to boat and visual aerial surveys for bird distributions at sea are well-established and largely standardised across the off-shore windfarm industry by the adoption of methodologies outlined in Camphuysen et al. (2004), which was the result of work funded by Collaborative Offshore Wind Research into the Environment (COWRIE). This standardisation in approach is highlighted in a forthcoming CEFAS/MMO-funded review which addresses whether post-consent off-shore wind farm monitoring in the UK meets the conditions set by the site specific Food and Environment Protection Act (FEPA) licences (although the focus is largely on England and Wales). MacLean et al. (2009) highlighted that it can be unclear from looking at industry reports whether all the guidelines are strictly followed due to insufficient detail being presented, and stressed the need for this information to be provided. This lack of detailed information presents a challenge when

assessing the extent to which field protocols have been adopted. Draft guidance provided by Scottish Natural Heritage (SNH) and Marine Scotland (Jackson and Whitfield, 2011) discussed the data collection protocols relating to boat and visual aerial surveys and suggested a number of minor revisions to Camphuysen et al. (2004).

For digital aerial surveys, another COWRIE-funded project provided a review of these techniques and initial protocols with respect to technical issues and survey design and analysis (Thaxter and Burton, 2009). Jackson and Whitfield (2011) reviewed in brief the field protocols relating to digital aerial surveys but did not provide any new recommendations or insights as their guidance was based on work that was discussed in more detail by Thaxter and Burton (2009).

Standardisation of vantage point (VP) methodology in the context of the marine renewable industry has not yet taken place. Although Jackson and Whitfield (2011) provided a number of recommendations, it is not yet apparent whether the methods are being adopted.

3.1 Marine mammal surveys

Much of this document will deal with data collection and analysis for seabirds. However most of the analysis methods described can also be applied to marine mammal data. However the collection of marine mammal data may differ slightly because the frequency with which marine mammals are detected may be much less than seabirds. The methodology commonly employed for collecting and recording marine mammal sightings is attributed to the Small Cetaceans in the North Sea (SCANS) surveys conducted in 1994 and again in 2005. Because detectability of marine mammals is imperfect, distance sampling methods are commonly employed and consequently distances and angles are recorded to the sightings via the use of thing such as angle boards, distances sticks, binocular reticles and inclinometers. Protocols for collection of such data for marine mammals is described by Hammond et al. (2013) with subsequent references to Gillespie et al. (2010) for computerised data recording systems and Gilles et al. (2009) for methodology associated with aerial surveys. We allocate the remainder of this data collection section to field protocols associated with seabird sightings.

3.2 Boat surveys

Boat surveys (otherwise referred to as ship-based surveys) have to date been the most commonly used census technique by the offshore wind farm (OWF) industry. The key advantages of boat surveys compared to all types of aerial surveys include the following: ships can be used to sample bio-physical data (e.g. temperature and salinity) at the same time as bird data (the use of this information

in analyses is discussed in section 4); they provide the opportunity to collect additional information on the behaviour of the birds (e.g. it is possible to distinguish between birds that are engaged in foraging flight as opposed to commuting flight or they can be used to collect information on migration events); and the probability of detecting birds that are diving is far higher, because any point on the trackline is in view for longer due to the relatively slow travel speed (Buckland et al., 2012). At one time the species classification success was higher for boat surveys compared to aerial surveys but as digital aerial surveys have improved, a comparison of classification accuracy by these methods is warranted.

As outlined in Camphuysen et al. (2004) the methodology for boat surveys is essentially a combination of several surveys which are run concurrently: line transect methodology for birds sitting on water; snapshot methodology for birds in flight and ad hoc record of birds in flight (not considered further here as these do not contribute to the estimates of at-sea numbers and distribution).

3.2.1 Line-transect methodology

For marine mammals accepted practice for gathering and recording sighting data were developed in anticipation of the SCANS II (small cetaceans in the North Sea) surveys conducted in 2005 (Hammond et al., 2013).

As boats travel along pre-defined survey transect routes[2], a 300m wide band is defined to one side and ahead of the ship which is regarded as the spatial extent of the area to be surveyed. Birds on the sea surface are more detectable when they are closer to the transect line followed by the ship. To overcome the problem of decreasing detectability of birds, as their distance from the trackline increases, the transect strip width is subdivided into distance bands, effectively defining strips perpendicular to the transect line. Camphuysen et al. (2004) recommended the use of the following distance bands: A = 0-50m, B = 50-100m; C = 100-200m, D = 200-300m and E >300m (although records from the latter do not contribute to the overall counts) and this appears to be universally used (MacLean et al., 2009).

[2] a transect is a pre-defined path along which counts or occurrences of an object are made

Corrected counts have been generated either by the use of published species specific correction factors (e.g. Stone et al. (1995)) or by the application of distance analysis (see Buckland et al. (2001) for details). Detection functions are likely to vary by location for a number of reasons; for instance, sites close to shore may more commonly exhibit flat sea conditions with relatively high detectability, whereas sites in more open sea may have greater swell and be exposed to weather conditions which reduce detectability further from the boat. Location-specific estimates of detectability from a distance analysis are therefore preferred as they account for these location-specific differences in detectability.

Observations of birds are allocated into set time intervals ranging

from 1 minute to 10 minutes (depending on the required resolution of the data) and the location of the boat is recorded at the start or the mid-point of each interval. MacLean et al. (2009) highlighted an off-shore development where surveyors had used an alternate methodology of recording the exact timing of bird observations rather than assigning them to a time interval. At high resolutions, assigning a time to an observation could give false accuracy with respect to the geographical location of the bird, unless the relative bearing to the observer is also given.

Camphuysen et al. (2004) recommended that observers should scan ahead using binoculars (as opposed to the default of detection by naked eye) to pick up on diving birds before they are flushed by the presence of the moving ship. This observation technique appears to have been expanded by observers who look ahead for seaducks (e.g. Lincs OWF – Centrica (2007)) and auks. It is important however, to consider that scanning ahead a long distance may detect birds which are technically in the next time interval, particularly if using short time intervals. These observations should where appropriate be denoted as occurring in the subsequent timed recording interval.

Jackson and Whitfield (2011) emphasised the need for surveyors to place priority on the detection of birds in distance band A (0-50m). One of the key assumptions of distance sampling is that all birds on the transect line are detected and therefore all surveyors should strive to meet this requirement.

3.2.2 *Snapshot methodology*

The snapshot methodology was developed to reduce the likelihood of overestimating the number of birds in flight (Camphuysen et al., 2004). In effect, the sampling procedure should be visualised as setting up a series of contiguous boxes which fit within the 300m transect width and observations of birds are allocated to individual boxes. To achieve this, the sampling frequency should take into account the ship's travelling speed (hence the distance travelled between each snapshot count) and the distance ahead to which observations are made. For example, a snapshot sampling frequency of 1 minute is based on the ship travelling at a speed of 9.7 knots (which equates to 300m travelled every minute) and if the observer looks ahead to a distance of 300m, the area encompassed by the snapshot is a box of 300m x 300m. A common error is for the snapshot sampling frequency and the distance ahead to be scanned to be determined regardless of how fast the ship is travelling.

MacLean et al. (2009) proposed that distance intervals for snapshot counts would be an improvement to the commonly used time intervals. The advantage being any variation in ship speed (e.g. due to the conditions of the wind, state of the tides or current direction) would be explicitly encompassed in the estimate of the distance covered. In theory it would be possible to change over from using

a stopwatch to denote the times at which the snapshots should be carried to a using an onboard GPS system and a portable receiver which could be set up to flag up when the appropriate location is reached. In terms of the analyses, however, effects are likely to be minimal if a relatively short time interval is used (e.g. a ship travelling at 9.7 knots will travel 300 metres in a 1 minute period).

For the snapshot methodology there need not be allocation of the birds to distance bands (Camphuysen et al., 2004). Attempts to allocate birds in flight to distance bands are likely to violate distance sampling assumptions, as the birds are moving (Buckland et al., 2001) and observers concentrating on distance bands may also lead to a greater number of errors in other recorded information such as species or location. Ronconi and Burger (2009) call for development of methods to overcome issues of bird movement in distance sampling protocols. However, with current approaches, snapshot counts should not allocate birds to distance bands. This assumes that there are no issues of detectability for birds in flight within the 300m transect, which is generally regarded as a reasonable assumption.

3.2.3 Collection of data on environmental covariates

In Camphuysen et al. (2004), it was suggested that the introduction of an aquaflow (for the measurement of temperature, chlorophyll and salinity) would improve cost effectiveness of the bird surveys by having environmental data gathered by the survey vessel. There has been no further guidance or recommendation in relation to this aspect (e.g. no reference in the draft guidance provided by SNH and Marine Scotland) and does not address the matter of needing environmental data at all locations throughout the study area, not merely along surveyed transects.

As number of seabirds and marine mammals fluctuate greatly at any given location, it may be difficult to detect change in numbers or distributions which may have occurred as a result of the presence of the wind farm development (Maclean et al. (2013), Lapeña et al. (2011)). Despite this inherent variability, there is an increasing body of work which demonstrates that foraging, and even migratory behaviour of marine birds and mammals, can be highly predictable over tidal cycles and seasonal currents (Scott et al., 2013). The reason being that these underlying physical processes can often explain the variation in primary production or plankton activity, which in turn attracts potential prey items for marine predators such as seabirds and mammals. For example, variation in spatial and temporal distributions of foraging seabirds has been significantly correlated with tidal processes and/or interactions with the topography of the seafloor which can be manifested as relatively high chlorophyll levels or relatively stratified waters (e.g. Embling et al. (2012); Scott et al. (2013)).

Relationships between biophysical features and the numbers and

distribution of marine predators are relatively difficult to tease out without a survey design which repeatedly samples the same area to control for variation in additional key factors e.g. different stages of the tidal cycle (Cox et al., 2013). At present, renewable industry development locations are typically surveyed on a monthly basis.

It is also worth working noting that the focus of the scientific studies discussed above has been to explain the distributions of birds which are actively foraging (see Camphuysen and Garthe (2004) for coding) rather than the distributions of all birds present in the study area, as is required for the impact assessment process for the renewable industry. Nevertheless, what is needed is a better understanding of the mechanistic processes which determine the biological and physical variables which in turn determine the relative amounts of prey available and will therefore reliably predict numbers and distribution of seabirds (Scott, in press). The inclusion of temporally varying covariates rather than solely static covariates (ideally environmental data that is collected synoptically to the timing of the bird surveys) would greatly help in this respect. Further discussion is given in section 9.2 in relation to covariates to be used in the modelling.

3.2.4 Instantaneous data recording

Camphuysen et al. (2004) stated that there should be no immediate computerisation of data during the surveys in order to maximise attention on detection, identification and recording. However, a number of software packages which allow the digitisation of observations from the ship's platform on to a laptop or tablet, without taking the observers' attention from bird detection have now been developed, and are in current use. Depending on the type of program used, the geographic location of the ship is derived directly from the ship's system communication port (and therefore additional information on other parameters wind speed, wind direction and surface water temperature can be transferred) or from an internal GPS within the computer itself. There are several major benefits to digital inputting of data which bypasses the need for paper forms including: the time spent inputting data is reduced as the recording of the time of observation is automated; relative survey effort is recorded by default; minor typographic errors are highly unlikely and; data verification tends to be quicker. However consideration would need to be given to issues over how to back up data sets and ensuring sufficient battery life.

3.2.5 Accreditation

Camphuysen et al. (2004) stated that observers must be trained and have adequate identification skills. More recently in the MS/SNH guidance, Jackson and Whitfield (2011) recommended that all surveyors should be ESAS (European Seabirds at Sea) accredited and have over 100 hours survey time experience or at least be paired

with an accredited surveyor until the minimum time is achieved. Although the rationale for the minimum required amount of survey hours experience is not given, the recommendation that all surveyors should be formally assessed is sensible. Such formal accreditation is likely to ensure further standardisation in approaches and Joint Nature Conservation Committee (JNCC), who oversee the ESAS database, are now currently in the process of preparing formal guidance for field protocols and how they should be applied. It should be noted that although ESAS methods also include the recording of mammals, this does not mean that the need for separate marine mammal observer who carries out the appropriate species specific survey methodology is redundant.

3.3 Visual aerial surveys

In comparison to boat-based surveys, aerial survey methods are relatively quick to carry out and are a cheaper option to cover the same area (Buckland et al., 2012). Given their capability for covering large areas in a short period of time they reduce the potential for under or over recording birds that may move around within the study area within the survey period. Another benefit of aerial surveys is that aeroplanes do not attract certain species of seabirds, which can be a particular problem for the boat surveys (e.g. Spear et al. (2004)). Whilst boat surveys also suffer from potential biases associated birds being disturbed by the presence of vessels (e.g. Borberg et al. (2005), Schwemmer et al. (2012)), disturbance issues are far less acute for visual aerial surveys, though may still exist (Buckland et al., 2012).

Camphuysen et al. (2004) acknowledged that there was relatively little published in this particular area at the time of their review and stated that the methods that they proposed were largely derivative of those used by the Danish National Environment Research Institute (NERI). The method for visual aerial surveys is based on line transect methodology and is similar to boat surveys since detectability from the track line of the aeroplane decreases with increasing perpendicular distance. However unlike boat surveys the distance bands used are not standardized for all aerial surveys, in part as flight heights may vary. This is not a problem as long as distance sampling techniques are applied appropriately. Some surveys have been carried out using strip transect methodology, which assumes that all birds within a given distance are detected and this assumption is likely to be violated for any reasonable size of strip width.

3.4 Digital aerial surveys

Digital aerial surveys, to a large extent, have superseded visual aerial surveys, in the UK at least, and are also beginning to be used elsewhere (e.g. in the U.S.A.). Visual aerial surveys have not been possible at a number of UK OWF sites for the post-construction

period due to the flight height restrictions and the potential risk of collision with turbines. Digital surveys offer a number of advantages including: aircraft fly higher so they are less likely to disturb birds; recording is less subject to the vagaries of human error since a permanent record is available for independent verification at a later stage and detection rates in the transects are considered to be 100% which negates the need for the allocation of individuals to distance bands and correction of count data at the analysis stage.

Further consideration of the relative merits of visual and digital aerial (still and video-based) surveys and how best to compare estimates obtained from different approaches is provided in Buckland et al. (2012). Whilst they reported that higher abundance estimates were produced using digital aerial rather than visual aerial surveys, they were unable to recommend either the use of still or video-based surveys in favour of the other. They stated that the relative precision of the estimates were largely driven by the proportion of the survey area which is surveyed and the relative spacing of the sampling units.

There have been notable advancements in digital aerial survey technology since the survey work which featured in Buckland et al. (2012) and Thaxter and Burton (2009). Considerable improvements in the ability to identify individuals to the species level have been achieved, which was previously a major weakness of the digital aerial surveys. Similarly, the strip width of the cameras used by digital surveys has previously been limited, meaning that, for a given number of transects, coverage of the development area would be less than for visual aerial or boat surveys. Because of rapid developments in digital technology in recent years, there is a need for updated best practice guidance.

3.5 Vantage point surveys

Vantage point surveys (VPs) collectively describe a number of methods for recording observational data from the area of interest from a fixed elevated position, generally on-shore. These types of surveys, when used in the appropriate context, have the following advantages; they are relatively simple to carry out and; VPs do not necessitate the use of relatively expensive of survey ships or aeroplanes.

Jackson and Whitfield (2011) recommended the use of VPs for relatively small sites that are located within 1.5km of the coast (although there was no rationale provided for this distance). In more general terms, the use of VPs is likely to be best suited to situations when the area of interest can be scanned in its entirety from a single elevated location. However, even from the elevated position observers ought to recognise diminution in detectability as function of distance from observation point.

Although there appears to be a lack of standardized data collection protocols for VPs in the context of the marine renewable indus-

try, Jackson and Whitfield (2011) argued that the methods should be reasonably flexible to allow adaptation to the specific monitoring requirements of the site. They use the term Bird Snapshot Scans (BSS) to describe the method best suited to provide an assessment of the numbers and distribution of birds using the site. In brief, the observer is required to scan in a single sweep a visible arc of sea and record all observations of birds seen on the water and in flight which are actively using the area (birds which are commuting through are not considered as part of the total count and are best counted using a separate methodology which they referred to as Flying Bird Watches).

Jackson and Whitfield (2011) argued in their guidance that distance sampling approaches could not be used to correct for detection bias associated with VPs carried out in the context of the off-shore environment. Therefore their methodology did not include the recording of any measure of distance relative to the observers. Although it is very difficult to separate the underlying distribution of animals which are unlikely to be uniformly distributed (and maybe influenced by oceanographic factors such as tides, currents or biophysical factors) from imperfect detection, there are methods which can cope with this (e.g. Cox et al. (2013), by collecting information on the radial distance and the angle for each detected animal it is possible to both estimate a detection function and animal densities; see literature review: http://creem2.st-andrews.ac.uk/software/Literature_Review.pdf). Methods that adjust for imperfect detectability should be employed to prevent confounding of diminished animal detection at distance from the vantage point with possible changes in animal density.

4 *Statistical modelling methods undergoing evaluation*

The utility of a statistical modelling method depends on its ability to characterise the average behaviour of the site (either as a single abundance estimate or as geo-referenced predictions) and its ability to provide realistic uncertainty about model predictions (including 95% confidence intervals).

For instance, it is insufficient for a modelling method to provide good (geo-referenced) predictions if the uncertainty reported about these predictions is unrealistic. This is of particular concern when the model portrays over-confidence in model predictions and therefore natural (and small) fluctuations in animal numbers are mistaken for genuine changes in animal numbers. The converse situation is also a problem (when uncertainty is overestimated) since genuine changes in animal numbers over time can be mistaken for natural changes and therefore overlooked.

It is also crucial for a modelling method in this setting to identify which covariates have a genuine relationship with the response, particularly if pre and post impact comparisons are of interest. For example, falsely identifying an impact effect (e.g. an average increase or decrease, or as a redistribution) might have serious consequences for the operation of marine renewable developments, while overlooking a genuine impact effect is also a problem for the species in focus.

For these reasons, the ability of the methods undergoing evaluation to characterise site behaviour *and* provide realistic measures of precision (e.g. 95% confidence intervals) about site behaviour will be considered.

The modelling methods chosen for evaluation and comparison in this document included those already being used by the marine renewables industry to analyse baseline monitoring and assessment data and modelling methods from the wider science base considered to be appropriate for data of this kind. Literature in this area was extensively reviewed and this report can be found at:

http://creem2.st-andrews.ac.uk/software/Literature_Review.pdf.

For brevity, the description of the statistical modelling methods considered here is brief and for more details, the reader is referred to relevant literature in each section.

4.1 Generalized Additive Models (GAMs)

GAMs respect the broad nature of the input data (e.g. animal counts) which, for example, ensures model predictions are within plausible limits. For count data this means predicted animal numbers are non-negative and for proportional/presence-absence data predictions must lie between 0 and 1. The limits on the predicted values depend on the so-called 'link' function employed and there are logical choices for different types of input data (e.g. counts and

presence/absence data). GAMs are also flexible by permitting the covariates to have nonlinear relationships with the response of interest (even on the link scale). This can be very important if the covariate relationships are to be described accurately in monitoring and impact assessment data. For details on GAMs, good references are Hastie and Tibshirani (1990) and Wood (2006).

Consider the model through an example in which animal counts are collected from a plane along transect i (where $i = 1, ..s$) and transects are divided into segments (where $j = 1, ...J$) which are visited at time point t (where $t = 1, ...n_i$). In this case, the response data y_{ijt} are modelled using a Poisson distribution with mean μ_{ijt}. This mean is modelled using *Season* (with 4 levels), a smooth function of *Depth* and the spatial co-ordinates ($XPos$ and $YPos$).

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\ &\quad \beta_4 Impact_{ijt} + s_1(Depth_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \quad (1)$$

The intercept (β_0) includes the baseline season (1) while the remaining levels of Season (2, 3 and 4) have associated coefficients $\beta_1, \beta_2, \beta_3$. The effect of impact (as affecting average numbers overall) is modelled here using β_4 while s_1 and s_2 represent smooth functions of *Depth* and the spatial terms respectively.

Note, s_2 relates to an interaction term between $XPos$ and $YPos$ which permits the relationship between the XPos coordinate and the response to change with the $YPos$ coordinate. The linear predictor ($\beta_0 + ...s_2(XPos_{ijt}, YPos_{ijt})$) is fitted inside the exponential function (exp) to ensure the predicted numbers of animals cannot be negative. Formally, a Poisson-based GAM with a log-link function is described here.

Typically the GAM described is fitted using penalized splines for the one dimensional smooth terms (e.g. s_1) and employs MGCV (Wood (2003), R Development Core Team (2009)) based model selection to choose the flexibility of the smooth. Additionally, the two dimensional smooth terms (e.g. s_2) are typically implemented using thin-plate splines with a global smoothing parameter also chosen using MGCV.

GAMs are population average models which relate to, and return predictions for, the average of a population. So in this example, given some values for Season, Impact, Depth, and spatial location the GAM returns an expected count for the population at that location.

GAMs (e.g Equation 1) are typically fitted using Maximum Likelihood (ML) which returns coefficients and associated estimates of uncertainty about these coefficients. These estimates of uncertainty are then used to calculate p-values for model covariates and very often 95% confidence intervals about parameters and georeferenced model predictions. Conclusions about model covariates (and their potential relationship with animal numbers) are then

often based on the p-values returned.

GAM-based p-values are calculated assuming there is no spatio-temporal correlation in model residuals (see page 7). This is a problem if there are patterns in the data which are unexplained by the model (which is common due to missing covariates, for example) since this can result in overconfidence in model results (and p-values which are too small) and potentially false identification of impact-related effects. A wider discussion about model selection can be found in section 5.3.1.

4.2 Generalized Additive Mixed Models (GAMMs)

Generalized Additive Mixed Models (GAMMs; Brown and Prescott (1999)) are an extension of GAMs which permit both flexible covariate relationships *and* spatio-temporal correlation within some user-specified blocks/panels (e.g. transects).

GAMMs accommodate positive correlation within blocks by allowing the blocks to attract values of a 'random effect'. These random effect predictions describe the way the blocks differ from some 'population average' coefficient. For example, baseline levels (e.g. based on the intercept coefficient) for each transect might be assumed to vary from each other and from the population average baseline in a particular way:

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}) \\ \mu_{ijt} &= \exp(\beta_0 + u_i + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\ &\quad \beta_4 Impact_{ijt} + s_1(Depth_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \tag{2}$$

where $u_i \sim Normal(0, \sigma_u^2)$ and all other terms are as described previously.

The inclusion of random effects in these models require assumptions to be made about the way the blocks (e.g. transects) vary from the population average (and with each other). Almost invariably, these terms are typically assumed to be normally distributed about the population mean parameter with some variance which is estimated from the data. For more complicated models with multiple random effects, the normal distribution is specified to have at least two dimensions (e.g. multivariate normal) and allow the random effects specified to vary with each other. This extra complexity can be an issue for estimation however when many random effects are specified.

In contrast to GAMs, GAMMs are 'so-called' conditional models which return predictions which relate more closely to the average block/panel (e.g. transect). So, given values for Season, Impact, Depth, and spatial location the GAMM returns the expected number of animals for something akin to the 'average' transect.

While this might appear to be a detail, the user must average predictions for a population of transects in order to obtain predictions which are in line with population averaged model

(such as GAMs). Additionally, if predictions to a grid based on blocks/transects not visited are required (and thus predictions for the random intercept ($\hat{u}_i$ in Equation 2) for these transects are not available) the user must undertake some kind of post-processing to obtain predictions under the model.

Fortunately, this post-processing can easily be achieved by sampling 'transect coefficients' from the (multivariate) normal distribution (with mean zero and estimated variance) and averaging over model predictions obtained for the sampled population of transects. However like any model, the quality of these model predictions and the associated uncertainty about the parameter estimates depends on how valid the model assumptions are. For instance, if the transect level coefficients (random effects) are not well described by a normal distribution then the model predictions obtained in this way will be systematically too high or too low (and thus biased), and the random effects assumption is very difficult to check in practice. Unfortunately, incorrect random effects specification can also affect the quality of the uncertainty about the coefficients which can invalidate model p-values.

Model fitting for GAMMs usually involves, ML or Penalized Quasi-Likelihood (PQL) depending on model specification (e.g. if a dispersion parameter requires estimation) and choosing the terms to include in a model either involves fit criteria based on these fitting engines or p-values returned by the model. Choosing terms for GAMMs is discussed in section 5.3.1.

4.3 Complex Region Spatial Smoother (CReSS)

The CReSS model appears similar to the GAM, however the smoothing methods which underpin CReSS can differ greatly from the GAM implementation (Scott-Hayward et al., 2013). Specifically, under the CReSS model described here, the one dimensional smooth term (e.g. s_3) is implemented using quadratic B-splines with flexibility chosen using the Spatially Adaptive Local Smoothing Algorithm (SALSA) model selection (Walker et al., 2010). Additionally, the two dimensional smooth term (e.g. s_4) differs from the GAM implementation and uses radial exponential basis functions with flexibility also targeted using SALSA. This approach can return similar results to GAMs when the underlying surface is uniformly smooth but can return very different results when the flexibility required varies a great deal across the surface (Scott-Hayward et al., 2013).

$$
\begin{aligned}
Y_{ijt} &\sim Poisson(\mu_{ijt}) \\
\mu_{ijt} &= \exp(\beta_0 + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\
&\quad \beta_4 Impact_{ijt} + s_3(Depth_{ijt}) + s_4(XPos_{ijt}, YPos_{ijt}))
\end{aligned} \tag{3}
$$

The starting point for SALSA-based model selection involves distributing a specified number of knots approximately evenly across

the covariate range (or spatial surface) to fill the range/space and adaptively move these knot locations to areas requiring the flexibility. These moves are either large or small depending on where in the covariate range/spatial surface the largest residuals are located, and if the moves being considered improve the fit statistic of interest (e.g. BIC).

While the smooth terms can produce relationships which are highly nonlinear on the link scale, the smoother-based terms are linear in their parameters which means a Generalized Estimating Equation (GEE; Hardin and Hilbe (2002)) fitting framework can be employed to provide coefficients and estimates of precision. GEEs permit correlation within panels/blocks and the type of correlation structure can either be chosen in advance by the user or the residuals used more directly; for example, empirical 'robust' estimates of variance can be used to return the uncertainty about model coefficients and associated predictions.

Model fitting for CReSS-based models can involve, ML, QL or GEEs depending on model specification (e.g. if a dispersion parameter requires estimation and/or temporal correlation is present in model residuals). Therefore, choosing the terms to include in a model either involves fit criteria based on these fitting engines or p-values returned by the model. This is discussed in more detail in section 5.3.1.

4.4 Methods chosen for comparison

An extensive review of the literature revealed only a small number of statistical modelling methods have been used for data of this sort, however some readers may note that regression kriging (which combines a deterministic (regression) model to estimate the global trend, with a stochastic component; (Hengl, 2007)) is not considered here. The reasons for this are described in the literature review, however in summary, only a partial comparison would have been possible (at best). For instance, while this method can accommodate spatial autocorrelation (as a combined process) the model selection tools are unadjusted for this correlation. Specifically, the associated GAM-based p-values are likely to be too small (and thus are likely to identify impact effects which are not genuine) and GAM-based information criterion are also not immune to residual correlation and are likely to result in overly-complex models (see the GAM results in section 6.4).

There are also different ways to implement this combined approach which would make a partial comparison problematic in the absence of off-the-shelf code with associated default specifications (or detailed guidance about alternatives). For instance, in some implementations, if the GAM fit measure was below some user-chosen value then a model would not be fitted at all and the raw/input data (rather than residuals) would undergo kriging-based interpolation instead. Decisions like this, and those related to the resolution

of the grid, or the nature of the spatial correlation chosen for the stochastic component (which often involves 'expert judgment') are likely to vary considerably across users.

It is also worth noting here that while GAMs were included here for comparison with GAMMs and CReSS, GAMs would normally have been ruled out for use on these data purely on the basis they do not permit residual correlation. In this case, GAMs were considered for the comparison due to their widespread use by the renewables industry.

5 The methods comparison process

5.1 Data generation

The data for the comparison process were generated based on data collected near-shore and off-shore. The near-shore data were based on data collected from a single vantage point while the off-shore data was generated based on aerial survey data collected from pre-defined transects. Three scenarios for near-shore and off-shore were manufactured:

1. no change post-impact
2. a post-impact decrease of 30%, on average, in animal numbers
3. a redistribution post-impact from the impact site into areas already popular with the animals. There was no change in abundance for this scenario.

To ensure no method was favoured by the manufacturing process two smoother-based methods were used to generate the data and to look at methodological performance in the long run, 100 data sets were generated for each scenario. Specifically, both GAMs and CReSS were used to supply the details for the smooth functions to generate the simulated data, since the smoothing approach used for the GAMMs is similar in nature to GAMs. For each of the off-shore and near-shore scenarios, three impact types were simulated and in each case these were either GAM-based or CReSS-based which resulted in 12 data set types, each with 100 realisations. All 1200 data sets included within block correlation and over-dispersion.

5.1.1 Off-shore scenarios

The data were assumed to come from an overdispersed Poisson distribution with mean μ_{ijt} (for transect i, segment j at time t) and dispersion parameter, ϕ. The mean of the distribution was assumed to be related to model terms related to season, water depth and a smooth function of the spatial co-ordinates (XPos and YPos). The

coefficients for these terms describing the covariate relationships were obtained by fitting the GAM-based and CReSS-based models to real data.

For the off-shore scenarios, three models were used to manufacture the data:

1. Model I: No change post-impact

> **Model I** describes a **'no change'** scenario post impact. Animal numbers recorded throughout the whole survey period (pre and post impact) are described *only* using season, water depth and the spatial co-ordinates.
>
> There is no 'impact' effect built into the model (or alternatively any impact terms have zero valued coefficients).

In summary, the process generating the data was the same pre and post impact and the covariate values were also the same pre and post impact. This can be written as:

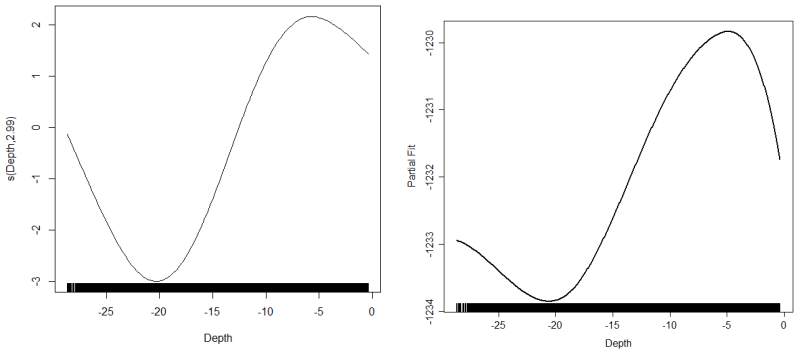
$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\ &\quad + s_1(Depth_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \tag{4}$$

Here, y_{ijt} describes animal counts collected on segment j ($j = 1, ..., J$), transect i ($i = 1, ...s$) at time point t ($t = 1, ...n_i$) and the number of observations per transect can differ across transects. Additionally, β_0 represents the intercept term (which 'applies' when $Season = 1$ and when $Depth$, and the $XPos$ & $YPos$ co-ordinates are all zero).

$Season2_{ijt}, Season3_{ijt}$ and $Seasion4_{ijt}$ are 'dummy' variables which are either effectively present or absent from the model depending on the season of interest. For instance, if the baseline season (1) is of interest then $Season2_{ijt} = 0, Season3_{ijt} = 0$ and $Season4_{4ijt} = 0$. However if season '3' is of interest then $Season2_{ijt} = 0, Season3_{ijt} = 1$ and $Season4_{ijt} = 0$. $\beta_1, ..., \beta_3$ are coefficients which describe the difference between Season 1 and Seasons 2, 3 and 4 respectively.

The smooth term for water depth, $s_1(Depth_{ijt})$, describes how the response responds to changing values of water depth and the spatial term $s_2(XPos_{ijt}, YPos_{ijt})$ is a two-dimensional smooth spatial surface which involves the $XPos$ and $YPos$ co-ordinates in a flexible way.

The smooth term for depth was generated in two ways to avoid favouring any particular approach at the evaluation stage. Specifically, GAMs and CReSS were used to generate the depth relationship, however in practice the covariate relationships were very similar across the two approaches (Figure 1).



(a) (b)

Figure 1: Depth relationships for the GAM (a) and CReSS (b) generated scenarios.

2. Model II: Post-impact decrease

Model II describes a **'decrease'** scenario post impact. Animal numbers recorded throughout the whole survey period (pre and post impact) are described using season, water depth, the spatial co-ordinates and impact. There is a decrease 'impact' effect built into the model to induce a 30% decrease overall in animal numbers (Figure 2).

Model II includes the terms in Model I with an additional 'Impact' coefficient (β_4) to describe a 30% decrease in animal numbers pre and post impact:

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\ &\quad \beta_4 Impact_{ijt} + s_1(Depth_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \quad (5)$$

Here, β_4 is an 'impact' coefficient and pre-impact observations have $Impact = 0$ and post-impact observations have $Impact_{ijt} = 1$. All other terms are as described for Model I. A spatial representation of the surfaces generated under the two smoothing approaches are shown in Figures 2 and 3 (page 27).

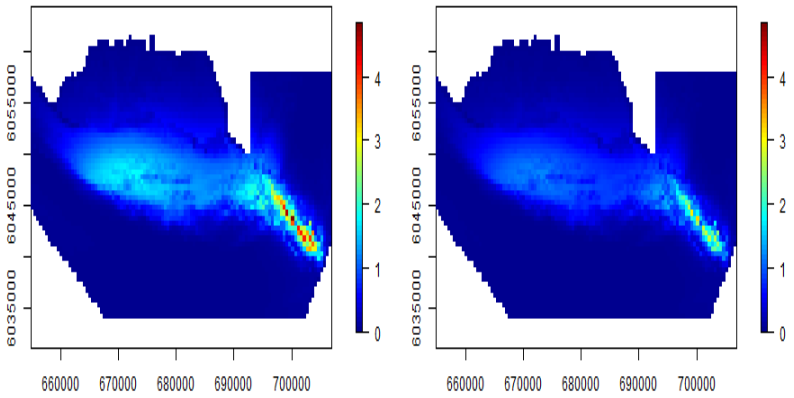


Figure 2: Underlying GAM-generated (smooth) surface pre and post impact for the **decrease** post-impact off-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The X and Y axes are in UTMs.

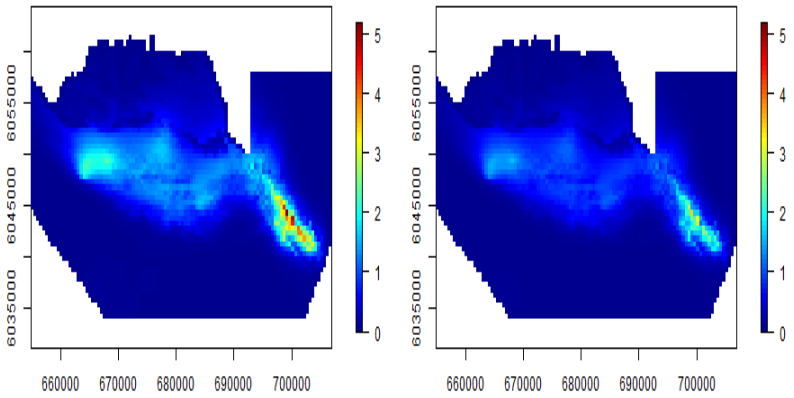


Figure 3: Underlying CReSS-generated (smooth) surface pre and post impact for the **decrease** post-impact off-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The X and Y axes are in UTMs.

3. Model III: Post-impact redistribution

Model III describes a **'redistribution'** scenario post impact. Animal numbers recorded throughout the whole survey period (pre and post impact) are described using season, water depth, and a spatial relationship with animal numbers which differs before and after impact.

There is no overall change in animal numbers pre and post impact.

Model III extends Model II by including a redistribution term post impact. This redistribution term is implemented using an interaction effect between the spatial element and the 'Impact' factor term ($s_3(XPos_{ijt}, YPos_{ijt})Impact_{ijt}$):

$$\begin{aligned} Y_{ijt} \sim & Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} = & \exp(\beta_0 + \beta_1 Season2_{ijt} + \beta_2 Season3_{ijt} + \beta_3 Season4_{ijt} + \\ & \beta_4 Impact_{ijt} + s_1(Depth_{ijt}) + \\ & s_2(XPos_{ijt}, YPos_{ijt}) + s_3(XPos_{ijt}, YPos_{ijt}) Impact_{ijt}) \end{aligned} \tag{6}$$

This interaction term attracts coefficients describing the difference between the spatial surface pre ($s_2(XPos_{ijt}, YPos_{ijt})$) and post $s_3(XPos_{ijt}, YPos_{ijt})Impact_{ijt})$ impact. All other terms are as described for Model I and II.

For the redistribution, a radial decrease was applied centrally to the impact region and a concurrent radial increase in an area popular with the animals pre-impact (e.g. Figures 4 and 5 for the GAM and CReSS generated surfaces respectively, page 29). The pre and post impact differences in the GAM and CReSS surfaces are shown in Figures 6 and 7 respectively (page 30).

The pre and post impact surfaces (and thus any differences) are very similar based on either the GAM or CReSS generation (pages 29 and 30).

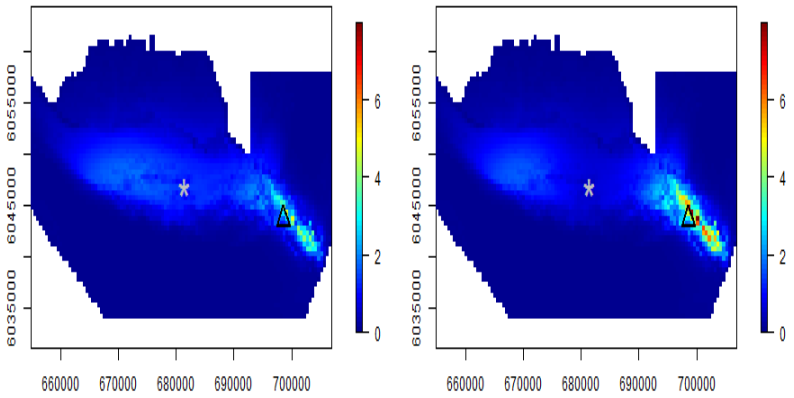


Figure 4: Underlying GAM-generated (smooth) surface pre and post impact for the **redistribution** post-impact offshore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The grey asterisk represents the centre of the impact and the black triangle represents the centre of the site for the redistribution post impact.

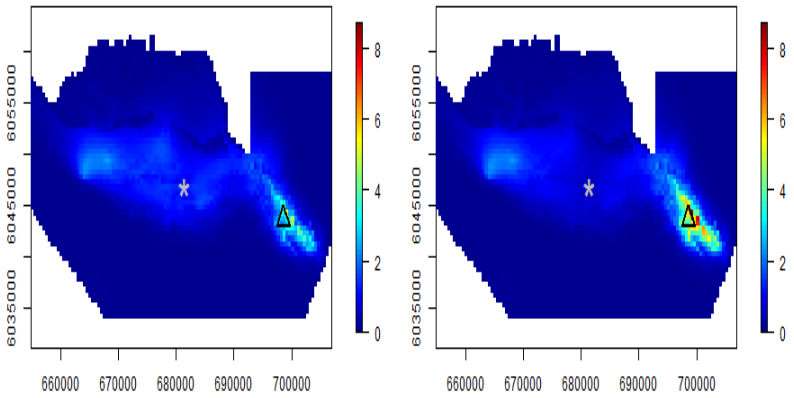


Figure 5: Underlying CReSS-generated (smooth) surface pre and post impact for the **redistribution** post-impact offshore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The grey asterisk represents the centre of the impact and the black triangle represents the centre of the site for the redistribution post impact.

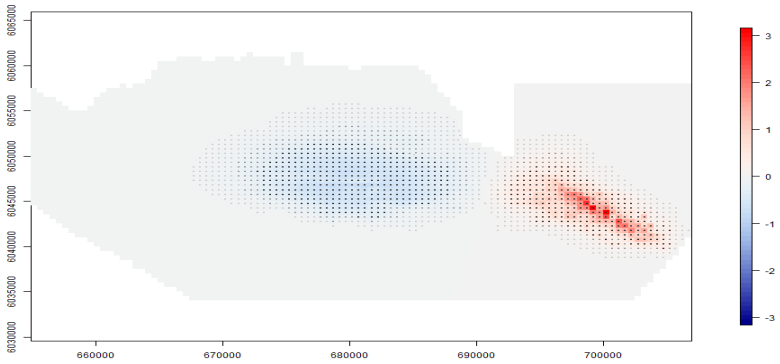


Figure 6: Difference between the underlying GAM-generated (smooth) surfaces pre and post impact for the **redistribution** post-impact off-shore scenario. The blue colour represents a post-impact decrease, while the red indicates a post impact increase. The grey colour indicates no change.

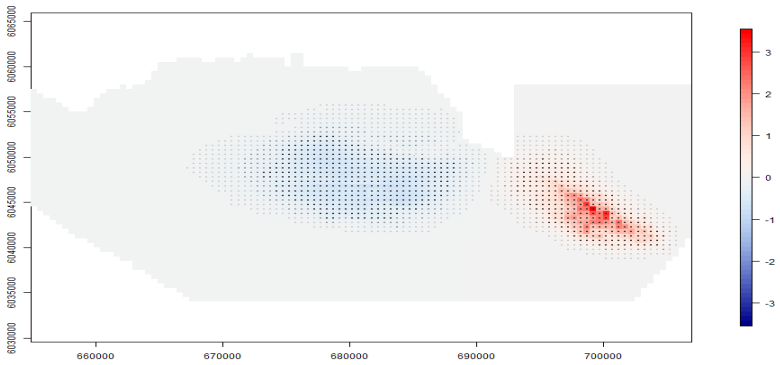


Figure 7: Difference between the underlying CReSS-generated (smooth) surfaces pre and post impact for the **redistribution** post-impact off-shore scenario. The blue colour represents a post-impact decrease, while the red indicates a post impact increase. The grey colour indicates no change.

5.1.2 *Near-shore scenarios*

The data manufacturing process was very similar as that used to generate the off-shore data however, tide state (rather than season) and observation hour (rather than water depth) were used as co-variates. For the near-shore scenarios, the data are collected over time from a predefined set of grid locations, on different days and different time points, and so y_{ijt} represents the count associated with the i-th grid location, on the j-th day at time t.

1. Model I: No change post-impact

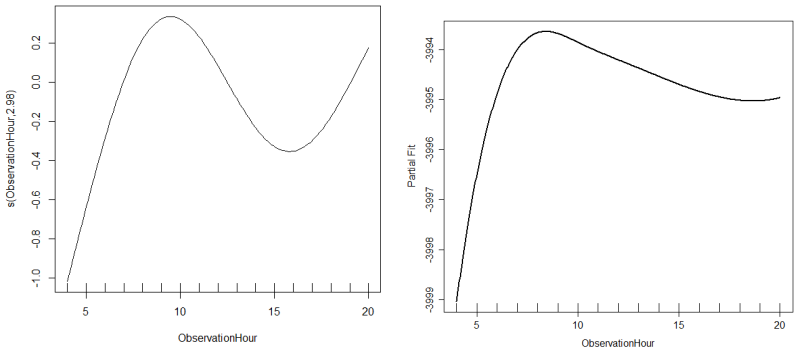
> **Model I** describes a **'no change'** scenario post impact. Animal numbers recorded throughout the whole survey period (pre and post impact) are described *only* using tide state, observation hour and the spatial co-ordinates.
>
> There is no 'impact' effect built into the model (or alternatively any impact terms have zero valued coefficients).

In summary, the process generating the data was the same pre and post impact and the covariate values were also the same pre and post impact. This can be written as:

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 FloodEbb2_{ijt} + \beta_2 FloodEbb3_{ijt} \\ &\quad + s_1(Hour_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \tag{7}$$

Here, y_{ijt} is as described above, β_0 represents the intercept term (which 'applies' when FloodEbb=1 and when Hour, XPos, YPos co-ordinates are all zero). Additionally, $FloodEbb2_{ijt}$ and $FloodEbb3_{ijt}$ are 'dummy' variables which are either switched on or off depending on the state of the tide. If the baseline tide ($FloodEbb = 1$) is of interest then $FloodEbb2_{ijt} = 0$ and $FloodEbb3_{ijt} = 0$. However if FloodEbb '3' is of interest then $FloodEbb2_{ijt} = 0$. β_1 and β_2 are coefficients which describe the difference when FloodEbb=1 and FloodEbb=2 respectively.

The smooth term for observation hour, $s_1(Hour_{ijt})$, describes how the response changes throughout the day (Figure 8) and the spatial term $s_2(XPos_{ijt}, YPos_{ijt})$ is a two-dimensional smooth spatial surface which involves the $XPos$ and $YPos$ co-ordinates in a flexible way.



(a) (b)

Figure 8: Observation hour relationships for the GAM (a) and CReSS (b) generated data.

2. Model II: Post-impact decrease

> **Model II** describes a **'decrease'** scenario post impact. Here, animal numbers recorded throughout the whole survey period (pre and post impact) are described using tide state, observation hour, the spatial co-ordinates and impact. There is a decrease 'impact' effect built into the model to induce a 30% decrease in animal numbers post impact (Figure 9).

Model II includes the terms in Model I with an additional 'Impact' coefficient (β_3) to result in a 30% decrease in animal numbers post impact:

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 FloodEbb2_{ijt} + \beta_2 FloodEbb3_{ijt} + \beta_3 Impact_{ijt} \\ &\quad + s_1(Hour_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt})) \end{aligned} \tag{8}$$

Here, β_3 is an 'impact' coefficient and pre-impact has $Impact_{ijt} = 0$ and post-impact has $Impact_{ijt} = 1$. All other terms are as described for Model I.

The pre and post impact surfaces for the near shore scenarios differ more across GAM and CReSS generation than for the off-shore scenarios – there are increased numbers at distance from the observation point in the CReSS manufactured data (Figures 9 and 10; page 33).

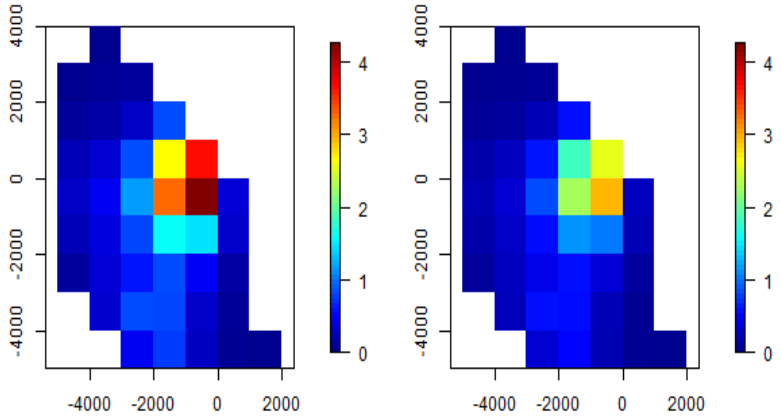


Figure 9: Underlying GAM-generated (smooth) surface pre and post impact for the **decrease** post-impact near-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact.

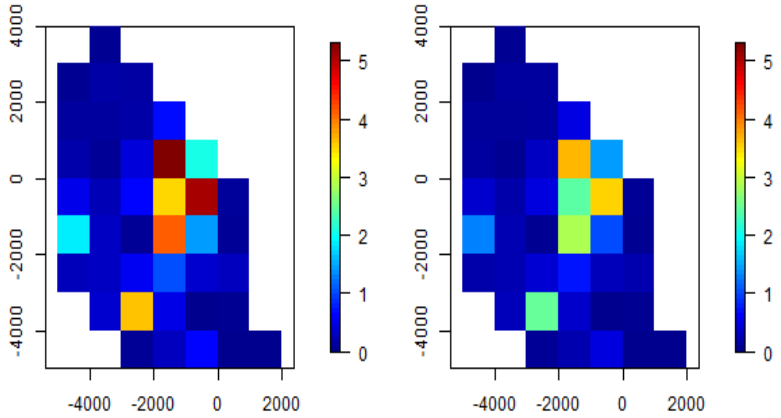


Figure 10: Underlying CReSS-generated (smooth) surface pre and post impact for the **decrease** post-impact near-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact.

3. Model III: Post-impact redistribution

Model III describes a **'redistribution'** scenario post impact. Animal numbers recorded throughout the whole survey period (pre and post impact) are described using tide state, observation hour, and a spatial relationship with animal numbers which differs before and after impact.

There is no overall change in animal numbers pre and post impact.

Model III extends Model II by including a redistribution term post impact. This redistribution term is implemented using an interaction effect between the spatial element and the 'Impact' factor term ($s_3(XPos_{ijt}, YPos_{ijt})Impact_{ijt}$):

$$\begin{aligned} Y_{ijt} &\sim Poisson(\mu_{ijt}, \phi) \\ \mu_{ijt} &= \exp(\beta_0 + \beta_1 FloodEbb2_{ijt} + \beta_2 FloodEbb3_{ijt} + \beta_3 Impact_{ijt} \\ &\quad + s_1(Hour_{ijt}) + s_2(XPos_{ijt}, YPos_{ijt}) + s_3(XPos_{ijt}, YPos_{ijt})Impact_{ijt})) \end{aligned} \tag{9}$$

This interaction term ($s_3(XPos_{ijt}, YPos_{ijt})Impact_{ijt}$)) attracts coefficients describing the difference between the spatial surface pre and post impact. All other terms are as described for Model I and II.

For the redistribution, a radial decrease was applied centrally to the impact region and a concurrent radial increase in an area popular with the animals pre-impact (Figures 11– 14, pages 35 & 36).

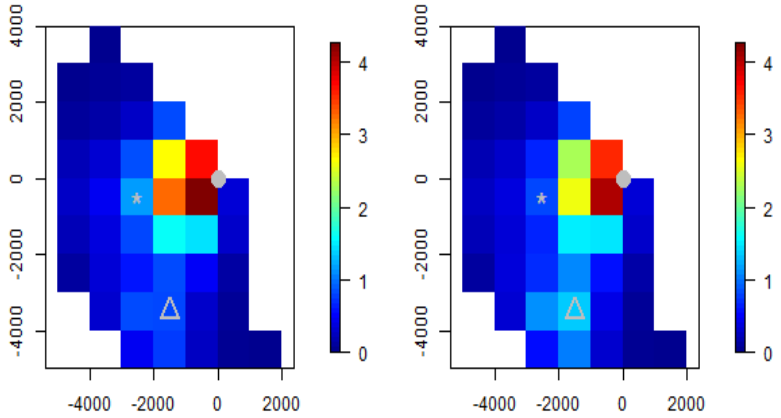


Figure 11: Underlying GAM-generated (smooth) surface pre and post impact for the **redistribution** post-impact near-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The solid grey circle represents the vantage point, the grey asterisk represents the centre of the impact and the grey triangle represents the centre of the site for the redistribution post impact.

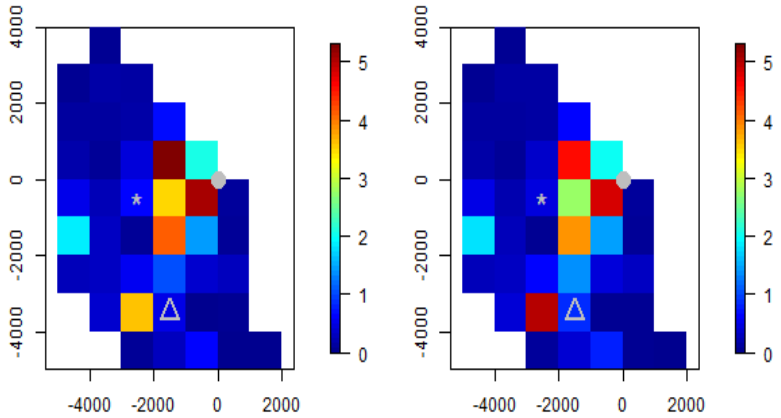


Figure 12: Underlying CReSS-generated (smooth) surface pre and post impact for the **redistribution** post-impact near-shore scenario. The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact. The solid grey circle represents the vantage point, the grey asterisk represents the centre of the impact and the grey triangle represents the centre of the site for the redistribution post impact.

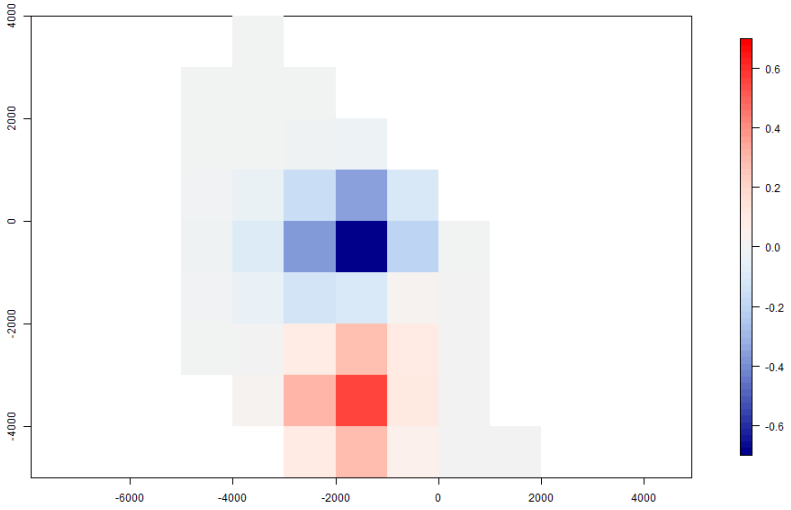


Figure 13: Difference between the underlying GAM-generated (smooth) surfaces pre and post impact for the **decrease** post-impact near-shore scenario. The blue colour represents a post-impact decrease, while the red indicates a post impact increase. The grey colour indicates no change.

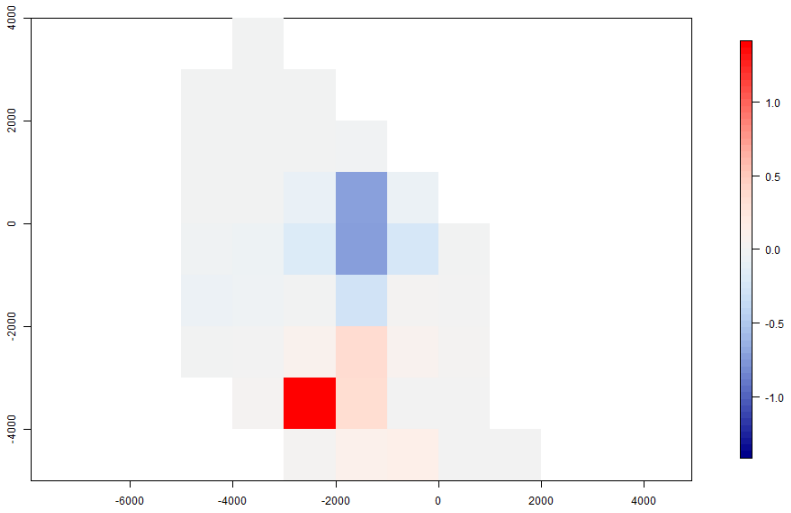


Figure 14: Difference between the underlying GAM-generated (smooth) surfaces pre and post impact for the **decrease** post-impact near-shore scenario. The blue colour represents a post-impact decrease, while the red indicates a post impact increase. The grey colour indicates no change.

5.1.3 Data generation details

Under GAM and CReSS based generation for the off-shore scenarios, 7,025 and 6,992 animals were present before impact (across the 12 months) respectively. For the decrease scenarios, 30% of the animals were removed resulting in a loss of 2,341 and 2,331 animals across the 12 months post-impact for the GAM and CReSS methods respectively. For the redistribution setting, 15% of the animals were moved from the central region to the south east.

Under GAM and CReSS based generation for the near-shore scenarios 26,934 and 11,901 animals were present before impact respectively for the 12 months simulated. Under the decrease scenario, 30% of the animals were removed resulting in a loss of 8,359 and 3,693 animals under the GAM and CReSS generation methods respectively. For the redistribution setting, 2,556 and 1,043 animals (9% of total) were moved from the central region to the south.

It must be noted that it was hoped to include data poor scenarios in the scope of the study, however due to the computing time involved in fitting models using some methods it was not possible to include data-poor scenarios in this piece of work. While we do not anticipate the relative performance of the methods to change with the move from data rich to data poor, this is speculation without additional work.

5.2 *Sampling data from the surface*

For each of the off-shore and near-shore scenarios, models I, II and III were used to generate data from the underlying surface, and a relevant sampling process used to lift points from these surfaces.

It is not realistic to assume that all animals available to be seen by the observers (from the boat/plane/vantage point) are actually seen and recorded with any real distance from the observation point. For this reason, the observation process formed a part of the simulation regime. Unfortunately, this was only possible here for the off-shore scenarios because no information was available about the detection process for the near-shore data. This is almost invariably the case; the data recorded from the vantage point is a mix of the underlying animal distribution surface and the imperfect detection process and there is no way to disentangle these two processes without independent information about the detection process (Cox et al., 2013).

Currently, data about the observation process for vantage point data is rarely (or never) collected and this could affect the results returned from any models. For instance, animals will also be missed from vantage point surveys due to imperfect detection and this is likely to be worse for grid cells farthest from the observation point. It is crucial to collect data about the observation process from vantage point surveys or the user runs the risk of concluding impact-related effects are present if the detection process changes during the survey period.

5.2.1 *Off-shore scenarios*

For the off-shore data, transects were sampled from the surface at 2 km intervals based on the transect spacing of the original survey, and at 0.5 km spacing along the transects. This returned 26 transects in a north-south direction from the study region.

To induce the correlation which is normally present due to covariates related to animal numbers missing from the model, correlation was added directly to the additive predictor on the scale of the link function. This involved generating ar(1) based correlated noise from pre-defined blocks, which were specified to be 208 transect-days for the off-shore data.

As mentioned above, the off-shore data were lifted from the underlying (simulated) surfaces and an imperfect detection process assumed to result in sets of simulated observed counts. For this reason, the observed number of animals (for the manufactured transects) was obtained in the following way:

1. The true number of animals were randomly placed within each grid cell surveyed; the grid cells were specified to be 0.5 km x 0.5 km and thus the maximum distance an animal could be from the trackline was 0.25 km.

2. The perpendicular distance to the transect line was determined for each animal and a detection probability obtained based on a half-normal detection function with a known scale parameter of 120. The furthest possible distance was the half-width of the grid cell, i.e. 0.25 km.

3. Each animal was either deemed to be recorded or overlooked based on a random binomial random process. Here, each animal was considered a trial and the assigned probability of detection was used for the probability of success.

4. In keeping with distance sampling analyses, the data returned by this process comprised the total number of detections for each line segment and the perpendicular distances to the detected animals.

5.2.2 Near-shore scenarios

The near-shore data was sampled 17 hours a day, for 11 days per month over a period of 24 calendar months (12 months each pre and post impact). Additionally, to induce the correlation which is normally present due to covariates related to animal numbers missing from the model, correlation was added directly to the additive predictor on the scale of the link function. This involved generating temporally correlated noise from pre-defined blocks – in this case, grid-cell days; this resulted in 5576 blocks for the near-shore scenarios.

There was no account for imperfect detection in the near-shore scenarios for reasons stated earlier.

5.3 *Evaluation of spatially explicit change*

Environmental impact assessment is largely a spatial question, and the conclusions drawn about any potential impacts should be made with reference to the spatial locations of any post-impact changes.

For instance, it may well be the case that the abundance and distribution of animals changes over time in an area without the introduction of any renewable-related equipment. In these cases, we would expect any spatially explicit changes post-impact to be distributed without major reference to, or at least not restricted to, locations in and around the point of interest. On the contrary, impact-related changes are most likely to manifest in and around the impacted site and thus significant changes centered about the impact site provide more compelling evidence for impact-related effects.

Sound decision making based on geo-referenced changes post-impact relies on models which accurately describe the magnitude of any changes, and in the correct locations. Additionally, the models must be able to distinguish change which is within the bounds of natural fluctuations (noise), from genuine change (the signal) – changes which may or may not be impact related. Furthermore, this signal to noise ratio might well change across the surveyed area in line with (amongst other things) survey effort.

For these reasons, we evaluate the ability of each method to return the magnitude and location of any spatially explicit changes. Additionally, we examine the performance of each method at correctly identifying genuine geo-referenced from background noise.

Quantifying spatially explicit changes pre and post impact involve the following steps:

1. Model selection
2. Generating model predictions and 95% confidence intervals
3. Calculating pre and post impact differences with 95% confidence intervals for these differences

and the methods for carrying these tasks are detailed in sections 5.3.1, 5.3.2 and 5.3.3.

During this process we also evaluate the ability of each method to return the generated surfaces pre and post impact (section 5.3.4), and so even if baseline characterisation is the current focus these metrics tell the user how effective each method is at returning the

generated surfaces – even when the differences across time are not impact related.

5.3.1 Model choice

Choosing appropriate complexity for a model is crucial for good results. Models which are too simple don't characterise the data well and have poor predictive power, while models which are too complex can produce predictions which are very close to the response data but would perform poorly for data unseen by the model (e.g. count data from slightly different transects/sites). Models which are too complex can also make it difficult to estimate model parameters precisely (or at all) which can make assessing impact-related effects problematic.

In quantitative environmental impact assessment, the explicit testing of an impact-related effect (either an overall increase/decrease or a re-distribution) is often one of the main objectives of post-impact analysis. Further, while the nature of the relationship between non-impact related covariates and animal numbers might also be of interest, the inclusion of additional covariates in a model also helps ensure any changes in animal numbers, due to these covariates, are not incorrectly attributed to an impact-related effect.

In the results that follow, the model selection approach which is routinely used for each method was employed where possible. Where alternative approaches are also common for a particular method, this information was also retained for comparison.

There are a variety of approaches used to choose between candidate models containing different covariates and the process of including or excluding covariates from a model typically depends on the estimation routine used to return parameter estimates. For example, information criteria (IC) such as AIC and BIC statistics are routinely used to select model covariates for those based on likelihood estimation (e.g. GLMs and GAMs with Normal, Binomial and Poisson errors), while models which also estimate over-dispersion (e.g. extra Poisson variability) rely on quasi-likelihood estimation and typically include the estimate of this dispersion parameter in the fit statistic (e.g. the QAIC). These IC-based fit scores (AIC, BIC and QAIC) are often used to govern model selection and models with smaller scores (for a given fit score) are favoured.

Choosing model terms based on the p-values associated with each is also a popular method of model selection for models with a small number of candidate covariates. In these cases, a backwards selection approach is often used which involves fitting the full model (with all candidate terms included) and omitting candidate

variables with large p-values one-by-one until the remaining covariates have associated p-values which are smaller than some nominal value (e.g. 0.05 or 0.01).

A p-value based approach can be particularly useful if one or more covariates are of particular importance (e.g. impact-related effects in a post-impact analysis). For instance, IC based selection might select a model which includes an impact-related effect, but the p-value associated with these terms might be relatively large (i.e. a p-value which is larger than some nominal value, e.g. 0.05 or 0.01) and as high as 0.1573 (Atkinson (1980), Lindsey and Jones (1998)). In this case, many will not find these impact-related effects compelling and a p-value based selection scheme (which excludes terms with large p-values) might be more suitable.

In the comparison results which follow, model selection for the candidate GAMs was carried out using the QAIC scores which involved comparing QAIC scores for models I, II and III. Model based p-values were not used in this case since it was thought *a priori* that these would likely be too small (due to the correlation known to be present in the manufactured data) and would therefore lead to the retention of impact-related covariates which weren't genuine. While no assumption was made that the QAIC would be immune to residual non-independence, this is likely to be one of the models selection measures used in practice by analysts in the renewables industry. For comparison with QAIC results, p-value based results were also retained.

GEEs use quasi-likelihood fitting but permits correlated errors and so model selection can proceed based on the QICu statistic (for IC-based selection) or the associated p-values for each term. If model-based p-values based on user specified error structures are desired, then a correlation structure can be chosen using QIC(R) values. In this work however, we chose to use p-values based on the empirical 'robust' estimates of precision which use model residuals more directly, rather than rely on a user-specified correlation model. This option is particularly useful when only a small number of correlation structures can be chosen (inside the software available) and the residual correlation within transects/grid locations is unlikely to follow the available correlation models exactly.

In the comparison results which follow, model selection for the candidate CReSS models was carried out using GEE-based p-values, since these incorporate any within-transect autocorrelation. Specifically, a backwards selection approach based on the impact term was used. Specifically, model III was fitted and if the impact-related interaction effect was significant this was retained (resulting in model III being chosen), otherwise the interaction term was omitted from the model leading to the fitting of model II. If the 'impact' term in model II was then not statistically significant, this term was also omitted from the model resulting in model I being 'chosen'.

The model selection problem is larger for mixed models (e.g.

GLMMs and GAMMs) than for GLMs & GAMs since fixed effects, random effects and (potentially) non-independent error structures must be explicitly specified. For GLMMs and GAMMs which employ ML estimation, the AIC or BIC can be used to choose between models with different fixed effects, random effects and error structures. However, since ML has been shown to give variance estimates which are too small, REML is often used to fit normal-errors models instead. When REML is used, care must be taken not to compare AIC (or BIC) scores across models with different fixed effects (covariates) because AIC/BIC scores should only be compared across models with the same random effects/error structures. In practice rather than use the (RE)ML, GAMMs are often estimated using Penalised Quasi Likelihood (PQL) because there are many often encountered features of real data (such as overdispersion, several random effects, and correlated non-normal errors) which makes estimation by ML either impossible or impractical. Unfortunately, these features are very often present in the process generating the data and some account of these features is required to obtain a realistic model and thus, realistic associated results.

It must be noted that PQL has been shown to give biased estimates (coefficients which are systematically too large or systematically too small) when the mean for Poisson data is low (as is the case here). Despite this, PQL is the most commonly used estimation routine for overdispersed Poisson data fitted using GAMMs. While it is indeed possible to specify the splines manually (e.g. using the `bs` function in R) and use the GLMM fitting engines (which are not necessarily PQL based), this would be unlikely to emulate what has been done in practice by those working in the renewables industry and would therefore shed little light on the reliability of GAMM-based results produced thus far. Further, this more manual approach would involve substantially more time investigating the many smoother-based alternatives considering the time required to fit these methods to the data available here. For these reasons, the widely available `mgcv:gamm` function was used to fit the GAMMs to these data.

Model selection for PQL-based GAMMs is, at best, problematic. There is a great deal of discussion in the literature about which methods should not be used for model selection but there are currently no agreed alternatives, and certainly none which are coded inside readily available software. In this case, the data are overdispersed (in line with the real data) and PQL estimation was necessary if the estimates of precision are to be believed. While the use of PQL estimation makes model selection complicated, GAMMs have been used for analysing baseline monitoring and impact data and are seen by some in the industry as a good way to incorporate the correlation within transects often seen for data of this sort.

The problems with PQL-based model selection in this case are two fold. The first issue is that IC based selection requires a likelihood based measure and PQL is only an approximation to the

likelihood. The second issue is that while p-value based selection might be sufficient for model selection purposes in some situations, in this case p-value based model selection is only possible for Models I and II (no change post-impact and decrease post-impact). For these models approximate Wald tests can be used to compare models with and without each (smooth) term, and in particular, the p-value for the decrease-post impact smooth term is directly supplied.

The inclusion of a smoother-based interaction term in the `mgcv:gamm` function however, supplies p-values with a different interpretation. For the specified interaction term, the model returns p-values associated with testing the smooth components for equality to zero for the spatial surface before and after impact. While this is informative about the nature of the spatial surface before and after impact (i.e. whether it is nonlinear) this does not help the user decide whether to include the interaction term in the model, or not (because the surfaces may be nonlinear and yet identical before and after impact). For some GAMMs an F-test would also be used to compare nested models (e.g. Model I and Model III), however the estimation of the dispersion parameter invalidates this test in this case and the associated testing function fails to return a result.

These issues leave the user with a model selection problem, and so a pragmatic solution was found in this case. For instance, while PQL is only an approximation to a likelihood AIC-style measures have been used to carry out GAMM selection for PQL-based models using the `gamm` function. While it is indeed stated in the `gamm` help files that the log-likelihood reported is not that of the fitted GAMM, comparing alternative models shows some evidence that it may be still appropriate for `gamm` (https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/inst/doc/

gamm.pdf?revision=91&root=mumin&pathrev=91). As a consequence, the `MuMIn` package (also available in `R`) allows model selection based on AIC scores (using a `gamm` wrapper function (http://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf) and we expect this score is what has been used to date in the renewables industry, in the absence of a good, or any, alternative. For this reason, the AIC-style score was used to discriminate between GAMM models fitted to the data manufactured for comparison purposes. In the comparison results that follow, the final GAMMs were chosen using this AIC-style score; models I, II and III were fitted and the AIC scores compared for each model.

In some cases, K-fold cross-validation (e.g. 10-fold CV)[3] might be a useful model selection alternative for GAMMs when practical. However, in this case the length of time taken to fit each model precluded this approach.

[3] This involves folding the data evenly into 10 mutually exclusive sets and using 9 of these to fit the data and 1 of these to 'validate' it. This is done until each set has been used for validation.

5.3.2 *Generating model predictions and 95% confidence intervals*

Predicted animal numbers from the three modelling approaches were projected onto a grid to compare relative model performance across the three methods and to measure model performance in absolute terms; the latter is found by comparing model predictions with 'truth' (the underlying surface).

The uncertainty in these model predictions was also obtained by considering the uncertainty in the detection function fitting process (where applicable) and the uncertainty in the spatial modelling fitting process.

Geo-referenced 95% confidence intervals were generated based on this (combined) uncertainty which should have an associated 'success' rate of 95%. This means the true value (on the grid) should lie somewhere within these intervals 95% of the time and so the proportion of the 100 intervals (for each scenario) that contain the true value was used to assess the reliability of the confidence intervals for each approach.

Model predictions based on the GAM and CReSS models were obtained in the standard way using the estimated coefficients based on the fitted models and the covariate data supplied for the prediction grid. Obtaining predictions from the GAMM required some post processing however, because estimated random effects are only available for the blocks/transects observed and predictions were required on a grid. This involved interpolating between transects. Additionally, mixed effects models are conditional models and we are interested in predicting average animal numbers for a population of blocks/transects onto the grid (to give predictions comparable with the other two population-averaged methods) and so there was a need to 'marginalise' model results. This marginalisation process for the 'random intercept' GAMM involved calculating the average predicted value (for each grid cell) based on sampling 5000 random effects from the random effects distribution with mean zero and the estimated standard deviation under the model.

The 95% confidence intervals for each grid cell were obtained by combining the uncertainty at the detection function fitting stage (in the case of the off-shore scenario results) and the uncertainty in model parameters at the model fitting stage (for both the off-shore and nearshore scenarios). For the off-shore scenarios, the blocks/transects were resampled (with replacement) and the detection function re-fitted each time to the resampled data using the detection function chosen using the actual data. In each case, the abundance estimates for each block were used as input to the model chosen (using the actual data), the model refitted, and a set

of predictions based on a set of parametric bootstrap coefficients were obtained. Percentile based 95% confidence intervals for the 100 sets of the predictions obtained for each cell were then calculated.

Model uncertainty was not considered as a part of this process; the model chosen for the detection function and spatial model (with additional covariates) was used in this procedure rather than chosen each time as a part of the resampling process. While this would make the confidence intervals at least as wide, a pilot study using the GAM method showed that model choice almost never changed at the spatial modelling stage using bootstrap replicates from the detection function process. This suggests that GAMs were stable with respect to model choice even given perturbations in the input data after correcting for imperfect detection.

5.3.3 Spatially explicit post-impact differences

The difference between model predictions pre and post impact and 95% confidence intervals for the pre/post impact differences were calculated for each cell on the prediction grid using the fitted models in each case.

Each method was also used to generate model predictions and associated 95% confidence intervals (section 5.3.2) for each grid cell. This process was followed for each of the 100 realisations and so 100 confidence intervals were available for each grid cell for each method type. The proportion of the confidence intervals in each case that contain the true value (based on the known surface) was calculated.

Therefore, if a high proportion of intervals contain zero (indicated by the red colour in the associated results; section 6.3) and thus 'no-difference' is a plausible value for the pre/post impact difference, then there are few differences identified under the model. Conversely, if a low proportion of intervals contain zero (indicated by the blue colour in the results; section 6.3) then a large number of significant differences pre and post impact are detected.

5.3.4 Evaluating model fit

In this piece of work, the data were simulated from a known pre and post impact surface and we are able to compare model predictions from three different methods with the underlying surface values. In practice however, the surface is always unknown and the user needs to rely on the agreement between the input data and each set of model predictions to discriminate between candidate models. For this reason, we will outline some evaluation methods

measuring the agreement between each set of model predictions and both the underlying surfaces and the modelled data.

The discrepancy between the underlying surface(s) and the model predictions were quantified using the Mean Squared Error (MSE). This is essentially a measure of overall lack-of-fit and thus lower MSE scores indicate better model performance. A statistical test was also employed to test for genuine differences in model performance since equivalent modelling approaches will never return exactly the same predictions and thus slightly different MSE scores.

The lack-of-fit scores (MSEs) were also viewed spatially to examine model performance in key areas (such as the impact and redistribution zones) across the three methods.

Additionally, the discrepancy between the input data and model predictions (as opposed to the underlying surface and model predictions) were viewed spatially to examine if these reflect the spatial position of the larger lack-of-fit scores. This is useful because the underlying surface is never known, and the user must rely on the discrepancies between the input data and model predictions to assess the available model(s).

The performance of each method was evaluated by quantifying the sum of the squared differences[4] between model predictions and the corresponding values from the underlying surface (i.e. from Model I, II or III). The mean across the 100 sums of squared differences was calculated (Mean Squared Error; MSE) for the 100 realisations to give an overall measure of difference for each model type.

[4] using the squared differences ensures positive and negative differences count equally

The fidelity to the underlying function was also compared statistically across two models for each realisation using a pairwise test. Specifically, a Wilcoxon paired signed rank test (Wilcoxon, 1945) was used to determine if there is a difference in the average MSE score between the methods, considering two at a time; e.g. GAMs vs GAMMs, GAMs vs CReSS and CReSS vs GAMMs. This test is non-parametric and uses both the magnitude and sign of the paired difference ranks to determine the difference. This tests for no-difference in the average MSE scores between two methods and a small p-value for each comparison means there is compelling evidence for a difference between the two approaches under consideration.

The MSEs can also be viewed spatially, to inspect the spatial locations of these squared differences for each model type. Here, the range of the MSEs for each grid location can be compared across methods and we can visually examine if the discrepancies between the underlying function and model predictions are tightly clustered

or widespread, and in particular their location with reference to the impacted site.

In reality, the underlying surface is unknown and so the user can only make a spatial examination of the input data compared with model predictions (i.e. residual plots). However this comparison should still be valuable if the input data is a good reflection of the underlying surface. For this reason, we present some examples here with reference to the underlying surface (which is unknown to the user) and the input data available to the user.

The agreement between the input (response) data (pre and post impact) and the models undergoing evaluation can be quantified using numerical measures. Two such measures are the marginal R^2 and the concordance correlation; the former of which is widely used to assess models fitted to independent and correlated data, while the latter is recommended for correlated data.

The marginal R^2 (R^2_{MARG}) assesses the predictive power of the model, and can be written as:

$$R^2_{MARG} = 1 - \frac{\sum_{i=1}^{s}\sum_{j=1}^{J}\sum_{t=1}^{n_i}(y_{ijt} - \hat{y}_{ijt})^2}{\sum_{i=1}^{s}\sum_{j=1}^{J}\sum_{t=1}^{n_i}(y_{ijt} - \bar{y}_{..})^2} \tag{10}$$

Here, $\bar{y}_{..}$ is the mean of the response data across all correlated blocks and time points.

The concordance correlation (r_c) can also be used to compare the response data to the fitted model and measures the agreement of two sets of values (e.g. the input data (y_{ijt}) and fitted values from the model ($\hat{y}_{ijt}$)) and is constrained to give values between zero and one:

$$r_c = \frac{2\sum_{i=1}^{s}\sum_{j=1}^{J}\sum_{t=1}^{n_i}(y_{ijt} - \bar{y}..)(\hat{y}_{ijt} - \bar{\hat{y}}_{..})}{\sum_{i=1}^{s}\sum_{j=1}^{J}\sum_{t=1}^{n_i}(y_{ijt} - \bar{y}_{..})^2} \tag{11}$$

Here, $\bar{\hat{y}}_{..}$ is the mean of the fitted values across all segments/grid locations ($i = 1, ..., s$), transects/days ($j = 1, ..., J$) and time points ($t = 1, ..., n_i$) and the remaining components are as described for R^2_{MARG}. A value close to one indicates good agreement between the input data and the model and thus, that the model fits the data well.

6 *Model comparison results*

The methods of evaluation were both numerical and visual to give the reader an appreciation of overall and geo-referenced model performance. There were five measures used to assess model performance:

1. Model choice (section 6.4)
2. Fit to the underlying process: mean squared error (section 6.5)
3. Spatially explicit bias: assessing the accuracy of the geo-referenced predictions (section 6.2)
4. Spatially explicit coverage: assessing the reliability of the reported geo-referenced precision (section 6.3)
5. Spatially explicit post-impact differences (section 6.1)

and the qualitative assessment of each method is provided in section 7

Choosing between model types and methods using standard fit measures (e.g. marginal R^2 and concordance criterion) is also discussed with reference to the processes known to be generating the data.

6.1 *Spatially explicit post-impact differences*

In this section, the ability of each method to detect spatially explicit change is assessed. The power of each method to detect genuine *overall* change in animal numbers and the prevalence of each method to falsely detect overall change is assessed in section 6.4.

CReSS and GAMMs produced broadly similar difference surfaces which were close to the manufactured data, however CReSS was clearly superior to the other methods at detecting genuine pre/post impact changes. Further, when CReSS falsely detected statistically significant change, the magnitude of these changes was so small as to remove any practical concern for the user.

GAMMs demonstrated some ability to detect spatially explicit change but the power to do so was often substantially poorer than CReSS.

GAMs generally produced over-fitted difference surfaces (which falsely indicated increases and decreases across the surface), however these differences were often not statistically significant even when the changes were genuine. For this reason, GAMs often demonstrated very low power at detecting spatially explicit change.

6.1.1 *Off-shore results*

The estimates for the pre/post impact differences and the prevalence of statistically significant pre/post impact differences are shown on pages 54– 60. For these figures, the plots on the left represent the pre and post change in the predicted number of animals (and a positive difference indicating more animals post impact), while the plots on the right represent the proportion of confidence intervals for these differences which include zero. Therefore, if a high proportion of intervals contain zero (indicated by the red colour in the associated figures) and 'no-difference' is a plausible value for the pre/post impact difference, then few statistically significant differences are identified under the model. Conversely, if a low proportion of intervals contain zero (indicated by the blue colour in the associated figures) then a large number of significant differences aross the realisations are detected under the model.

For the no-change scenarios, we expect the left-hand plots in Figures 15–19 to be green in colour (indicating minimal predicted change), and the right-hand plots to be red in colour in line indicating non-significant change (with 0.95 on the response scale).

For the no-change scenarios, the GAM difference surfaces (the point estimates for the pre-post impact differences) exhibited both positive and negative change in inappropriate places (Figures 15

and 18) however, these were not typically statistically significant. The GAMMs also predicted non-negligible change in inappropriate places, however these were rarely reported as statistically significant (Figures 17 and 20). The CReSS surfaces indicated negligible change post impact across the whole survey area (in keeping with the manufactured data), and when these changes were statistically significant the tiny magnitude of these differences was unlikely to cause any practical concern (Figures 16 and 19).

For the decrease scenarios, correct results would be signalled by blue colouration in both the left and right-hand plots in Figures 21–25 (demonstrating a decrease post-impact) and a small percentage of confidence intervals which contain zero (0.05 on the response scale).

For the decrease scenarios, the GAMs predicted increases, rather than decreases, post impact and these were largely in the south of the survey area (Figures 21 and 24).

The CReSS and GAMM difference surfaces were closer to the truth, however the decreases were largely centered about the impact and redistribution areas (Figures 22, 23, 25 and 26). While the CReSS and GAMMs difference surfaces were similar, CReSS located the genuine decreases with higher success (40%-80%; Figures 22 and 25).

The GAMs and GAMMs did a substantially poorer job of identifying the differences post-impact across the surface (Figures 21, 23, 24 and 26).

For the redistribution scenarios, we expect the left-hand plots in Figures 27–29 and 32–34 to be largely green in colour with blue colouration (representing a decrease) in and around the impact area and red colouration (representing an increase) in the redistribution area. The right-hand plots should show largely red colouration excepting blue areas in the impact and redistribution areas. For reference, the true pre-post impact differences are shown in Figures 30 and 31.

For the redistribution scenarios, the GAM-based difference results were highly variable and while it located decreases in the centre of the area and increases at the redistribution site, it also predicted other changes that weren't genuine (Figures 27, 30, 31 and 32), although some of these weren't statistically significant.

In contrast, the CReSS and GAMM surfaces were closer to the truth and located the decreases and increases in the appropriate areas (Figures 28, 29, 33, 34, 30 and 31).

The identification of significant differences across the surface was fairly similar across the three methods, however CReSS exhibited markedly better ability at detecting spatial change for both the decreases at the impact and the increases at the redistribution sites (Figures 28 and 33). Alongside this ability to detect genuine change, CReSS falsely detected differences about 30% of the time in the south of the area. However, the estimates for these differences were effectively zero and so are not practically important (Figures 28 and

33).

6.1.2 Near-shore results

CReSS was clearly superior to the other methods at detecting genuine post impact change and yet showed no tendency to falsely detect change (Figures 36 and 39). The GAMM difference surfaces were sometimes similar to the CReSS results (Figures 46, 49 and 54), however the GAMMs exhibited worse power to detect change when it was present, particularly for the post-impact decrease scenario (Figures 43 and 46). The GAMs tended to identify significant changes which weren't genuine.

For the no-change scenarios, we expect the left-hand plots in Figures 35–40 to be green in colour (indicating minimal predicted change), and the right-hand plots to be red in colour in line indicating non-significant change (with 0.95 on the response scale).

For the no-change scenarios, the GAM difference surfaces (the point estimates for the pre-post impact differences) exhibited both positive and negative change in inappropriate places (Figures 35 and 38) which were often statistically significant. The GAMMs also predicted non-negligible change in inappropriate places, however these weren't commonly reported as statistically significant (Figures 37 and 40). The CReSS surfaces indicated negligible change post impact across the whole survey area (in keeping with the manufactured data), and these changes were rarely, if ever, statistically significant (Figures 36 and 39).

For the decrease scenarios, correct results would be signalled by blue colouration in both the left and right-hand plots in Figures 41–46 (demonstrating a decrease post-impact) and a small percentage of confidence intervals which contain zero (0.05 on the response scale).

For the decrease scenarios, the GAMs tended to predict either no-change or increases, rather than decreases post-impact, and these increases were restricted to the south of the survey area (Figures 41 and 44). The CReSS surfaces were closer to the truth, however the decreases were predicted to be uneven and larger in the impact and redistribution areas. The CReSS method performed exceptionally well at detecting significant decreases across the survey area, particularly for the CReSS-based generated data (Figures 42 and 45). The GAMM difference surfaces failed to detect the decreases in most locations and when identified by the model, the decreases were almost never significant for either the GAM or CReSS generated data (Figures 43 and 46).

For the redistribution scenarios, we expect the left-hand plots in Figures 47–54 to be largely green in colour with blue colouration (representing a decrease) in and around the impact area and red colouration (representing an increase) in the redistribution area. The right-hand plots should show largely red colouration excepting blue areas in the impact and redistribution areas. For reference, the

true pre-post impact differences are shown in Figures 50 and 51.

The GAMs tended to detect the increases in the redistribution area with more success than the decreases near the impacted site (Figures 47 and 52) but commonly identified significant change in the central part of the surveyed area (including areas where no change occurred). CReSS tended to return surfaces which were closer to reality and tended to return the decreases in the impact zones and associated increases in the redistribution area. However, CReSS was better able to detect the increases in the redistribution area as significant, compared with the decreases in the impact site (Figures 48 and 53). GAMMs methods returned difference surfaces that were very similar to CReSS and the true redistribution however, GAMMs exhibited worse power at detecting change for the redistribution scenarios (Figures 49 and 54).

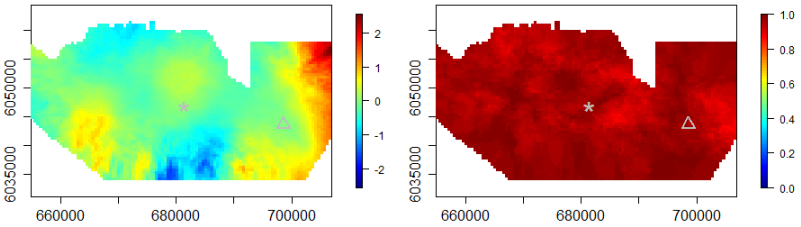


Figure 15: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

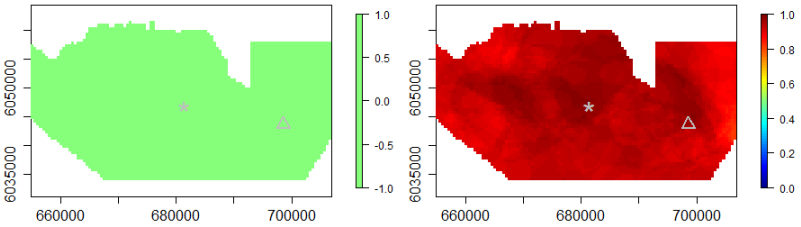


Figure 16: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

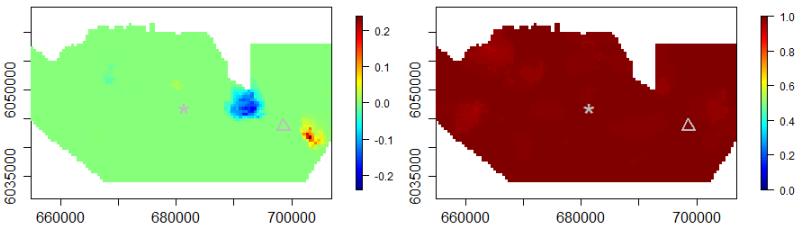


Figure 17: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

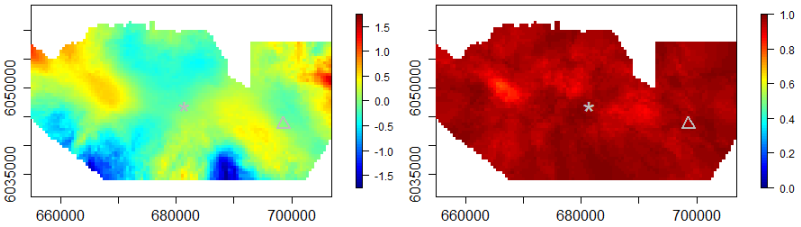


Figure 18: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

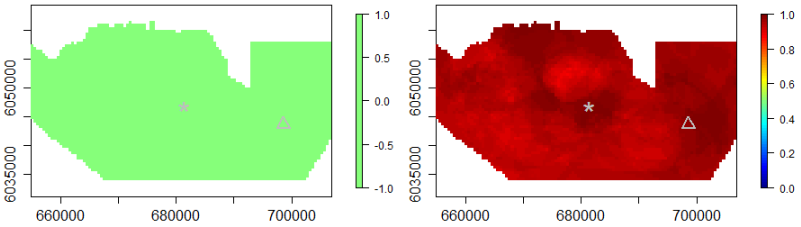


Figure 19: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

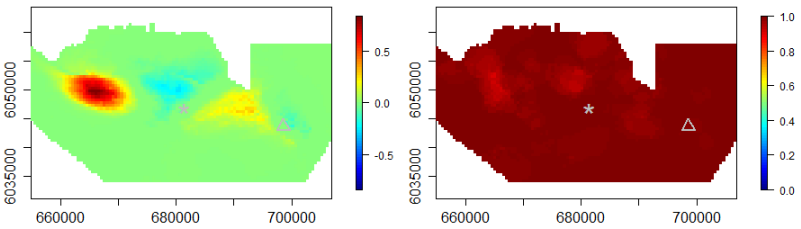


Figure 20: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

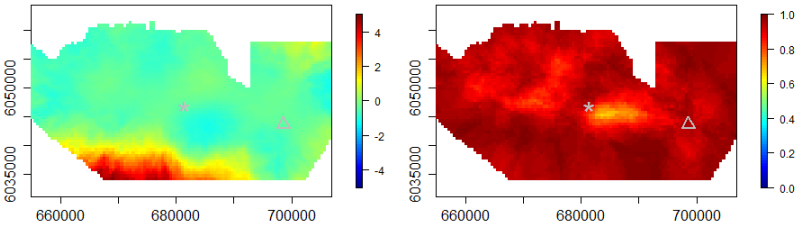


Figure 21: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

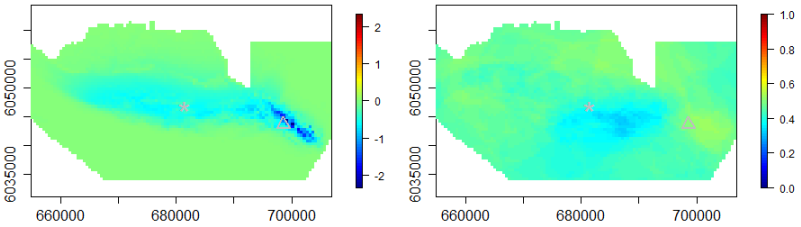


Figure 22: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

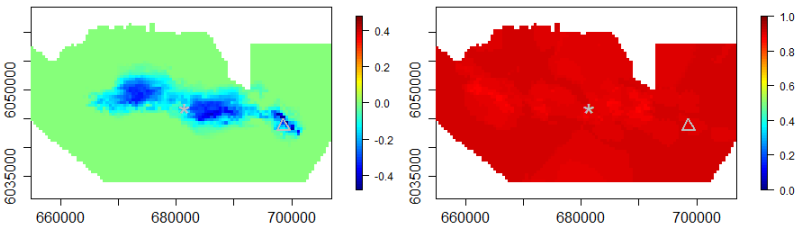


Figure 23: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

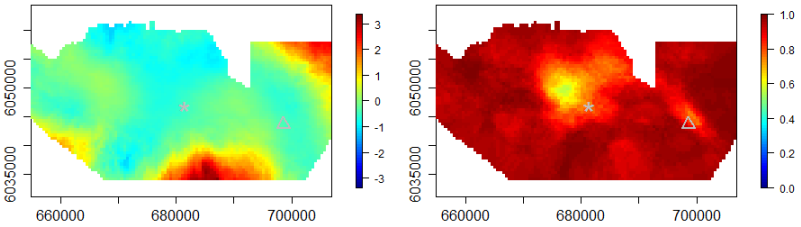


Figure 24: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

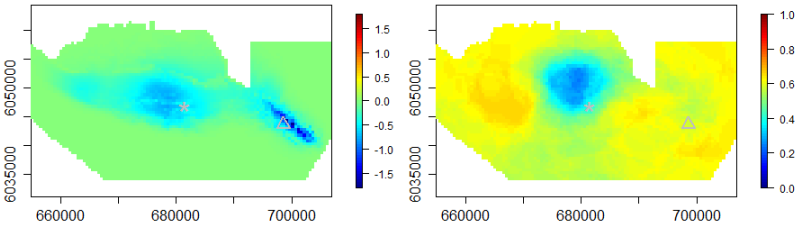


Figure 25: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

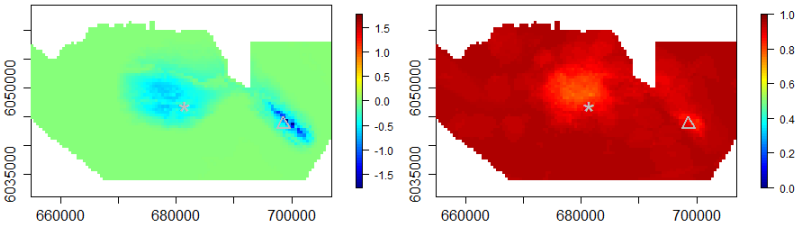


Figure 26: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

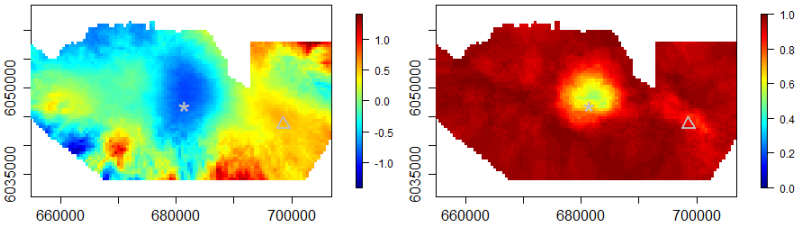


Figure 27: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

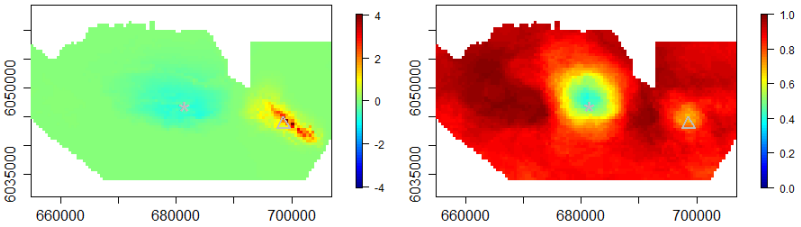


Figure 28: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

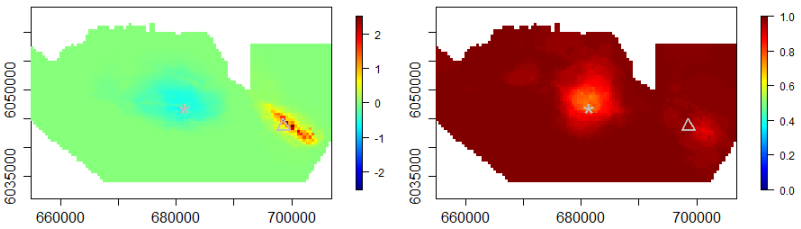


Figure 29: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

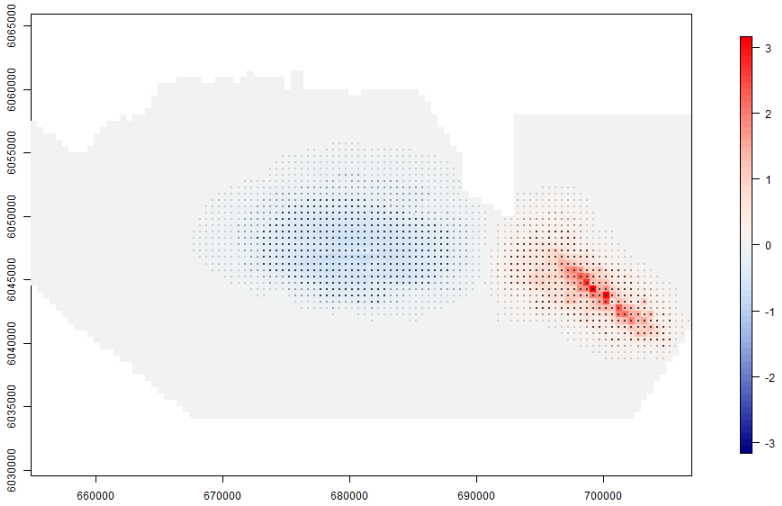


Figure 30: Difference plot showing the pre/post impact redistribution differences for the GAM generated data.

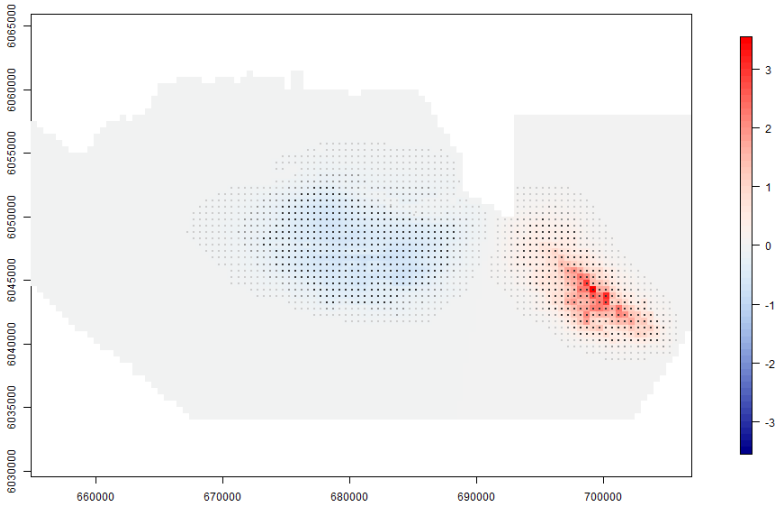


Figure 31: Difference plot showing the pre/post impact redistribution differences for the CReSS generated data.

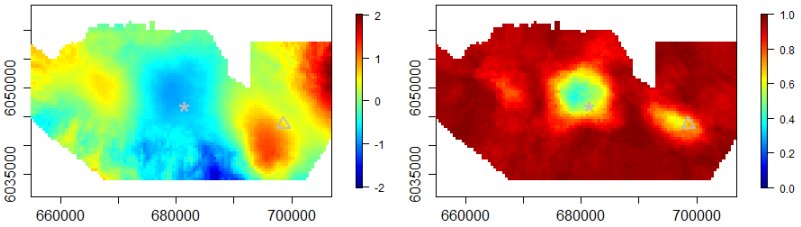


Figure 32: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

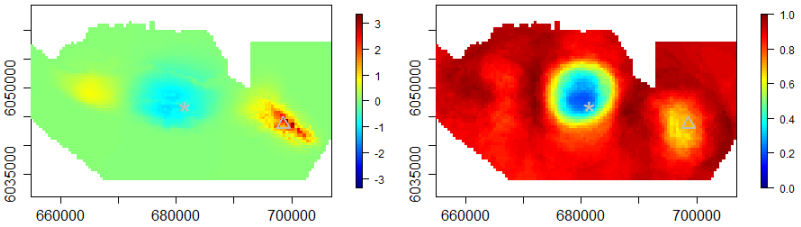


Figure 33: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

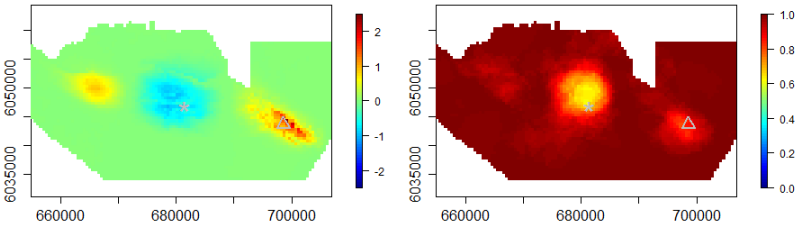


Figure 34: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

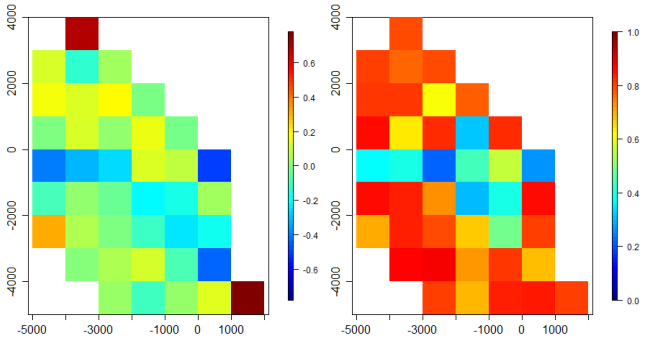


Figure 35: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

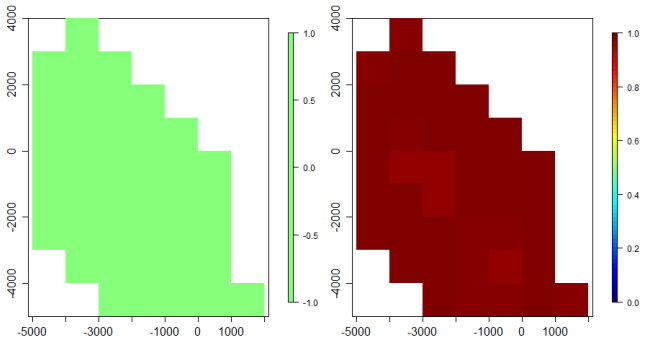


Figure 36: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

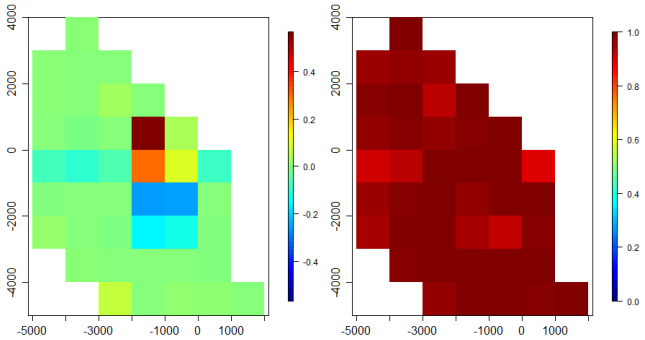


Figure 37: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with **no-change** post impact (Model I).

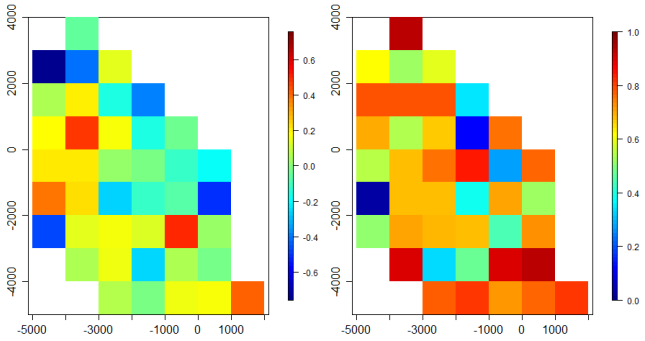


Figure 38: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

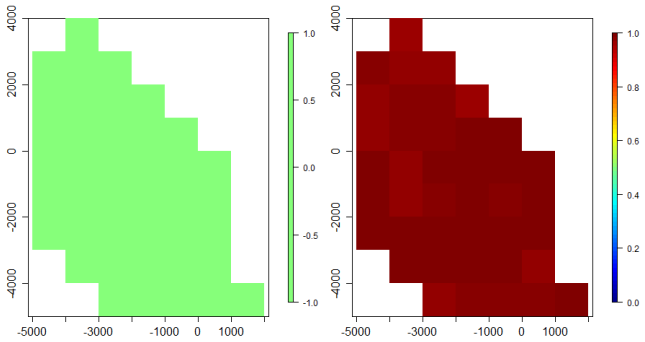


Figure 39: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

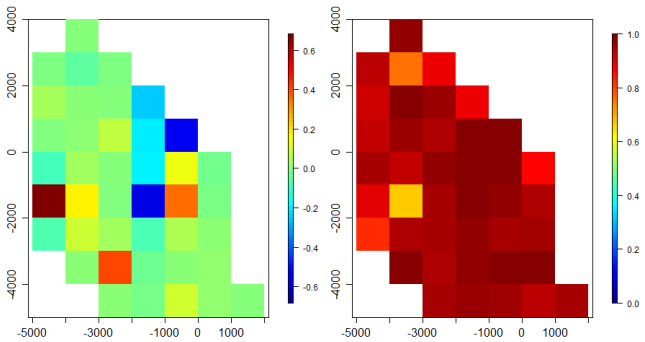


Figure 40: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with **no-change** post impact (Model I).

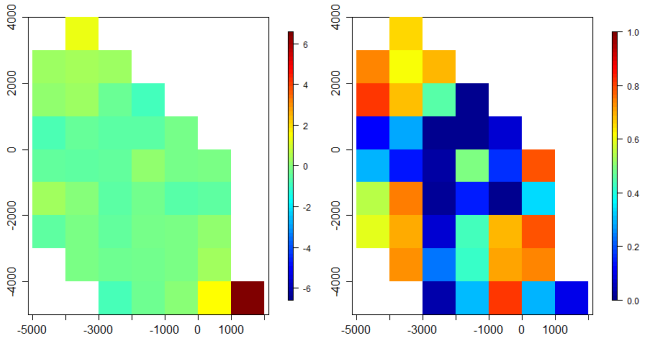


Figure 41: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

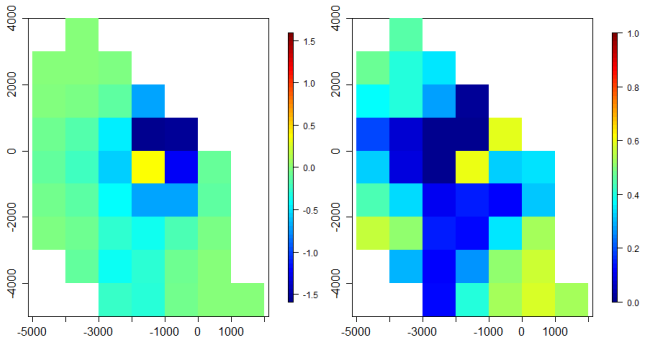


Figure 42: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

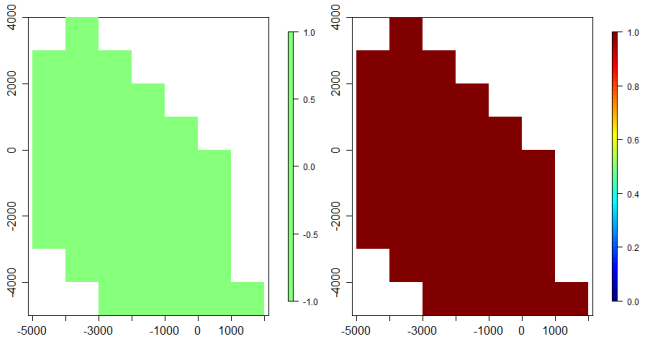


Figure 43: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **decrease** post impact (Model II).

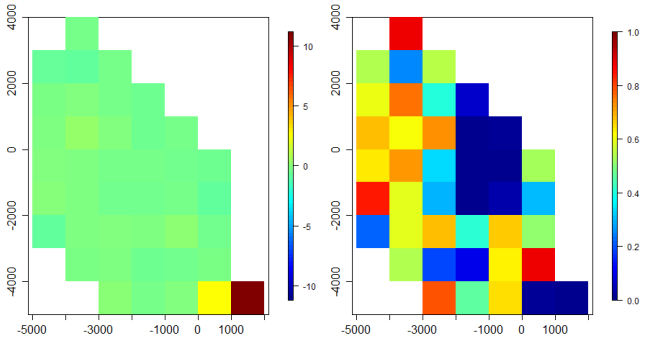


Figure 44: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

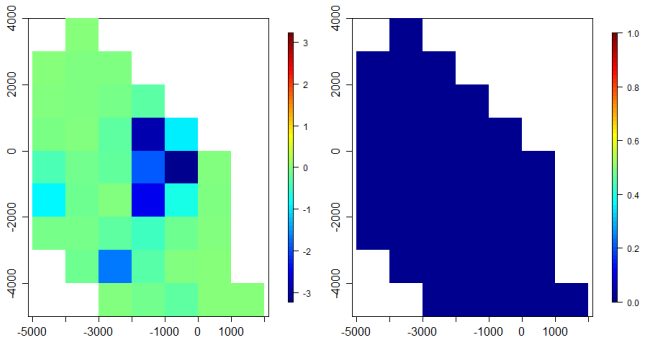


Figure 45: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

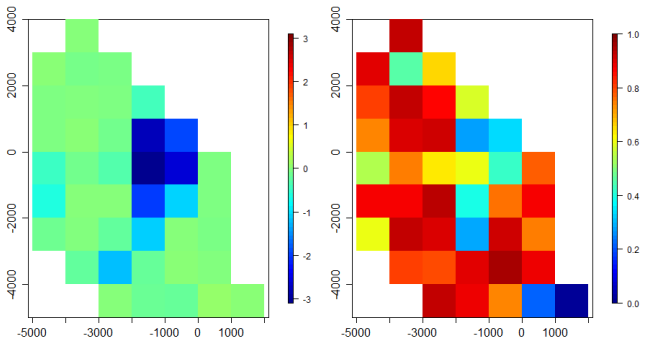


Figure 46: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **decrease** post impact (Model II).

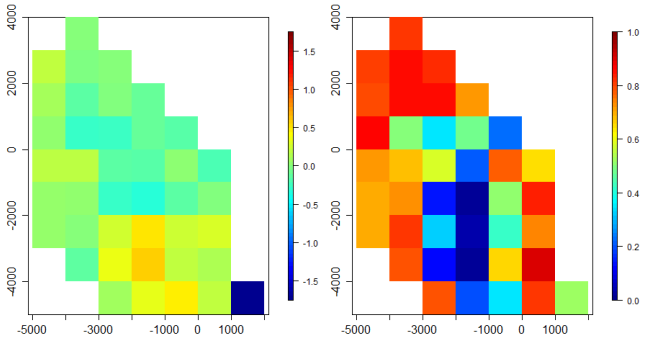


Figure 47: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

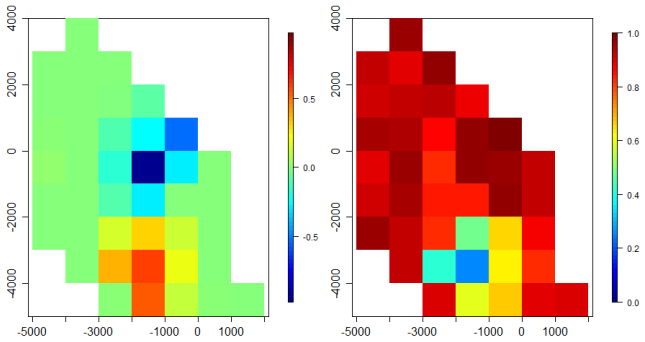


Figure 48: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

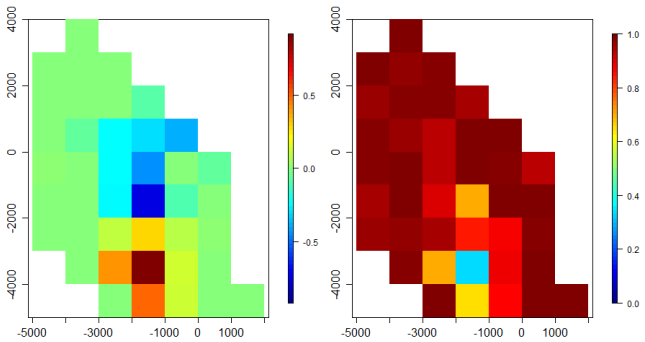


Figure 49: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the GAM generated data with a **redistribution** post impact (Model III).

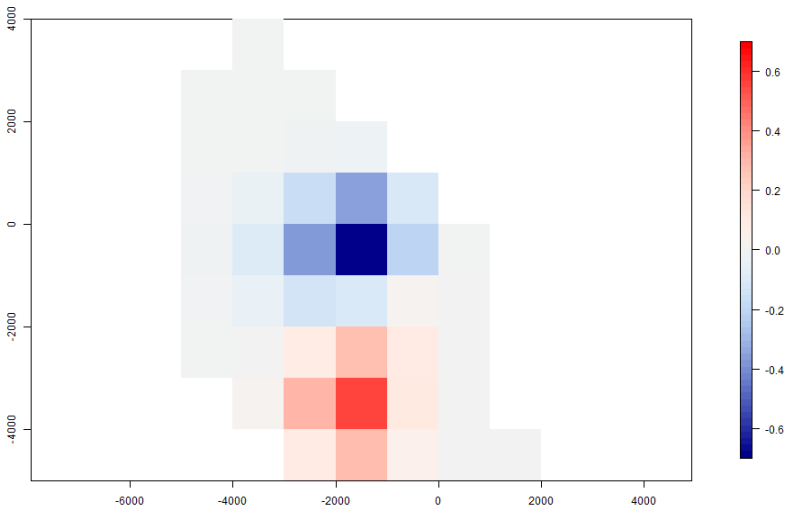


Figure 50: Difference plot showing the pre/post impact redistribution differences for the GAM generated data.

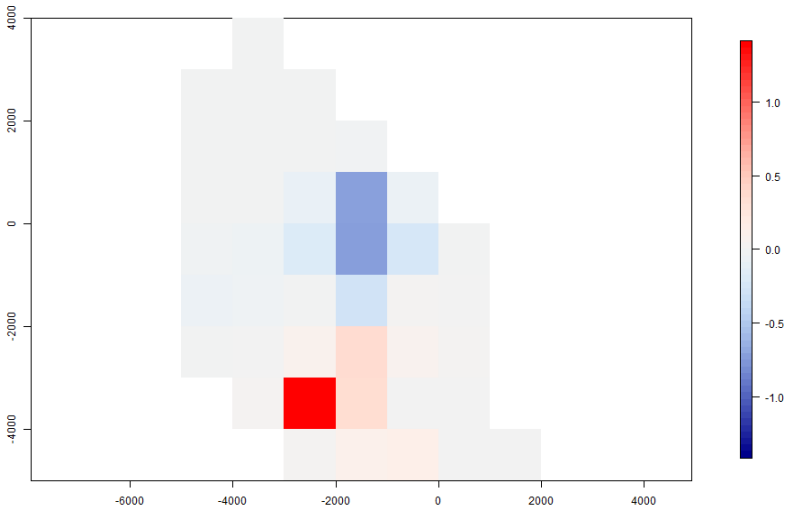


Figure 51: Difference plot showing the pre/post impact redistribution differences for the CReSS generated data.

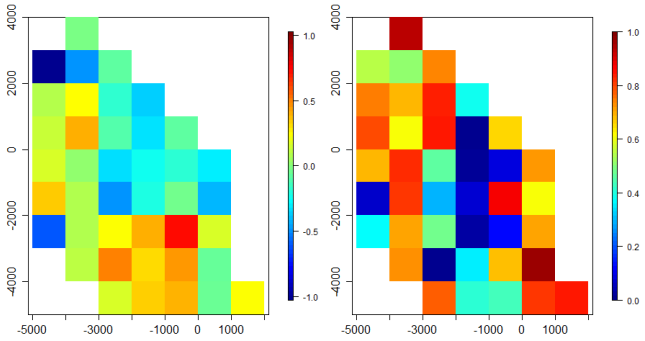


Figure 52: GAM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

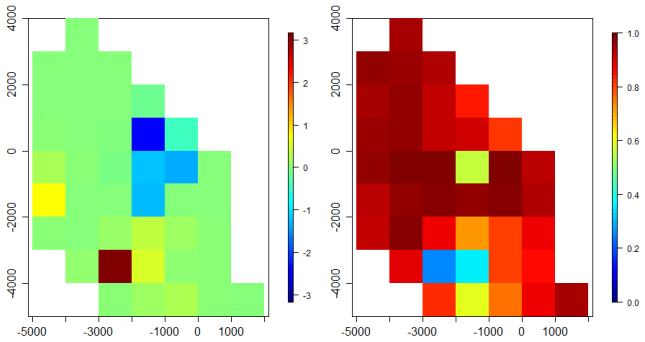


Figure 53: CReSS-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

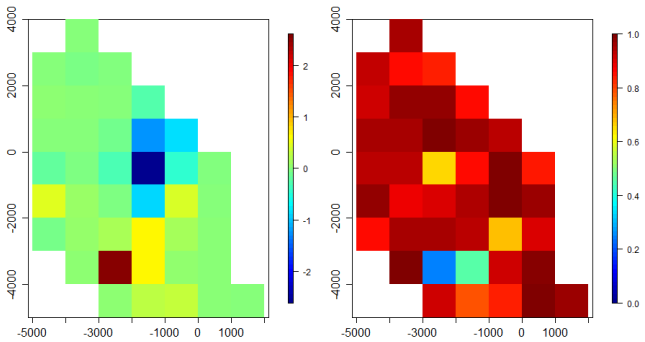


Figure 54: GAMM-based point estimates for the pre/post impact difference (left-hand plot) and the proportion of the 95% confidence intervals which include zero (right-hand plot) for the CReSS generated data with a **redistribution** post impact (Model III).

6.2 *Spatially explicit bias: assessing the accuracy of the geo-referenced predictions*

The differences in the location and magnitude of the spatially explicit bias between the competing methods for the off-shore results were small, and thus there was little to choose between them. There was, however, substantial bias in the near-shore results for the GAMM method, suggesting GAMMs were a particularly poor choice for the near-shore scenarios.

Reassuringly, the spatial locations of the discrepancies between model predictions and the data were similar to the differences between model predictions and the underlying surface. This illustrates that visualising geo-referenced residuals can be a good way to examine if any hotspots exhibited by a model are simply artefacts of an ill-fitting model, or genuinely features in the data.

Please note: using residual plots alone to choose a model however, would be ill-advised because an overly complex model would fit the data more closely than other models/methods. While this delivers smaller differences between the data and the model, it is not necessarily an accurate reflection of the underlying surface (e.g. the GAM results for the off-shore scenarios).

6.2.1 *Off-shore results*

The GAM bias maps look similar to the CReSS and GAMM equivalents (e.g. Figures 188–190), however the GAMM bias appeared to be more widely dispersed, while still being centered about the impact and redistribution zones. A full range of bias maps can be found in section 12.3).

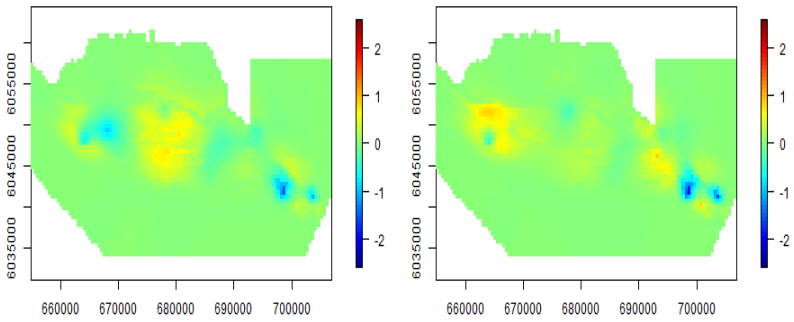


Figure 55: GAM based average bias for CReSS-generated data pre and post impact for the **redistribution** off-shore scenario (Model III). The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact.

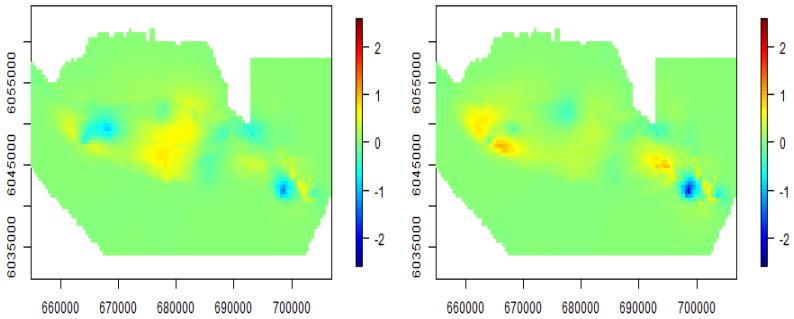


Figure 56: CReSS based average bias for CReSS-generated data pre and post impact for the **redistribution** off-shore scenario (Model III). The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact.

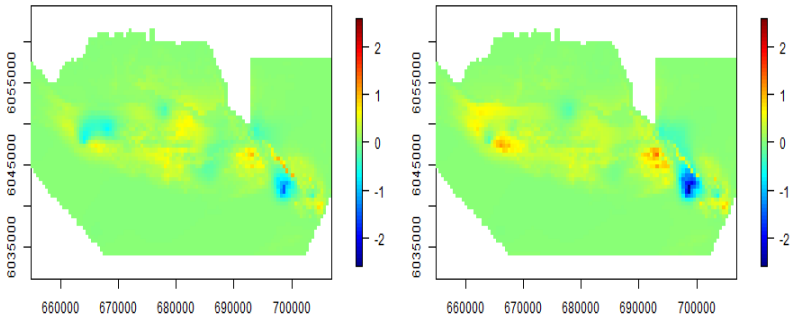


Figure 57: GAMM based average bias for CReSS-generated data pre and post impact for the **redistribution** off-shore scenario (Model III). The surface on the left is pre-impact (baseline data) while the right-hand surface is post impact.

The location of the largest model residuals was very similar to the location of the bias, however there was very little difference in the magnitude and location of model residuals across the three methods undergoing comparison (section 12.6). In general, the residuals for the no-change, decrease and redistribution scenarios (under either the smooth or more flexible surface) were largest in and around the centre and south-west of the survey area (e.g. Figure 58).

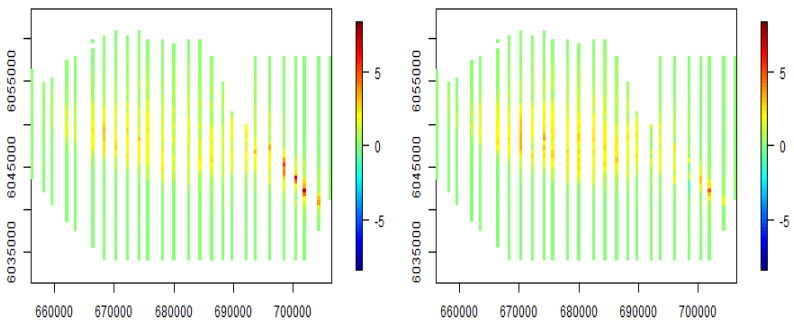


Figure 58: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

6.2.2 *Near-shore results*

All methods tended to overpredict animal numbers located between the impact site and the vantage point (Figures 59– 61) but GAMM bias was more prominent in this area (Figure 61). The GAMM bias was also typically more localised than the other methods (near the impact site) and the bias was more widely dispersed for all three methods under the redistribution scenarios (Model III), particularly for the smoother GAM-generated surface (Figure 61). A full range of bias maps for the near shore data can be found in section 12.4.

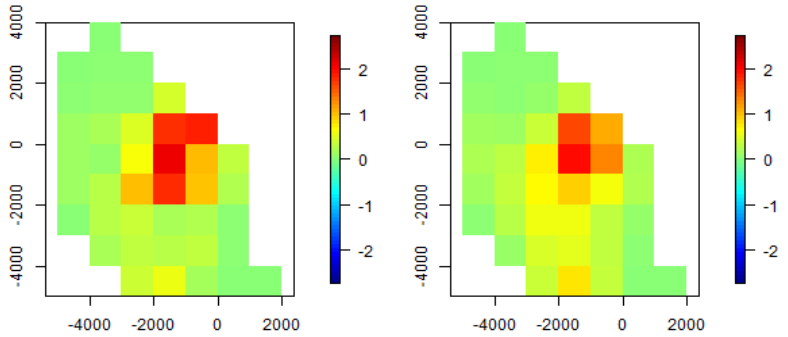


Figure 59: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean bias while the right-hand plot represents post-impact mean bias.

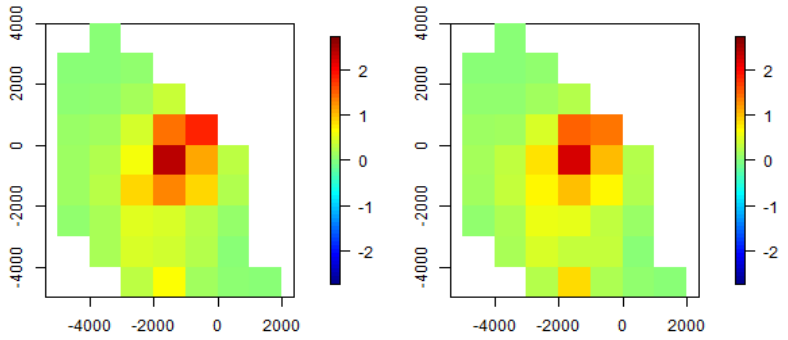


Figure 60: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean bias while the right-hand plot represents post-impact mean bias.

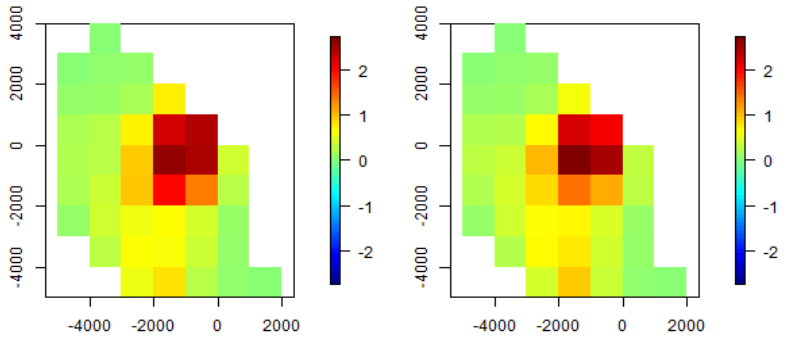


Figure 61: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean bias while the right-hand plot represents post-impact mean bias.

The relatively high bias in the GAMMs (compared with other methods) is also exhibited in the model residuals; the otherwise small residuals seen with other methods (e.g Figure 62) are much larger for the GAMM models, particularly in and around the impact site (e.g. Figure 62).

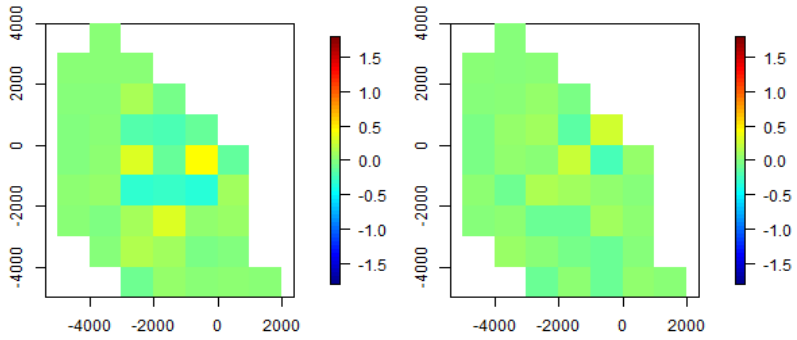


Figure 62: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

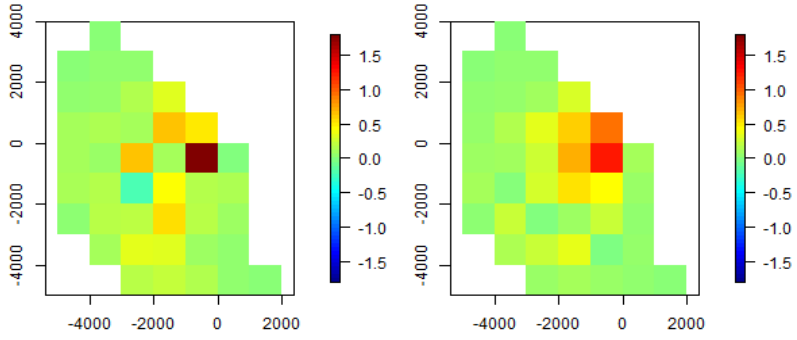


Figure 63: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

6.3 *Spatially explicit coverage: assessing the reliability of the reported geo-referenced precision*

The results for the three methods were mixed across the off-shore and near-shore scenarios. GAMMs demonstrated the best coverage rates for the off-shore results, however CReSS performed better than GAMMs for the near shore scenarios. GAMs exhibited particularly poor coverage around the edges of the survey area for the off-shore scenarios (due to confidence intervals which were too narrow) but this improved for the near-shore scenarios.

As expected, all methods performed badly in the parts of the surface where model predictions were prone to being systematically too high, or systematically too low (i.e. in locations with high bias)

Poor geo-referenced coverage can be because the associated confidence intervals are unbiased but too small, or due to bias in these locations. For instance, in high-bias areas the centre of the confidence intervals are systematically too high or too low and so the true value likely lies outside the confidence intervals unless they are particularly wide. Additionally, since there is no reason to expect that areas of the surface with high bias will also have high variance, the coverage plots are best considered alongside the bias maps.

6.3.1 *Off-shore results*

The GAMs exhibited confidence intervals which were too narrow in areas to the south and east of the survey area (with coverage probability ≈ 0.5), however the often pronounced poor coverage to the south of the survey area improved for the post-impact surfaces (e.g Figure 64 and section 12.7). The CReSS coverage for the smoother surfaces (for Models I, II and II) was largely good except in the far-east along the edge of the survey area (e.g. Figure 65). We expect this may be due to the 'edge' effect generated under the GAM surface which is difficult to approximate using local radial basis functions available to CReSS - though this is speculation without further work.

The coverage for the more flexible surfaces (generated using CReSS) was more patchy under all approaches, although this patchiness was more severe for CReSS and GAMM based models (e.g. Figures 67 – 69). In particular, the CReSS and GAMM approaches returned coverage rates that were very low (with coverage probability ≈ 0.2) in areas of high bias. The GAMM approach tended to produce better coverage rates overall compared with the other methods (e.g. Figure 66).

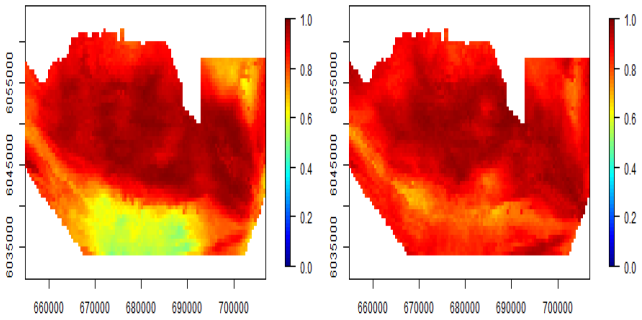


Figure 64: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

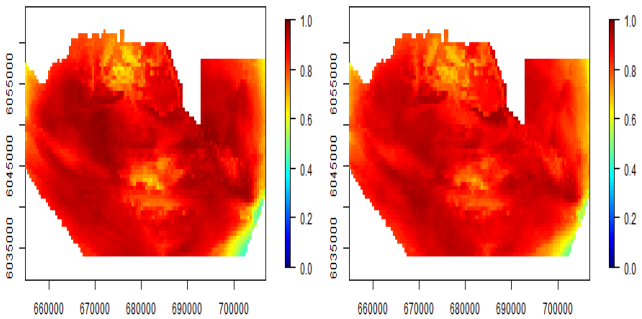


Figure 65: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

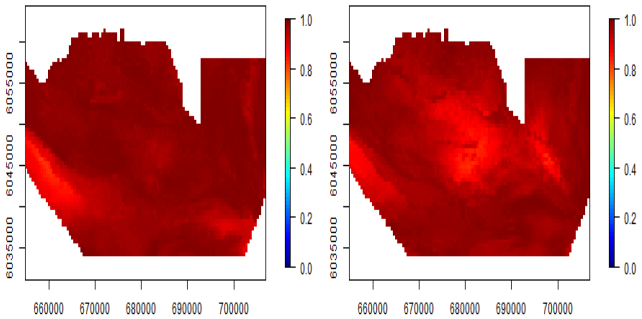


Figure 66: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

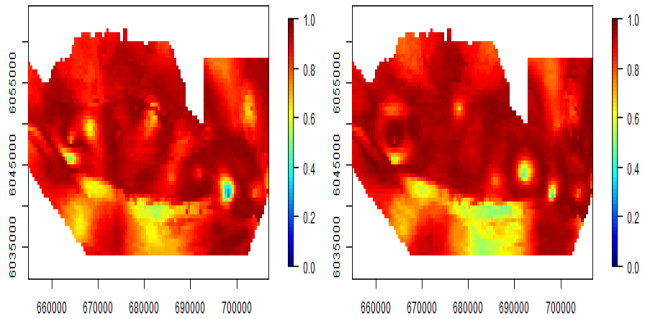


Figure 67: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

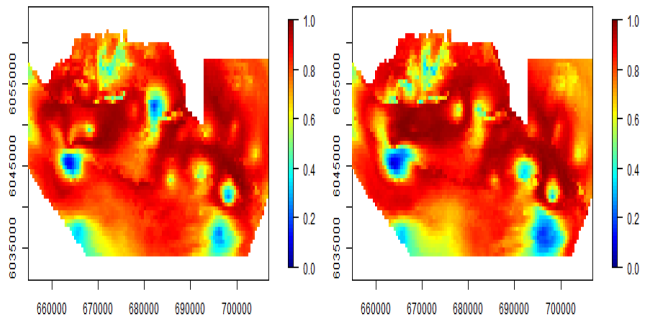


Figure 68: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

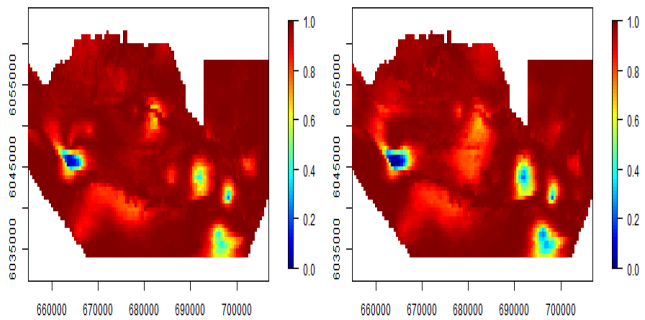


Figure 69: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

6.3.2 *Near-shore results*

The often high bias in the model predictions resulted in extremely poor coverage in these areas, particularly for the GAM generated surfaces (e.g. Figures 70–72). All three methods gave poor results, but the GAMMs were noticeably poorer for models I, II and III. In line with the off-shore results, the more flexible surfaces gave patchier results ranging from almost zero to the nominal 95% coverage.

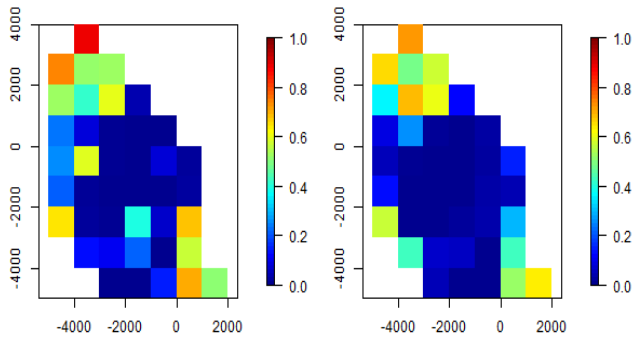


Figure 70: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

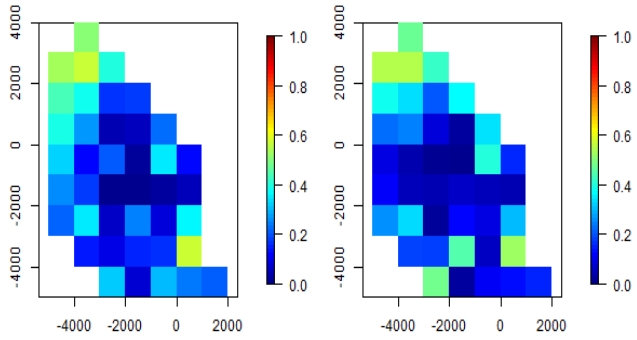


Figure 71: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

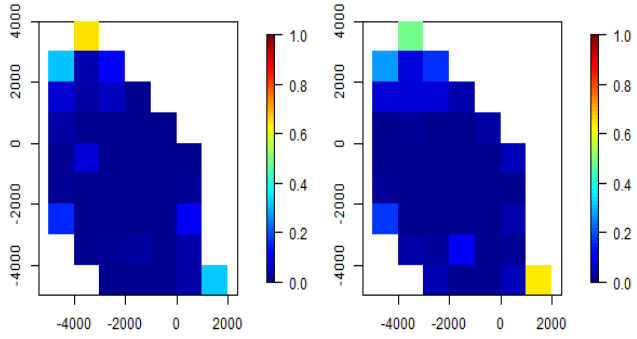


Figure 72: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

6.4 Model choice

The ability of each modelling approach to choose the appropriate model was of interest because model choice is integral to accurate impact assessment – the reader must be confident that any impact effects identified by the models are genuine and confident that when any impact effects exist they are not overlooked. The model selection process followed for each method is described in section 5.3.1.

In summary, all methods had (at least) a tendency to choose the redistribution post-impact model, even if there was no change post impact or that change was an overall decrease.

The model choice results show that CReSS performed markedly better than GAMs or GAMMs and that GAMs performed so badly as to invalidate their use for model selection (this is also arguably true for GAMMs).

GAMs always chose the (most complicated) redistribution post-impact model even when there was no impact at all, or the genuine impact was an average decrease.

6.4.1 Off-shore results

The results for GAMs suggest overfitting was widespread (Table 1) and the most complicated model was always chosen even when the process generating the data had no impact effect at all, or the post-impact was an overall decrease. GAM selection performance did not improve if p-values were used; the model selection results were identical under the QAIC and model-based p-values.

This overfitting was entirely expected – GAMs are not appropriate for positively correlated data (when the covariates are unavailable to model this correlation in full) and in these cases return p-values which are too small.

CReSS also tended to overfit, although less dramatically than the GAMs. While CReSS chose a redistribution model more often than it chose the correct, 'no impact', model (Table 1 and Figures 143 & 144) the performance was much better for the post-impact decrease (Figures 147 & 148) and markedly better for the post-impact redistribution (Figures 153 & 154).

GAMMs invariably performed substantially worse than CReSS but uniformly better than GAMs regarding model choice. GAMMs tended to overfit and choose the redistribution model even when an impact was not genuinely present or a average decrease was in effect (Table 1 and Figures 145 146, 149 & 150). Of course, when a re-distribution was in effect, this model type more often than not chose the correct model (Figures 155 & 156), but since choosing

model III was most likely under GAMMs anyway (whatever the true model) the identification of a redistribution impact might well leave the user wondering if this effect is genuine.

Surface	GAM	CReSS	GAMM
GAM generated: Model I	0%	43%	30%
CReSS generated: Model I	0%	51%	12%
GAM generated: Model II	0%	60%	15%
CReSS generated: Model II	0%	42%	9%
GAM generated: Model III	100%	87%	64%
CReSS generated: Model III	100%	97%	78%

Table 1: Percentage of fitted models that correctly chose the true model (i.e. I, II or III) for the off shore scenarios using the model selection criteria outlined in section 5.3.1.

6.4.2 Near-shore results

In keeping with the off-shore results, the redistribution model was always chosen using GAMs for the near shore scenarios (Table 2 and Figures 157, 158, 167, 168) even when the process generating the data had no impact effect at all or the post-impact was an overall decrease. The results for the GAMs were also the same if p-values, rather than the QAIC scores, were used.

In contrast to the off-shore results, CReSS only had a tendency to overfit for the data generated with a smooth surface and a post-impact decrease (Table 2 and Figures 163 & 164), but otherwise performed remarkably well in absolute terms and when compared to the other modelling approaches (Figures 159, 160, 171 & 172).

In contrast to the off-shore results, GAMMs performed almost as poorly as the GAMs and substantially worse than CReSS. GAMMs tended to overfit and choose the redistribution model even when an impact was not genuinely present or a average decrease was in effect with one exception (Table 2 and Figures 161 162, 165 & 166).

Surface	GAM	CReSS	GAMM
GAM generated: Model I	0%	96%	21%
CReSS generated: Model I	0%	96%	3%
GAM generated: Model II	0%	40%	0%
CReSS generated: Model II	0%	100%	4%
GAM generated: Model III	100%	76%	69%
CReSS generated: Model III	100%	76%	94%

Table 2: Percentage of fitted models that correctly chose the true model (i.e. I, II or III) for the near shore scenarios using the model selection criteria outlined in section 5.3.1.

6.5 Fit to the underlying surfaces

In summary, no single method was clearly superior to the others for both the off-shore and near-shore scenarios.

For example, GAMMs performed better for some of the off-shore scenarios but markedly worse in all of the near-shore scenarios.

In contrast the GAMs performed the worst overall in nearly all of the off-shore scenarios but the best for all of the near-shore scenarios.

The CReSS approach performed as well as the GAMMs (first-equal) for some of the off-shore scenarios and practically similar for the remaining, and there was very little difference between the performance of the CReSS and GAM approaches for the near shore scenarios.

6.5.1 Off-shore results

GAMMs generally seemed to give the best results when compared with the underlying function (Figures 73–78) although the fit results for CReSS were statistically indistinguishable for the decrease post-impact models for both the GAM and CReSS generated surface types (Table 3, page 81). Notably, there was a great deal of overlap in MSE scores across the methods undergoing comparison, indicating any practical differences in model performance were relatively small.

CReSS performed worse on average than GAMM in 4 of the 6 cases (Table 3), equivalent to GAMM in 2 cases, but better than GAMs in 5 of the 6 cases (Table 3). In particular, CReSS tended to overfit to the underlying function; i.e. the predictions based on the model were closer to the data than GAMMs (Table 4, page 82), but these predictions were further from the underlying function. While there was some evidence of overfitting with CReSS, this method overfitted much less than the corresponding GAMs (also Table 4).

GAMs performed equivalent to GAMM in 1 of the 6 cases, equivalent to CReSS in 2 of the 6 cases, and had the worst performance of the three methods in 3 of the 6 cases (Table 3). Specifically, GAMs tended to overfit to the underlying function; i.e. the predictions based on the model were closer to the data than GAMMs or CReSS (Table 4), but these predictions were further from the underlying function.

Notably due to overfitting, the marginal R^2 value would not choose the best model for these scenarios. Here, GAMs appeared to fit the data best (Table 4) but GAMs were most often the poorest

choice when considering the process generating the data (Table 3). Further, while GAMMs fitted the worst to the data (according to the marginal R^2) they approximated the underlying function better in most cases. The fit of the CReSS models to the data was intermediate between the two alternative methods (Table 4) and performed equivalently to GAMMs in 1/3 of the scenarios but worse in the others when considering closeness to the underlying functions (Table 3).

These models report r_c values between 0.2 and 0.3 (Table 5) but due to overfitting by GAMs, this measure (and the marginal R^2 already described) also suggests the GAM is the best model of the three options. This is not the case: GAMs fit better to the input data but do not return values closest to the underlying surface.

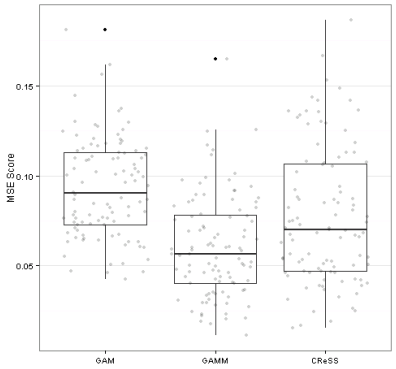


Figure 73: MSE scores for the GAM-generated (smooth surface) with mean defined using Model I (no-change post impact) across the three model types.

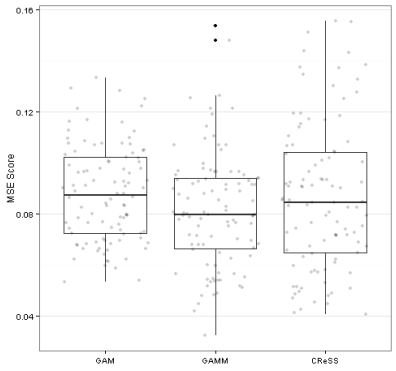


Figure 74: MSE scores for the CReSS-generated (flexible surface) with mean defined using Model I (no-change post impact) across the three model types.

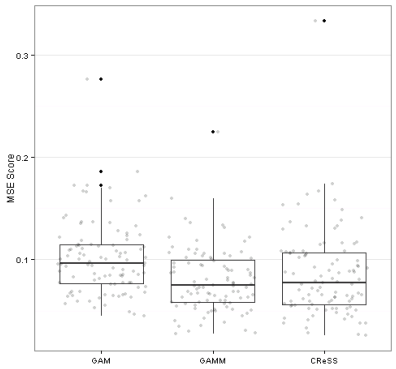


Figure 75: MSE scores for the GAM-generated (smooth surface) with mean defined using Model II (decrease post impact) across the three model types.

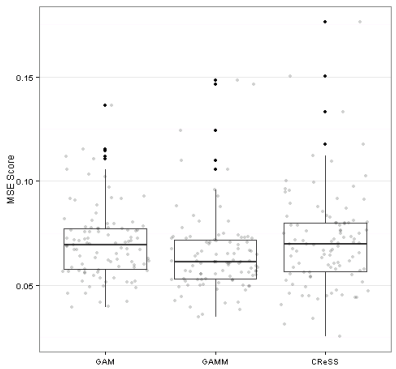


Figure 76: MSE scores for the CReSS-generated (flexible surface) with mean defined using Model II (decrease post impact) across the three model types.

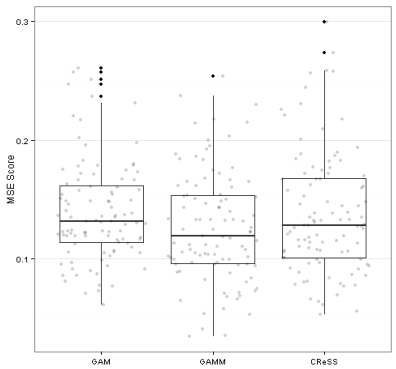


Figure 77: MSE scores for the GAM-generated (smooth surface) with mean defined using Model III (redistribution post impact) across the three model types.

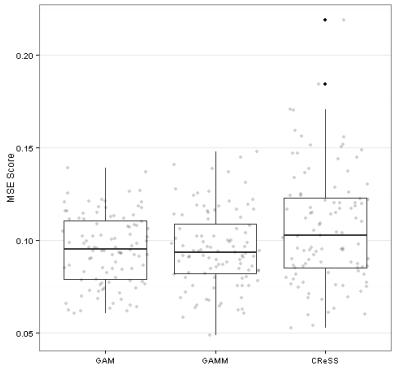


Figure 78: MSE scores for the CReSS-generated (flexible surface) with mean defined using Model III (redistribution post impact) across the three model types.

Surface	Best Model	Mean Difference in MSE
Comparison	GAMM vs GAM	
GAM generated: Model I	GAMM	0.0328
CReSS generated: Model I	GAMM	0.0059
GAM generated: Model II	GAMM	0.0212
CReSS generated: Model II	GAMM	0.0056
GAM generated: Model III	GAMM	0.0135
CReSS generated: Model III	NS	0.0005
Comparison	GAMM vs CReSS	
GAM generated: Model I	GAMM	0.0156
CReSS generated: Model I	NS	0.0052
GAM generated: Model II	NS	0.0043
CReSS generated: Model II	GAMM	0.0066
GAM generated: Model III	GAMM	0.0102
CReSS generated: Model III	GAMM	0.0104
Comparison	GAM vs CReSS	
GAM generated: Model I	CReSS	0.0166
CReSS generated: Model I	NS	0.0017
GAM generated: Model II	CReSS	0.0158
CReSS generated: Model II	NS	0.0006
GAM generated: Model III	NS	0.0020
CReSS generated: Model III	GAM	0.0106

Table 3: Pairwise comparison results for MSE scores across the three model types for the off-shore scenarios. The model type which fits significantly better (according to the Wilcoxon paired signed rank test, see above) is listed in the Best Model column, when NS is shown this indicates there was no significant difference between the model types.

Surface	GAM	CReSS	GAMM
GAM generated: Model I	**0.1186**	0.1111	0.1064
CReSS generated: Model I	**0.1521**	0.1441	0.1403
GAM generated: Model II	**0.1280**	0.1196	0.1125
CReSS generated: Model II	**0.1417**	0.1345	0.1265
GAM generated: Model III	**0.1610**	0.1530	0.1447
CReSS generated: Model III	**0.1749**	0.1691	0.1598

Table 4: Average Marginal R^2 values for models chosen for the data generated using model I, II or III for the off-shore scenarios. The method returning the highest score for a particular scenario is highlighted in bold.

Surface	GAM	CReSS	GAMM
GAM generated: Model I	**0.2105**	0.2022	0.2004
CReSS generated: Model I	**0.2653**	0.2557	0.2555
GAM generated: Model II	**0.2216**	0.2053	0.2101
CReSS generated: Model II	**0.2482**	0.2342	0.2393
GAM generated: Model III	**0.2744**	0.2617	0.2645
CReSS generated: Model III	**0.2984**	0.2851	0.2938

Table 5: Average r_c values for models chosen for the data generated using model I, II or III for the off-shore scenarios. The method returning the highest score for a particular scenario is highlighted in bold.

6.5.2 *Near-shore results*

In stark contrast to the off-shore results, GAMMs generally seemed to give the worst results compared with the underlying function (Figures 79–84). Further, while there was little practical difference in model performance between CReSS and GAMs (Figures 79–84), GAMs fitted significantly closer to the underlying function when interrogated using the Wilcoxon tests (Table 6, page 85). Notably, there was typically a great deal of overlap in MSE scores obtained by CReSS and GAMs.

In contrast to the off-shore results, the methods appeared to be underfitting (i.e. choosing models with less flexiblity than required by the underlying surface). GAMMs in particular exhibited both poor fit to the underlying surface and poor fit to the data (Table 7). The fit to the data and underlying function improved under CReSS and GAMs, in that order; Tables 6 and 7. Interestingly, the concordance criterion alternated between GAMMs and CReSS as exhibiting the best fit to the data (Table 8, page 85).

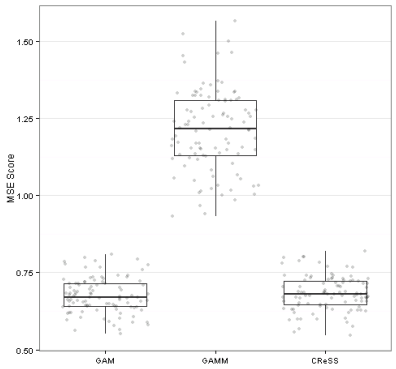


Figure 79: MSE scores for the GAM-generated (smooth surface) with mean defined using Model I (no-change post impact) across the three model types.

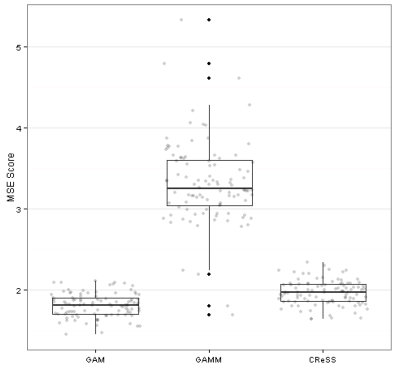


Figure 80: MSE scores for the CReSS-generated (flexible surface) with mean defined using Model I (no-change post impact) across the three model types.

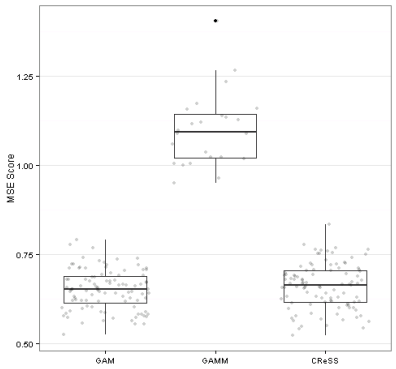


Figure 81: MSE scores for the GAM-generated (smooth surface) with mean defined using Model II (decrease post impact) across the three model types.

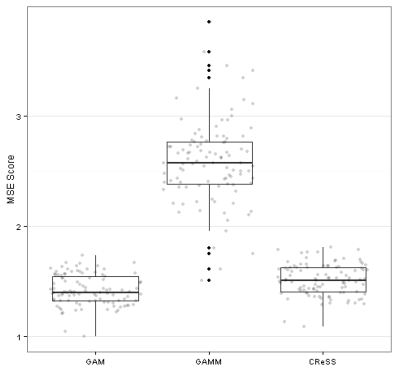


Figure 82: MSE scores for the GAM-generated (flexible surface) with mean defined using Model II (decrease post impact) across the three model types.

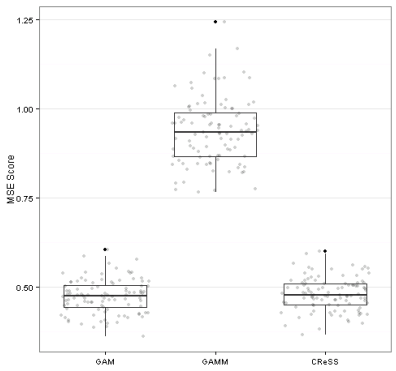


Figure 83: MSE scores for the GAM-generated (smooth surface) with mean defined using Model III (redistribution post impact) across the three model types.

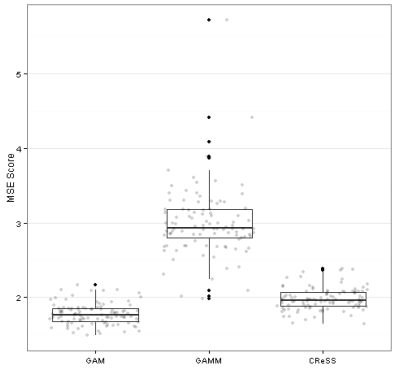


Figure 84: MSE scores for the CReSS-generated (flexible surface) with mean defined using Model III (redistribution post impact) across the three model types.

Surface	Best Model	Mean Difference in MSE
Comparison	GAMM vs GAM	
GAM generated: Model I	GAM	0.54
CReSS generated: Model I	GAM	1.49
GAM generated: Model II	GAM	0.45
CReSS generated: Model II	GAM	1.18
GAM generated: Model III	GAM	0.46
CReSS generated: Model III	GAM	1.20
Comparison	GAMM vs CReSS	
GAM generated: Model I	CReSS	0.53
CReSS generated: Model I	CReSS	1.34
GAM generated: Model II	CReSS	0.45
CReSS generated: Model II	CReSS	1.09
GAM generated: Model III	CReSS	0.45
CReSS generated: Model III	CReSS	1.00
Comparison	GAM vs CReSS	
GAM generated: Model I	GAM	0.009
CReSS generated: Model I	GAM	0.152
GAM generated: Model II	GAM	0.013
CReSS generated: Model II	GAM	0.090
GAM generated: Model III	GAM	0.007
CReSS generated: Model III	GAM	0.200

Table 6: Pairwise comparison results for MSE scores across the three model types for the near-shore scenarios. The model type which fits significantly better (according to the Wilcoxon paired signed rank test, see above) is listed in the Best Model column, when NS is shown this indicates there was no significant difference between the model types.

Surface	GAM	CReSS	GAMM
GAM generated: Model I	**0.1737**	0.1719	0.1611
CReSS generated: Model I	**0.2462**	0.2427	0.1921
GAM generated: Model II	**0.1520**	0.1495	0.1404
CReSS generated: Model II	**0.2371**	0.2347	0.1857
GAM generated: Model III	**0.1523**	0.1501	0.1417
CReSS generated: Model III	**0.2446**	0.2421	0.1950

Table 7: Average Marginal R^2 values for models chosen for the data generated using model I, II or III for the near-shore scenarios. The method returning the highest score for a particular scenario is highlighted in bold.

Surface	GAM	CReSS	GAMM
GAM generated: Model I	0.2937	0.2919	**0.3175**
CReSS generated: Model I	0.3871	**0.3879**	0.3652
GAM generated: Model II	0.2628	0.2611	**0.2855**
CReSS generated: Model II	0.3758	**0.3764**	0.3571
GAM generated: Model III	0.2618	0.2601	**0.2891**
CReSS generated: Model III	0.3852	**0.3874**	0.3596

Table 8: r_c values for models chosen for the data generated using model I, II or III for the near-shore scenarios. The method returning the highest score for a particular scenario is highlighted in bold.

7 Comparison summary

In summary, GAMs fit the underlying surfaces well (measures 2 &3; Table 9) but were very poor at assessing if impact-related effects were present and at locating spatially explicit change (measures 1 & 5; Table 9).

CReSS performed the best of the three methods at assessing if impact-related effects were present and at identifying spatially explicit differences (measures 1 & 5; Table 9). CReSS also performed well, and similar to GAMs, at approximating the underlying process (measures 2 & 3; Table 9).

GAMMs performed very poorly when assessing if impact-related effects were present but peformed well at locating spatially explicit change – at least for the off-shore scenarios (measures 1 & 5; Table 9); GAMMs performed poorly when attempting to approximate the underlying surface for the near-shore scenarios (measures 2 & 3; Table 9) and for this reason attracted a red rating for these measures.

All three methods showed mixed performance regarding spatially explicit coverage of the 95% confidence intervals.

Measure	GAM	CReSS	GAMM
(1) Model Choice			
(2) Fit to the underlying process			
(3) Spatially explicit bias			
(4) Spatially explicit coverage			
(5) Spatially explicit change			

Table 9: Qualitative summary of the comparison results across the three methods. The green colour indicates the method performed well, the red colour indicates the method performed poorly, while the amber colour indicates the results were mixed and performed much better in some cases than others.

8 Practical considerations

8.1 Computational issues

GAMMs proved to be very difficult to fit to the data generated here. The computational time was much longer than the other two model types and the frequently encountered convergence problems often meant results were unable to be obtained for some realisations – and there was no apparent reason why these realisations were particularly different to any of the others. GAMMs are also conditional models and thus, predicting to a grid requires some post-processing; this adds to the time involved in their use. There are also still unresolved model selection issues regarding GAMMs which makes model choice problematic in practice unless an empirical and computationally intensive approach (e.g. cross validation) is adopted for this purpose.

8.2 Surface fitting in areas with complex topography

Users should take care that GAMs (or similar methods) are not used to model areas with internal exclusion zones of any conse-

quential size in the survey area (since the smoothing method is underpinned by straight-line point to point distances). For instance, using GAMs (as they are typically implemented) in these situations could result in hotspots leaking across islands and coastline areas. The SOAP basis can be used instead of the default smoother inside the GAM fitting engine for sites with complex topography, but this methodology doesn't account for temporal autocorrelation and there are also many practical issues involved with their fitting (see Scott-Hayward et al. (2013) for details).

CReSS models are spatially adaptive and therefore potentially more flexible than GAM surfaces. The CReSS approach also allows the smoother to be based on geodesic distances which are more appropriate when the survey area includes internal exclusion zones - or complex topography. While these features can be advantages, geodesic distances must be provided to the software in order to generate the smoother basis and this requires specialist user input. There is also some care required when setting up the initial grid containing the anchor points (or knots) for basis generation, however code is available and continually being improved to make this step more automatic. CReSS computer fitting time was similar in general to GAM fitting, however generation of a geodesic distance matrix can take some time depending on the resolution of the distances calculated.

8.3 Surface fitting for spatially-patchy (highly uneven) distributions

In this work we have simulated data based on an off-shore and near-shore scenario and while it is speculation about how exactly these methods will perform for new data, we have published results regarding CReSS and GAM comparisons for spatially adaptive surfaces. In particular, based on the spatially adaptive capabilities of the CReSS method, CReSS was shown to outperform GAM-based surface fitting for distributions which are highly uneven across the surface (see results in (Scott-Hayward et al., 2013)).

9 Recommendations

9.1 Minimum survey design criteria

The importance of spatial sampling design could be driven by finding locations where the distribution of covariate values mimics the distribution of covariate values throughout the study area.

The guiding principle of concern with placing sampling effort to be analysed with model-based methods is to ensure that the range of possible covariate values is included in the sample. Likewise, and more difficult to achieve, a good range of combinations of covariate values should also be included in the sample, particularly if interactions are included in the density surface models (e.g. interaction of depth and chlorophyll index).

Note however, if sampling represents the range of the covariate over which animal density is invariant, there will be the presumption of no relationship between the covariate and animal density. For this reason, the covariate may not be selected for inclusion because its range is not represented in the sample.

Practical advice for placement of survey design, when model-based inference will be used for analysis, is very similar to advice for design-based inference; place transects along gradients in animal density (such that high and low density areas are sampled). Hopefully this will capture a range of covariates; in the absence of detailed information about the shape of relationships between covariates and animal density.

There is little published literature on design recommendations for model-based estimates of animal abundance. Some guidance regarding the role of design in model-based inference is provided in Borchers et al. (2002, Section 3.2.5). A hybrid of model- and design-based sampling is called "model-assisted survey sampling." It is the subject of a text by Särndal and Swensson (2003). A recent paper by Peel et al. (2013) discusses ways in which survey effort can be allocated so as to improve estimates of fish abundance produced by model-based inference. In a contrasting vein, Shibata et al. (2013) discusses the consequence of using predictive covariates that do not influence animal distribution. This demonstrates that model-based inference is a more *fragile* mode of inference because if the model is wrong, then the inference about animals will be incorrect.

9.2 Identification of predictive covariates

A successful model selection procedure depends on a comprehensive identification of covariates. Within the framework of impact assessments and monitoring related to offshore renewables it is especially important in order to assess the impact from the renewables project that covariates reflect both habitat features and existing pressures on model targets.

Both habitat features and pressures may be important structuring elements in the distribution of the animals recorded prior to the construction of an offshore wind farm or an ocean energy device. Obviously, both habitat and pressure covariates reflect processes at different scale (see Covariate relevance), and accordingly it is critical to consider the scale of the impact study or monitoring, and requirements for spatial resolution of predicted distributions.

Because most processes in the marine environment are dynamic, prediction at a high resolution requires covariates to be available at a high spatio-temporal resolution. Data in high spatio-temporal resolution on physical oceanographic processes are generally available as hindcasts of hydrodynamic models in which distribution of currents, water levels and water masses is estimated for time steps from hours to days for the whole water column. Statistical uncertainty of model predictions is routinely determined by cross

validation against measurements. However, data are not freely available, and acquisition of such data can be resource demanding. Adding to this, hydrodynamic data are normally not readily available in the form habitat features (e.g. eddy or frontal activity), and these features require post-processing of flow parameters.

Coupled to hydrodynamic models, water quality models may provide important dynamic covariates like water transparency, oxygen and chlorophyll. Hindcasts of water quality parameters are provided by only few institutes worldwide and, like for the hydrodynamic data, acquisition of such time series can be resource demanding.

Data on prey distributions are rarely available for larger areas at the required spatio-temporal resolution. If available, data on benthic invertebrates and plants are typically describing mean distributions with no or limited information about statistical uncertainty, and can only be used as static covariates.

Recently, in response to the legal requirements of the EU Habitats Directive, coarse-scale benthic habitats or landscapes have been mapped according to the EUNIS classification system[5] and are, or will become, freely available. Although these maps provide a new source of information about the geography of the seabed they mainly reflect geomorphological differences, and convey only limited information about the distribution of plant and animal communities. Statistical uncertainty associated with these data has been estimated for some but not all areas.

[5] www.eunis.eea.europa.eu

Data on the distribution of fish abundance is available at a coarse resolution from seasonal fish resource surveys coordinated by ICES, and undertaken in most European shelf seas. Although the data covers a wide range of species variable catchability means that data on several potentially important prey species is limited, for example sandeels *Ammodytidae* is limited. Information on fish distribution is generally not available at higher resolutions, and some resources are generally needed to acquire the data.

The data layers now freely available and showing coarse-scale benthic habitats and landscapes include information on surface sediments and bathymetry. In many cases, surface sediment distribution has been estimated based on interpretation of available data, and so the quality of these data varies depending on the degree of habitat mapping. Unfortunately, indications of statistical uncertainty are generally not provided with the geological data layers.

Detailed bathymetric measurements are available from most European seas, yet even for areas like the North Sea and the Celtic Sea, regions exist in which rather few depth samples have been obtained. Accordingly, available GIS maps showing water depths for extended areas of sea bed are interpolations from depth measurements, and hence estimated depths possess a variable and often undisclosed amount of uncertainty.

Data on the distribution of pressures can be split into dynamic

and static data. AIS (Automatic Identification System) monitoring now provides data on the density of ships in many areas, for various temporal and spatial scales. Provision of large amounts of data may be resource demanding. The coverage of AIS stations is far from complete, and even for areas like the North Sea coverage of ship activity is highly variable. Yet, for coastal areas AIS measurements are available in rather high spatial resolution, and constitute a useful indicator of potential disturbance from ships. Static data on many pressure covariates are easily obtained, including location of coastal developments and coastal and offshore infrastructures.

9.2.1 Covariate relevance

As covariates reflect processes at different scales it is critical at an early stage when conceptualising the model, to assess the resolution of predicted distributions required to meet the aim of the study. In most impact assessments related to offshore renewables the project footprint and impacted area comprise less than 20 km^2. This means that covariates involved in processes affecting the distribution of target species at a relatively small scale should be identified. In addition, in order to investigate distributions at the scale of an individual renewables project, it is necessary to partition the variance into appropriate scales of ecological organisation. This is because a pattern generated by an ecological process at a particular scale will be masked by patterns generated by other processes at both larger and smaller scales (Ciannelli et al., 2008). Thus, to describe a small-scale process such as the concentration of animals at a hydrographical shelf front in proximity to a renewables project site, one has to remove the masking effect of large-scale processes by identifying and controlling for them in the analyses. Similarly, in order to describe a large-scale process such as a seasonal effect reflected by sea temperature, one has to remove the noise from small-scale processes by aggregating or smoothing the data or by employing a hierarchy of models with covariates reflecting processes at different scales.

To describe the effects of spatial gradients associated with marine renewables and marine habitats over less than 10 km it is necessary to achieve a spatial resolution of covariates at as fine a resolution as possible. The covariates which have been found by a large number of studies to be important structuring elements controlling the aggregation of marine animals are bathymetry, surface sediments and flow dynamics.

Bathymetry (and derived covariates like seabed slope and complexity) is an important driver in the distribution of many benthic and pelagic species, and at a fine scale discontinuities in the bathymetry are often related to shallows, boundaries of water masses and areas of frontal activity, which make them suitable foraging areas for predators feeding on the sea bed as well as for predators feeding on the higher densities of pelagic prey typically

concentrated here. If flow data are available the intensity of processes responsible for the enhancement of pelagic prey can be more precisely described and used as covariates:

1. Hydrographic frontal activity (enhanced primary and secondary production, ecotone effect);
2. Upwelling/downwelling (enhanced primary and secondary production);
3. Eddy activity (prey retention);
4. Stratification of water column (enhanced primary and secondary production).

Instantaneous data on salinity and temperature generally describe water masses while time series describe seasonal trends. In both cases the effects on animal distribution will be coarse scale, and hence should be used with care as covariates in studies related to renewables.

Despite the aggregation of prey at frontal features. a spatial mismatch between the distribution of marine mammals and that of their prey is often observed locally (Torres and Read (2008), Fauchald et al. (2011)). Thus, even if scientific data on the distribution of prey abundance is only available at a coarse resolution, this may actually be more useful than information on prey available from surveys undertaken at a higher resolution. Contrary to the situation for pelagic predators, animals feeding on benthic invertebrates and plants show strong correlations with the distribution of their food sources. Thus, fine-scale data on the distribution of benthic plant and animal prey can provide important covariates in models for benthic predators.

In the same way as for habitat and coarse scale covariates, the scaling of pressure covariates should be assessed carefully at the conceptual stage. Especially, one has to ensure that any masking effect of pressures is identified and controlled for in the analyses.

9.3 Modelling recommendations

As with all analyses, the modelling approach should be chosen with the objectives of the analysis in mind, however based on the work contained in this report we make the following recommendations.

If pre and post impact type comparisons are of interest then CReSS-based modelling (as implemented here for instance) is recommended. This approach was best able to identify spatially explicit impacts - concerning both the magnitude and location of post-impact change. While model selection results regarding post-impact change were disappointing across all three approaches for the off-shore data, these results drastically improved under CReSS based selection for the near shore data. Further, when CReSS chose

incorrectly there were no serious adverse effects of doing so. For instance, when CReSS failed to correctly choose the no-impact model, redistribution (rather than a decrease) was typically chosen instead and the results of these analyses clearly demonstrated no statistically significant differences when examined spatially. In contrast, GAMs always chose redistribution (regardless of truth) and due to the tendency to overfit, GAMs often systematically predicted increases and decreases which didn't exist and these were identified as significant changes up to 30% of the time.

If baseline characterisation is the main focus (and post-impact data is not, yet, available for instance) then CReSS would also be the preferred methodological choice. CReSS performed similarly, or better than GAMs, at approximating the underlying surfaces but GAMs tend to under-report the uncertainty for models fitted to data of this sort, and so any confidence intervals associated with geo-referenced predictions (or covariates) are likely to be too narrow. Additionally, model selection results may also be flawed resulting in the retention of unrelated covariates, in models fitted to baseline data.

The remaining parts of this section outline a general set of recommendations for modelling baseline monitoring and impact assessment data. It is not intended to provide a comprehensive guide to statistical modelling but instead covers crucial aspects of the process. This process is also, often, iterative and the models are updated in light of insights gained about the covariate relationships and the noise component assumed for the model.

This section covers some of the main issues, and more details are provided in the worked examples in sections 10 and 11.

9.3.1 Specifying a model

Asking a few general questions about the survey design and data available can assist the analyst when specifying a model. While this is by no means an exhaustive list, the following questions are useful:

1. What is the nature of the response variable? Is the response of interest continuous or discrete? Are there natural boundaries to the response variable?
   For example, are the response data counts bounded by zero? and if the data are counts, are there large numbers of zeros? (e.g. Figure 85)

   Alternatively, are the response data presence/absence records, bounded by 0 and 1? (Figure 86) The nature of the response variable helps guide model choice, since some models explicitly respect the natural boundaries of the response data and ensure

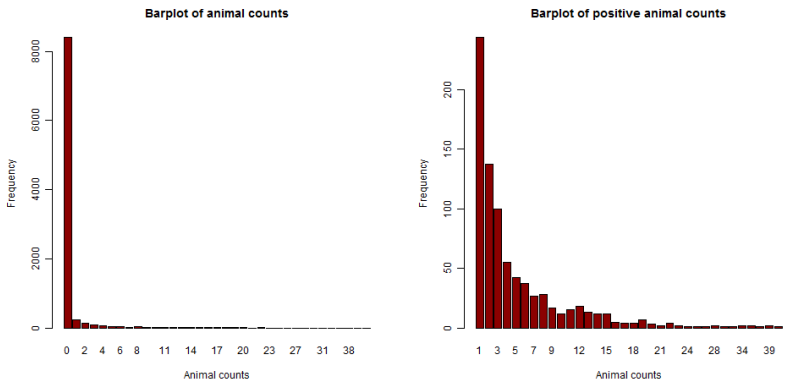


Figure 85: Distribution of count response data with the zero counts included (left) or excluded (right).

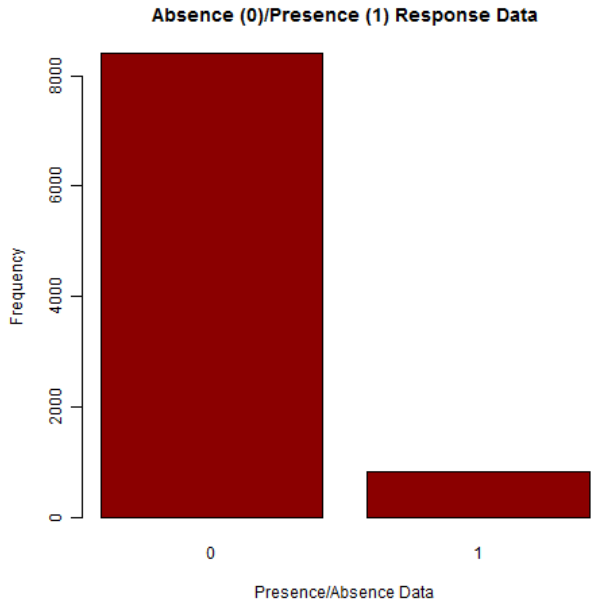


Figure 86: Distribution of binary response data

model predictions are returned within these boundaries. For example, some models employ 'link' functions which can ensure that impossible predictions are not returned by the model (e.g. negative numbers of animals) and these functions are typically dictated by the nature of the response variable and these natural boundaries.

Additionally, if the data are counts and there are large numbers of zeros (e.g. Figure 85) then the response data are likely to be more variable than assumed under some model types (e.g. overdispersed) and this variability must be permitted under the model for credible results.

In some cases, dedicated approaches to response data with large numbers of zeros can be used (e.g. so-called 'zero inflated' (ZI) models) which treat the zeros in the response differently from non-zeros depending on the assumed origin of the zero counts.

Zero inflated models were not considered here for two reasons: there are currently no off-the-shelf ZI models that include either smoother based terms (for the environmental covariates or a spatial term) or residual correlation (along transects for instance). It is, of course, possible to manually implement smoother-based terms on the scale of the link function to address the smoother-based issue (e.g. using the `bs` function in the `splines` package in `R` for the environmental covariates) and even perhaps, develop code to fit zero inflated mixed models, but these methods are not widely available and therefore unlikely to be used in the renewables industry.

2. What is the observation process like? Are all the animals at the surface likely to be seen by the observer, or are they likely to be harder to see at distance from the boat/plane? Are the focal animals ever underwater and therefore unavailable to be seen at the surface by the observer?

If some animals are overlooked at the surface when they are available to be seen, or they are unavailable to be seen because they are underwater, then some attempt at inflating the observed numbers of animals (by estimating the numbers that are missed) should be made. This is *crucial* even if the user is only interested in detecting changes pre and post impact. For instance, if these detection issues are ignored then it's possible that a change in the detectability of the animals pre and post impact might be mistaken for an impact-related effect. i.e. animal numbers might appear to be substantially lower post-impact solely because animals are harder to see post-impact (due to poorer sampling conditions, for example). Distance sampling (Thomas et al., 2010) is a widely used method developed to correct observed counts

for the ones that are missed and can be implemented in a wide variety of survey situations.

In relatively recent surveys, photographic or video technology has been employed to capture images of the ocean surface in the survey area. In these cases, while it is assumed that all animals at the surface are correctly identified and detected, correction may still be necessary to account for those animals underwater when the images are captured.

3. How were the data collected? Were the observations sampled randomly from a set of independent locations or is there some natural order to the response data in space and/or time?
For example, were the counts sampled from a boat or plane travelling along transects across the ocean on a set of surveyed days?

Observations collected close together in space/time are often more similar than observations distant in space/time. For example, consider Figure 87. The survey design consists of a set of transects, and for a given X-coordinate consecutive animal counts along the transects (in the Y direction) are typically similar. Here, we see that locations with relatively large numbers of animals are located close to each other.

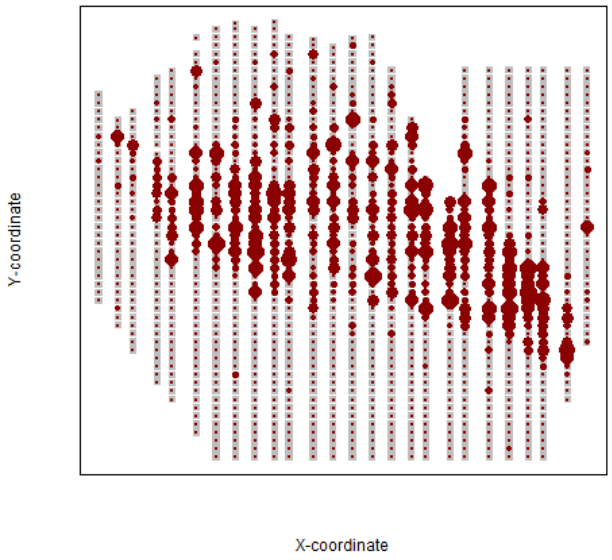


Figure 87: Example count data represented spatially. The relative size of the circles reflect the size of the animal counts.

While it is possible that the covariate data (which may have been collected alongside the counts) could explain the similarities in the counts along the transect lines, it is unlikely that this covariate data will explain this correlation in the counts in full. These patterns in the data, which are unexplained by the model, will necessarily be allocated to the error (noise) term in the model, which violates a crucial assumption for many widely used methods (e.g. GAMs). Some alternative methods (e.g. GEEs) explicitly permit these patterns across space/time (e.g. along transects) as a part of the noise component and thus should be considered for response data of this sort.

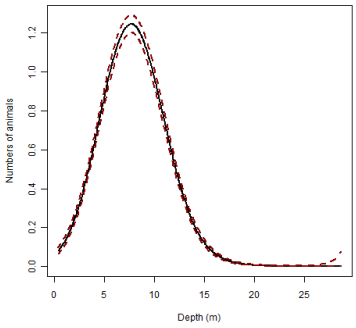


Figure 88: GAM-based depth relationship with animal numbers

While details about the survey design can assist the analyst in choosing a modelling framework, the analyst does not need to rely on this information alone; there are graphical ways and formal statistical tests to assess if residual autocorrelation is an issue for the model at hand.

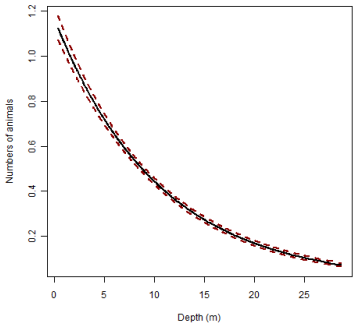


Figure 89: GLM-based depth relationship with animal numbers

4. What is the broad nature of the relationships between the continuous covariates and the response data? Are the numbers of animals (or the presence/absence of animals) likely to rise and fall (or fall and rise) as the covariate increases in value?
For example, animal numbers might tend to be low for shallow depths, most common at moderate depths, and then found in lower numbers again in the deepest waters (e.g. Figure 88).

Basic information about the covariate relationships helps the user choose between linear and nonlinear models (e.g GLM/GLMMs and GAM/GAMMs). Linear models, by default, do not accommodate covariate relationships which rise and fall with the response, however nonlinear models permit *both* nonlinear and linear relationships (e.g. Figures 88 and 89) since the linear relationship forms a special case; when it doesn't make biological sense to expect a covariate relationship to rise and fall (or vice-versa) with the response, this covariate relationship can be specified to be linear (on the link scale).

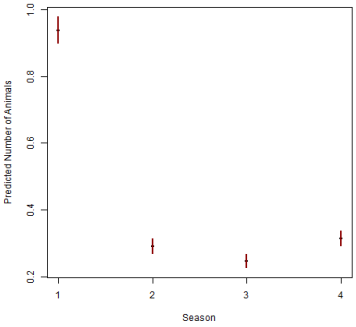
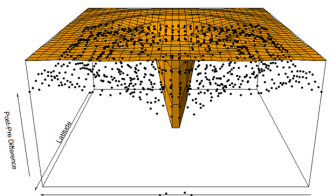


Figure 90: Factor level (e.g. seasonal) relationship with animal numbers

Note: the linear/nonlinear model choice is not an issue for factor-level covariates since 'linear' models are linear in their parameters and thus estimate different coefficients for each level of the factor variable (excepting the baseline level), e.g. Figure 90. GLMs and GLMMs can easily accommodate factor-level covariate relationships that rise and fall with the response (or any shape) since the coefficients rise and fall with each factor-based level.



5. Is there a spatial element to the data? Is there a desire to include a spatial term in the model as a proxy for unmeasured covariates?

   For example, animal numbers (or the presence/absence of animals) at each spatial location will likely be due to a complex mix of covariates, and many of these covariates will not be available for inclusion in a model. This is likely to result in unmodelled patterns in the response data across the surface which can be approximated by a spatial smooth term.

There are a variety of ways to include spatial information in a model ranging from the simple, and restrictive, inclusion of the spatial coordinates as-is (separately) in a model, to including a potentially complex surface which considers the coordinates together. For instance, including spatial information separately in a model (e.g. XPos and YPos) assumes that the relationship between one of the coordinates (e.g. latitude) and the response, is unaffected by changing values of the other (e.g. longitude) and this is often unreasonable in practice. For this reason, more complex surfaces are often considered which, for example, allow the nature of the relationship between latitude and the response to depend on values of longitude. This is typically closer to reality and this added flexibility also aids more localised surface features in animal distribution (such as 'hotspots' and impact-related effects) to be identified.

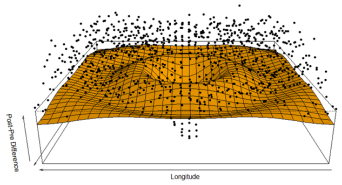


Figure 92: Example of underfitting; the fitted surface understates the magnitude of the impact and overstates the range of the impact.

If an interaction-type surface for the spatial component is desired there are basic choices about the nature of the surface flexibility, worthy of consideration. For example, is the surface likely to be evenly smooth? Or could the surface be highly uneven and exhibit some areas which are relatively flat and other areas which are highly structured (e.g. Figure 91)? Evenly smooth surfaces can be well approximated using 'global smoothing parameter' methods which allocate a single parameter to surface flexibility which applies across the whole surface. Highly uneven surfaces, however, are better approximated using 'spatially adaptive' methods, which allow the flexibility to vary across the surface.

The specification of the spatial surface can widen the model selection task considerably, since spatially adaptive methods typically require more parameters. However, choosing between 'global smoothing parameter' and 'spatially adaptive' methods is a serious consideration. For instance, surfaces which have localised impact-related effects (e.g. in and around the installation, Figure 91) but are fitted using models without sufficient flexibility (i.e. the surface is under-fitted) are likely to return model predictions which understate the magnitude at the impact site and overstate the range of the impact (Figure 92).

Conversely surfaces that are too flexible about the impact site

(or are overfitted to the response data), might overstate both the magnitude and the range of the impact (Figure 93). Both underfitting and overfitting are a problem in these cases, and so attention must be paid to the performance of the surface, particularly in and around the potentially impacted site(s).

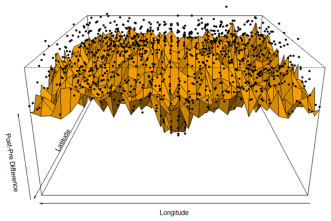


Figure 93: Example of over-fitting; the fitted surface overstates the range and magnitude of the impact.

Choosing to employ a globally smooth or spatially adaptive method can be considered as one part of the model selection process. For instance, both approaches could be trialled and their respective fits to the data compared. While this is a reasonable approach, any over-fitting tendencies by either method will result in one of the approaches fitting better to the data, but worse to the underlying function and this will be unknown to the user (as seen in the work presented here). This could easily result in an over-fitted surface being chosen.

If fitting two sets of spatial models is not an attractive option, then one could view a global smoothing method as a special case of a spatially adaptive method (since the latter can also return a uniformly smooth fitted surface) and only fit the latter. For example, the starting position of CReSS is a model with flexibility allocated evenly across the fitted surface (using a space-filled design). Viewing a global smoothing method as a special case of a spatially adaptive method means that the user still must rely on the model selection routine employed by the spatially adaptive smoothing method to choose appropriate model flexibility. However, the results shown here demonstrate that CReSS method (coupled with SALSA model selection) was shown to perform well at returning surfaces which were close to the simulated truth.

In some cases, it may not be necessary or desirable to include a spatial term in the model. For example, if only a small number of key drivers dictate animal numbers, and these are included correctly in a model, then the unexplained part of the response across the spatial surface (post model fitting) will be patternless noise. In that case, including a spatial term would not be preferable and a model selection procedure ought to reflect this (by not 'choosing' the spatial term). If there is some concern that including a spatial element in a model will mask genuine covariate relationships (and therefore the opportunity to gain biological insights will be lost), trialling the spatial element in a model after already considering the other covariates as candidates might be preferred.

Another factor to consider for some monitoring and impact assessment sites is the presence (and location) of any exclusion zones which may exist within the survey area. These are sub-areas inside the area of interest which are unavailable to the animals, and typically include land forms (for marine mammals and some birds). In these situations, recently developed smoothing methods based on geodesic (as-the-animal-swims) rather

than euclidean/straight line (as-the-crow-flies) distance should be considered for the spatial element. The CReSS method also allows for this (along with other newly developed methods, e.g. SOAP) and should be implemented based on distances relevant for the animals in these situations.

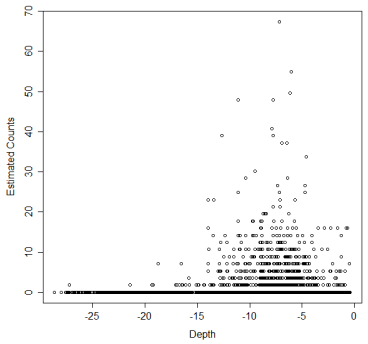


Figure 94: Example of a scatterplot showing the estimated counts (obtained by Distance sampling) with water depth; here zero represents land and negative values represent depth in metres.

9.3.2 *Exploratory Data Analysis; EDA*

Exploration of the data is necessary to familiarise the user with the data available for modelling. Even simple plots can help the analyst identify any obvious errors in the data entry/download, any large gaps in the covariate range and any outlying values in the response data.
The user should also ensure the candidate covariates considered are not too similar ('collinear') to avoid issues such as model instability.

Once the input data have been corrected for imperfect detection (when necessary), a modelling framework has been chosen (e.g. CReSS coupled with GEEs), a set of candidate covariates identified, and some thought dedicated to the broad nature of these covariate relationships, some data exploration work is useful.

Typically the analyst has a set of covariates available as candidates for the model – some of which may be categorical (i.e. defined classes) and be trialled in a model as factor-level covariates[6] or continuous in nature and be trialled as a linear or nonlinear term. The creation of simple plots (e.g. scatterplots (Figure 94) and boxplots (Figure 95)) can help the analyst identify any errors, any large gaps in the covariate range and any large gaps in the response data.

As a part of the exploratory process some checks should be made to ensure that the covariates in the model don't share too much information with each other, regarding their relationship with the response variable. Specifically, model covariates should not be 'collinear' which results in unstable model predictions when essentially the same information enters the model twice. The practical consequence of including collinear variables together in a model is that parameter uncertainty can be very high and indicate important covariates should be excluded from a model (based on large p-values). For this reason, objective measures to detect prohibitive levels of collinearity (e.g. Variance Inflation Factors; VIFs (Fox, 2002)) ought to be used.

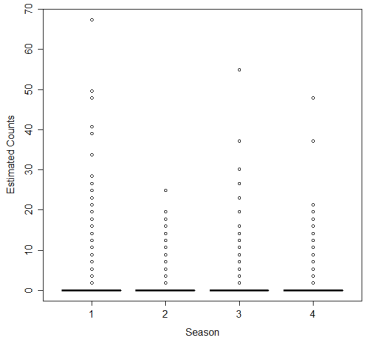


Figure 95: Example of a boxplot showing the estimated counts (obtained by Distance sampling) with season

[6] Note, for a model to return coefficients for factor-level covariates, there typically has be either non-zero counts (in the case of count data models) or both zeros and ones (in the case of presence/absence data) in each level of the factor. This can easily be checked using simple numerical summaries.

9.3.3 *Choosing model covariates*

Choosing which variables to include in a model is an important part of the modelling process; including too few variables makes for poor predictions and including too many variables can make predictions highly uncertain.

Sometimes, the user has an interest in choosing between a select set of models chosen in advance (e.g. based on prior information/analysis) but often models are chosen using automated methods. If the automated approach is employed, efforts should be made to choose from the full set of models available; this enables the user to see how well all candidate models fit the data and either present the best model or some aggregate of similarly performing models.

Model selection involves fitting a set of preliminary models to the data and discriminating between them using some predetermined criteria. Often a 'full' model is fitted to the data which contains all of the covariates considered worthy of inclusion and some type of model selection process is carried out. This can take the form of 'backwards selection' (by fitting a full model and dropping covariates deemed to be unimportant, e.g. based on p-values, one by one) or the process can work in reverse and a 'forwards selection' process adopted. The latter process starts with a model containing a small number of 'core' covariates (or only the intercept term) and covariates are added one-by-one as they are deemed to be important to the response. Importance in this context can be determined by objective fit criteria (e.g. QIC for GEEs) or is determined by associated p-values (for each covariate) which are smaller than some chosen value (e.g. 0.05).

Stepwise selection combines the forwards and backwards methods, uses an algorithmic approach and can trial a large number of models as a results. This process does not, however, guarantee to trial every possible model and therefore some good model combinations might be overlooked, just by chance. In addition to this, stepwise methods often only return the 'best' model and there may be other models which *were* trialled and were almost as good; the user is typically blind to this and valuable insights might be lost.

For this reason, all possible subsets selection can be used instead which involves comparing the fit criteria for all possible models given a set of (carefully considered) covariates. For instance, for a model with three candidate covariates, the fit of every one-covariate model is compared with the fit of the possible two-covariate models and the fit of the full model (containing all three). Owing to its utility, this type of model selection is becoming commonplace as an option in widely available software packages (including the `MuMIn`

library in R).

An advantage of the all possible subsets procedure is that the user is immediately aware of the performance of other (competing) models and is able to make informed decisions about model choice accordingly. For example, if the 'best' model has a very similar (virtually indistinguishable) fit to an alternative model (with different covariates) and the close fitting alternative model has more desirable covariates for practical reasons, then this model might be used instead.

The all possible subsets procedure also lends itself easily to model averaging. Here, if a single model is not a requirement (or desirable), and some weighted combination of a set of 'good' models is permitted (or preferred) then model predictions can be obtained under each model of interest and averaged in line with their weights (determined by the objective fit criteria).

9.3.4 *Evaluating the performance of the fitted model*

The performance of the chosen model can be assessed by examining how close the model predictions are to the response data. A good model should produce predictions which are close to the input data and high fit scores.

Following the selection of the working model, the performance must be examined and simple plots comparing the observed versus fitted values (on the vertical and horizontal axes respectively) can assess model fit. Here, a good fit is reflected by a one-to-one relationship between the observed data and fitted values based on the model (Figure 96).

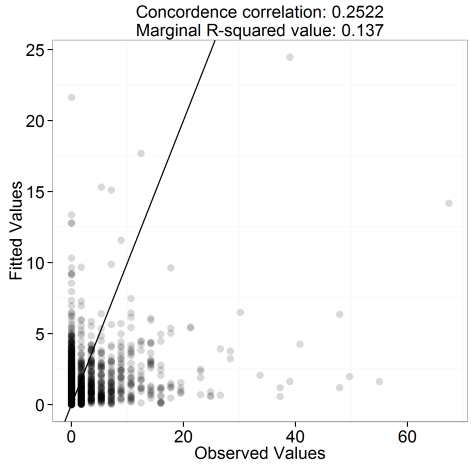


Figure 96: Example of a scatterplot showing the response (corrected counts) with the fitted values under a model.

In particular, model failings such as under or over-prediction are evidenced by clusters of positive or negative residuals either with respect to each model covariate or with reference to a prediction

grid (Figure 97). In this case, a good mix of negative and positive residuals across the survey area and a distinct lack of pattern indicates the model fit is spatially adequate.

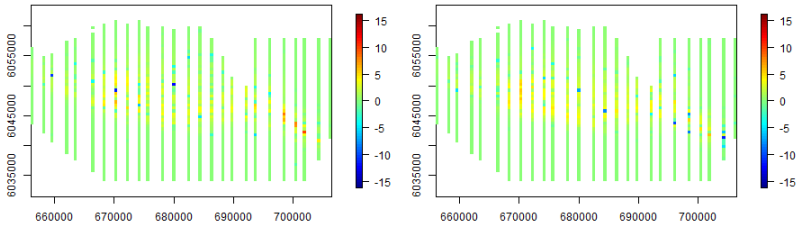


Figure 97: Example of a scatterplot showing the residuals pre (left-hand side) and post impact (right-hand side).

9.3.5 *Diagnosing model problems*

> Modelling is an iterative process and statistical models must be updated, where possible, in light of any inadequacies uncovered. Specifically, it is important to interrogate the working model to see if model assumptions are reasonable and the associated results are defensible.
>
> For any particular model, the user needs to understand under what circumstances, and in what ways, does the model not describe the data well and convey these limitations to the reader along with model results.

In particular, the fitted relationships should be examined for parts of the covariate range which exhibit predictions which are systematically too high (over-prediction) or too low (under-prediction). This can easily be done using model residuals, post model fitting.

Ill-fitting covariate relationships can be diagnosed graphically using cumulative residual plots. Specifically, 'raw' residuals (observed values-predicted values) ordered by covariate value are summed across the covariate range and compared with a horizontal reference line at zero. Systematic under (or over) prediction is signalled by cumulative sums which are well separated from zero and are persistently positive (or negative). For GLMs fitted to independent data, the track of the cumulative sum observed for a particular data set/model combination can be compared with a reference set of cumulative sum tracks likely to be found under a 'good' model (e.g. using the `gof` library in R), but the theory that underpins this comparison has only been recently developed for correlated data (Lin et al., 2002) and has not yet been programmed into R.

While a reference set of cumulative sum tracks is not yet available in R and it might be difficult to evaluate the form of a covariate

relationship in absolute terms, this diagnostic can be used to compare covariate relationships across competing models.

For example, the improvements made by allowing depth to be nonlinear on the link scale (fitted as a GAM rather than a GLM) are evidenced by substantially smaller cumulative residuals and more 'mixing' of positive and negative residuals (Figures 98 and 99). Cumulative residual plots can also be used to look for runs of positive and negative residuals in observation order which aids in the diagnosis of serial correlation (Figure 100, for example).

The details of the assumption checking that is appropriate depends on the exact model fitted, however for GLMs and GAMs (which are most commonly used for data of this sort), independence in the noise component (the errors) is assumed and this *must* be checked.

Correlation within surveys/transects/grid-cells can be be assessed visually using empirical autocorrelation function plots ('acf' plots) or more formally diagnosed numerically using a 'runs test' (Mendenhall, 1982). The runs test is an numerical test for 'randomness' that has multiple uses in a modelling context, including the diagnosis of ill-fitting covariate relationships. Put simply, the sign (rather than the magnitude) of model residuals are tracked in order and the number of uninterrupted sequences ('runs') of positive and negative residuals are calculated (right hand plot in Figure 100). The number of runs observed are then compared with the number of runs expected[7] when the values are random and too few (or too many) runs signal the residuals are positively (or negatively) correlated. In particular, compelling evidence of positive correlation is evidenced by a negative runs test statistic and a small p-value (e.g. $p < 0.01$).

[7]

$$E(T) = \frac{2n_p n_n}{n_p + n_n} + 1$$

$$V(T) = \frac{2n_p n_n (2n_p n_n - n_p - n_n)}{(n_p + n_n)^2 (n_p + n_n - 1)}$$

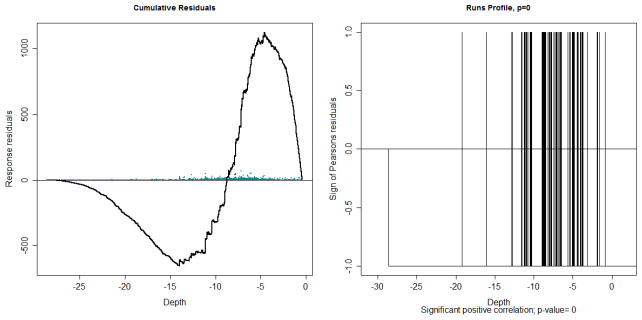


Figure 98: Cumulative residuals in depth order (left-hand plot) and runs test profile (right-hand plot) for a depth relationship fitted using a GLM.

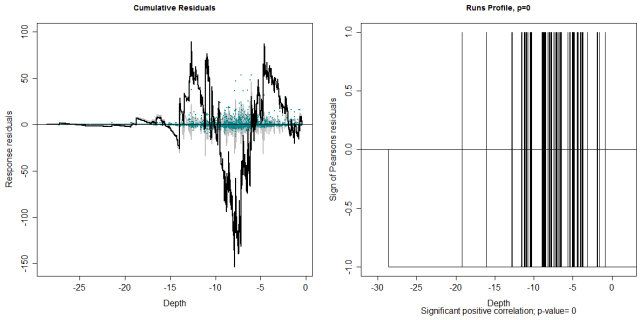


Figure 99: Cumulative residuals in depth order (left-hand plot) and runs test profile (right-hand plot) for a depth relationship fitted using a GAM.

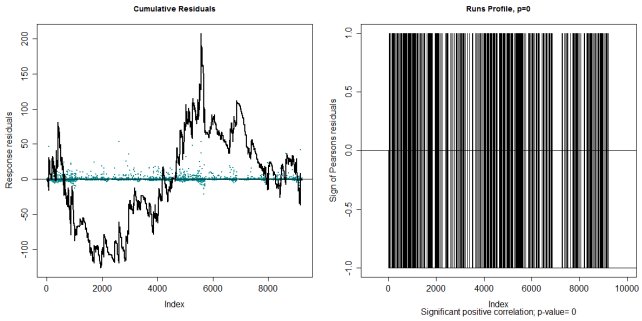


Figure 100: Cumulative residuals in observation order (left-hand plot) and runs test profile for time-ordered model residuals (right-hand plot).

The cumulative residual plots and associated runs test are useful for one-dimensional covariates (e.g. Depth) but the user may benefit from a visual representation of how the model fits to the data spatially. For example, viewing the residuals on the prediction grid (across the survey area) can also be represented spatially to enable the user to ascertain if there are some areas of the surface which are over (or under) predicted by the model. In particular, this will help the user assess how well any 'hotspots' (areas of high density) predicted by the model are supported by the data, or if there are correspondingly large residuals in these areas (e.g. Figure 97).

In addition to the diagnosis of systematic over or under prediction, it is important to ensure model conclusions do not depend on a small number of correlated clusters/blocks (e.g. transects) in the data. For instance, it would be unwise to rely on a model which returns different results when a small number of blocks/transects are omitted from the analysis. For this reason, we should examine how aspects of the model change when individual observations and/or correlated blocks/clusters are removed from the analysis.

The PRESS statistic quantifies the sensitivity of model predictions to removing each subject/observations. Here, model coefficients are re-estimated when each subject/observation is omitted (one-by-one) and the sum of the squared differences between the response data and the predicted values (when that subject is removed) are found:

$$PRESS_i = \sum_{j=1}^{J} \sum_{t=1}^{n_i} (y_{ijt} - \hat{y}_{ijt,-i})^2 \tag{12}$$

where y_{ijt} represents the response values for transect i, on segment j at time point t and $\hat{y}_{ijt,-i}$ represents the predictions when the i-th transect is omitted. Relatively large values signal the model is sensitive to these subjects (e.g. Figure 101).

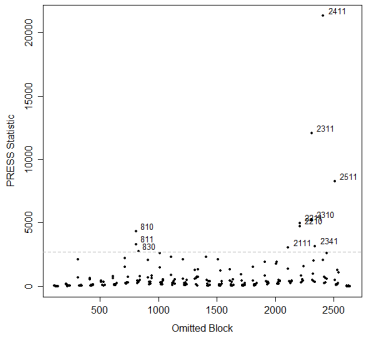


Figure 101: PRESS scores for a fitted model. The labels refer to individual blocks/cluster numbers.

In contrast to the PRESS statistic, the COVRATIO statistic signals the change in the precision of the parameter estimates when each subject/observation is omitted. Values greater than one signal removing the subject inflates model standard errors while values less than one signal standard errors are smaller when that individual is excluded (e.g. Figure 102).

The PRESS and COVRATIO statistics are relative measures and in the event that model predictions or measures of precision appear particularly sensitive to omitted blocks, it would be prudent to examine model conclusions based on models with and without the potentially problematic blocks.

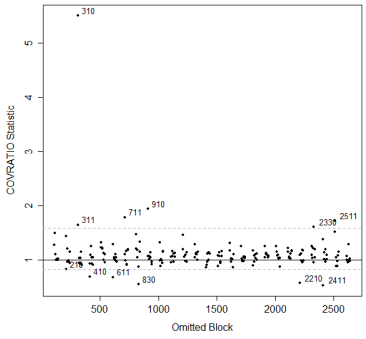


Figure 102: COVRATIO scores for a fitted model.

9.3.6 *Interpreting modelling results*

After the user is satisfied that model assumptions have been met and the model is safe to interpret, there are some graphical and numerical measures which should be examined. In particular, the fitted covariate relationships should be examined and model coeffi-

cients interpreted appropriately.

> The uncertainty about the estimated parameters is *crucial* when drawing conclusions based on a model. For instance, while some model (impact-related) terms might be included in a model chosen using objective fit criteria (e.g. the QAIC/QIC) there might be considerable uncertainty about these estimates and it is important to view these coefficients in light of the associated uncertainty.
>
> For example the magnitude of the estimated pre/post impact difference might raise considerable concerns. However, if 'no-change' post-impact is also plausible in light of the data (as indicated by a 95% confidence interval about the impact parameter) then this will not likely provide compelling evidence for a pre/post impact difference.
>
> In this case, the 95% confidence intervals might better serve as 'best' and 'worst' case scenarios for any impact-related differences in animal numbers.

The fitted covariate relationships in a model can be identified visually using 'partial' plots post model fitting. These partial plots illustrate how the model relates each covariate to the response with some perspective provided by confidence limits about the fitted curve. The values on the vertical axis for these graphics are not necessarily on the scale of the response data but reflects how the response (on a 'link' scale) responds in relation to changing covariate values under the model. The confidence limits about these fitted relationships are almost as important as the curves themselves since these enable the user to consider any curvature/patterns along with the uncertainty in these relationships. Specifically, the details of fitted curves which are highly uncertain are purely speculative compared with fitted curves which are very precise. In this example (Figure 103) animal numbers are predicted to increase as the water becomes shallower, and peak in number at approximately 5 metres in depth. The confidence intervals are also very narrow about this fitted relationship enabling this interpretation.

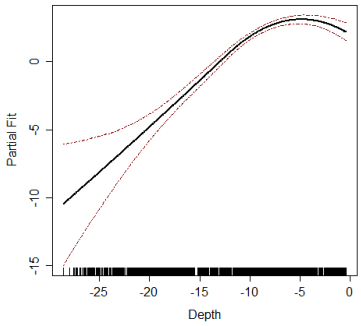


Figure 103: Example of a fitted depth relationship in a model.

The coefficients for factor variables (with reference to a baseline coefficient) can also be viewed using partial plots. Here the confidence intervals about the fitted coefficients tell us about plausible values for the size of the coefficients for each factor level (e.g. different seasons) differences between each non-baseline coefficient and the baseline at zero (e.g. Figure 104).

Parameter estimates, estimates of uncertainty and p-values associated with each model term are returned by default, and for simple (usually linear) terms, just one coefficient and associated p-value is associated and so, the assessment about statistical significance for that term/parameter is easily ascertained.

However, in some cases more than one coefficient is attributed

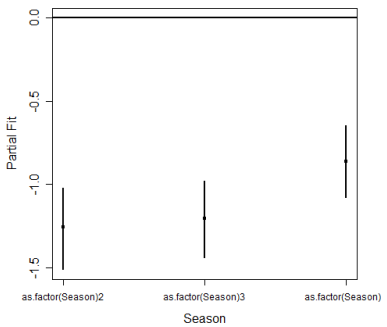


Figure 104: Example of a fitted factor variable in a model.

to a particular term (e.g. there are many coefficients involved in a spatial term) and so assessing the statistical significance of a model term requires considering the associated coefficients as a collection (e.g. using Wald tests).

9.4 Quantifying the power to detect change

The inability to detect genuine post-impact changes in numbers and distributions is a real concern for the marine renewables industry. For example, if a model returns large but highly uncertain impact-related effects (which are not statistically significant due to the high uncertainty) the user is left wondering if this result is simply due to lack of power, or there is no underlying post-impact change.

Alternatively only baseline data may be available, and the question remains as to how likely the current sampling regime is to detect future post impact change.

A 'power analysis' approach could be used to quantify the chance that a genuine impact effect is detected. This could involve, for example, manufacturing post-impact data based on what is known about the survey design, the sampling frequency and the response data pre-impact, and fitting a model which includes an impact-related effect to the manufactured data. The success rate of the data and modelling approach therefore at detecting a significant impact effect can be determined based on repeated sets of manufactured data.

While simple in concept, reliable power analysis results need to be based on realistic features of the data such as over-dispersion, nonlinearities and autocorrelation. In this piece of work, we did not develop methods (and associated code) to quantify the power to detect change in this setting but this could be considered in the future and build on the work carried out here.

10 *Impact assessment for off-shore data; worked example*

This section contains a worked example for a quantitative impact assessment based on simulated off-shore data. We begin by describing how the data was manufactured and then detail the analysis process. The analysis process involves:

1. correcting the observed counts for imperfect detection (to account for animals missed by the observers)
2. exploration of the data
3. model specification & fitting
4. diagnosing model faults
5. using the chosen model to make predictions & quantifying any differences
6. expressing the uncertainty about covariate relationships and associated model predictions

Finally, since the data in this case was manufactured we are able to compare the results with the process used to generate the data – the 'truth'.

10.1 *Manufacturing the data*

The data were simulated based on off-shore survey data collected before construction. An impact effect was then imposed which reduced animal numbers in the impacted area and re-distributed these animals to the south east of the study region (Figure 105); the total number of animals before and after impact was constant.

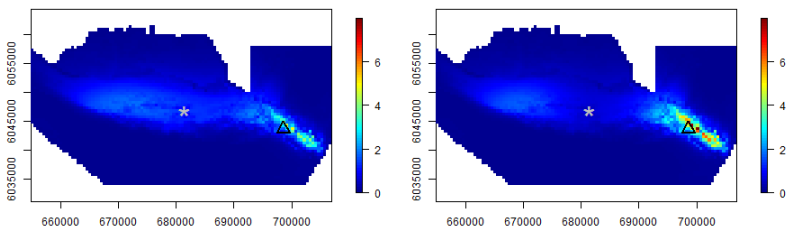


Figure 105: Simulated densities of birds (per km^2) before impact (left) and after impact (right) for the redistribution scenarios. The grey star indicates the centre point of the impact and the black triangle, the centre point of the re-distribution.

Observed counts were then lifted from the simulated surfaces in the form of line-transects, which were repeated seasonally (four times), both before and after the 'development' (Figure 106). Imperfect detection was then imposed on these lifted counts to mimic the observation process. This returned counts from 26 transects, repeated 8 times with 9232 segments (each approx. 0.25 km^2). Of those segments, 654 contained detections and there were 2373 detections in total.

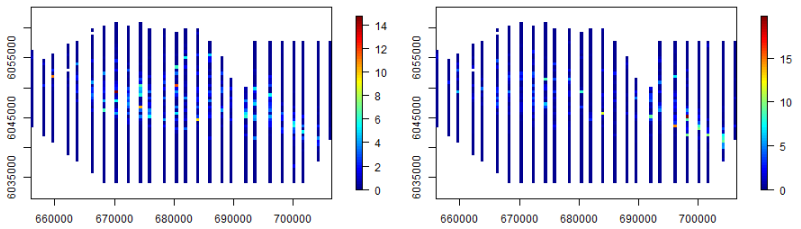


Figure 106: Observed data before (left) and after impact (right). Each cell is 0.5 km^2 and the colour represents mean bird count.

10.2 Statistical analysis

The data analysis for the worked example involves correcting the observed counts for imperfect detection, exploring the data, model specification and fitting followed by some diagnostics. Following the selection of an appropriate model, the model is used to make predictions, generate ranges of plausible values for these predictions (95% confidence intervals) and on this basis identify any spatially explicit differences. The results produced in this section are then compared with the surface used to generate the data in this case.

10.2.1 Correcting for imperfect detection: Distance sampling

A half normal detection function ($g(y)$) was estimated based on the distance sampling data (Figure 107) and a Cramer-von Mises (CvM) goodness of fit test (H_0: the fitted model is the true model) and a qq-plot used to evaluate the quality of the fit. The fit was adequate in this case; Figure 108 and the CvM test (p-value=0.7) confirm the estimated detection function fits the data well.

From these results we can make the following interpretation. "We are 95% confident that the scale parameter (which describes the shape of the detection function) lies somewhere between 112.7 and 122.4, and that the average probability of detection is somewhere between 0.568 and 0.571". So, approximately 60% of the birds were detected and approximately 40% were missed.

Based on this result, the raw counts were inflated. The estimated counts for each cell before and after impact are shown in Figure 109.

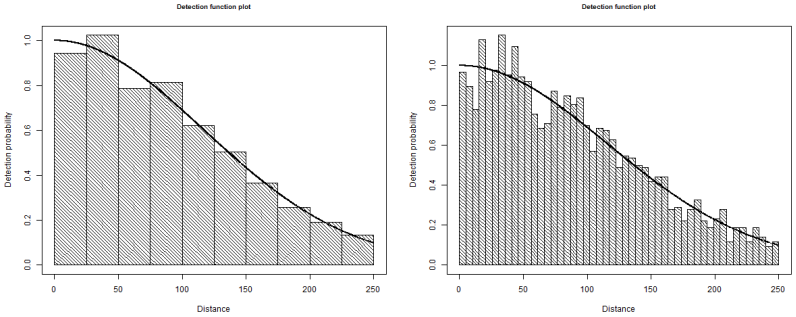


Figure 107: Histograms of raw distance data with the estimated half normal detection function overlaid. The figures show an example of two different histogram bin widths.

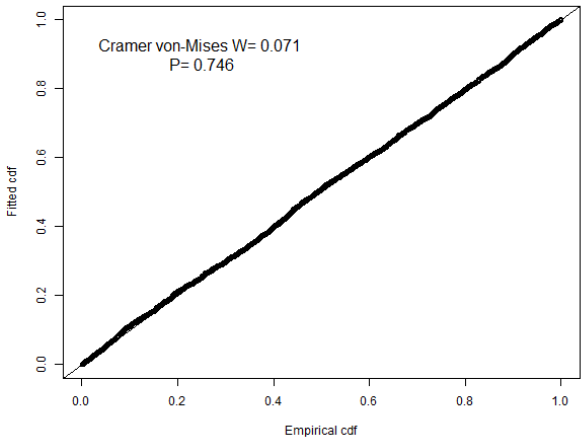


Figure 108: QQ plot showing goodness of fit for the fitted detection function.

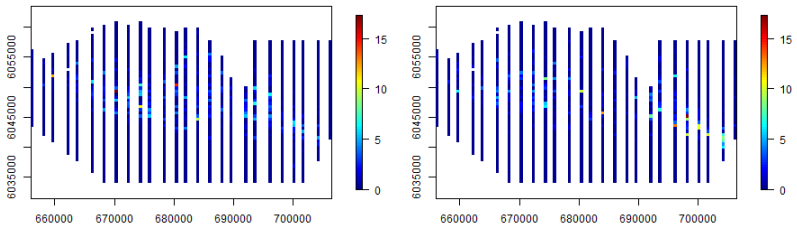
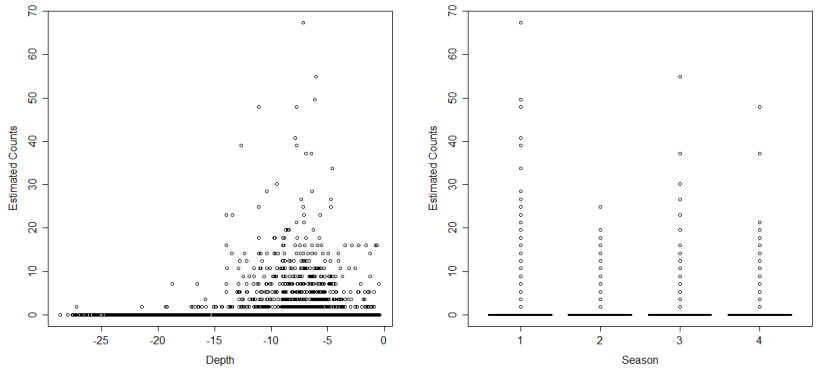


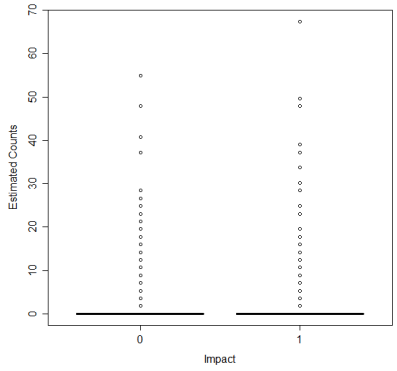
Figure 109: Mean bird counts estimated from a distance sampling analysis for before (left) and after (right) an impact event. Each cell is 0.25 km^2 and the colour represents mean bird count.

10.2.2 *Data exploration*

Prior to model fitting, it is always a good idea to examine the relationships between the available covariates and the response (e.g. bird counts; Figure 110). In this case, birds were predominantly seen in shallow waters, and few were seen in waters deeper than 15m. Further, the relationship between depth and bird numbers appears to be non-linear. It is difficult to identify the nature of the relationship between either season or impact and bird numbers due to the large number of zeros in the data. In particular, there are no outlying points evident – no large gaps in the range of any of the candidate covariates or any outrageously large response values.



(a) (b)



(c)

Figure 110: Plots of (a) depth, (b) Season and (c) Impact against the estimated bird counts.

10.2.3 *Model specification*

The count data were manufactured in this case, and so we know the data are well described by an overdispersed Poisson distribution (with a dispersion parameter of 7.03). Typically, the data are not synthetic and thus the nature of the process generating the counts is unknown, however it is often appropriate to model count data using a quasi-poisson error structure and this is the approach used here. The main candidate terms considered for the model were:

- a smooth term for depth (*s(Depth)*)

- a spatial term (*s(XPos,YPos)*)
- a four-level factor variable term for season
- a two-level factor variable term for impact (e.g. windfarm installation/operation)

In order to quantitatively assess if a post-impact re-distribution had occurred, a spatial-impact interaction term was also included. An 'offset' was also specified to allow for variable search effort across the estimated counts. For example, most segments were 0.25 km^2, but some contained land resulting in a smaller search area. For prediction, the area was given as $1km^2$ and so any predictions will be a bird density, per km^2.

Variance Inflation Factors (VIFs) were used to assess co-linearity between covariates and measure the extent of standard error inflated incurred by fitting covariates together in the model. Here, Generalised VIFs (GVIFs; Fox and Monette (1992)) are appropriate since some of the candidate covariates attract more than one parameter, and GVIFs are adjusted for the number of parameters (technically the degrees of freedom) fitted for each covariate ($GVIF_{adj} = GVIF^{1/2*Df}$). Each $GVIF_{adj}$ quantifies the decrease in precision of estimation due to collinearity (equivalent to $\sqrt{VIF}$). For example, a $GVIF_{adj}$ of 2 means that the confidence intervals are twice as wide as they would be for uncorrelated predictors.

In this case, the variables are not highly collinear (Table 10); 'Impact' has the highest GVIF score and this is not prohibitively high. Remedial action is recommended if values are as high as $\sqrt{5}$.

Variable	**GVIF**	**Df**	**$GVIF_{adj}$**
Depth	6.54	3	1.37
Season	1.00	3	1.00
s(XPos,YPos)	3670	10	1.51
Impact	7.52	1	2.74
s(XPos,YPos):Impact	5000	10	1.53

Table 10: Variance inflation factors to assess co-linearity amongst covariates. GVIF is a generalised variance inflation factor, and $GVIF_{adj}$ is GVIF adjusted for the degrees of freedom ($GVIF^{1/2*Df}$).

10.2.4 Model Fitting and model selection

A CReSS-GEE framework was employed here to estimate the smooth terms in the model whilst allowing for positive autocorrelation in model residuals. The fitted model included a relatively smooth function specified for depth (with just 3 parameters) and a smooth spatial term which employed SALSA-based model selection to determine flexibility. SALSA can also be used to choose the smoothness of the depth relationship (see the user guide for the `MRSea` package for more details). The season and impact terms were fitted as factor variables. In this case, we have very few terms in the model and it was decided to use p-value based model selection (since these are corrected for residual correlation under a GEE approach).

10.2.5 Results

A very low dimensional smooth term was chosen for the spatial element ($df = 6$) and the smooth surface was spatially adaptive; the surface comprised a mixture of local and globally acting basis functions.

Despite the covariates included in the model, there was compelling evidence for time-based correlation in (Pearson) residuals; a runs test returned a very small p-value ($p << 0.05$). As a result, a GEE fitting framework was employed using a *transect-season-impact* blocking structure – here residuals are permitted to be correlated within transects, within seasons, and either pre or post impact. This structure returned 208 unique *transect-season-impact* blocks with up to 26 segments in each block. For some blocks the correlation is low at any lag, whilst others have high correlation even for residuals 10 time points apart (Figure 111).

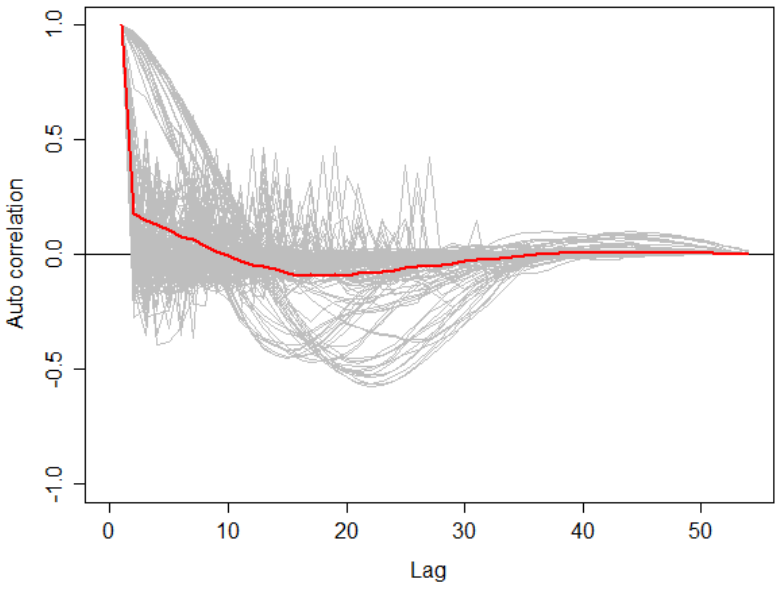


Figure 111: Plot of the estimated correlation in residuals at various lags for the set of blocks (grey lines). The estimated mean correlation at each lag is indicated in red.

The output (Figure 112) shows p-values associated with each coefficient which collectively make up the smooth terms for depth ($df = 3$), the spatial term ($df = 6$), and the factor variables for impact ($df = 1$) and season ($df = 3$). While useful in some cases, backwards model selection using p-values requires a single (Wald test based) p-value for each candidate covariate which does not depend on the order in which the terms are fitted in the model (i.e. they are marginal p-values). Based on these marginal p-values, Depth, Season, the spatial-impact interaction term are all highly significant (Table 11) and thus were retained in the model.

To investigate if there was a change in average numbers pre and post impact, the spatial-impact interaction term was removed and

the remaining terms re-fitted in the model; there was no evidence for an overall difference ($p = 0.5460$) pre and post impact.

```
Coefficients:
                                                 Estimate   Std.err  Wald  Pr(>|W|)
(Intercept)                                     -1.04e+01  1.75e+00 35.68  2.3e-09 ***
bs(Depth)1                                       5.28e+00  2.00e+00  6.93   0.0085 **
bs(Depth)2                                       1.52e+01  1.72e+00 77.72  < 2e-16 ***
bs(Depth)3                                       1.25e+01  1.92e+00 42.42  7.4e-11 ***
as.factor(Season)2                              -1.26e+00  1.45e-01 75.14  < 2e-16 ***
as.factor(Season)3                              -1.20e+00  1.79e-01 45.04  1.9e-11 ***
as.factor(Season)4                              -8.63e-01  1.40e-01 37.98  7.1e-10 ***
LocalRadialFunction()b1                         -5.07e+05  3.36e+05  2.28   0.1310
LocalRadialFunction()b2                          3.80e+06  4.61e+05 67.71  2.2e-16 ***
LocalRadialFunction()b3                         -6.06e+00  2.16e+00  7.89   0.0050 **
LocalRadialFunction()b4                         -9.13e+06  2.22e+06 16.99  3.8e-05 ***
LocalRadialFunction()b5                          4.29e+02  7.46e+01 33.15  8.5e-09 ***
LocalRadialFunction()b6                          7.24e+00  2.39e+00  9.19   0.0024 **
as.factor(Impact)1                              -1.78e-01  2.44e-01  0.53   0.4651
LocalRadialFunction()b1:as.factor(Impact)1 -6.93e+05  4.54e+05  2.33   0.1269
LocalRadialFunction()b2:as.factor(Impact)1  7.39e+05  6.24e+05  1.40   0.2368
LocalRadialFunction()b3:as.factor(Impact)1 -2.26e+00  3.11e+00  0.53   0.4679
LocalRadialFunction()b4:as.factor(Impact)1  1.22e+06  3.33e+06  0.13   0.7139
LocalRadialFunction()b5:as.factor(Impact)1 -2.82e+01  1.08e+02  0.07   0.7937
LocalRadialFunction()b6:as.factor(Impact)1  2.67e-02  3.26e+00  0.00   0.9935
---
```

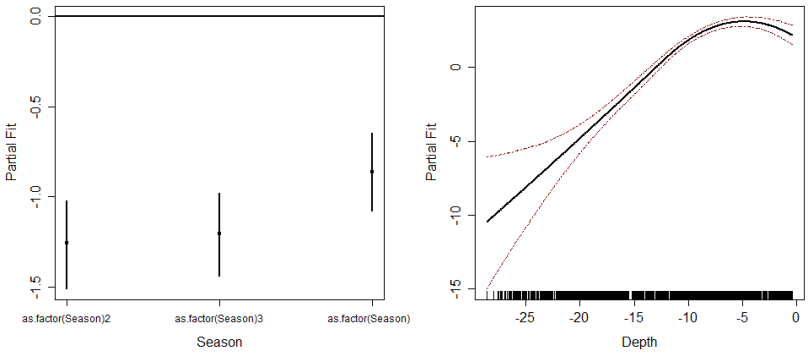
Figure 112: Model output from our fitted GEE model. Note: the covariate labels have been shortened so as to fit on the page.

Variable	p**-value**
Depth	< 0.0001
as.factor(Season)	< 0.0001
s(X,Y)	< 0.0001
as.factor(Impact)	0.5468
s(X,Y):as.factor(Impact)	0.0081

Table 11: GEE based p-values (to 4 decimal places) for a CReSS model with SALSA knot placement. $p < 0.05$ suggests that the covariate should be retained in the model.

Partial plots (Figure 113) allow the examination of how the one-dimensional (non-interaction based) covariates (e.g. depth & season) feature in the model, and based on the extent of the associated uncertainty (95% confidence intervals) the user can decide whether the smooth functions are justified. In this case, the season coefficients (Figure 113) demonstrate that, on average, significantly fewer animals are present in seasons 2, 3 & 4, compared with season 1. The confidence intervals on the season coefficients indicate little difference in the differences between the three levels and the baseline season, given the other covariates in the model.

The partial plot for depth (Figure 113) indicates a declining non-linear relationship with increased depth (i.e. fewer birds present in deeper waters) and that birds are more likely to be found in waters approximately 3-8m deep. The tight pointwise confidence intervals about the fitted function suggest a relatively precise relationship with depth. If the confidence intervals for depth were, instead, very wide it would be reasonable to assume depth could be adequately modelled as a linear term.



(a) (b)

Figure 113: Partial Plots for (a) season and (b) depth.

10.2.6 Diagnostics

The plot of observed versus fitted values suggests the fit of the model is not fantastic (Figure 114) and that the very high numbers of animals are difficult to predict. This is very common for very noisy data; the largest values tend to be under-predicted and many of the observed zeros are over-predicted. The relatively low fit scores ($R^2_{MARG} = 0.14$ and $r_c = 0.25$) confirms the data are noisy and difficult to predict.

Note, the covariates used to manufacture the data were also supplied to the model for fitting, and so these low values are due to the high noise in the data rather than signify that important covariates are missing.

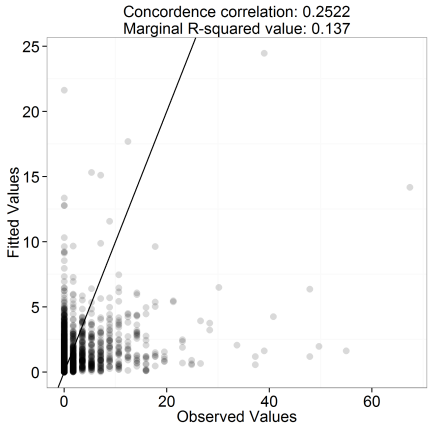


Figure 114: Assessing model fit: observed vs fitted values. The black line indicates where data should lie for a perfect fit. The data points are adjusted for over-plotting and the heavier the shading, the more over-plotting.

If a model is correctly specified (as it is here), we expect to see no pattern in a plot of fitted values against the residuals. Additionally, since we have a Poisson-based model (with extra-Poisson variation) the variance of the residuals is expected to increase with the fitted values and so the residuals must be adjusted for this if we are to see no patterns; for this reason scaled Pearsons residuals are used. In this case, while the raw points may appear to show patterns in the associated plot (due to large amounts of over-plotting), the locally weighted least squares regression line does not indicate any

unusual patterns (Figure 115).

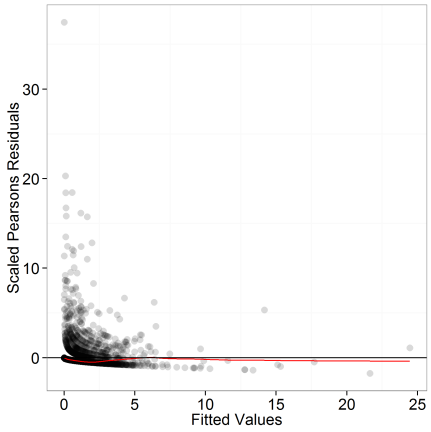


Figure 115: Assessing model fit: fitted vs scaled Pearsons residuals. The red line is a locally weighted least squares regression to indicate pattern in the plot, which might otherwise be hidden due to over-plotting. The data points are adjusted for over-plotting and the heavier the shading, the more over-plotting.

Systematic over or under prediction can be diagnosed using cumulative residual plots, and in this case depths around 6-9m appear to be systematically over-predicted while depths either side of this range are under-predicted (Figure 116). The grey line in the background of the plots indicates what we would expect if we were modelling depth with a very large number of parameters. In this case, more 'mixing' of positive and negative residuals around the zero line would be ideal, the improvements seen by modelling depth using a nonlinear function compared with a linear depth relationship are evident (Figures 116 and 117). Perhaps, implementing model selection for the depth relationship (to add flexibility perhaps) using SALSA (see the user guide for `MRSea`) would improve the associated cumulative residual plot.

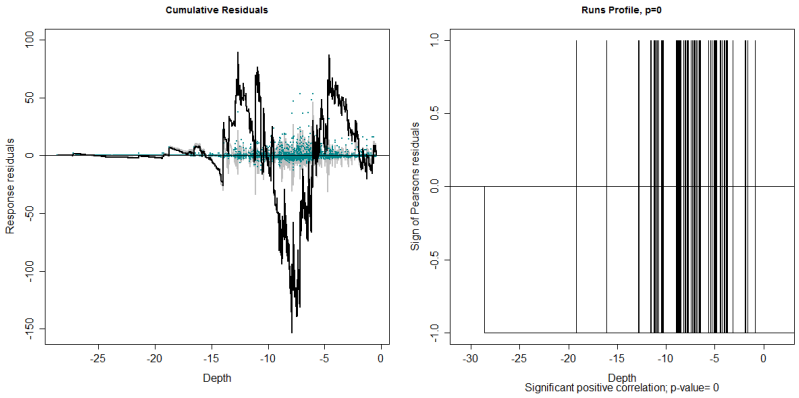


Figure 116: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by depth. The blue points are the residual values, the black line represents the cumulative residuals. The grey line in the background is what we would expect the cumulative residuals to be if depth was modelled correctly.

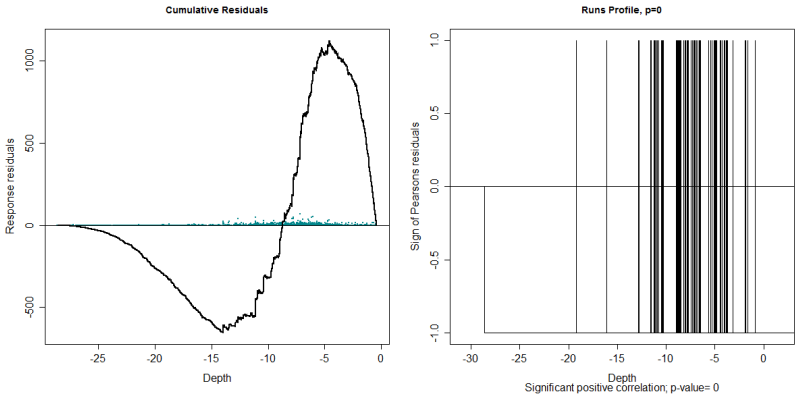


Figure 117: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by depth when depth is fitted as a linear term in the model.

Figure 133 indicates that the smaller predicted values are systematically over-predicted (a common issue since zeros can never be predicted exactly under the model) and the larger predicted values are under-predicted. This confirms what was seen in the plot of observed vs fitted values (Figure 114) where the very large values are hard to predict. Additionally, the earlier observations (pre-impact) are over predicted while the latter observations (post-impact) appear to be under-predicted (Figure 119).

The runs plots test for randomness of the residuals given the order of the data, and in all cases the p-values are < 0.05 and indicate positive correlation (right-hand plots in Figures 116 –119). The mixing is best for the residuals in time order (i.e. there are lots of runs), but this number is still fewer than would be expected if the residuals were random. This confirms that modelling temporal correlation through the use of GEEs is justified.

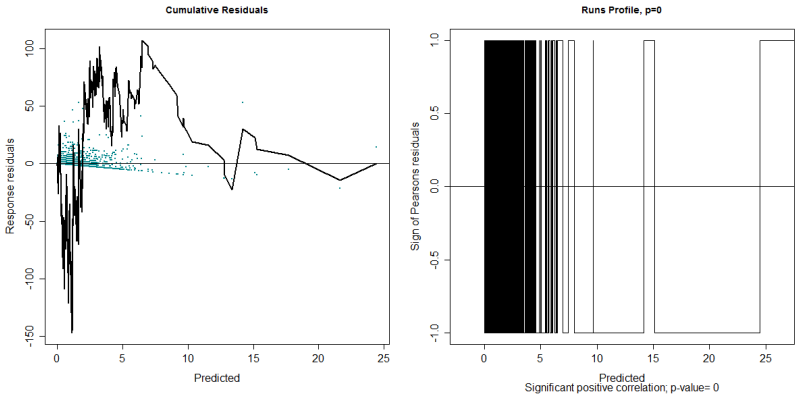


Figure 118: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by the predicted value. The blue points are the residual values and the black line represents the cumulative residuals.

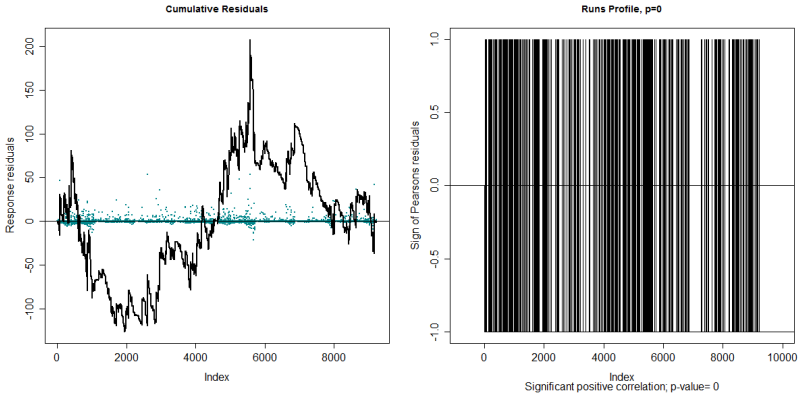


Figure 119: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by the index of observations (ordered temporally). The blue points are the residual values and the black line represents the cumulative residuals.

The quality of the spatial predictions can be assessed by examining model residuals spatially (Figure 120). This can help the user to ascertain if some areas of the surface are over or under-predicted. In this case, the highest residuals lie in the central area and down to the eastern side where the highest counts were observed; this is not surprising given the very large numbers were difficult to predict. There also appears to be some over-prediction in the area to the east of the area pre-impact but no corresponding systematic under-prediction.

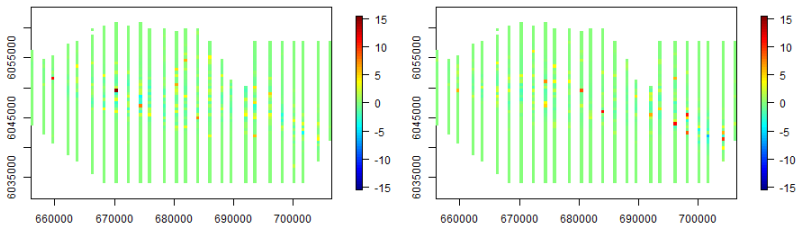


Figure 120: Raw residuals before impact (left) and after impact (right). These residuals are observed-fitted values and so positive residuals imply under-prediction and negative residuals imply over-prediction.

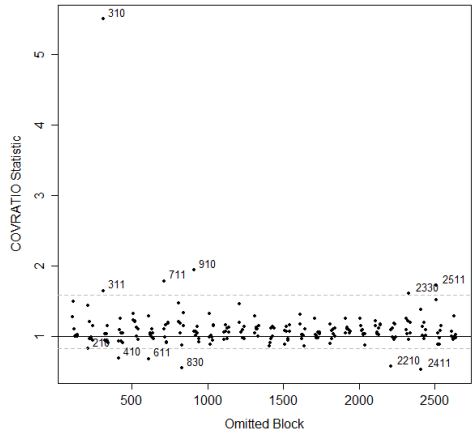
To ensure model conclusions do not depend on a small number of blocks in our data, the COVRATIO and PRESS statistics should also be examined (Figure 121). The COVRATIO plot (left hand plot)

indicates the change in precision of the parameter estimates when each block is removed and so scores less than one indicate that the removal of the block will increase precision of the parameter estimates (and vice versa for scores$>$ 1). In particular points which lie far below the lower quantile are worth investigating further.

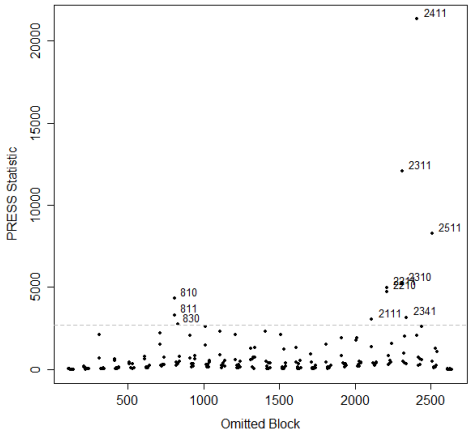
In this case there is a marked decrease in the standard errors when blocks '830', '2210' or '2411' are removed. The latter block contains the high numbers of birds as a part of the post-impact redistribution and removing blocks with large numbers (such as these) will naturally lower the variance. Blocks '830' and '2210' also both have relatively high counts in each, and so removing these blocks would also naturally lower the variance.

The PRESS statistic (right hand plot) quantifies the sensitivity of model predictions to removing each block, and model predictions are sensitive to blocks: '2511', '2311' and '2411' in particular, if removed. These are transects 23-25 in season 1 post-impact. The sensitivity of model predictions to these transects is unsurprising since these are the sites for the re-distribution post-impact (second to fourth transects in from the right hand side of Figure 109).

In this case, we are not alarmed by the blocks that the model is sensitive to and so we are happy to use the model for interpretation and prediction.



(a)



(b)

Figure 121: Plots of influence measures. (a) COVRATIO statistic; the dashed grey lines indicate the lower 2.5% and upper 97.5% quantiles of the statistics and (b) PRESS statistic; 95% of the statistics fall below the dashed grey lines. Labelled points on both plots are outside the grey dashed line(s) and are labelled with the identifier for the block that has been removed to create the statistic (not an observation number).

10.2.7 *Prediction and Inference*

Since we are satisfied with the model we can make predictions pre and post impact, and calculate any pre-post differences, across a grid in the area of interest. We can also make inference about these predictions and any differences, by combining the parameter uncertainty associated with the detection function and model fitting processes.

Based on the model, declines in bird density in the central region and increases in the south east of the study region are evident post-impact (Figure 122). Additionally, since the central region is where the impact was imposed the results suggest the birds have moved from the impact site to the south-eastern area post-impact.

Alongside model predictions, the upper and lower confidence intervals pre and post impact (Figure 123) are crucial in putting the predictions into perspective, and ultimately in determining genuine differences. In this case, the relatively high density of birds in the south-east persist even in the lower confidence limits. While the upper and lower confidence limits pre and post impact are useful, it is impossible to tell (from these figures alone) where any significant differences may occur, and for this reason the mean differences pre and post impact are also of interest (Figure 124. In this case, the mean differences in birds/km^2 pre and post impact are clear; there is a significant decline in animals around the impact site (even though the location of the impact was unknown to the model) and a significant increase in animals in the south-east, implying redistribution.

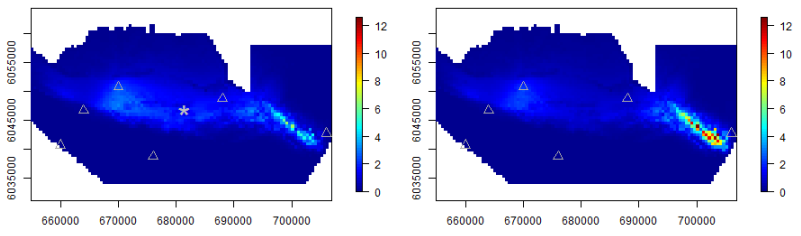
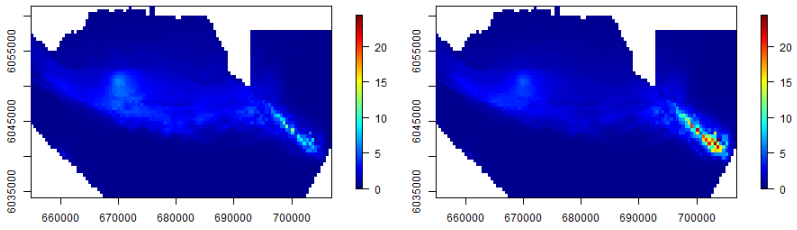
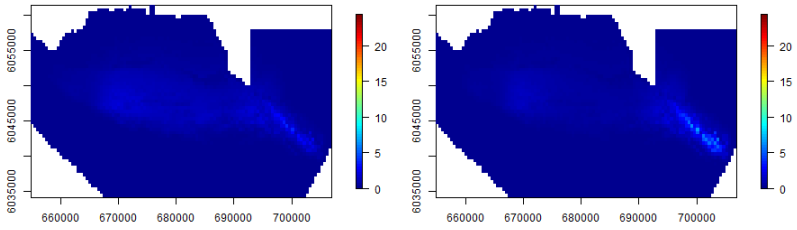


Figure 122: Predictions of bird density (birds/km^2) from the fitted model for before (left) and after (right) an impact event. The grey triangles indicate the location of the 6 knot locations for the smooth of space and the grey star is the site of the impact event (e.g. a wind turbine development)



(a)



(b)

Figure 123: (a) upper and (b) lower 95 percentile confidence intervals of bird density (birds/km^2) from the fitted model for before (left) and after (right) an impact event.

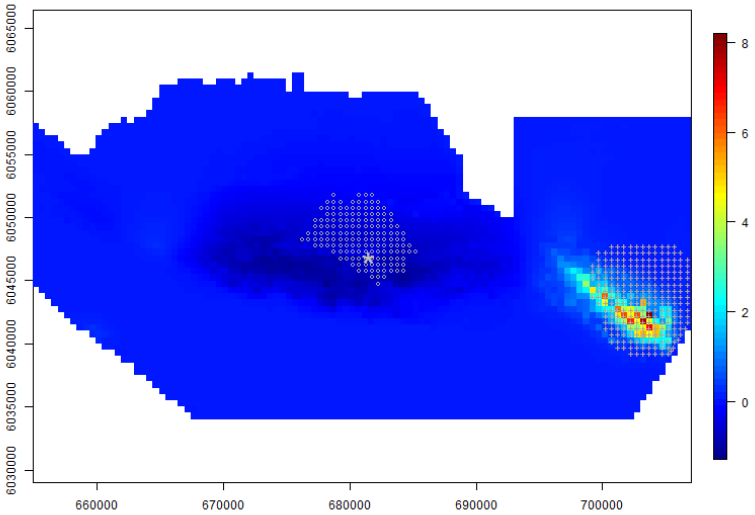


Figure 124: A plot of the mean difference in predicted bird density (birds/km^2) before and after impact. Positive values indicate more birds post impact and negative values fewer birds post impact. Significant differences were calculated using percentile confidence intervals: '+' indicates a significant positive difference and 'o' a significant negative one. The grey star is the site of the impact event.

10.3 Comparison to the Truth

In this case, we manufactured the data and so are able to compare our modelling results with the surface(s) which generated the data.

The results obtained by the model were in line with the surfaces generated. For instance, the detection function scale parameter (120) was contained within the 95% confidence interval (112.7, 122.4) and the generated data were over-dispersed and positively correlated in blocks – both features which were recovered by the model.

The simulated redistribution post-impact was also recovered (via the significant interaction term), and the redistribution patterns were consistent with data generation (Figure 124). Further, there was no estimated change in average numbers post impact (as evidenced by the impact related p-value for the non-interaction model), a feature that is also true under the model.

11 *Impact assessment for off-shore data; worked example*

This section contains a worked example for a quantitative impact assessment based on simulated near-shore data. We begin by describing how the data was manufactured and then detail the analysis process. The analysis process involves:

1. exploration of the data
2. model specification & fitting
3. diagnosing model faults
4. using the chosen model to make predictions & quantifying any differences
5. expressing the uncertainty about covariate relationships and associated model predictions

Finally, since the data in this case was manufactured we are able to compare the results with the process used to generate the data – the 'truth'.

Note: for brevity we have assumed the reader is familiar with the off-shore example.

11.1 *Manufacturing the data*

The near-shore data were generated using pre-impact observations and augmented by imposing an impact in the central of the study region. The impact caused a reduction in animal numbers in the area which was matched by a re-distribution into the southern part of the study area (Figures 125 and 126); there was no change in the total number of birds pre and post-impact.

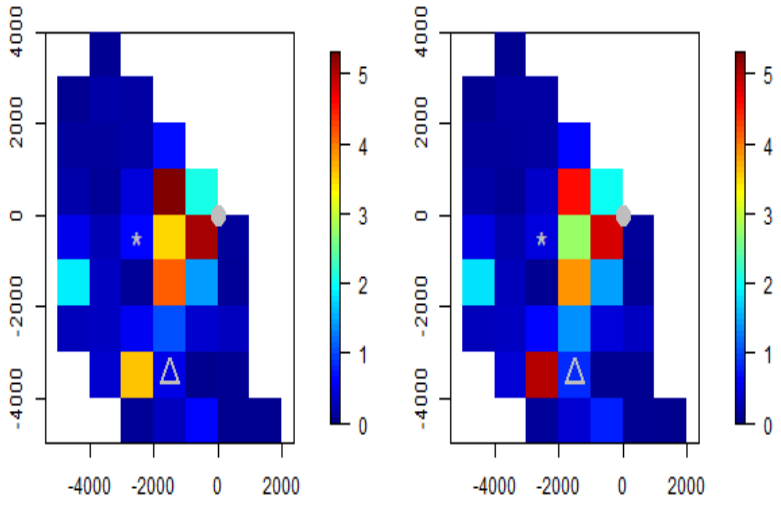


Figure 125: Simulated truth before impact (left) and after impact (right). The grey star indicates the centre point of the impact and the grey triangle, the centre point of the re-distribution. The grey circle is the location of a cliff-top observer.

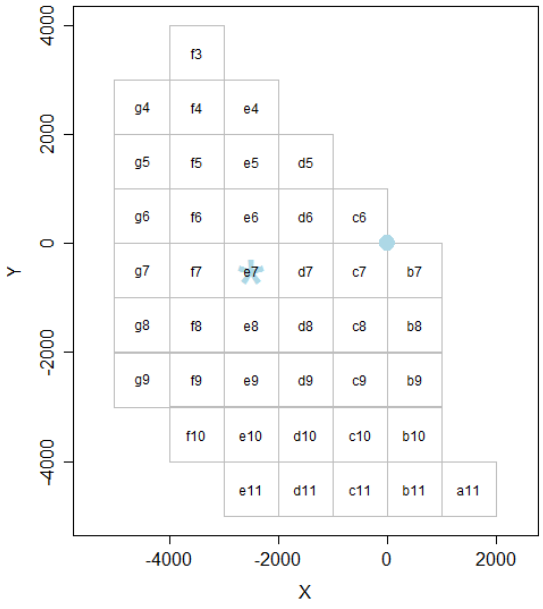


Figure 126: Grid cell identifiers for the study region. The grey circle is the location of a cliff-top observer and the star the site of the impact event.

The observed data (e.g. Figure 127) were lifted from this simulated reality and correlated noise (with extra-Poisson variability) was added. Each observation cell was approximately 1 km^2 except in cases with land, where the effort in those cells was less than one.

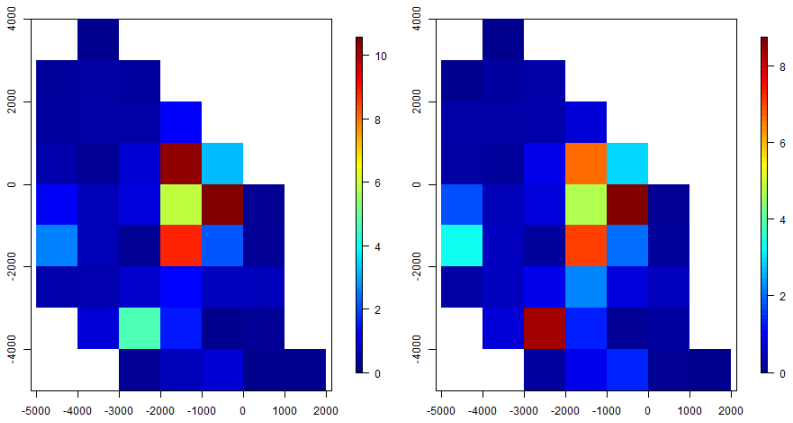


Figure 127: Observed data before (left) and after impact (right). Each cell is 0.5 km^2 and the colour represents mean bird count.

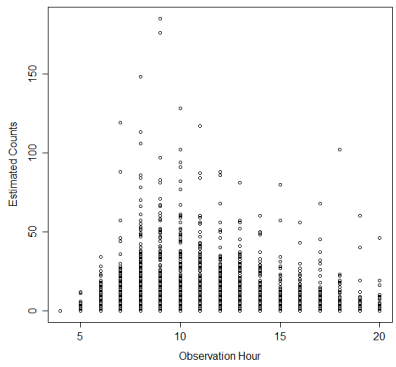
11.2 *Statistical analysis*

The data analysis for the worked example involves exploring the data, model specification and fitting followed by some diagnostics. Following the selection of an appropriate model, the model is used to make predictions, generate ranges of plausible values for these predictions (95% confidence intervals) and on this basis identify any spatially explicit differences. The results produced in this sec-

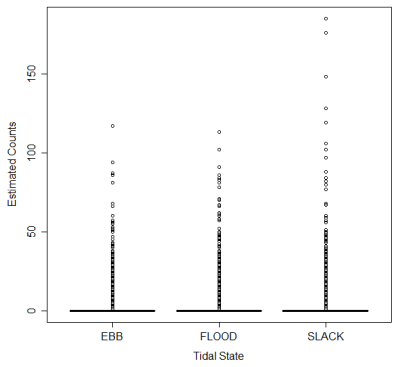
tion are then compared with the surface used to generate the data in this case.

11.2.1 Data exploration

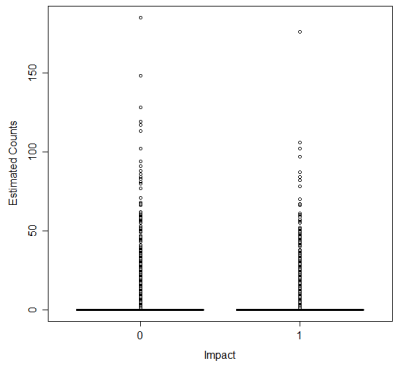
Animals were predominantly seen in the morning hours (7am-12pm; Figure 128) while fewer birds were seen very early or very late in the day. Further, the relationship between observation hour and bird count appears to be non-linear. It is difficult to identify any relationship between tidal state or impact and bird counts due to the large number of zeros in the data.



(a)



(b)



(c)

Figure 128: Plots of (a) observation hour, (b) tidal state and (c) Impact against the observed bird counts.

11.2.2 *Model specification*

The data are over-dispersed counts (and generated using a dispersion parameter of 9.19) and so a model with quasipoisson errors was fitted.

The main candidate terms considered for the model were:

- a smooth term for Observation Hour (*s(ObservationHour)*)
- a spatial term (*s(XPos,YPos)*)
- a three-level factor variable term for tidal state (*FloodEbb*)
- a two-level factor variable term for impact (e.g. windfarm installation/operation)

In order to quantitatively assess if a post-impact re-distribution had occurred, a spatial-impact interaction term was also included in the model. An 'offset' was also specified to allow for variable search effort across the estimated counts. For example, most segments were 1km^2, but some contained land resulting in a smaller search area. For prediction, the area was given as 1km^2 and so any predictions will be a bird density, per km^2.

In keeping with the off-shore example, VIFs were used to assess prohibitive levels of collinearity. In this case, however, there were no collinear variables (Table 12).

Variable	**GVIF**	**Df**	**GVIF$_{adj}$**
ObservationHour	1.05	3	1.01
FloodEbb (tide state)	1.05	2	1.01
s(X,Y)	5.46×10^4	15	1.44
Impact	5.54	1	2.35
s(X,Y):Impact	1.74×10^5	15	1.50

Table 12: Variance inflation factors to assess co-linearity amongst covariates. GVIF is a generalised variance inflation factor, and GVIF$_{adj}$ is GVIF adjusted for the degrees of freedom (GVIF$^{1/2*Df}$).

11.2.3 *Model Fitting and model selection*

A CReSS-GEE framework was employed here to estimate the smooth terms in the model whilst allowing for positive autocorrelation in model residuals. The fitted model included a relatively smooth function specified for observation hour (with just 3 parameters) and a smooth spatial term which employed SALSA-based model selection to determine flexibility. SALSA can also be used to choose the smoothness of the observation hour relationship (see the user guide for the `MRSea` package for more details). The tide state and impact terms were fitted as factor variables. In this case, we have very few terms in the model and it was decided to use p-value based model selection (since these are corrected for residual correlation under a GEE approach).

11.2.4 *Results*

A moderately low dimensional smooth term was chosen for the spatial element ($df = 15$) and the smooth surface was spatially adaptive; the surface comprised a mixture of local and globally acting basis functions.

Despite the covariates included in the model, there was compelling evidence for time-based correlation in (Pearson) residuals; a runs test returned a very small p-value ($p << 0.05$). As a result, a GEE fitting framework was employed using a *grid code-year-month-day* blocking structure – here residuals are permitted to be correlated within grid cell-days (e.g. days within months and years). This structure returned 5576 unique blocks with up to 11 observations in each block. For some blocks the correlation is low at any lag, whilst others have high correlation even for residuals 7 time points apart (Figure 129).

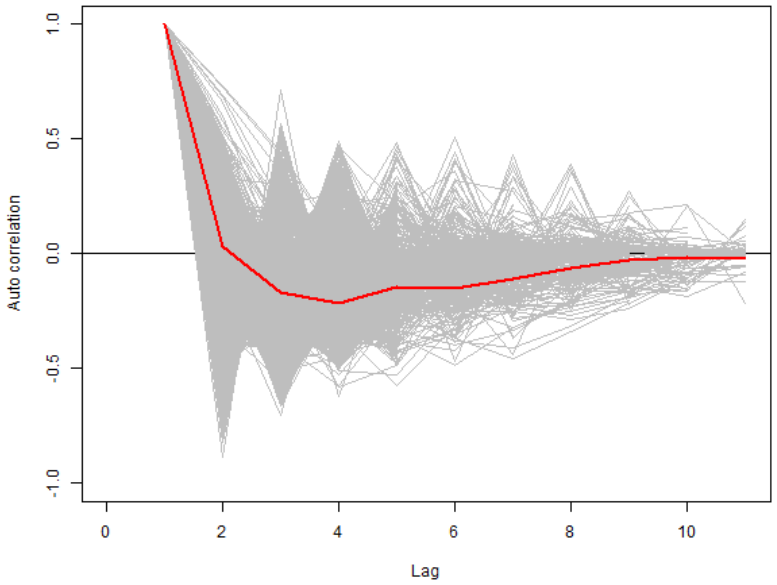


Figure 129: Plot of the correlation in residuals for each block (grey lines). The mean correlation at each lag is indicated in red.

The model returned p-values for observation hour ($df = 3$), tide state ($df = 2$), the spatial term ($df = 15$), and the factor variables for impact ($df = 1$). There are also 15 spatial parameters describing the difference between a pre and post-impact surface (Table 13). Based on these marginal p-values, Observation hour, tide state, and the spatial-impact interaction term are all highly significant (Table **??**) and thus were all retained in the model.

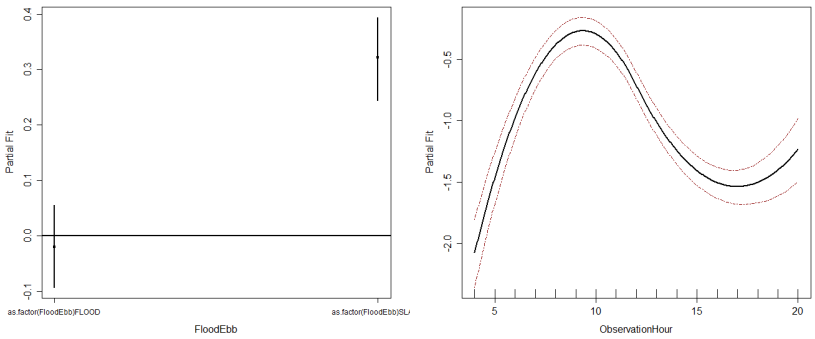
To investigate if there was a change in average numbers pre and post-impact, the spatial-impact interaction term was removed and the remaining terms re-fitted in the model; there was no evidence for an overall difference ($p = X$) pre and post-impact.

Based on this analysis, average numbers of animals appear to

Variable	p**-value**
Observation Hour	< 0.0001
FloodEbb (tide state)	< 0.0001
s(X,Y)	< 0.0001
Impact	0.561
s(X,Y):Impact	0.019

Table 13: GEE based p-values (to 3 decimal places) for a CReSS model with SALSA knot placement. $p < 0.05$ suggests that the covariate should be retained in the model.

be significantly higher for 'Slack' tidal state compared with the baseline ('Ebb') state (Figure 130(a)), however there is no significant difference between numbers in Flood and Ebb tide states. The relationship for observation hour (Figure 130(b)) indicates a non-linear relationship with bird counts (i.e. fewer birds present in the very early and later hours of the day) which peaks at about 10am. The confidence intervals about this relationship are also small which means some interpretation about the curve is possible – if the confidence intervals for observation hour were, instead, very wide and in particular a straight line would suffice, observation hour could instead be modelled as a linear term.



(a) (b)

Figure 130: Partial Plots for (a) tidal state and (b) Observation-Hour.

11.2.5 *Diagnostics*

The fit plot shows the model is not ideal (Figure 131; $R^2_{MARG} = 0.23$ & $r_c = 0.37$) and in particular, (but not unusually) the very large values are difficult to predict. Of note is the large number of points in the plot and thus, it is difficult to see what proportion of points is not well described by the model.

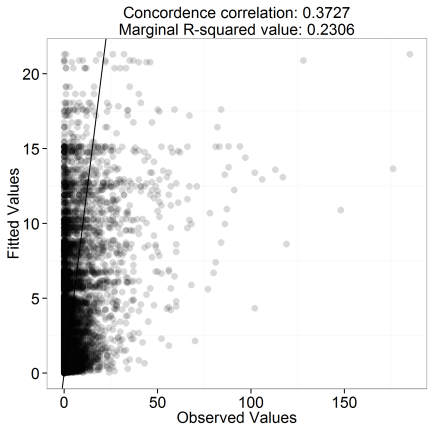


Figure 131: Assessing model fit: Observed vs fitted values. The black line indicates where data should lie for a perfect fit. The data points are adjusted for over-plotting and the heavier the shading, the more over-plotting.

There is no evidence for unmodelled patterns in the mean-variance relationship as evidenced by the locally weighted least squares regression line (Figure 132). For this reason, we are happy with the variance assumed under the model, in this case.

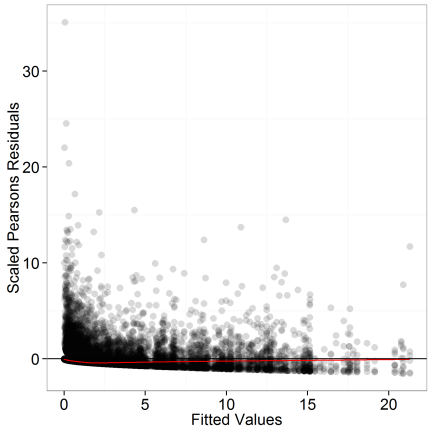


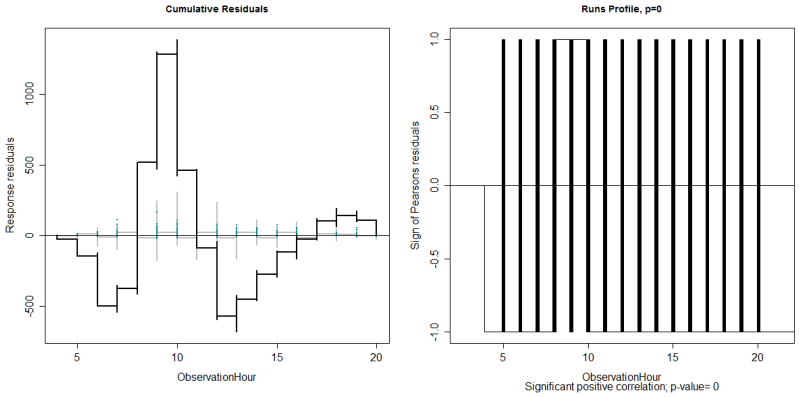
Figure 132: Assessing model fit: Fitted vs scaled Pearsons residuals. The red line is a locally weighted least squares regression to indicate pattern in the plot, which might otherwise be hidden due to over-plotting. The data points are adjusted for over-plotting and the heavier the shading, the more over-plotting.

A plot of cumulative residuals allows us to see if there is any systematic over or under-prediction. The left hand plots in Figures 133-135 show cumulative residuals across the range of observation hours (Figure 133), predictions (Figure 134) and, by ordering the data, temporally (Figure 135).

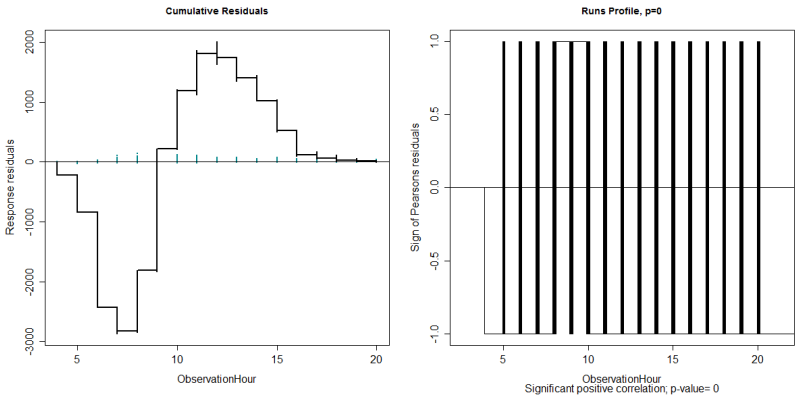
Under the model, observation hours between 8-11am are systematically under-predicted (Figure 133(a)) while animal numbers outside of this time period are over-predicted. This relationship might further be improved if model selection is used to choose the flexibility of the term, and while the improvements are clear over modelling observation hour as a linear term (Figure 133(b)) allow-

ing the relationship to be more flexible shows much better mixing about the line (Figure 133(c)). The runs profiles are more difficult to interpret for discrete variables (Figure 133(right)) and regardless of how observation hour was fitted in the model, the runs tests signal positive correlation.

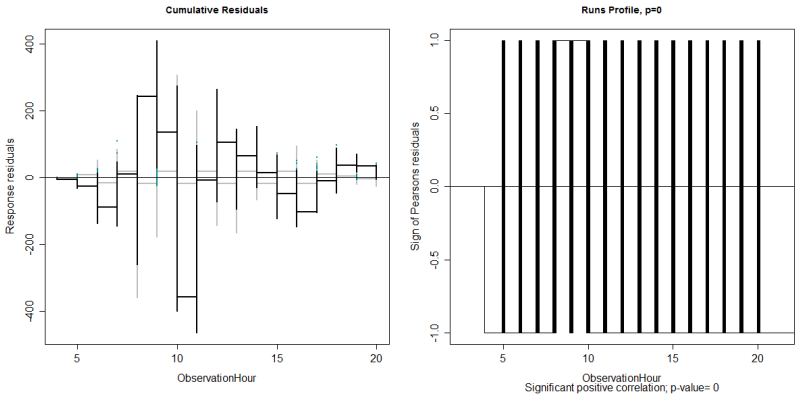
In summary, the smaller predictions (1-3 birds) appear to be too small, while over-prediction was the tendency up to approximately 13 birds (Figure 134). Thereafter both under and over-prediction are evident, but there are fewer points to consider. Temporally, there is a good deal of mixing about the horizontal line (Figure 135) and while there are a large number of corresponding runs, these are still fewer than expected for random residuals. We have modelled residual correlation using GEEs, and so are unconcerned about this result.



(a)



(b)



(c)

Figure 133: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by Observation Hour. (a) the results of the current model, (b) when Observation Hour is fitted as a linear term and (c) when it is fitted with three knots at the 25, 50 and 70^{th} quantiles. In the cumulative residual plots, the blue dots are the residual values, the black line the cumulative residuals and the grey line in the background is based on a curve with an extremely large number of parameters (an over-fitted model).

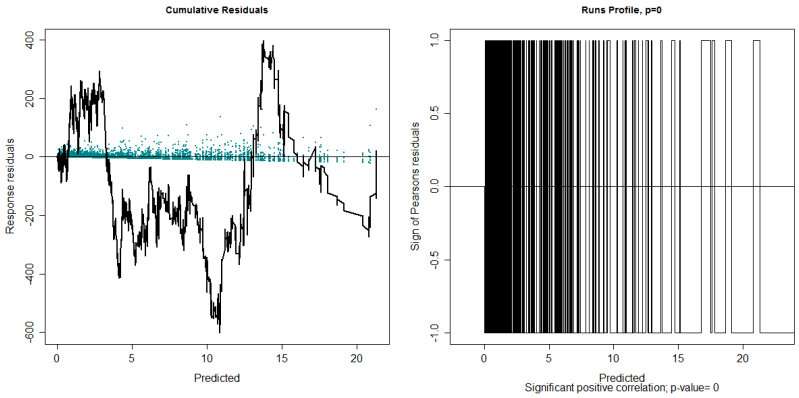


Figure 134: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by predicted value. In the cumulative residual plot, the blue dots are the residual values and the black line the cumulative residuals.

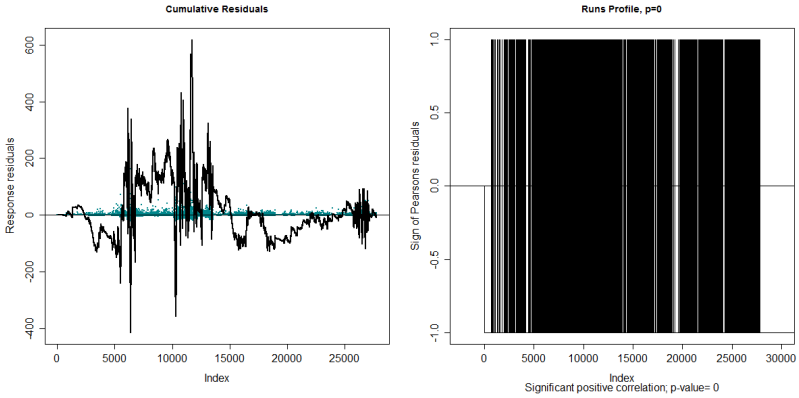


Figure 135: Cumulative residual plots (left) and runs profiles (right) for residuals ordered by index of observations (ordered temporally). In the cumulative residual plot, the blue dots are the residual values and the black line the cumulative residuals.

Spatially, the cell to the right of the cliff top observer is under-predicted by the model both pre and post-impact (Figure 136) but otherwise, there are no compelling spatial patterns and so does not give rise to any concerns.

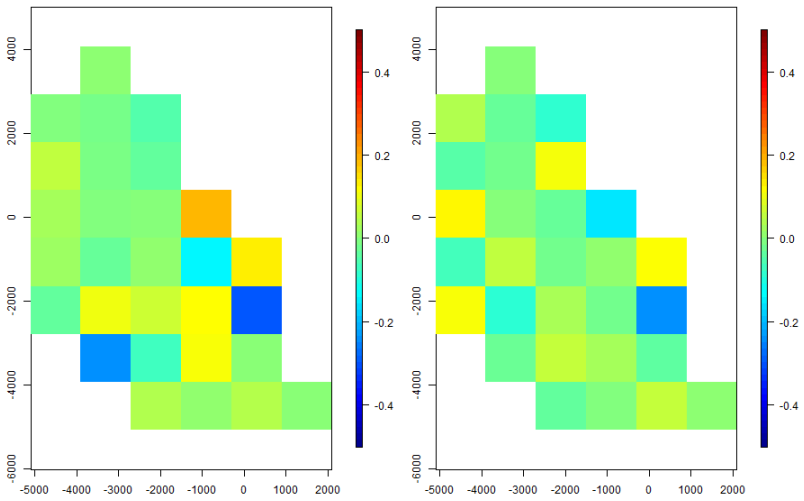
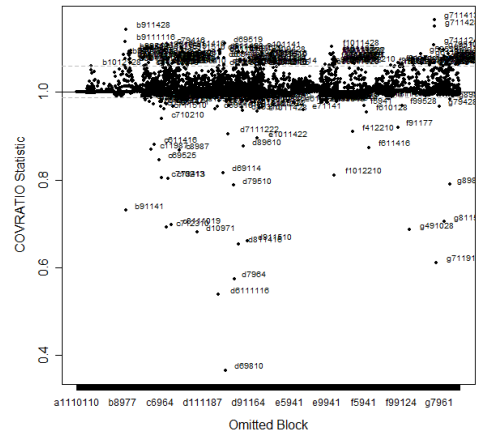
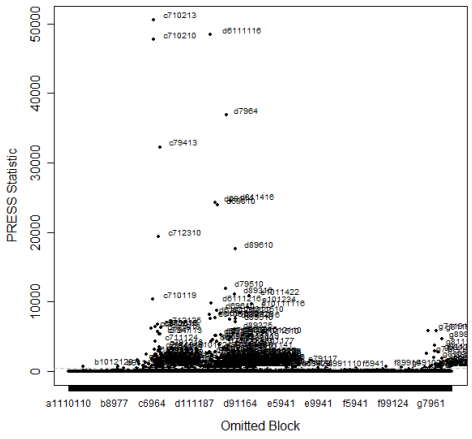


Figure 136: Raw residuals before impact (left) and after impact (right). These residuals are observed-fitted values and so positive residuals imply under-prediction and negative residuals imply over-prediction.

The COVRATIO statistics signal a marked decrease in the standard errors when block 'd69810' is removed (Figure 137) which corresponds to cell 'd6' in year 2009 (the first of the two years before impact), month 8 and the 10^{th} day of the month. The two cells above ('d6111116' and 'd7964') are also notable in this regard. High values for these blocks are unsurprising – they all have very high bird counts, compared with other blocks, and so thus add considerably to the variability in the data. This is routine for count data and so is of no concern here.

The PRESS statistic (Figure 137(b)) signals that model predictions are sensitive to blocks 'c710213', 'c710210' and 'd6111116'. These cells area all 'before' impact and all contain large counts and so it is unsurprising to find they stand out as influencing model predictions. In this case we can see why the blocks identified are notable and so accept they are not unduly affecting the model for the wrong reasons (e.g. a block with peculiar values).





(b)

Figure 137: Plots of influence measures. (a) COVRATIO statistic; the dashed grey lines indicate the lower 2.5% and upper 97.5% quantiles of the statistics and (b) PRESS statistic; 95% of the statistics fall below the dashed grey lines. Labelled points on both plots are outside the grey dashed line(s) and are labelled with the identifier for the block that has been removed to create the statistic (not an observation number).

11.2.6 *Prediction and Inference*

The prediction plots show a decline in bird density in the central region (Figure 138) and an increase in the south of the study region. This central region is where a known impact has taken place and so the results suggest that birds have moved from this region to the south post-impact. The lower and upper confidence interval for before impact (Figure 139, left-hand plot) shows a higher density of birds in the central and south region. Post-impact however, both the lower and upper confidence limits represent a decline in density in the central region and an increase in one of the southern grid cells.

This redistribution is clearer in Figure 140 (page 139) which shows the mean difference in birds/km^2 between before and after and in particular, where any significant differences lie. In this case, there is a significant decline in animals in three grid cells to the east and north of the impact site and a significant increase in animals in the south (four grid cells).

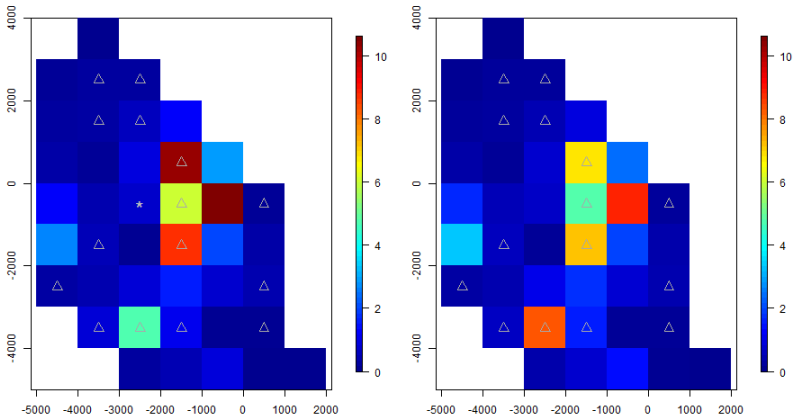
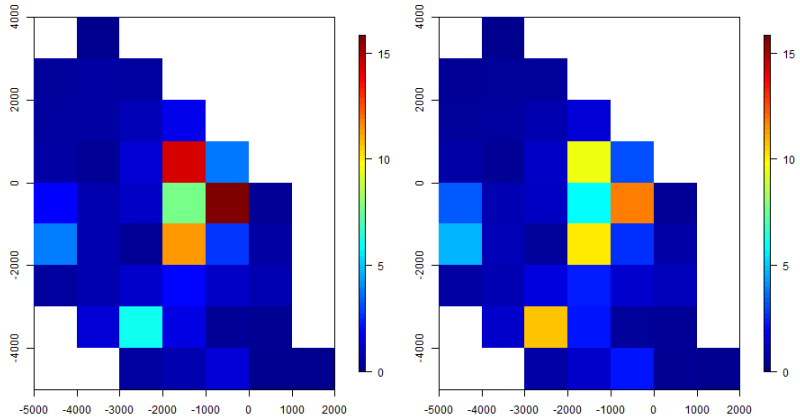
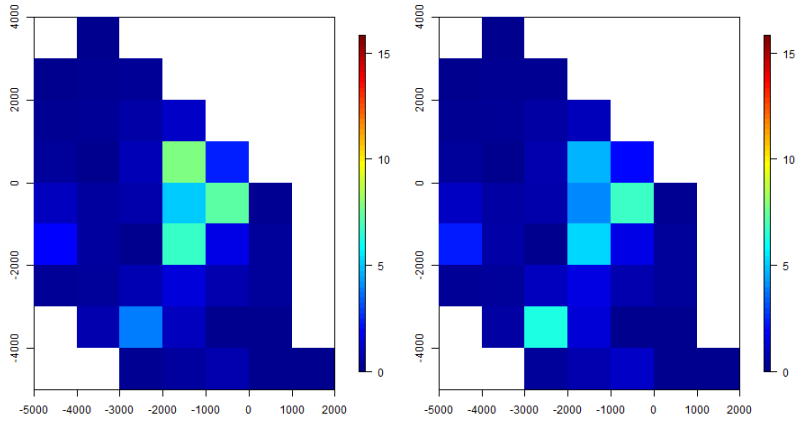


Figure 138: Predictions of bird density (birds/km^2) from the fitted model for before (left) and after (right) an impact event. The grey triangles indicate the location of the 6 knot locations for the smooth of space and the grey star is the site of the impact event (e.g. a wind turbine development)



(a)



(b)

Figure 139: (a) upper and (b) lower 95 percentile confidence intervals of bird density (birds/km^2) from the fitted model for before (left) and after (right) an impact event.

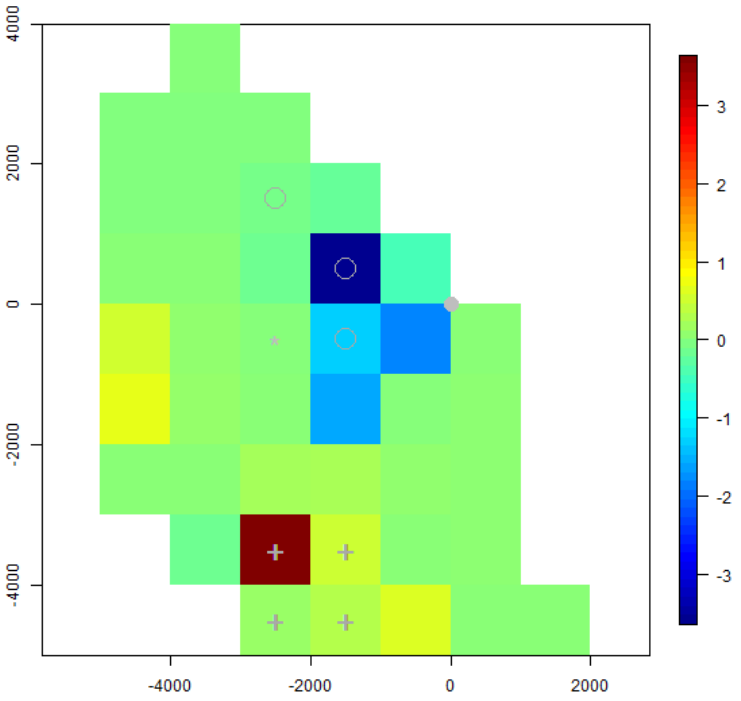


Figure 140: A plot of the mean difference in predicted bird density (birds/km^2) before and after impact. Positive values indicate more birds post-impact and negative values fewer birds post-impact. Significant differences were calculated using percentile confidence intervals: '+' indicates a significant positive difference and 'o' a significant negative one. The grey star is the site of the impact event.

11.3 Comparison to the Truth

Our conclusions were well aligned with the simulated reality. The generated data were over-dispersed and positively correlated in blocks, both of which were recovered by the model.

The impact effect was also recovered; the simulated impact effect was a redistribution of animals away from the impacted area into the south, while the overall abundance was constant pre and post-impact. This redistribution was picked up by the model with a significant interaction term and a refitted model without the interaction effect correctly signalled no underlying change in abundance pre and post-impact. While a significant reduction in birds was found around the impact site, this was not the case for the cell in the centre of the impact; we expect this is due to the fact that so few animals were present in the first instance. In general, the model adequately reduced and re-allocated animals to the correct locations (Figure 140), in the face of highly correlated and over-dispersed count data.

12 Appendix

12.1 Model choice results for the off-shore scenarios

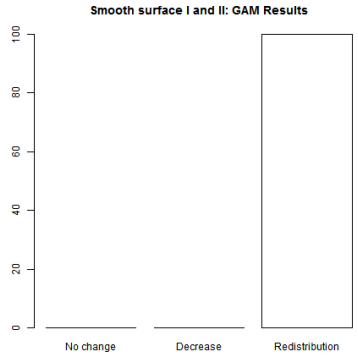


Figure 141: GAM-based model selection results for data generated with the GAM-based surface and Model I or II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

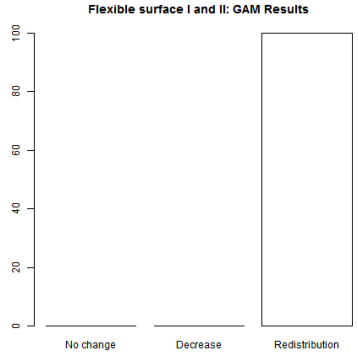


Figure 142: GAM-based model selection results for data generated with the CReSS-based surface and Model I or II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

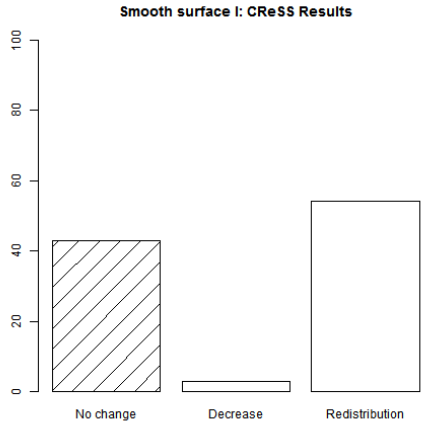


Figure 143: CReSS-based model selection results for data generated with the GAM-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

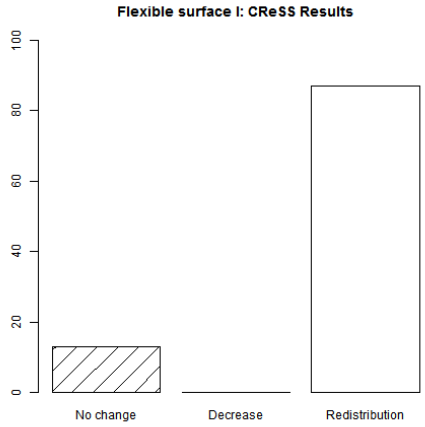


Figure 144: CReSS-based model selection results for data generated with the CReSS-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

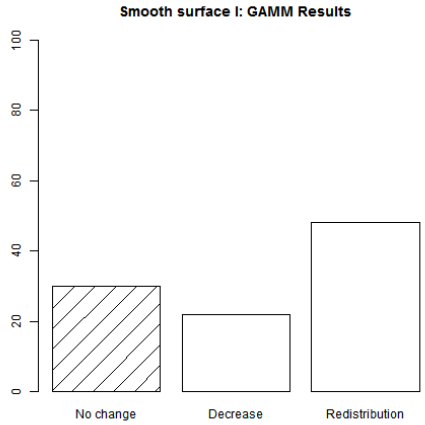


Figure 145: GAMM-based model selection results for data generated with the GAM-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

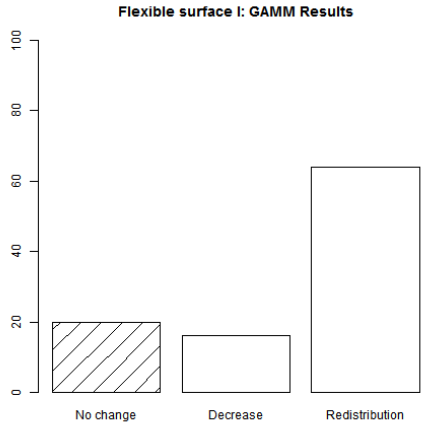


Figure 146: GAMM-based model selection results for data generated with the CReSS-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The red columns represents the models which are incorrectly chosen while the green column in each case represents the correct model.

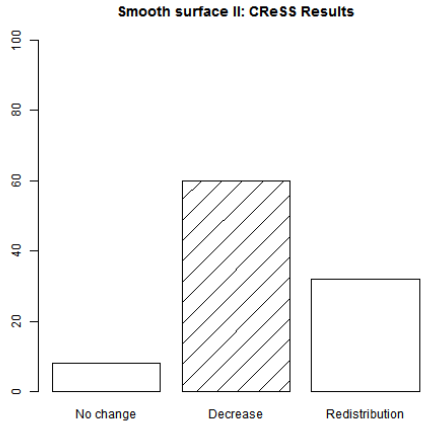


Figure 147: CReSS-based model selection results for data generated with the GAM-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

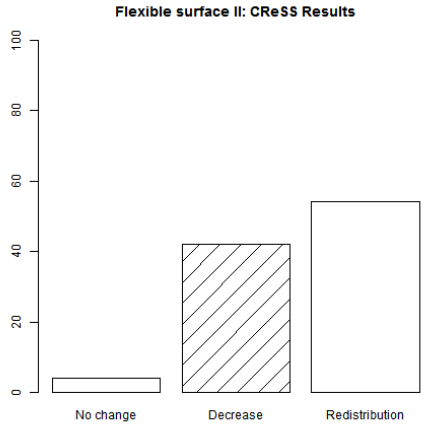


Figure 148: CReSS-based model selection results for data generated with the CReSS-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

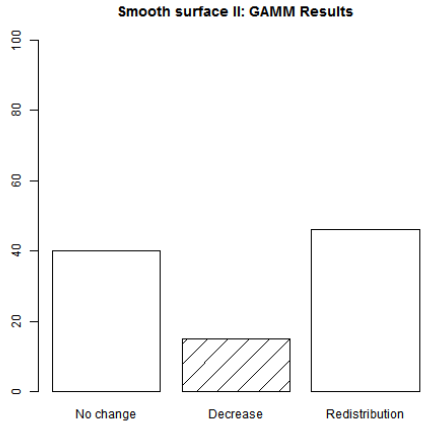


Figure 149: GAMM-based model selection results for data generated with the GAM-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The red columns represents the models which are incorrectly chosen while the green column in each case represents the correct model.

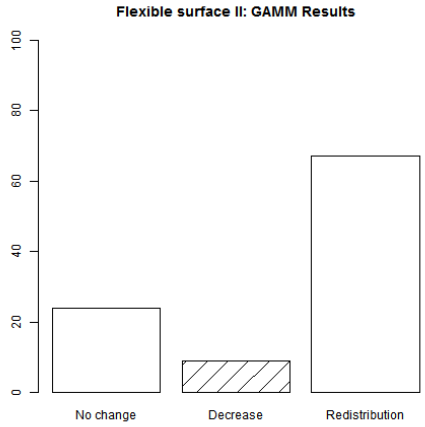


Figure 150: GAMM-based model selection results for data generated with the CReSS-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

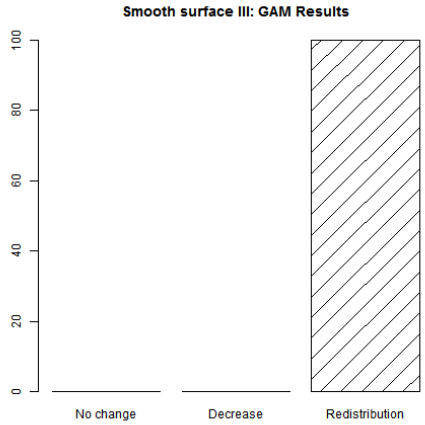


Figure 151: GAM-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct modell.

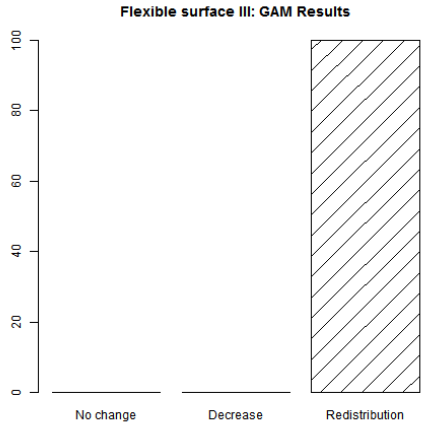


Figure 152: GAM-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

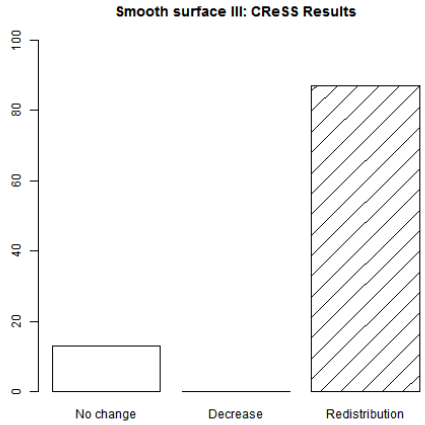


Figure 153: CReSS-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

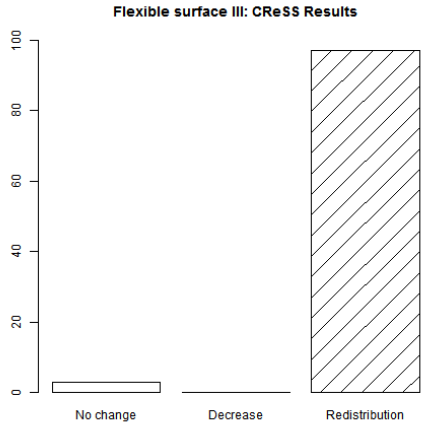


Figure 154: CReSS-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

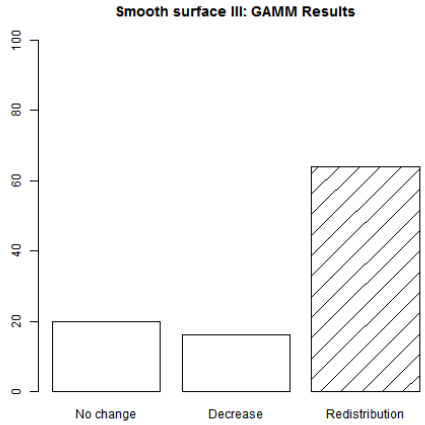


Figure 155: GAMM-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

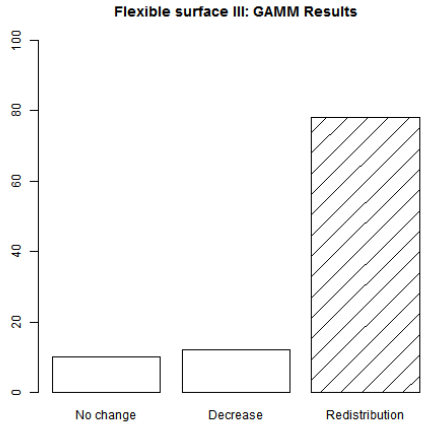


Figure 156: GAMM-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

12.2 *Model choice results for the near-shore scenarios*

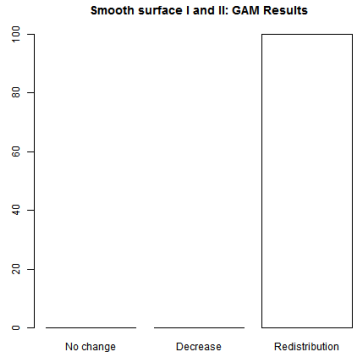


Figure 157: GAM-based model selection results for data generated with the GAM-based surface and Model I or II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

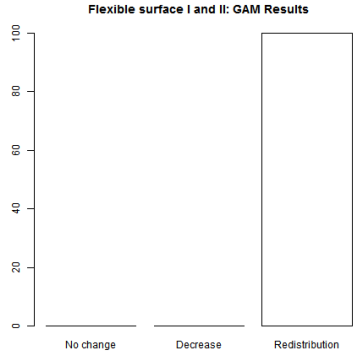


Figure 158: GAM-based model selection results for data generated with the CReSS-based surface and Model I or II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

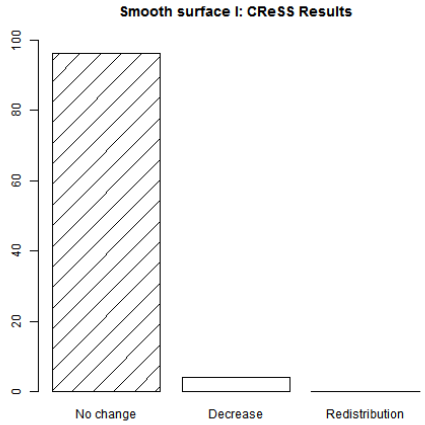


Figure 159: CReSS-based model selection results for data generated with the GAM-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

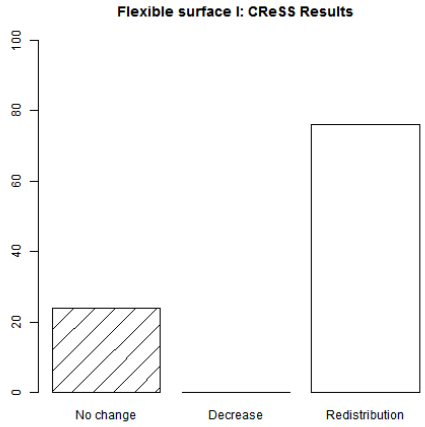


Figure 160: CReSS-based model selection results for data generated with the CReSS-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct modell.

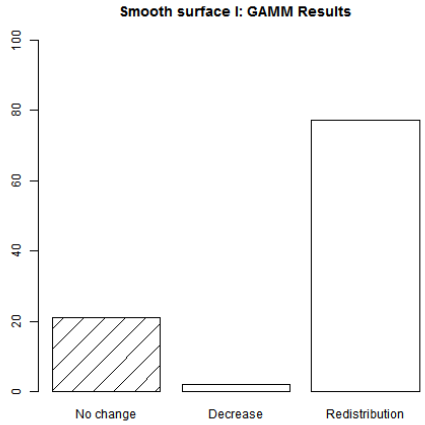


Figure 161: GAMM-based model selection results for data generated with the GAM-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The red columns represents the models which are incorrectly chosen while the green column in each case represents the correct model.

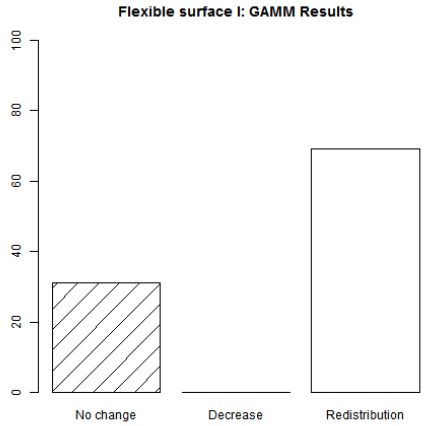


Figure 162: GAMM-based model selection results for data generated with the CReSS-based surface and Model I. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

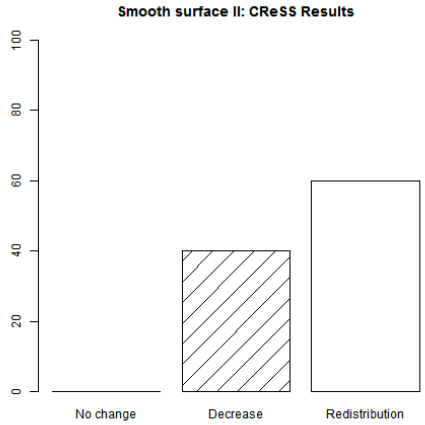


Figure 163: CReSS-based model selection results for data generated with the GAM-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

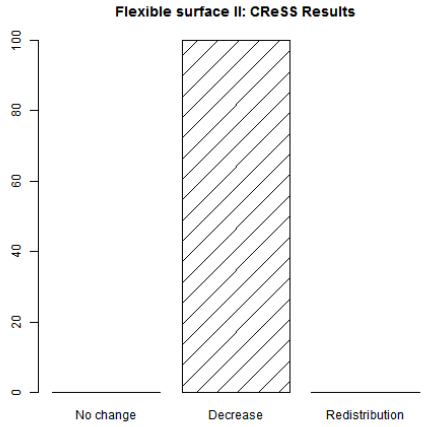


Figure 164: CReSS-based model selection results for data generated with the CReSS-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

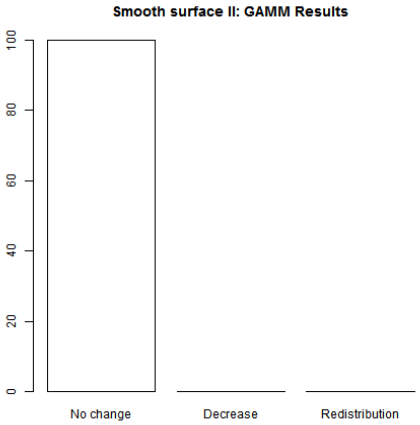


Figure 165: GAMM-based model selection results for data generated with the GAM-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

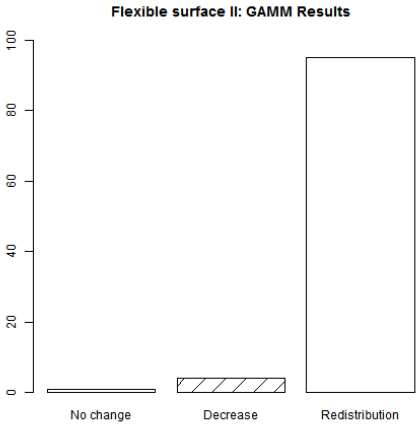


Figure 166: GAMM-based model selection results for data generated with the CReSS-based surface and Model II. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

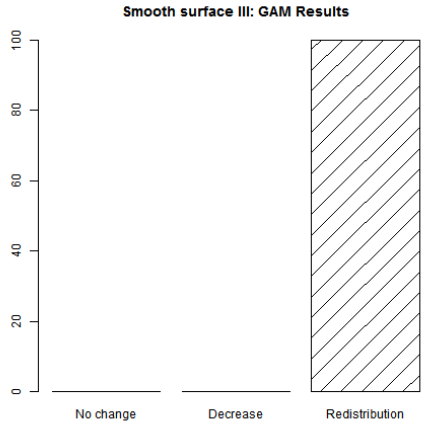


Figure 167: GAM-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

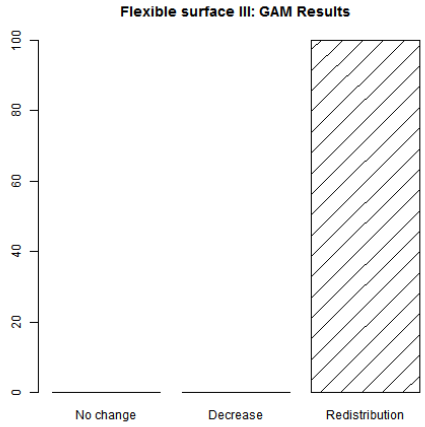


Figure 168: GAM-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

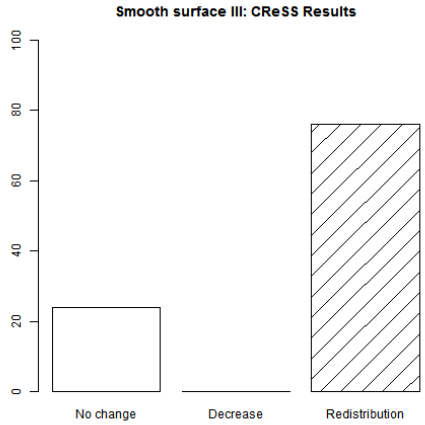


Figure 169: CReSS-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

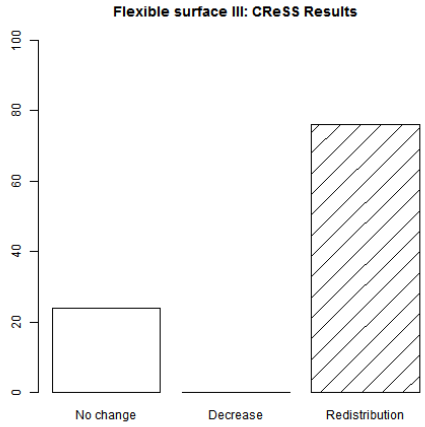


Figure 170: CReSS-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

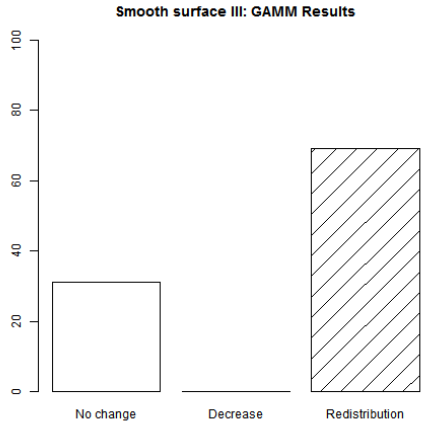


Figure 171: GAMM-based model selection results for data generated with the GAM-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

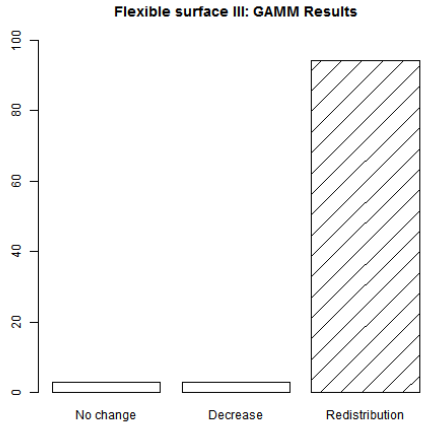


Figure 172: GAMM-based model selection results for data generated with the CReSS-based surface and Model III. The label on the horizontal axis details the candidate model (I, II or III) and the vertical axis represents the frequency of the 100 realisations which selected each model. The columns without pattern represents the models which are incorrectly chosen while the patterned column in each case represents the correct model.

12.3 *Spatially explicit performance for the off-shore scenarios*

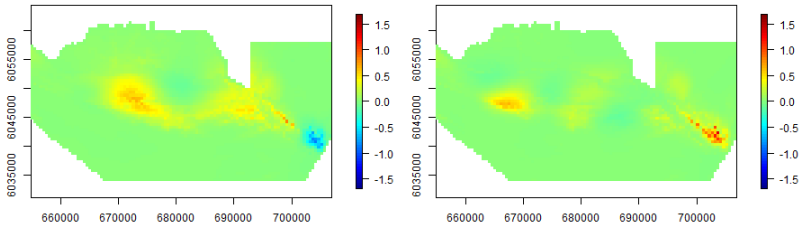


Figure 173: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

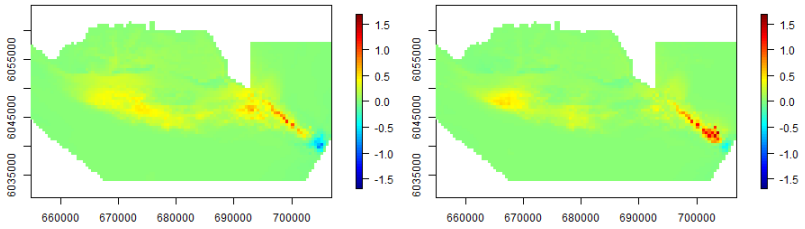


Figure 174: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

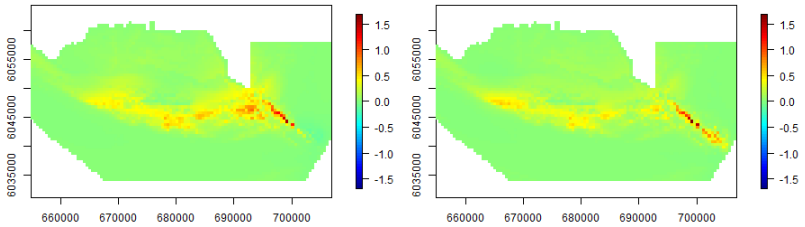


Figure 175: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

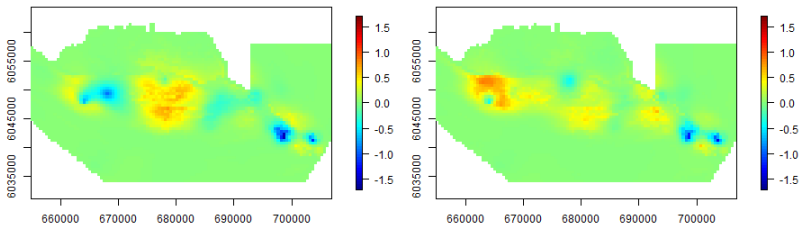


Figure 176: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

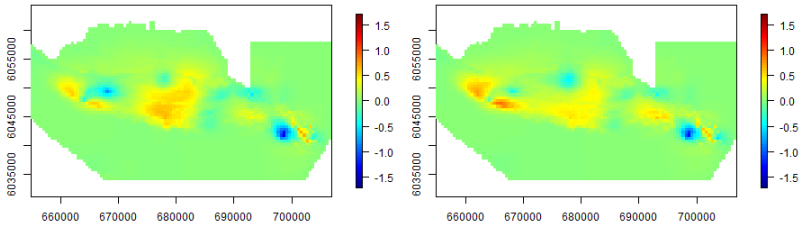


Figure 177: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

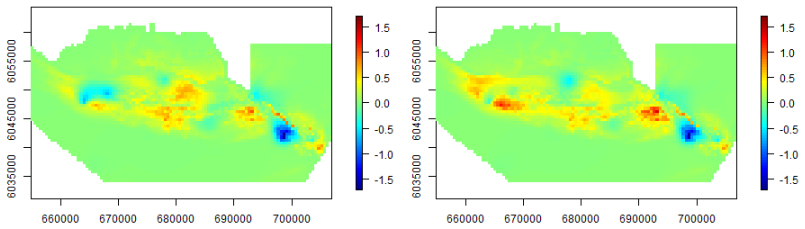


Figure 178: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

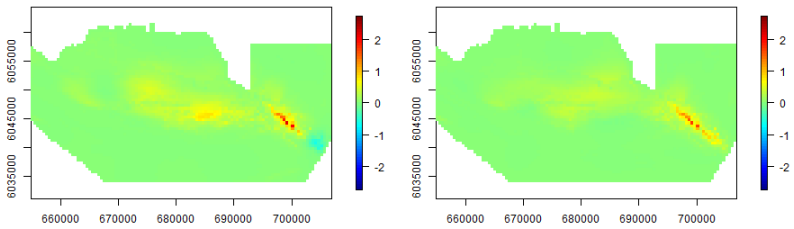


Figure 179: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

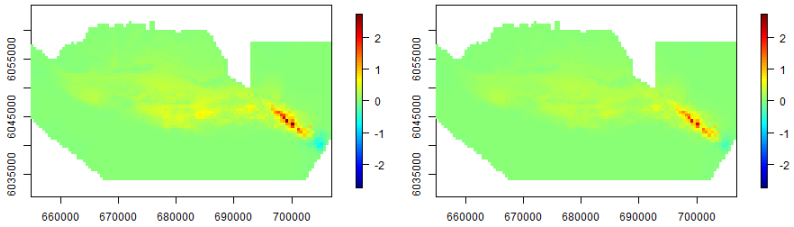


Figure 180: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

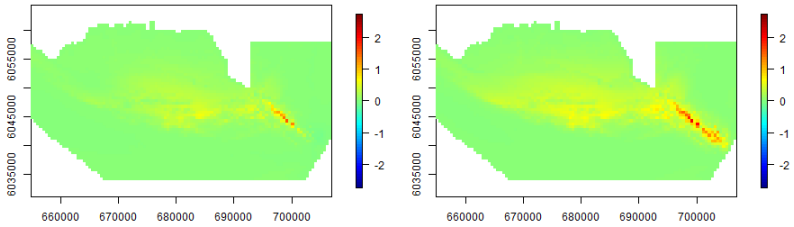


Figure 181: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

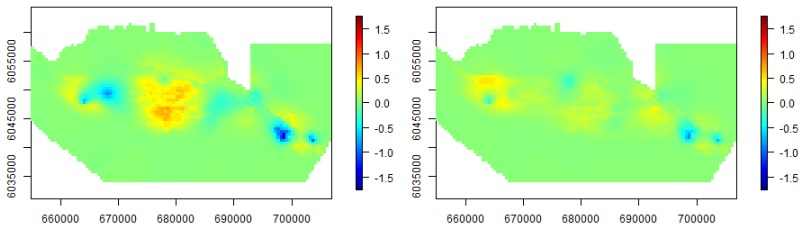


Figure 182: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

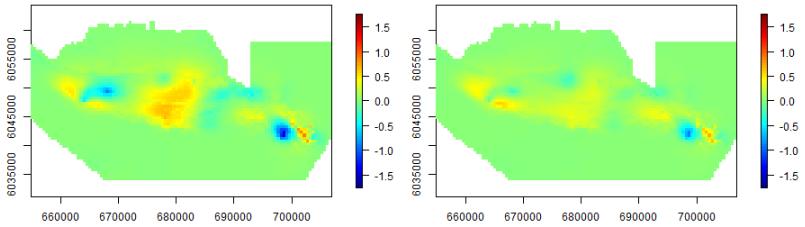


Figure 183: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

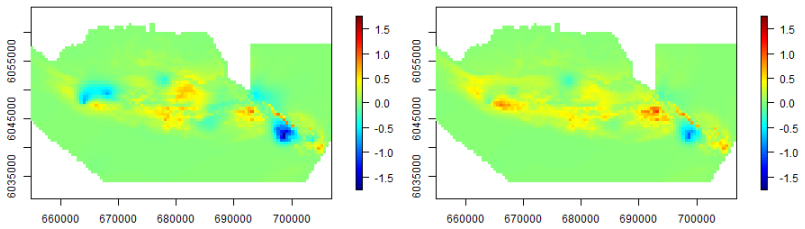


Figure 184: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

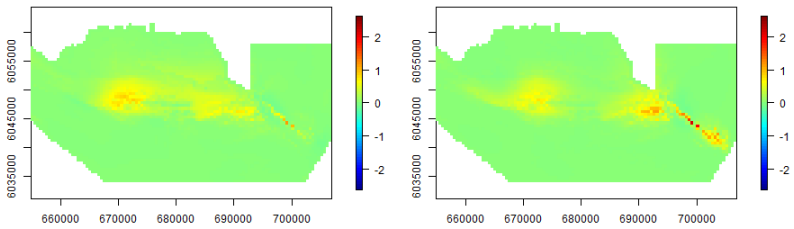


Figure 185: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

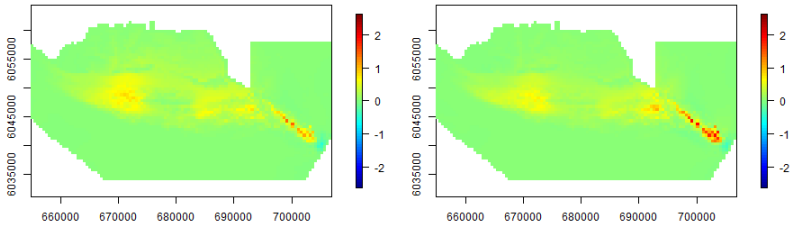


Figure 186: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

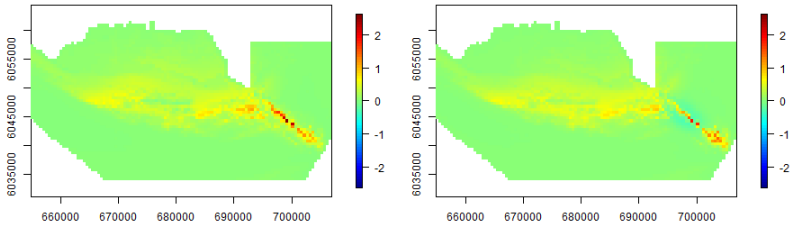


Figure 187: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

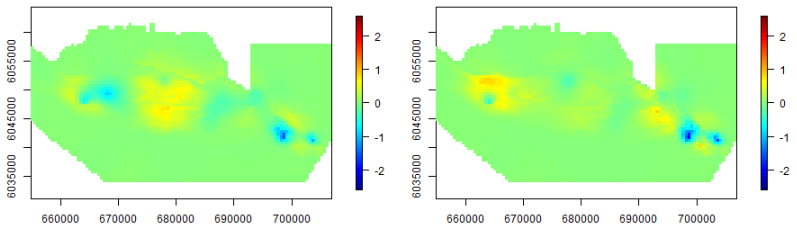


Figure 188: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

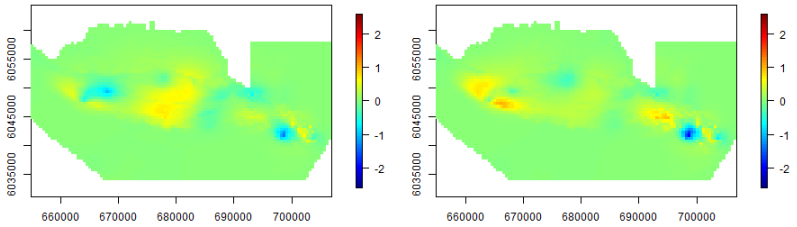


Figure 189: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

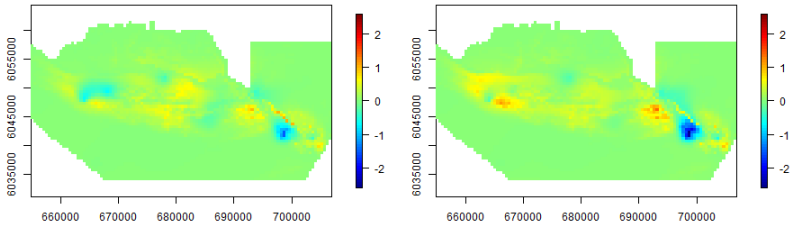


Figure 190: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

12.4 *Spatially explicit performance for the near-shore scenarios*

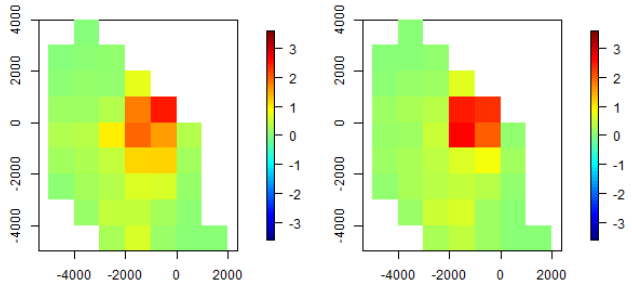


Figure 191: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

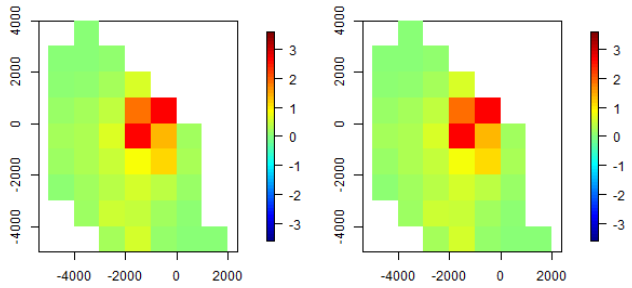


Figure 192: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

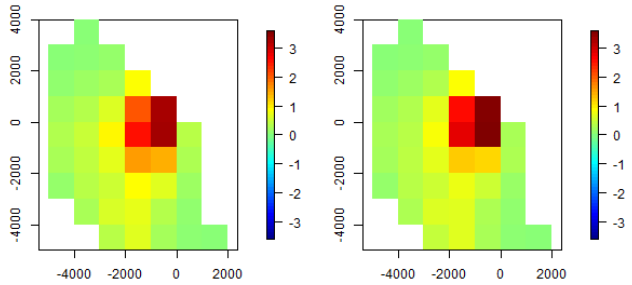


Figure 193: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

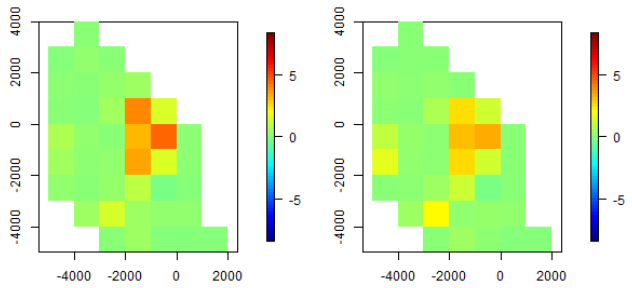


Figure 194: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

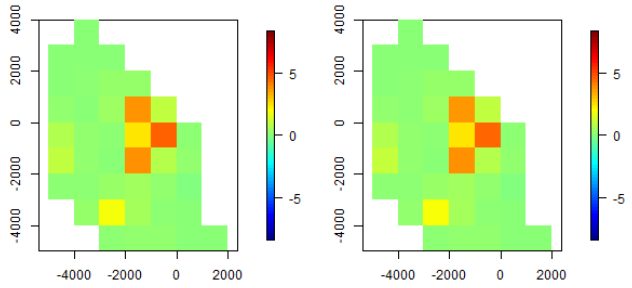


Figure 195: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

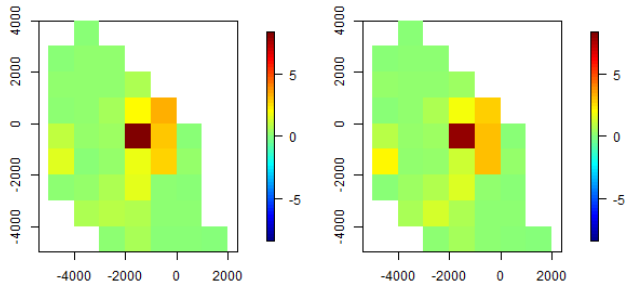


Figure 196: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

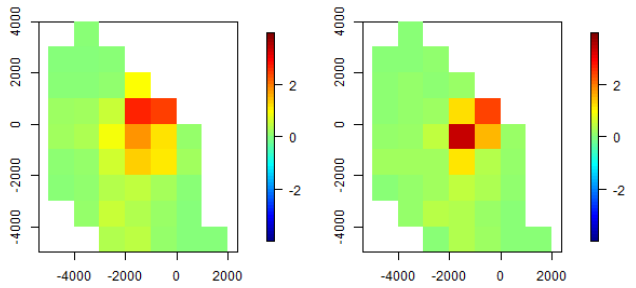


Figure 197: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

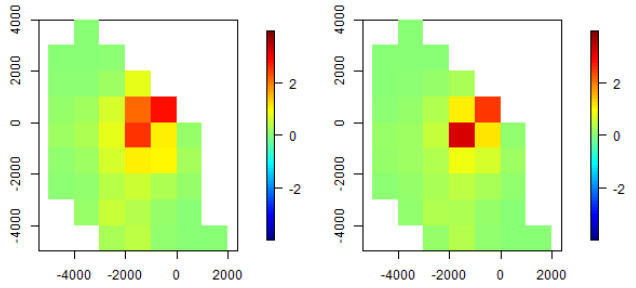


Figure 198: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

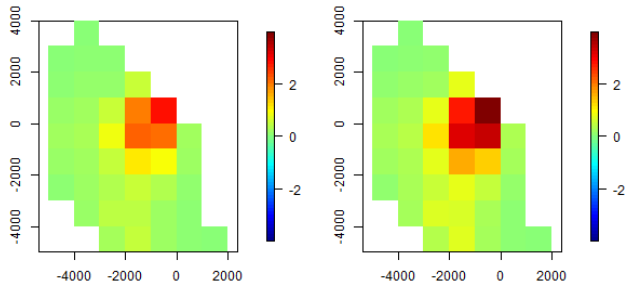


Figure 199: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

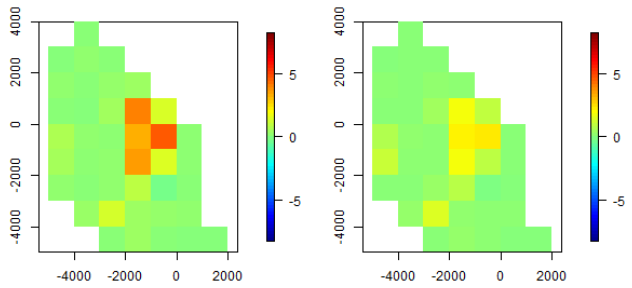


Figure 200: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

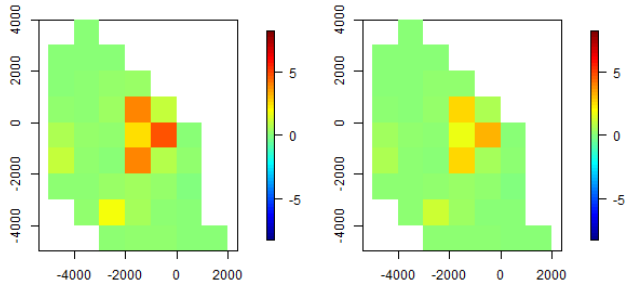


Figure 201: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

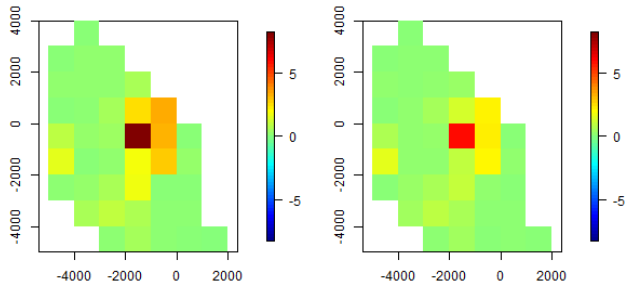


Figure 202: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

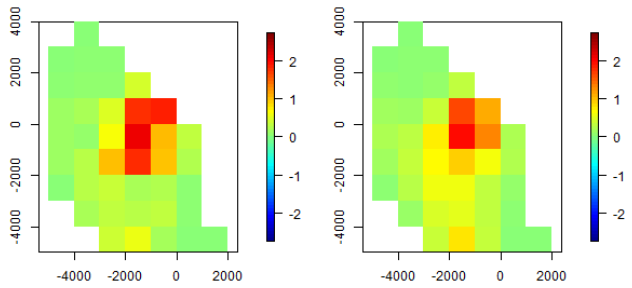


Figure 203: Average GAM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

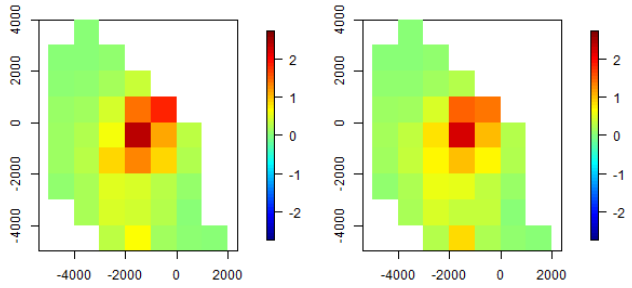


Figure 204: Average CReSS-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

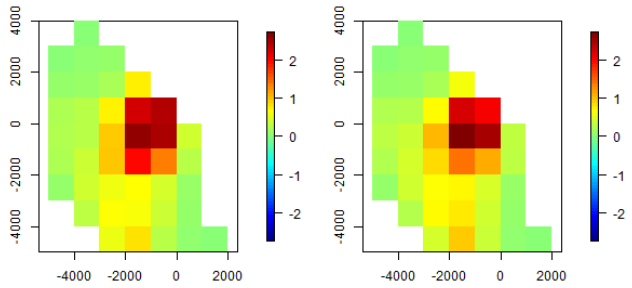


Figure 205: Average GAMM-based bias (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

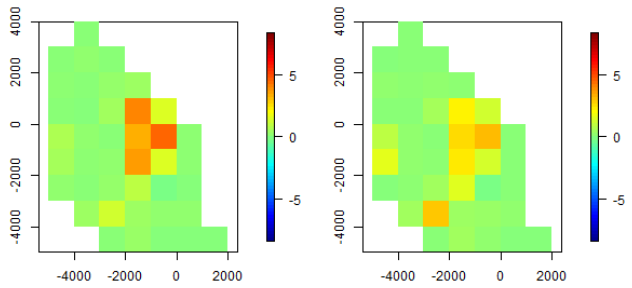


Figure 206: Average GAM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

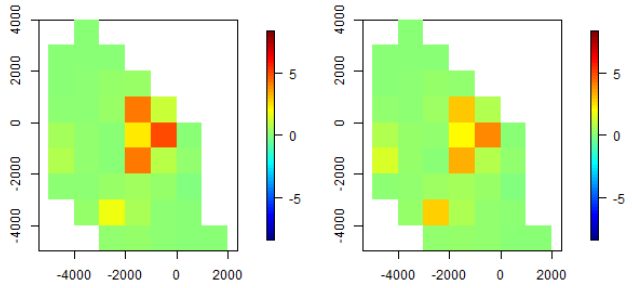


Figure 207: Average CReSS-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

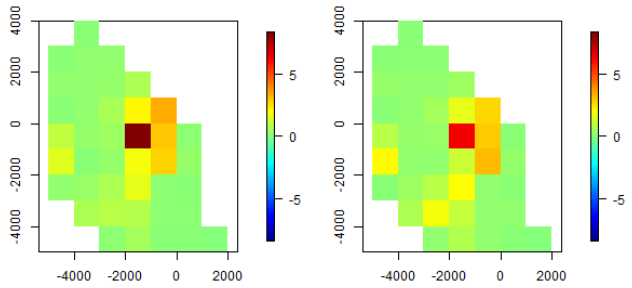


Figure 208: Average GAMM-based bias (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

12.5 *Residual analysis for the off-shore scenarios*

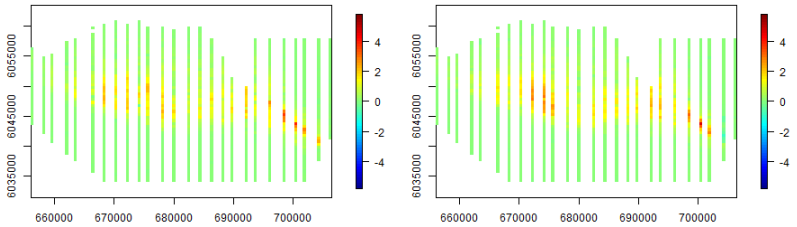


Figure 209: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

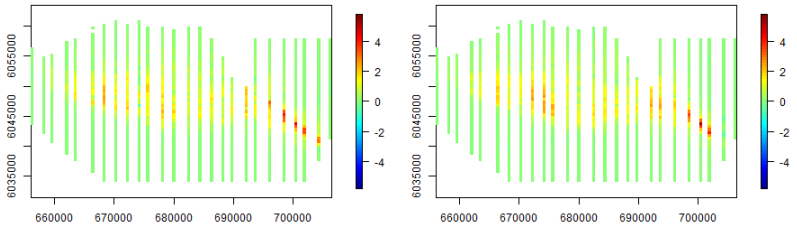


Figure 210: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

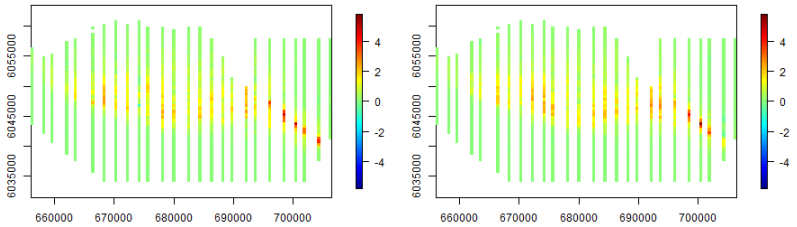


Figure 211: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

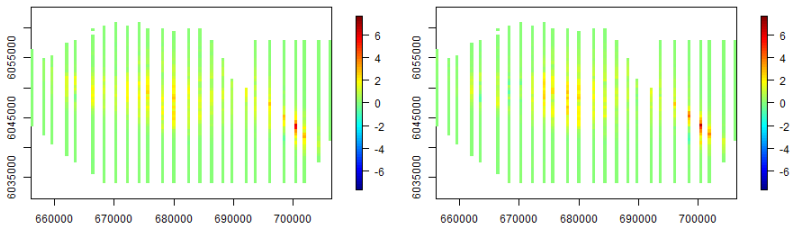


Figure 212: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

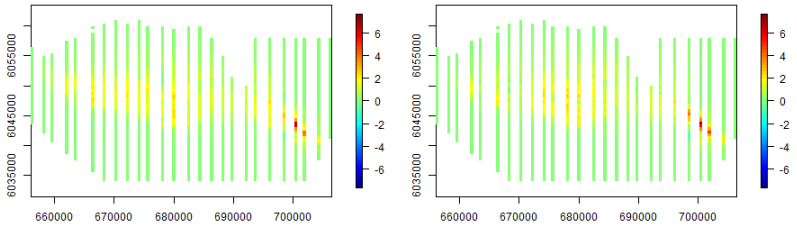


Figure 213: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

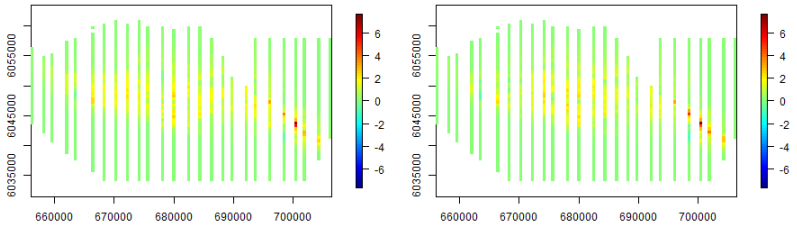


Figure 214: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

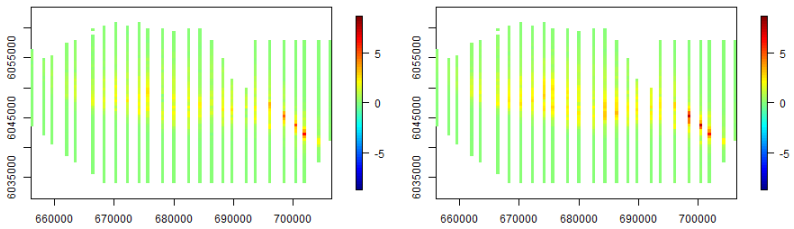


Figure 215: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

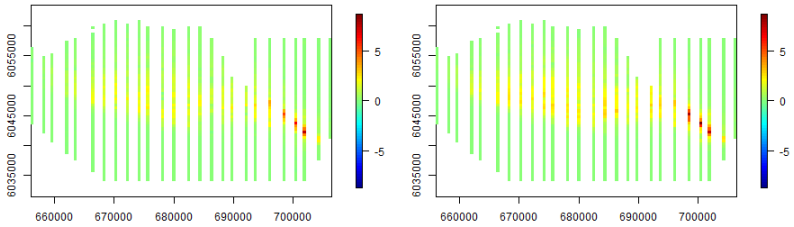


Figure 216: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

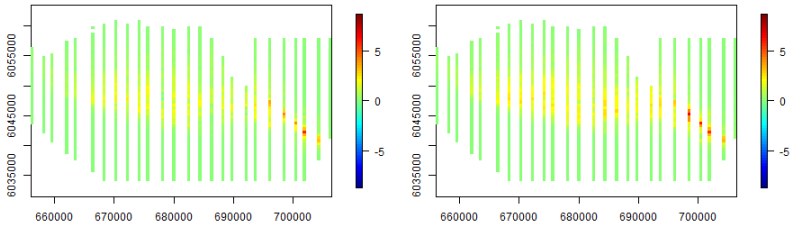


Figure 217: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

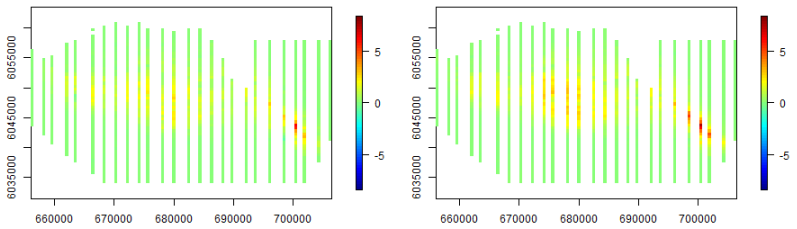


Figure 218: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

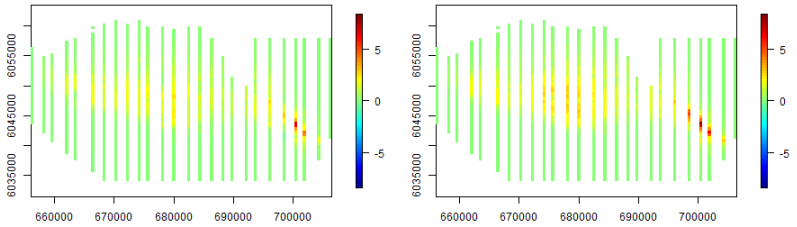


Figure 219: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

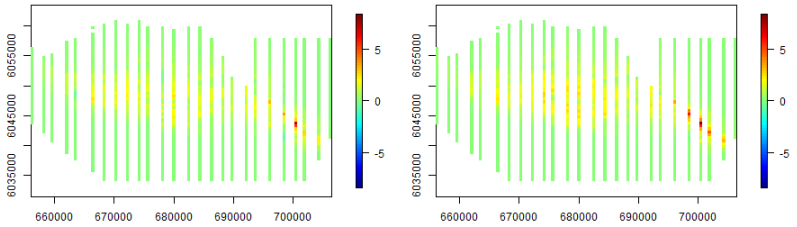


Figure 220: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

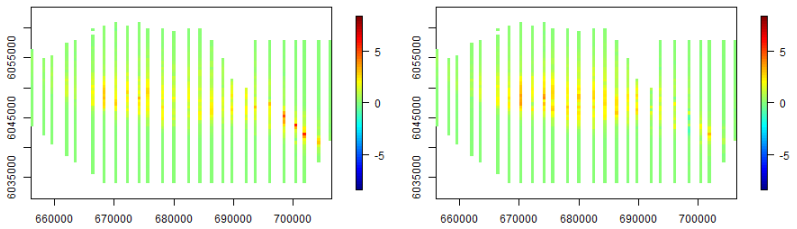


Figure 221: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

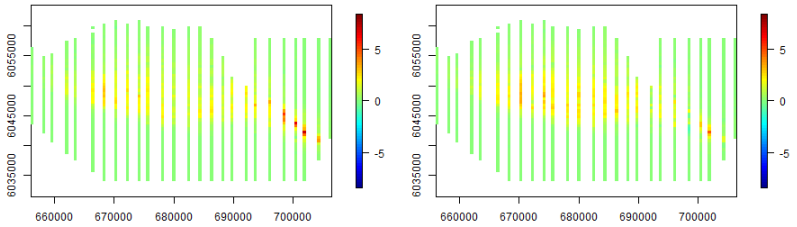


Figure 222: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

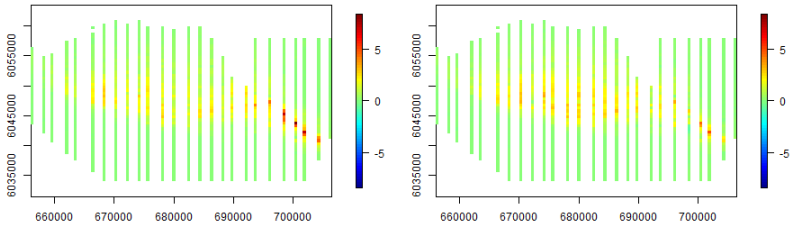


Figure 223: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

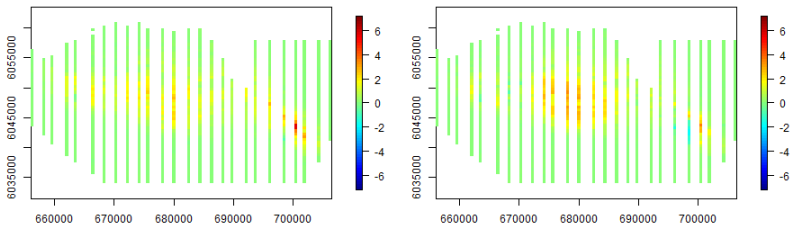


Figure 224: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

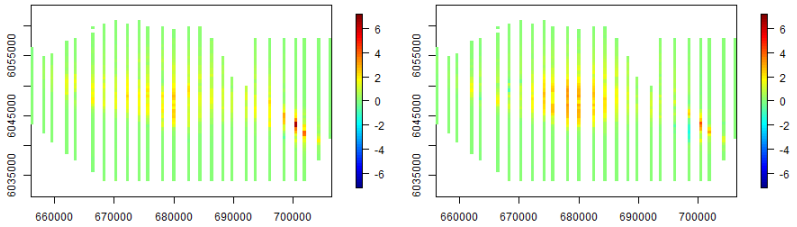


Figure 225: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

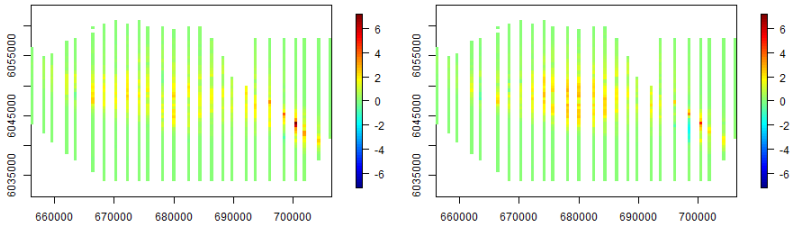


Figure 226: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

12.6 *Residual analysis for the near-shore scenarios*

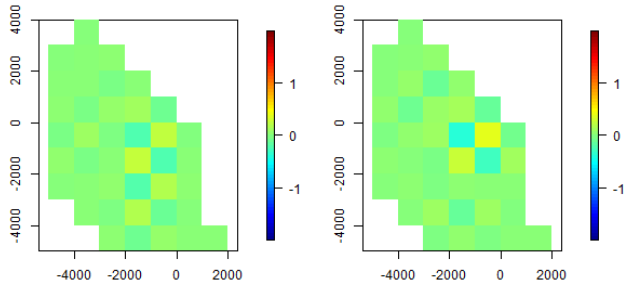


Figure 227: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

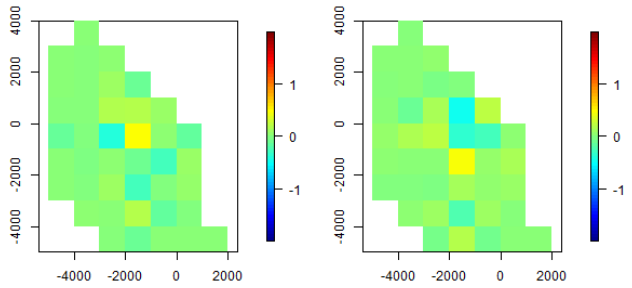


Figure 228: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

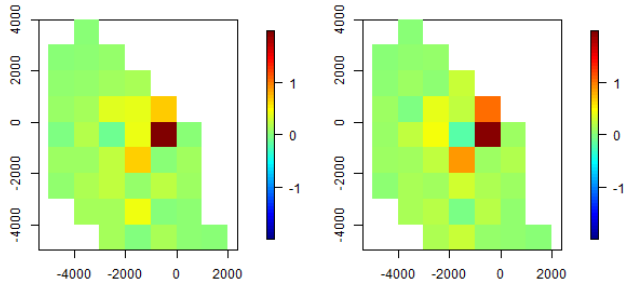


Figure 229: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact bias while the right-hand plot represents post-impact bias.

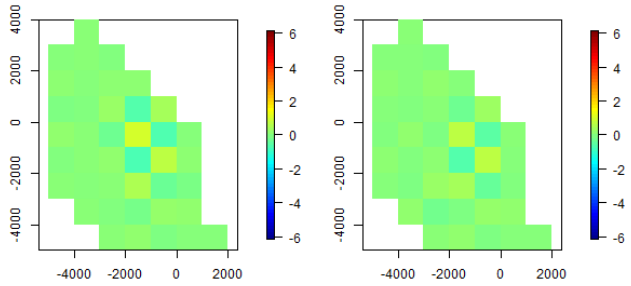


Figure 230: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

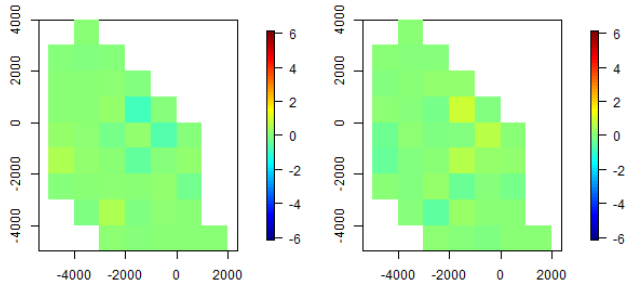


Figure 231: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

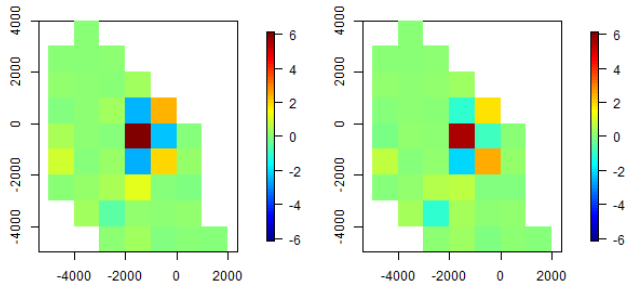


Figure 232: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

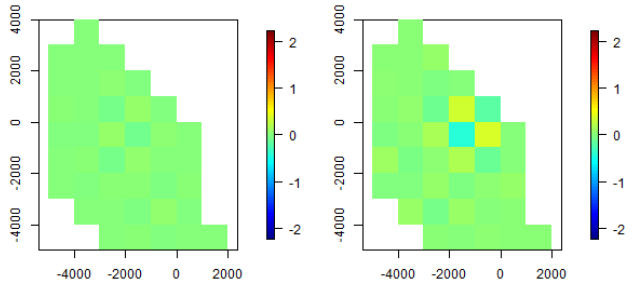


Figure 233: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

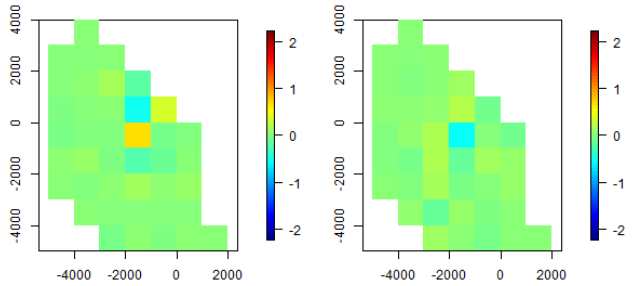


Figure 234: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

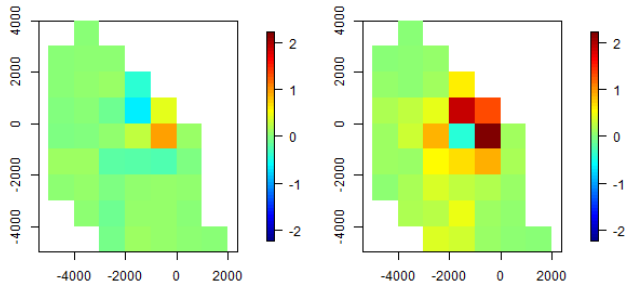


Figure 235: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

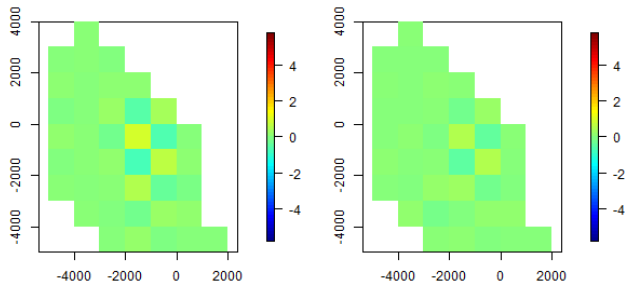


Figure 236: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

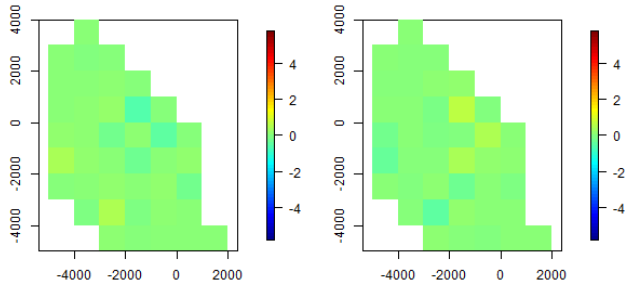


Figure 237: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

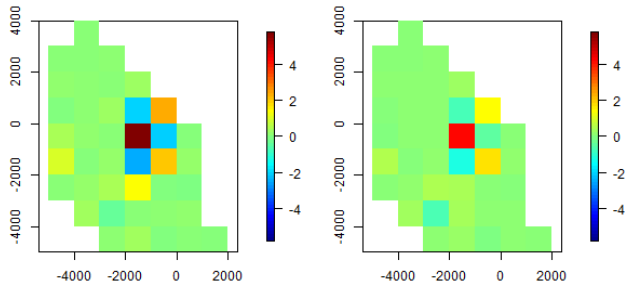


Figure 238: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

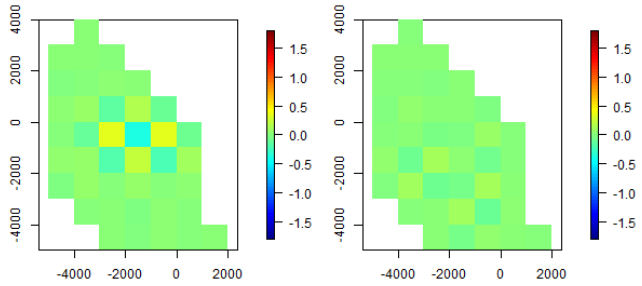


Figure 239: Average GAM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

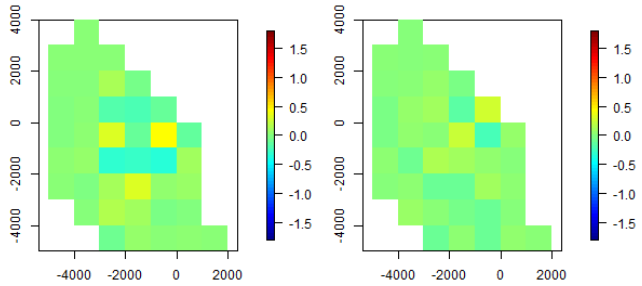


Figure 240: Average CReSS-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

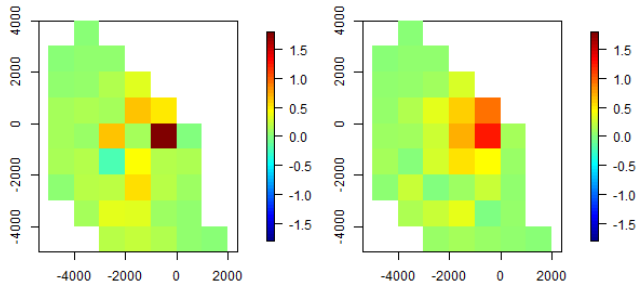


Figure 241: Average GAMM-based residuals (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

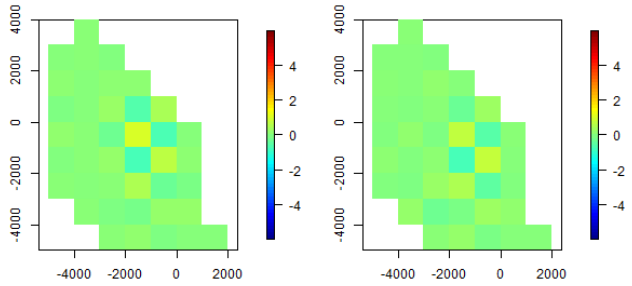


Figure 242: Average GAM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

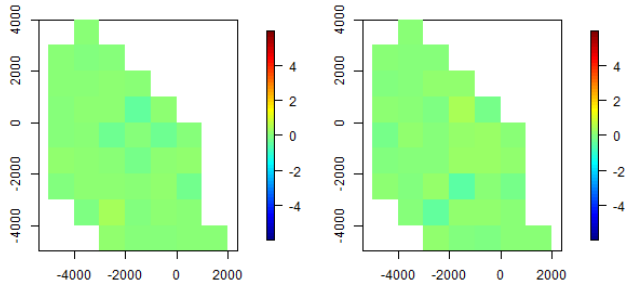


Figure 243: Average CReSS-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

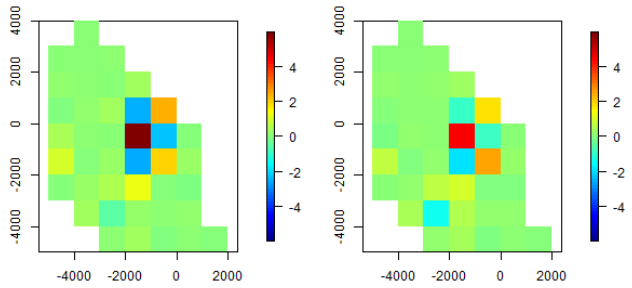


Figure 244: Average GAMM-based residuals (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact mean residuals while the right-hand plot represents post-impact mean residuals.

12.7 *Assessing the reliability of the reported geo-referenced precision for the off-shore scenarios*

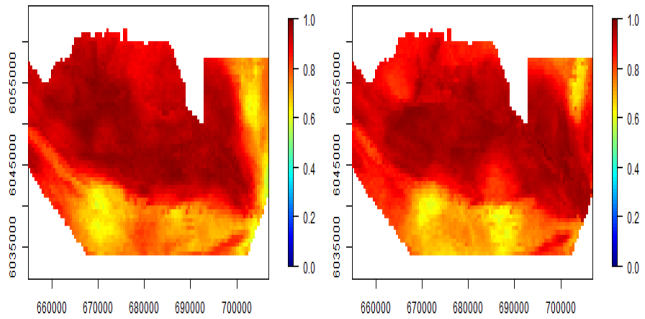


Figure 245: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

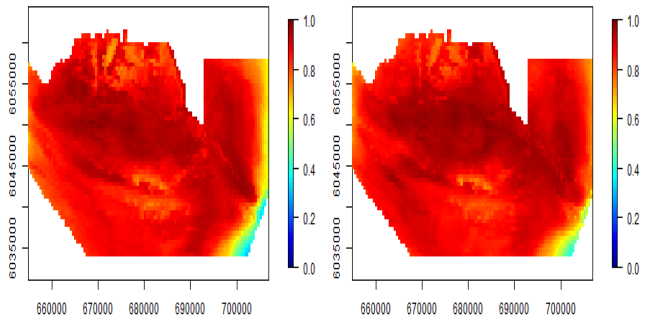


Figure 246: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

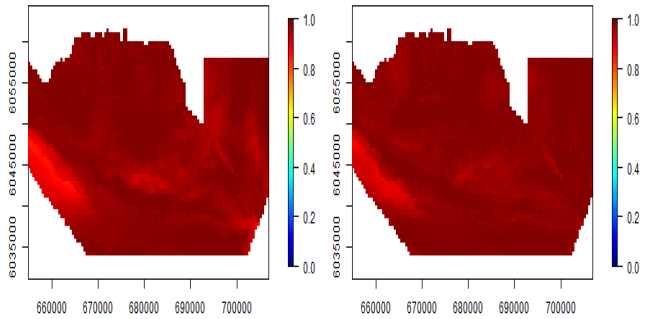


Figure 247: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

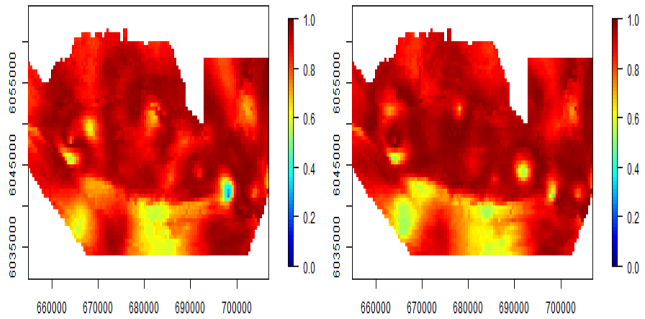


Figure 248: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

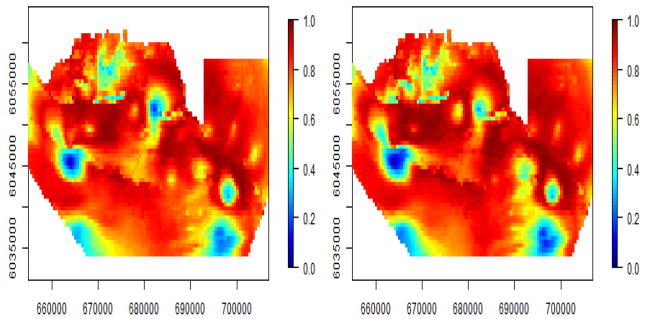


Figure 249: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

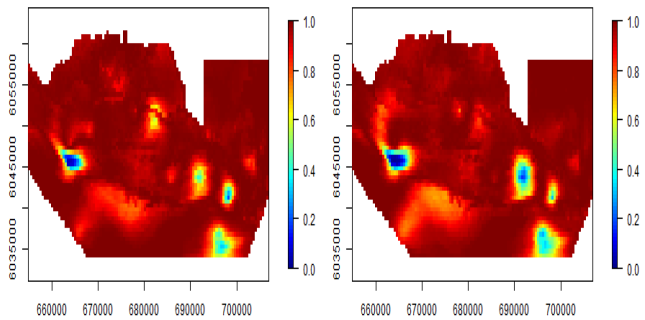


Figure 250: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

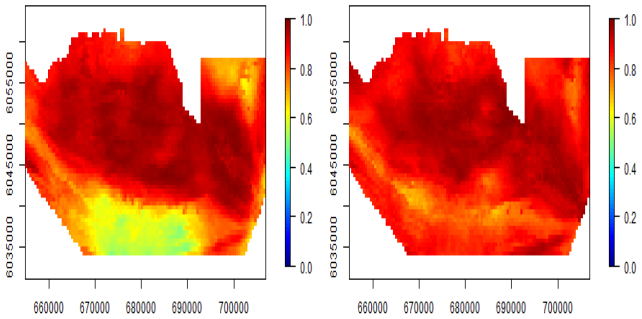


Figure 251: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

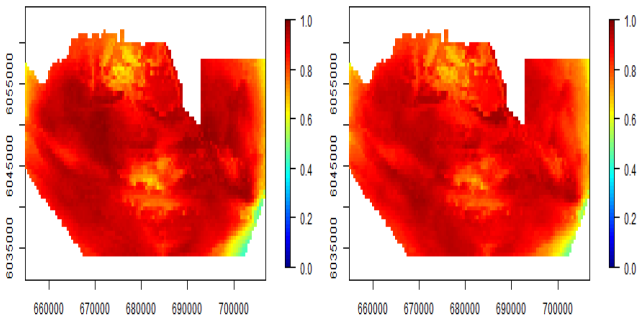


Figure 252: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

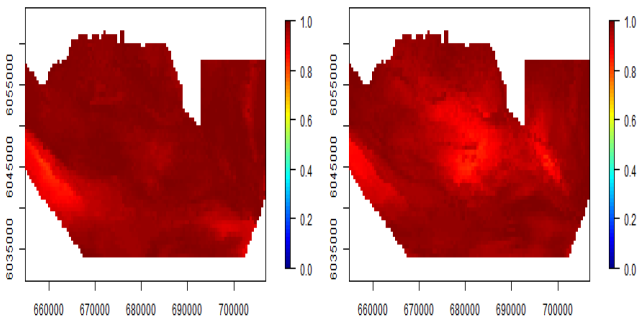


Figure 253: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

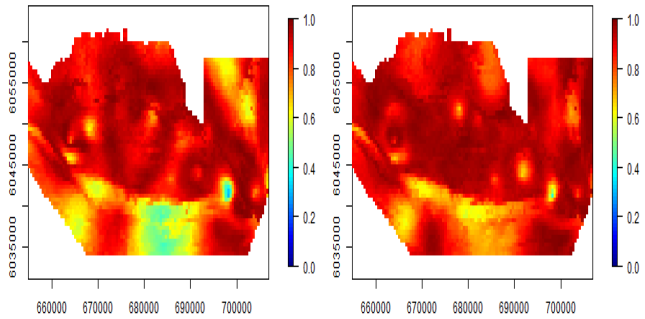


Figure 254: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

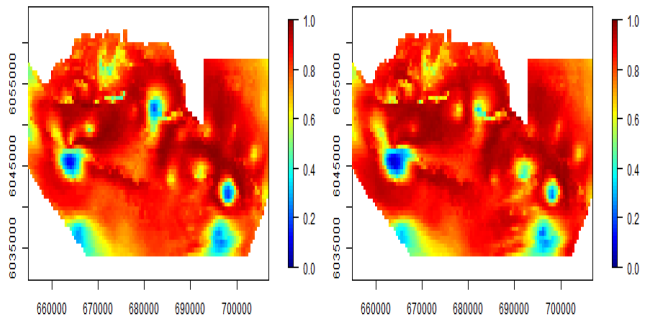


Figure 255: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

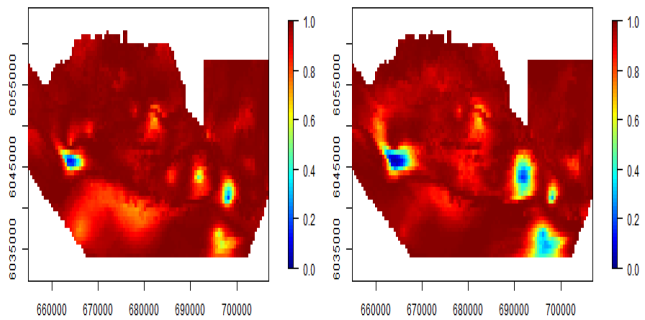


Figure 256: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

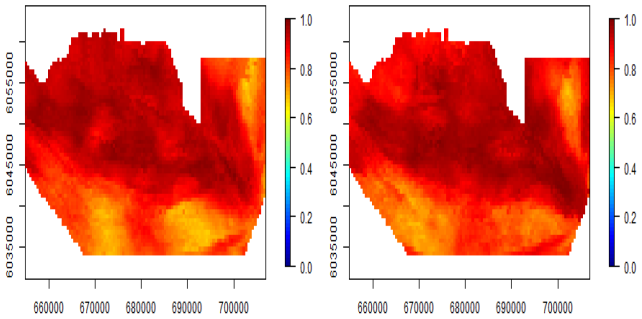


Figure 257: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

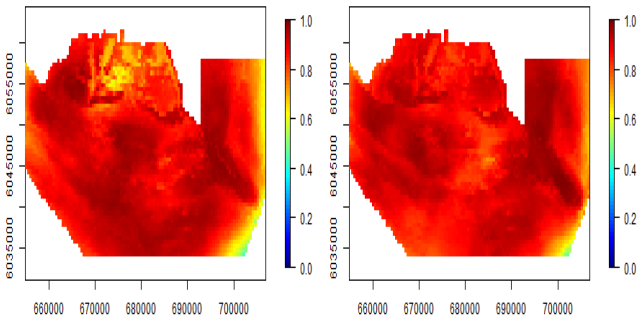


Figure 258: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

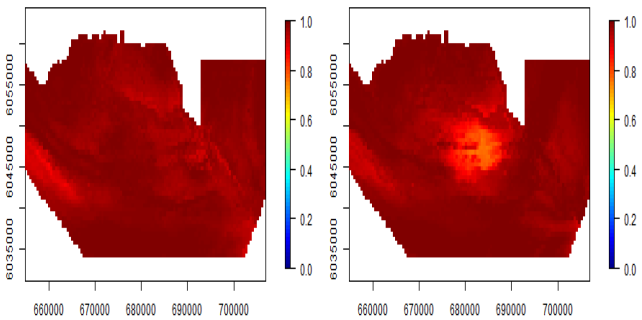


Figure 259: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

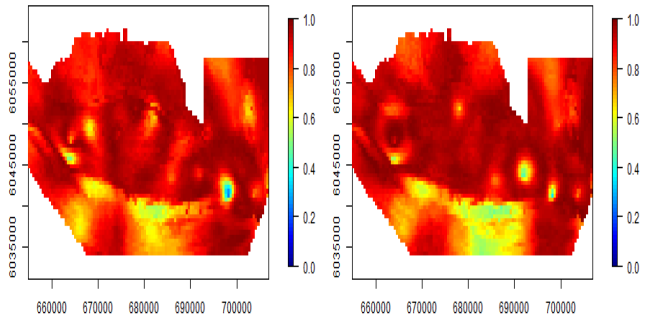


Figure 260: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

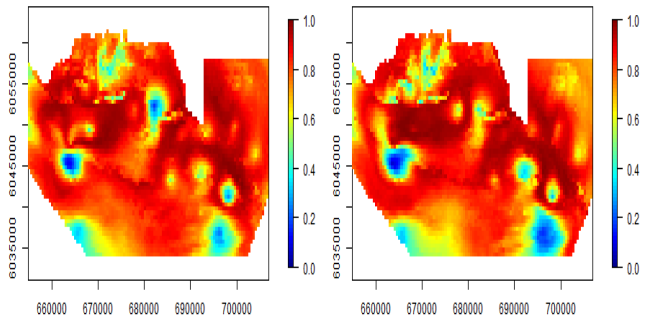


Figure 261: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

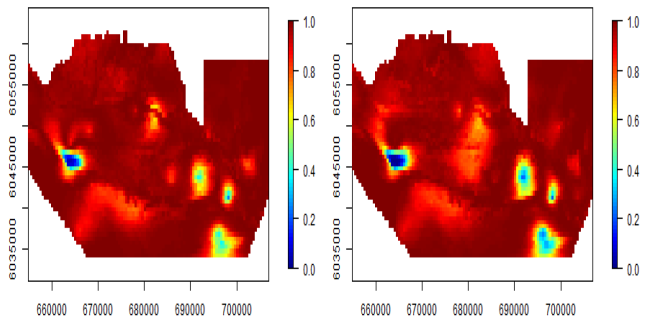


Figure 262: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

12.8 *Assessing the reliability of the reported geo-referenced precision for the near-shore scenarios*

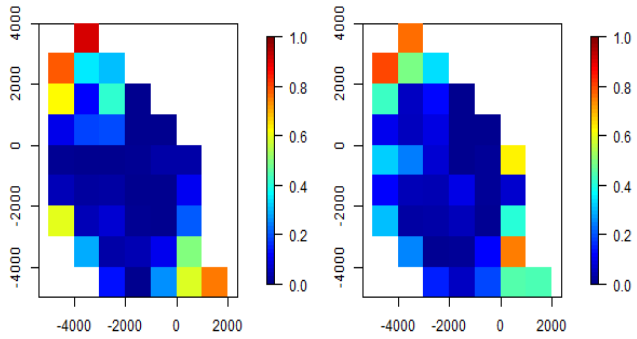


Figure 263: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

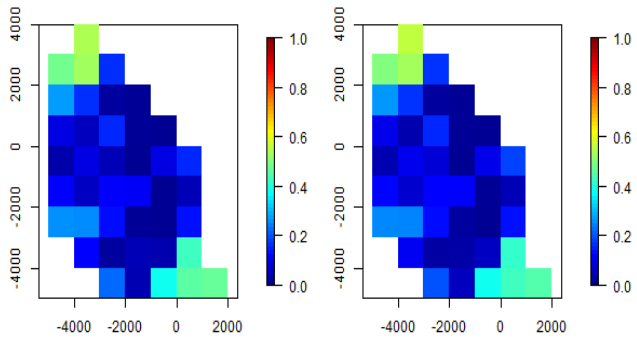


Figure 264: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

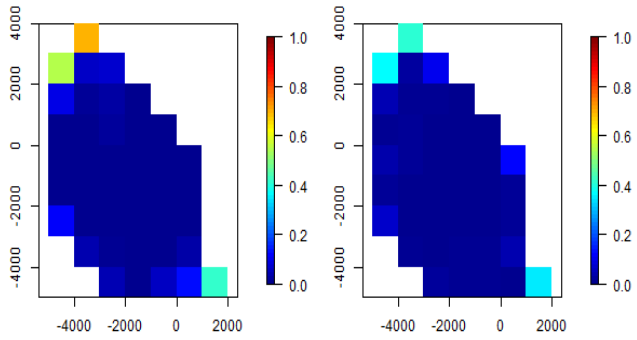


Figure 265: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

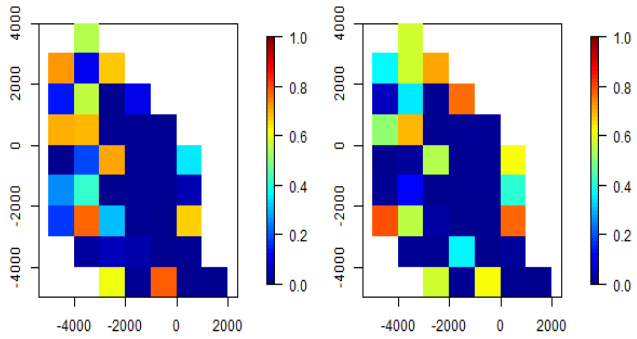


Figure 266: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

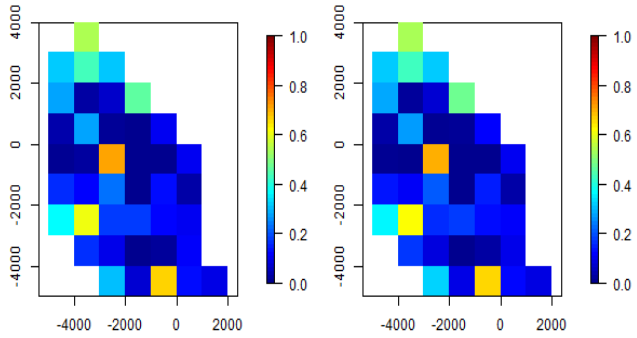


Figure 267: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

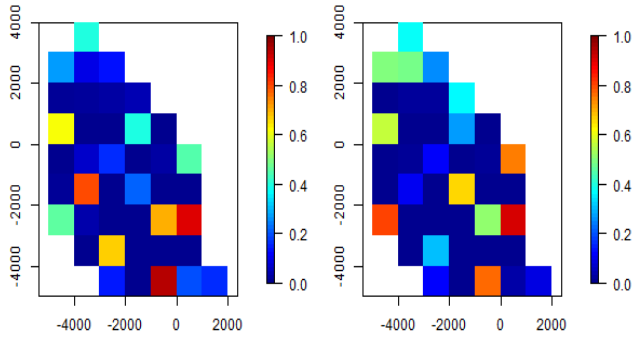


Figure 268: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **no-change** post impact (Model I). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

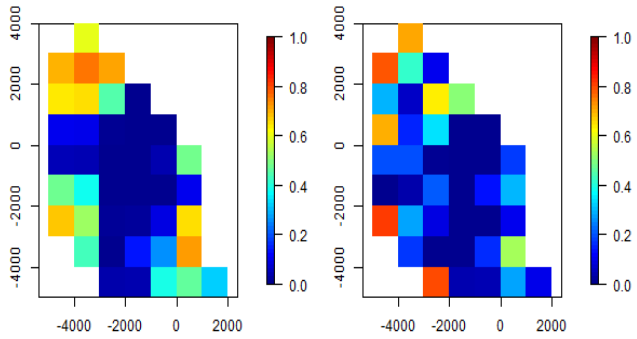


Figure 269: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

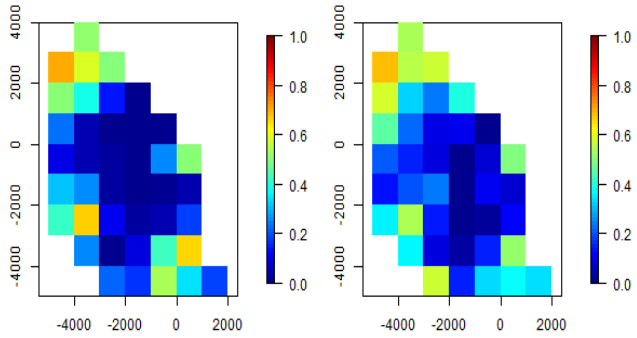


Figure 270: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

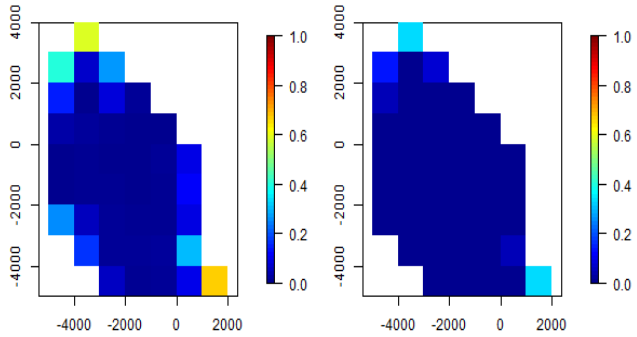


Figure 271: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

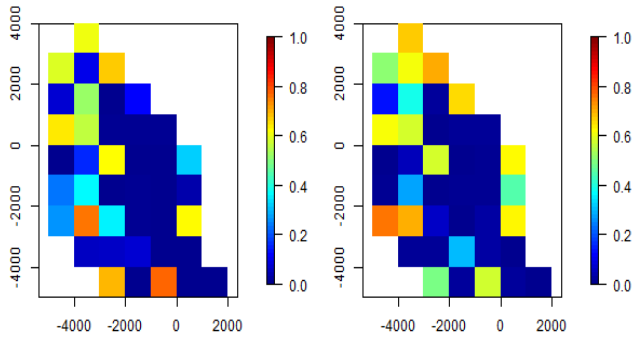


Figure 272: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

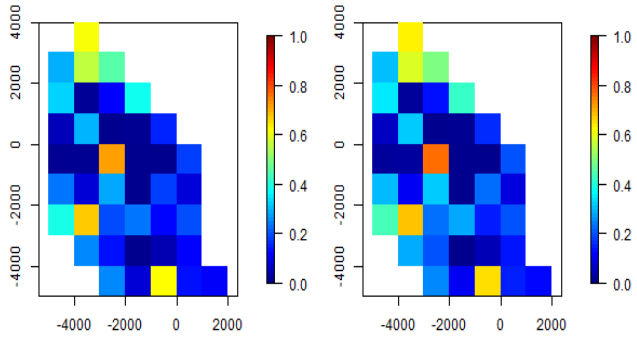


Figure 273: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

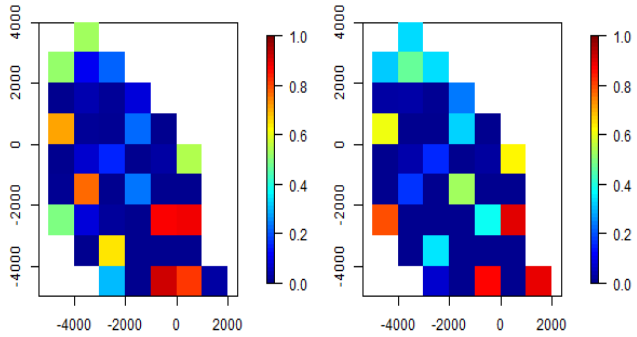


Figure 274: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **decrease** post impact (Model II). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

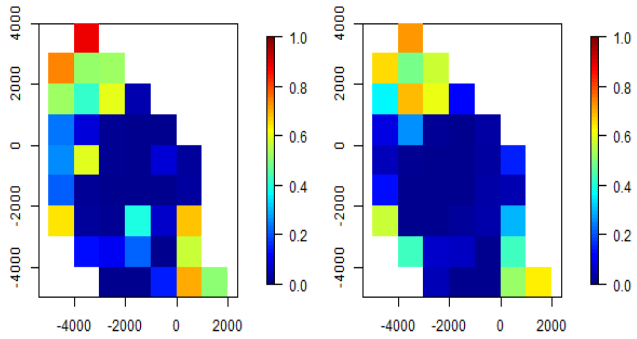


Figure 275: GAM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

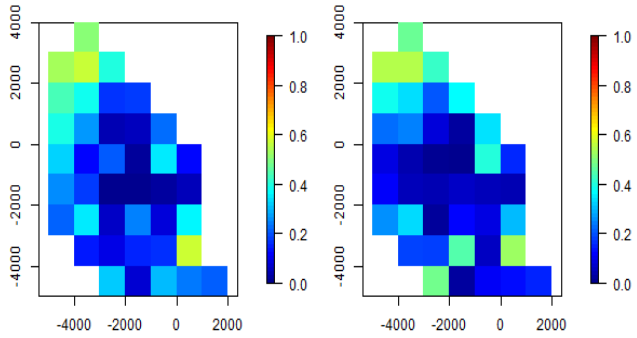


Figure 276: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

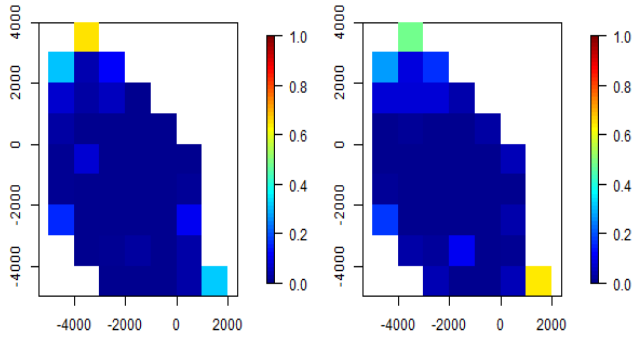


Figure 277: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the GAM generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

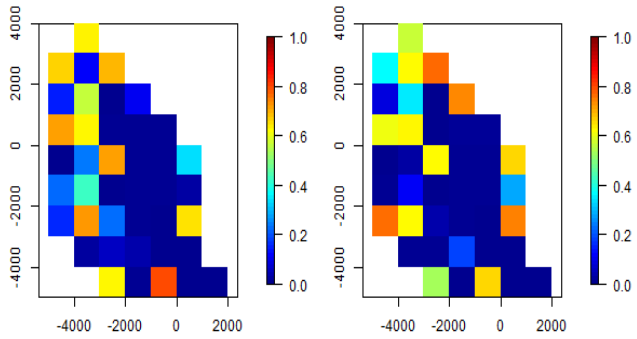


Figure 278: GAM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

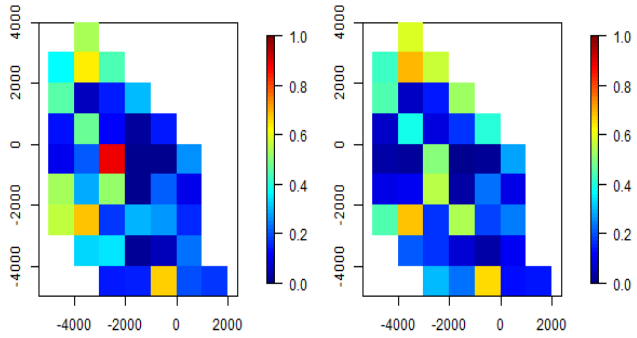


Figure 279: CReSS-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

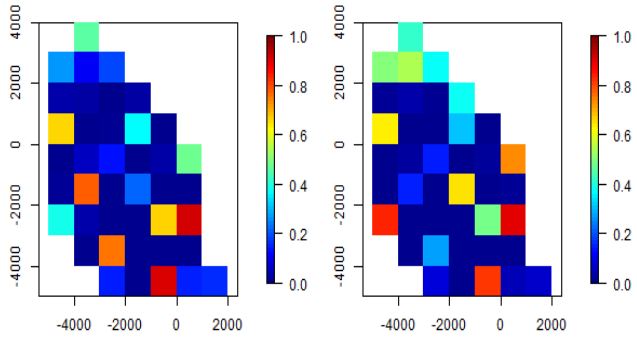


Figure 280: GAMM-based percentage coverage (across the 100 realisations) represented spatially for the CReSS generated data with a **redistribution** post impact (Model III). The left-hand plot represents the pre-impact coverage while the right-hand plot represents post-impact coverage.

13 *Literature cited*

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika* **67,** 413–418.

Borberg, J., Ballance, L., Pitman, R., and Ainley, D. (2005). A test for bias attributable to seabird avoidance of ships during surveys conducted in the tropical pacific. *Marine Ornithology* **33,** 173–179.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations*. Springer-Verlag.

Brown, H. and Prescott, R. (1999). *Applied Mixed Models in Medicine*. J. Wiley & Sons, West Sussex, England.

Buckland, S., Burt, M., Rexstad, E., Mellor, M., Williams, A., and Woodward, R. (2012). Aerial surveys of seabirds: the advent of digital methods. *Journal of Applied Ecology* **49,** 960–967.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001). *Introduction to Distance Sampling*. Oxford University Press.

Camphuysen, C. J. and Garthe, S. (2004). Recording foraging seabirds at sea: standardised recording and coding of foraging behaviour and multi-species associations. *Atlantic Seabirds* **6,** 1–32.

Camphuysen, K. J., Fox, A. D., Leopold, M. F., and Petersen, I. K. (2004). Towards standardised seabirds at sea census techniques in connection with environmental impact assessments for offshore wind farms in the U.K.: a comparison of ship and aerial sampling methods for marine birds, and their applicability to offshore wind farm assessments. Technical report, NIOZ report to COWRIE (BAM 02-2002), Texel, 37pp.

Centrica (2007). Lincs offshore wind farm environmental statement.

Ciannelli, L., Fauchald, P., Chan, K., Agostini, V., and Dings, G. (2008). Spatial fisheries ecology: Recent progress and future prospects. *Journal of Marine Systems* **71,** 223–236.

Cox, M. J., Borchers, D. L., and Kelly, N. (2013). nupoint: An r package for density estimation from point transects in the presence of nonuniform animal density. *Methods in Ecology and Evolution* **4,** 589–594.

Cox, S. L., Scott, B. E., and Camphuysen, C. J. (2013). Combined spatial and tidal processes identify links between pelagic prey species and seabirds. *Marine Ecology Progress Series* **479,** 203–221.

Embling, C., Illian, J., Armstrong, E., van der Kooij, J., Sharples, J., Camphuysen, K., and Scott, B. E. (2012). Investigating fine scale spatio-temporal predator-prey patterns in dynamic marine ecosystems: A functional data analysis approach. *Journal of Applied Ecology* **49,** 481–492.

Fauchald, P., Skov, H., Skern-Mauritzen, M., Hausner, V. H., Johns, D., and Tveraa, T. (2011). Scale-dependent response diversity of seabirds to prey in the North Sea. *Ecology* **92,** 228–239.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Sage.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association* **87,** 178–183.

Gilles, A., Scheidat, M., and Siebert, U. (2009). Seasonal distribution of harbour porpoises and possible interference of offshore wind farms in the German North Sea. *Marine Ecology Progress Series* **383,** 295–307.

Gillespie, D., Leaper, R., Gordon, J., and Macleod, K. (2010). An integrated data collection system for line transect surveys. *Journal of Cetacean Research and Management* **11,** 217–227.

Hammond, P. S., Macleod, K., Berggren, P., Borchers, D. L., Burt, L., Cañadas, A., et al. (2013). Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation* **164,** 107 – 122.

Hardin, J. and Hilbe, J. (2002). *Generalized Estimating Equations*. Chapman & Hall/CRC.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.

Hedley, S. L. and Buckland, S. T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological and Environmental Statistics* **9,** 181–199.

Hengl, T. (2007). A practical guide to geostatistical mapping of environmental variables. Technical report, JRC Scientific and Technical Reports, Joint Research Centre, European Commission.

Jackson, D. and Whitfield, P. (2011). Guidance on survey and monitoring in relation to marine renewables deployments in scotland, volume 4. birds. Technical report, Scottish Natural Heritage and Marine Scotland.

Johnson, D. S., Laake, J. L., and Ver Hoef, J. M. (2010). A model-based approach for making ecological inference from distance sampling data. *Biometrics* **66,** 310–318.

Lapeña, B. P., Wijnberg, K., Stein, A., and Hulscher, S. (2011). Spatial factors affecting statistical power in testing marine fauna displacement. *Ecological Applications* **21,** 2756–2769.

Lin, D., Wei, L., and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58,** 1–12.

Lindsey, J. K. and Jones, B. (1998). Choosing among generalized linear models applied to medical data. *Statistics in Medicine* **17,** 59–68.

Maclean, I. M. D., Rehfisch, M. M., Skov, H., and Thaxter, C. B. (2013). Evaluating the statistical power of detecting changes in the abundance of seabirds at sea. *Ibis* **155,** 113–126.

MacLean, I. M. D., Wright, L. J., Showler, D. A., and Rehfisch, M. M. (2009). A review of assessment methodologies for offshore wind farms. Technical report, Report commissioned by COWRIE Ltd.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, 2^{nd} edition.

Mendenhall, W. (1982). *Statistics for Management and Economics*. Duxbury Press, Boston, fourth edition.

Peel, D., Bravington, M. V., Kelly, N., Wood, S. N., and Knuckey, I. (2013). A model-based approach to designing a fishery-independent survey. *Journal of Agricultural, Biological and Environmental Statistics* **18,** 1–21.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ronconi, R. A. and Burger, A. (2009). Estimation of seabird densities from vessel transects: distance sampling and implications for strip transects. *Aquatic Biology* **4,** 297–309.

Royle, J. A., Dawson, D. K., and Bates, S. (2004). Modelling abundance effects in distance sampling. *Ecology* **85(6),** 1591–1597.

Särndal, C. and Swensson, B. (2003). *Model Assisted Survey Sampling*. Springer.

Schwemmer, P., Mendel, B., Sonntag, N., Dierschke, V., and Garthe, S. (2012). Effects of ship traffic on seabirds in offshore waters: implications for marine conservation and spatial planning. *Ecological Applications* **21,** 1851–1860.

Scott, B., Webb, A., Palmer, M., Embling, C., and Sharples, J. (2013). Fine scale bio-physical oceanographic characteristics predict the foraging occurrence of contrasting seabird species: Gannet *(Morus bassanus)* and storm petrel *(Hydrobates pelagicus)*. *Progress in Oceanography* .

Scott, B. E. (in press). Seabirds and marine renewables: are we asking the right questions about indirect effects. *Ibis* .

Scott-Hayward, L., Mackenzie, M. L., Donovan, C. R., Walker, C. G., and Ashe, E. (2013). Complex Region Spatial Smoother (CReSS). *Journal of Computational and Graphical Statistics* .

Shibata, Y., Matsuishi, T., Murase, H., Matsuoka, K., Hakamada, T., Kitakado, T., and Matsuda, H. (2013). Effects of stratification and misspecification of covariates on species distribution models for abundance estimation from virtual line transect survey data. *Fisheries Science* pages 1–10.

Stone, C., Webb, A., Barton, C., Ratcliffe, N., Reed, T., Tasker, M., Camphuysen, C., and Pienkowski, M. (1995). *An atlas of seabird distribution in northwest European waters*. Joint Nature Conservation Committee.

Thaxter, C. and Burton, N. (2009). High definition imagery for surveying seabirds and marine mammals: A review of recent trials and development of protocols. Technical report, British Trust for Ornithology report, commissioned by COWRIE Ltd.

Thomas, L., Buckland, S. T., Rexstad, E. A., Laake, J. L., Strindberg, S., Hedley, S. L., Bishop, J. R. B., Marques, T. A., and Burnham, K. P. (2010). Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* **47,** 5–14.

Torres, L. and Read, A. (2008). Fine-scale habitat modeling of a top marine predator: do prey data improve predictive capacity? *Ecological Applications* **18,** 1702–1717.

Walker, C., Mackenzie, M., Donovan, C., and O'Sullivan, M. (2010). SALSA - a Spatially Adaptive Local Smoothing Algorithm. *Journal of Statistical Computation and Simulation* **81,** 179–191.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1,** 80–83.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society B* **65(1),**.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.