Evaluating Statistical Models for Baseline Characterization and Measuring Change in Environmental Monitoring Data

Hannah L. Linder

A thesis

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2016

Committee:

Dr. John K. Horne

Dr. Timothy Essington

Dr. Eric Ward

Dr. Andrea Copping

Program Authorized to Offer Degree:

Aquatic and Fishery Sciences

University of Washington

Abstract

Evaluating Statistical Models for Baseline Characterization and Measuring Change in Environmental Monitoring Data

Hannah L. Linder

Chair of the Supervisory Committee:
Dr. John K. Horne
School of Aquatic and Fishery Sciences

In Before-After monitoring studies, statistical models are used to characterize baseline (i.e., pre-disturbance) conditions, and to detect, quantify, and forecast change during operational monitoring (i.e., post-disturbance). To establish best practices for analyzing monitoring data, a model evaluation was developed and applied using Marine Renewable Energy (MRE); a case study of a disturbance with no best practice monitoring methods. The evaluation was performed on normal and non-normal acoustic metrics representative of MRE monitoring data. Evaluated models included: generalized regression models, time series models, and nonparametric models. 10-fold Cross Validation was used to evaluate baseline model fit. Models were then fit to 5 simulated Before-After change scenarios using Intervention Analysis. A power analysis was used to evaluate model ability to detect change. Residual error diagnostics were used to quantify model fit and forecast accuracy. State-space models are recommended for baseline characterization. Deterministic Parametric models are recommended to detect change. Time series and semi-parametric models are recommended to quantify change. Nonparametric models are recommended to forecast change. These recommendations form best practices for analyzing MRE monitoring data, which enables comparisons among MRE sites and reduces uncertainty in environmental effects. The evaluation approach is applicable to any monitoring program.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to extend my acknowledgement and deepest thanks to:

My advisor, John Horne, for guiding me through this thesis work, and pushing me to be a better scientist in all regards along the way,

My committee for providing me with great support and insight; especially Eric Ward for tirelessly helping me understand and code a wide variety of statistical models, Tim Essington for being so helpful in steering me through the big picture of model evaluation studies, and Andrea Copping for sharing her abundant knowledge of Marine Renewable Energy monitoring with me,

The National Science Foundation Sustainable Energy Pathways for providing the funding for this research,

Dale Jacques and Lauren Wiesebron for conducting previous work that paved the way for my own, and being on call for any questions I may have had,

All members of the Fisheries Acoustics Lab, including: Dale Jacques, Lauren Wiesebron, Beth Phillips, Dave McGowan, Mei Sato, Silvana González, and Ross Hytnen for their ideas, friendships, and support,

Jim Faulkner, who was so generous for providing code and clear explanations to help develop the Hidden Markov simulation model,

Ian Oviatt for being such a considerate and encouraging partner during my thesis work and throughout life,

And lastly to my family; Mary-Megan, Robert, Megan, Riley, and Shane, for being my primary source of love, laughter, and support.

This work was in part facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

Chapter 1:

General Introduction

1.1 Introduction

Environmental monitoring programs are used to detect and measure change, and to inform environmental management decisions used to prevent or mitigate detrimental effects on natural resources (Gitzen 2012, Lindenmayer et al. 2012). Change may be caused by either natural or anthropogenic disturbance of pre-existing environmental conditions. Monitoring methods, including data collection, study designs, and analytic methods affect the definition, size, and shape of change measured in monitoring programs (Lovett et al. 2007, Magurran et al. 2010, Nuno et al. 2014). Consequently, it is critical to develop consistent and clearly-defined monitoring methods. Standardizing monitoring methods reduces uncertainty in the assessment of environmental change, and provides comparable data across time and monitoring sites to produce the most efficient environmental monitoring programs (Froján et al. 2016).

A successful monitoring program is dependent on the collection of relevant and informative data used to measure environmental change. The stressor-receptor monitoring construct is a standard approach used to select monitoring data that represents primary environmental effects of concern. Change in the environment is caused by a stressor (i.e., natural or anthropogenic disturbance of pre-existing natural conditions) that can be assessed by measuring change in components of that environment (i.e., biotic and abiotic elements that comprise the environment) (Jackson et al. 2000). A receptor is defined as a component of the ecosystem that responds to the stressor (Boehlert and Gill 2010). The condition of a receptor can be measured using environmental indicators. Environmental indicators are measurable proxies of

the state of the receptor. These indicators may be physical, chemical, or biological measurements (Kurtz et al. 2001). Biological indicators are used to measure change to the structure, function, and composition of a receptor, three key attributes of an ecosystem (Noss 1990). Structure is defined as the physical organization of a system. Function is defined as ecological and evolutionary processes. Composition is the identity and variety of elements in a collection. Abundance and distribution are conventional indicators of structure, diversity is a common indicator of composition, and behavior is an indicator of function (Noss 1990; Niemi and McDonald 2004). Change in these indicators represents change in the state of the environmental receptor, and therefore a change in the baseline of an environment. Biological indicator metrics (such as abundance counts, diversity indexes, location measurements) are commonly used as environmental monitoring data.

Before-After Intervention Analysis (IA) study designs are a standard approach to measuring change in monitoring data (Stewart-Oaten and Bence 2001, Magurran et al. 2010). This monitoring design uses baseline (i.e., pre-disturbance) data to estimate any change in operational monitoring (i.e., post-disturbance) data from pre- baseline conditions. Baseline monitoring data are used to characterize pre-existing trend, environmental relationships, and natural variability in the environment prior to any disturbance (Treweek 1996, 2009). Baseline characterization can be used to identify perceived effects of a specified disturbance, and enable an operational monitoring program that can effectively measure hypothesized change outside of the range of natural baseline variability (McCann 2012). Operational monitoring data are critical to quantify the size and shape of change, and to forecast future change to both inform environmental management decisions, and to minimize or mitigate any detrimental environmental effects of a disturbance (Schmitt and Osenberg 1996, Lindenmayer et al. 2012).

Statistical models are used to characterize baseline data and to detect, quantify, and forecast change in operational monitoring data. The choice of model impacts data characterization and estimates of change (e.g., Thomas 1996, Jones-Farrand et al. 2011, Nuno et al. 2014). Evaluations of models used to monitor change have been repeatedly conducted in ecology (e.g., Thomas 1996, Ward et al. 2014, Bell and Schlaepfer 2016) in an effort to identify models that accurately describe pattern in monitoring data. Previous model evaluation studies are often limited to a specific class of models, dataset, and/or a single monitoring objective. Any model evaluation must encompass a wide range of model classes and scenarios change to provide accurate and robust model recommendations. This type of model evaluation is necessary for monitoring programs that do not have standards for analyzing monitoring data, including baseline characterization and detecting, quantifying, and forecasting change.

Marine Renewable Energy (MRE) development exemplifies a planned anthropogenic disturbance and the need for a model evaluation to recommend standards for analyzing monitoring data. The MRE industry continues to grow as the mandated use of renewable energy continues to increase (Lewis et al. 2011). Currently, MRE development is at the demonstration stage (e.g., 1-2 devices are installed for testing and validation purposes) rather than commercial operations. In North America, the lack of commercial scale MRE projects is partially attributed to the uncertainty associated with environmental effects of MRE development. Environmental monitoring programs are required for permitting and licensing of MRE sites, but there are no standards for meeting monitoring requirements (McCann 2012). Development and standardization of environmental monitoring protocols will ensure effective monitoring programs and will expedite the MRE permitting process (Dubbs et al. 2013, Copping et al. 2016). In an attempt to standardize monitoring protocols, general MRE monitoring guidelines

recommend the use of the stressor-receptor construct and Before-After monitoring of receptor indicators to ensure comparable and accurate evaluation of environmental change (e.g., McCann 2012, Klure et al. 2012, Copping et al. 2014). There are no established best practices for characterizing baseline data, or detecting, quantifying, and forecasting change in operational monitoring. An evaluation of statistical models used to analyze monitoring data is required to recommend best practices that will enable cross-site comparisons among MRE sites and to reduce uncertainty in the evaluation of effects due to MRE development.

1.2 Objectives

Objectives of this study are to develop and execute an evaluation of statistical models that can characterize baseline conditions, and can detect, quantify, and forecast change in operational monitoring data. Results of the evaluation will be used to recommend best practices for analyzing monitoring data. An empirical baseline dataset that is considered representative of common MRE monitoring data is used for the model evaluation. A range of statistical models from three classes of monitoring models are included in the evaluation: generalized regression models, time series models, and nonparametric models. A scenario analysis is developed to evaluate model ability to detect, quantify, and forecast a range of shapes and amplitudes of change in simulated Before-After monitoring data. Results from the evaluation are used to recommend model(s) that most accurately characterize baseline data, and detect, quantify, and forecast change in operational monitoring data. The evaluation approach and model recommendations can be used as best practices for the analysis of monitoring data to enable accurate and efficient environmental monitoring programs.

Chapter 2:

Characterizing Baseline Environmental Conditions at Marine Renewable Energy Sites

2.1 Introduction

Marine Renewable Energy (MRE) provides alternate reliable energy that is cleaner than fossil fuels. As the mandated use of renewable energy increases, the MRE field continues to develop and expand (Lewis et al. 2011). There are numerous unknowns associated with development that impede advancements in the industry. Most notably is the effect of site development on the environment (Shumchenia et al. 2012). The environment is defined as abiotic and biotic components comprising a marine ecosystem. It is critical to understand effects of MRE on the environment to minimize, manage, and mitigate any harmful effects and enhance any positive effects on the ecosystem. Environmental monitoring of an MRE site before, and during operational stages (i.e., construction, operation, maintenance, and decommissioning), is designed to track changes relative to baseline conditions and to measure effects of MRE development on the environment.

To understand and document environmental effects, US and UK marine renewable energy developers are generally required to have plans accepted for pre-installation environmental assessments and post-installation monitoring as part of their application and permit licensing for project proposals (DOE 2009, Croll and Andina-Pendas 2009, Portman 2010). The objective of pre-installation environmental assessment is to evaluate the potential change to an environment and to prevent or minimize anticipated detrimental effects caused by MRE development (Jay et al. 2007). Baseline monitoring (i.e., pre-installation monitoring) is a key component of these assessments because it characterizes conditions prior to any alterations to the environment

(Treweek 2009). Characterization of baseline conditions includes the estimation of natural variability, preexisting trends, and relationships between biotic and abiotic components of the environment (Treweek 1996, 2009). The characterization of baseline conditions is used to structure the design of operational monitoring (i.e., post-installation monitoring) programs used to measure environmental change caused by MRE development (McCann 2012). Hypothesized effects of MRE development on the environment are formed from baseline characteristics. Operational monitoring programs are structured to assess the hypothesized effects (Polagye et al. 2011). Assessment of baseline variability is used to determine the appropriate frequency and scale of sampling to ensure accurate measure of change caused by MRE development outside the range of natural variability (McCann 2012). Baseline assessments are a critical first step to establish an efficient and effective environmental monitoring program.

In an effort to ensure efficient, comparable, and informative monitoring programs, protocol guidelines have been developed for MRE monitoring study design and data collection. These guidelines emphasize the use of the stressor-receptor construct to focus MRE monitoring efforts on primary effects of concern (Boehlert and Gill 2010, Klure et al. 2012, Boehlert et al. 2013). A stressor is defined as an external feature or event associated with renewable energy development that causes change in the environment (Boehlert and Gill 2010). Predominant forms of stressors in the MRE monitoring literature include: static (e.g., the device), dynamic (e.g., device movement), chemical, noise, energy removal, and electromagnetic fields (EMF) (Klure et al. 2012, McCann 2012, Boehlert et al. 2013). A receptor is defined as a component of the ecosystem that shows some form of response to the stressor (Boehlert and Gill 2010). Marine mammals, birds, fish, and habitat are ecosystem components commonly cited as important biological receptors (Klure et al. 2012, McCann 2012, Boehlert et al. 2013). The state of a

receptor can be measured using environmental indicators that serve as measurable proxies of the state of the receptor. Biological indicators are used to measure changes to the structure, function, and composition of a receptor, three key attributes of an ecosystem (Noss 1990). Change in an indicator represents change in the state of a biological receptor, and therefore a change relative to the baseline of an environment. Indicators recommended for measuring change in the state of ecosystem components include abundance, distribution, diversity, and behavior (Noss 1990, Nemi and McDonald 2004). Common methods of collecting metrics as proxies of these indicators include trawl surveys, trappings, acoustics, and optics (Polagye et al. 2014, Klure et al. 2012, McCann 2012).

At this time there are no established standard analytic methods to characterize baseline receptor data. Since the choice of statistical method used to model monitoring data can impact trend estimates (Thomas and Martin 1996), the choice of model used to characterize baseline data potentially impacts assessment of baseline conditions, the hypothesized effects of a stressor, and the monitoring used to investigate expected effects. It is important to use the same methods in baseline and operational monitoring to quantify change (McCann 2012). The choice of statistical model for characterizing baseline data also affects the measurement of change in baseline conditions caused by MRE development. An explicit evaluation of statistical models capable of characterizing baseline monitoring data is necessary to aid in the development of best practice models for quantifying baseline conditions.

The objective of this chapter is to evaluate statistical models applicable for analyzing monitoring data in an effort to establish best practices to characterize baseline MRE conditions. The evaluation was applied to empirical baseline case study data representative of common MRE monitoring data. A second objective is to recommend statistical model(s) that are most able to

characterize baseline conditions of biological indicators. Recommendations can be used to establish best practice statistical models for MRE environmental baseline characterization, which will reduce uncertainty in environmental assessments, decrease permitting costs, and enable comparison among monitoring sites to enable the most effective operational monitoring program.

2.2 Methods

2.2.1 Approach

The evaluation was developed to assess the ability of statistical models to characterize baseline environmental conditions to identify potential effects of MRE development on the environment, and to enable accurate measures of effects in operational monitoring. A wide range of statistical models representing model classes used to analyze empirical monitoring data were evaluated. Model selection using a quantitative measure of model interpolation accuracy, or ability to predict data within the range of the empirical data, was used to parameterize models and to evaluate a model's ability to estimate baseline variability and identify relevant predictors (i.e., covariates) of the data. This approach ensured an equal assessment of model accuracy across a range of statistical model classes, while at the same time parameterized all candidate models to have the greatest probability of success in accurately characterizing the data. Residual diagnostics were used to assess the validity of model error distribution and autocorrelation structure assumptions. Results from the evaluation were then used to recommend model(s) most capable of characterizing the baseline monitoring data. All analysis was conducted in the R statistical software environment (R Core Development Team, 2016).

2.2.2 MRE Baseline Case Study

Empirical data from a proposed MRE site was used for model evaluation. A tidal turbine pilot project proposed by the Snohomish County Public Utility District No. 1 in conjunction with

OpenHydro Systems would be located ~1 kilometer off Admiralty Head, Puget Sound Washington, at a depth of ~60 meters (Public Utility District No. 1 of Snohomish County 2012). The project, suspended due to financial constraints, would deploy two 6 meter turbines (http://www.openhydro.com/).

In 2011, acoustic and midwater trawl sampling was conducted to characterize baseline conditions at the site. This study uses the stationary acoustic data from the site (48.18˚ N, -122.73˚ W), which is assumed to be representative of a primary monitoring method that would be used throughout the life of an MRE project. Acoustic backscatter is representative of nekton (i.e., macro-invertebrates and fish that move independently of fluid motion) within the water column (MacLennan et al. 2002). Acoustic backscatter data were recorded using a bottom mounted BioSonics DTX echosounder mounted on a Sea Spider platform (http://www.oceanscience.com/Products/Seafloor-Platforms/Sea-Spiders.aspx) and operating at 120 kHz from May 11 to June 8, 2011 (Horne et al. 2013). The echosounder sampled at 5Hz for 12 minutes every 2 hours. Because of a 3rd surface echo, data values were constrained to 25 m from the bottom, a height corresponding to twice that of the proposed OpenHydro tidal turbine. A -75 dB re $1m^{-1}$ threshold was applied to remove noise (Horne et al. 2013)

A suite of metrics derived from the data was used to quantify nekton density and vertical distribution in the water column (Burgos and Horne 2007, Urmy et al. 2012). The metrics include: mean volume backscattering strength (i.e., Sv) (unit: dB re 1 m-1(Maclennan et al., 2002) hereafter dB), center of mass (units: M), inertia (units: m^2), and aggregation index (units: m^{-1}). Sv functions as a proxy for density, center of mass is a weighted measurement of distribution in the water column, and inertia and aggregation are measures of the spread and dispersion of fish. All metrics are continuous, display periodic autocorrelation (Jacques 2014),

and are trend stationary, assuming that the periodicity in the data is associated with environmental variability. Values of Sv, center of mass, and inertia metrics are normally distributed. Aggregation index (AI) values are right-skewed, non-normally distributed data composed primarily of low aggregation values with spikes of high aggregation (Fig. 2.1). The terms *low state* and *high state* will be used to refer to the two amplitudes of aggregation index values that comprise the empirical data. To maintain efficiency, the evaluation is only applied to the Sv data, illustrative of normally distributed metrics, and the non-normal AI data (Fig. 2.1). These datasets are considered representative of MRE baseline data, because they are collected using common MRE monitoring instrumentation, and are metrics of recommended indicators (i.e., abundance and behavior) of the state of fish (Polagye et al. 2011), a primary biological receptor of MRE stressors. The properties of these datasets are common for monitoring time series data (i.e., temporally continuous, autoregressive, normal and non-normal data) (Chandler and Scott 2011, Gitzen 2012).

Ancillary environmental measurements collected during baseline Admiralty Inlet surveys were used as potential covariates in the candidate models (Jacques 2014). Daily tidal range (m), tidal speed (m/s), and Julian day of year were matched to each time stamp from May 11th to June 8, 2011. Tidal range was calculated as integrated tidal speed through the day (Jacques 2014). A Fourier series defined by a 24 hour period was also included as an environmental variable to represent time-of-day.

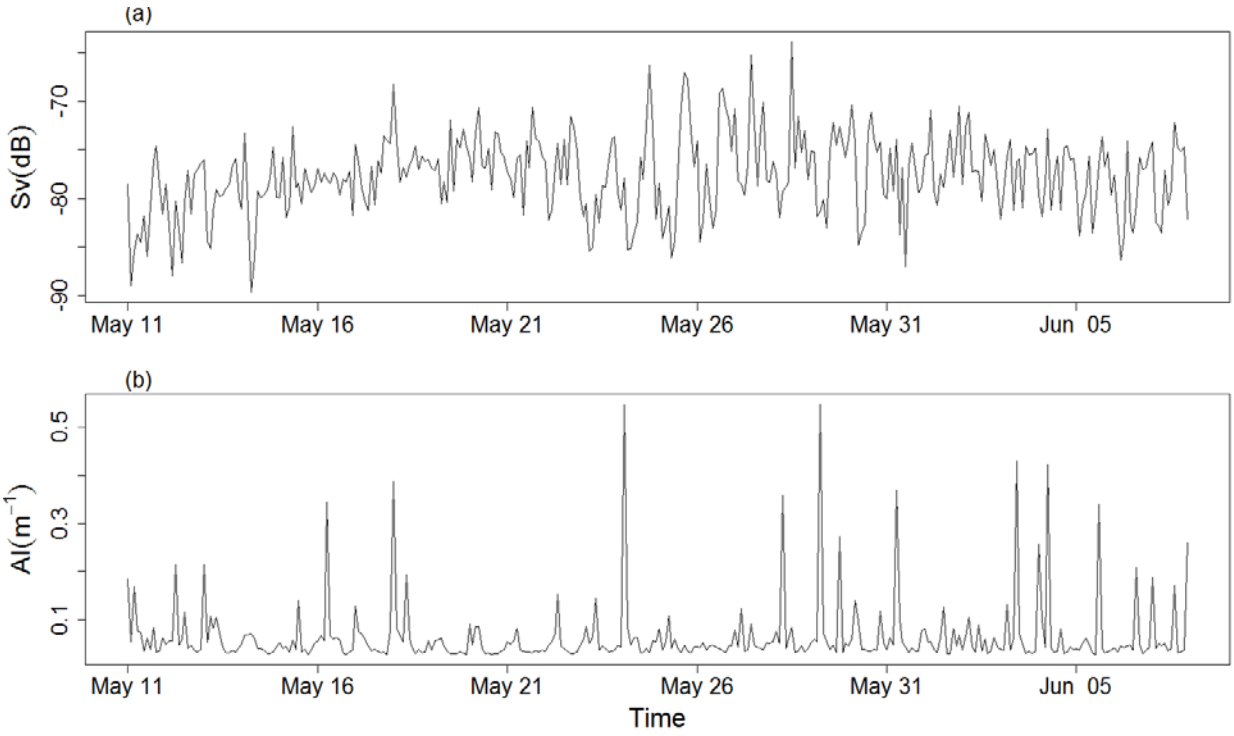

Figure 2.1. Acoustic-based index values derived from data collected from May 11th to June 8, 2011 in 2 hour intervals. (a) Nekton density (Sv) was normally distributed. (b) Nekton aggregation (AI) ranged from 0 to 1 and consisted of low aggregation values with higher values spikes across time.

2.2.3 Candidate Model Classes

All candidate models were selected based on the ability to characterize baseline conditions and to measure change during operational monitoring, because it is desirable to use the same models in all stages of an MRE monitoring program (McCann 2012). Statistical models included in the evaluation were: generalized regression models, time series regression models, and non-parametric regression models (Table 2.1). The model classes encompass models commonly used to measure change over time; including parametric, nonparametric, and stochastic statistical models (Chandler and Scott 2011). Generalized regression models include deterministic

parametric and semi-parametric regression, time series models include both deterministic and stochastic parametric models, and nonparametric models are considered stochastic because they included lagged dependent variables. Deterministic models are defined as assuming a fixed trend pattern, whereas a stochastic model does not assume a fixed trend and includes lagged dependent variables to model the process (Chandler and Scott 2011). Parametric models have a pre-defined functional model structure, nonparametric models do not have predetermined function forms and use the data to develop the variable relationships in the model, and semi-parametric models include parametric and nonparametric model components (Chandler and Scott 2011). All of these models are applicable for baseline characterization, prediction of MRE development effects on the environment, and measuring change in operational monitoring. Models within these classes were chosen based on current use in MRE monitoring studies or their potential applicability to fit the case study data characteristics. Candidate model equations and defined terms can be found in Appendix 2.1.

Generalized Regression Models

The generalized regression model class is composed of a linear regression and parametric and semi-parametric models that are generalized extensions of the linear regression model. Generalized regression models are most commonly used to analyze MRE monitoring data, because they are able to characterize baseline conditions of a biological receptor of interest and predict effects of MRE development on those conditions (e.g., Duck et al. 2006, Tollit et al. 2013, Viehman et al. 2015). These models are capable of measuring natural variability, trend, and predictors of an ecosystem component of interest, including external change. These models produce measures of uncertainty around estimates which are often used in monitoring studies to assess the confidence of baseline characterization or measured change in those conditions

Table 2.1. Description of evaluated candidate regression models including model class, linear or nonlinear form, parametric or nonparametric structure, error components (observation and/or process error) and distribution, and autocorrelation structure.

Model	**Class**	**Form**	**Parametric/ Nonparametric**	**Error Components**	**Error Distribution**	**Auto-correlation Structure**
Linear	GR	Linear	Parametric	Observation error	Normal	None
Generalized least squares (GLS)	GR	Linear	Parametric	Observation error	Normal	Residual correlation
Generalized linear model (GLM)	GR	Linear	Parametric	Observation error	Exponential family	None
Generalized linear mixed model (GLMM)	GR	Linear	Parametric	Observation error	Exponential family	Residual correlation
Generalized additive model (GAM)	GR	Nonlinear	Semi-parametric	Observation error	Exponential family	None
Generalized additive mixed model (GAMM)	GR	Nonlinear	Semi-parametric	Observation error	Exponential family	Residual correlation
Multivariate autoregressive state-space model (MARSS)	Time series	Linear	Parametric	Process and Observation error	Normal	AR-1 lagged variable
Regression - autoregressive moving average model (Reg-ARMA)	Time series	Linear	Parametric	Observation error	Normal	ARMA error
Regression - autoregressive moving average – generalized autoregressive conditional heteroscedasticity model (Reg-ARMA-GARCH)	Time series	Linear	Parametric	Observation error	Generalized normal	ARMA error; GARCH residual variance
Random forest (RF)	NP	Nonlinear	Nonparametric	N/A	None	Lagged variables
Support vector regression (SVR)	NP	Nonlinear	Nonparametric	N/A	None	Lagged variables

Note: The evaluated model classes are generalized regression (GR), time series, or nonparametric (NP) models.

(Schmitt and Osenberg 1996, James et al. 2015). These models have several restrictions that may constrain their ability to accurately characterize baseline data. Parametric and Semi-Parametric models require many *a priori* assumptions, including error distribution and a parametric relationship between predictors and the dependent variable. These models only account for observation error, and do not model the stochastic nature of time series data that is driven by dependence between successive observations through time (Chandler and Scott 2011).

Linear Regression/Generalized Least Squares

Linear regression and Generalized Least Squares (GLS) regressions are the traditional methods used to detect change in Before-After Control-Impact (BACI) monitoring studies (Stewart-Oaten and Bence 2001, Wagner et al. 2002). An analysis of variance (i.e., ANOVA) is a special case of linear regression most often used to detect change in BACI monitoring studies (Hewitt et al. 2001). A linear regression is the most standard form of a regression model. It is a linear, parametric regression model that assumes normal, homoscedastic, and independent errors. A Generalized Least Squares model (GLS) additionally can account for autocorrelation in the residual correlation structure (Pinheiro and Bates 2000).

The linear regression analysis was conducted using the R statistical package "stats" (version 3.4.0), which uses the least squares method for estimating model parameters. The GLS regression analysis was conducted using the R statistical package "nlme" (version 3.1-126), which uses Restricted Maximum Likelihood (REML) parameter estimation.

Generalized Linear (Mixed) Models

Generalized Linear (Mixed) Models have been used to characterize MRE baseline monitoring data and to detect change in data post-installation of MRE devices (e.g., Bergstrom et

al. 2013, Wade et al. 2014, Stenberg et al. 2015). Generalized Linear Models (GLM) are an extension of linear regression that do not assume that data are normally distributed. Generalized Linear Mixed Models (GLMM) are an extension of GLMs that can account for autocorrelation within the structure of a mixed-effects model (Pinheiro and Bates 2000).

The GLM(M)s were only applied to the non-normal dataset, because they do not improve modeling capabilities of a linear regression or GLS for a normal dataset. The GLM(M)s assume a Gamma distribution with an identity link for the non-normal AI data. The Gamma distribution is appropriate for modeling the AI data, because it is a skewed distribution with values >0 and the AI values range from >0-1. The Gamma distribution and link assumption of a constant coefficient of variation is reasonable for the AI dataset, because the data were collected over a full lunar cycle with consistent measurement across time. The identity link was chosen, because previous work has identified this link as the most accurate for parameter estimation when modeling incremental effects of change (Polgreen and Brooks 2012). Although the GLM(M) models are not measuring change in this study, models use to characterize baseline data may also be used to measure change in all stages of monitoring, and consequently are parameterized to be used for this general purpose. Model results were inspected to ensure that the use of an identity link did not produce negative predicted values.

In the GLMM the random effect was a 24 hour count index of time-of-day, used to account for variability among hour-of-day, and the residual correlation structure measured any remaining 24 hour cyclic autocorrelation. An ACF plot of AI data illustrated a 24 hour pattern of residual correlation, which was used to identify data blocks of the time series to define the random effect. The GLM analysis used Iterative Weighted Least Squares (IWLS) for the estimation process, while the GLMM used Penalized Quasi-Likelihood (PQL). The GLM

analysis was conducted using the statistical package "stats" (version 3.4.0), and the GLMM analysis was conducted using the statistical package "MASS" (version 7.3-45). The "MASS" R package was specifically used to fit the GLMM, because it is the only available package that contains options for estimating autocorrelated correlation structures within the traditional GLMM framework.

Generalized Additive (Mixed) Models

Generalized Additive (Mixed) Models have been applied to MRE monitoring studies to characterize data and detect change caused by MRE development (e.g., Petersen 2011, Mackenzie et al. 2013). These models are semi-parametric extensions of GLM(M)s (Wood 2006). GAM(M)s apply smoother functions of predictor variables to model nonlinear relationships (Wood 2006).

The identity link Gamma distribution was specified for the GAM(M)s for the non-normal dataset. Smoother splines were pre-specified for the purposes of this study. The time-of-day predictor consisted of 24 hour count variables formatted as a cyclic cubic regression spline, suggested for cyclic variables (Wood 2006). This structure was used to model time-of-day in the GAM(M)s to be consistent with the shape of a Fourier series, used to model time-of-day in all other candidate models, while permitting use of the smoother spline model structure of the GAM(M)s. All other variables used the thin-plate-regression spline with shrinkage. The thin plate regression splines are the default for the chosen R package, because they are computationally efficient, may include multiple covariates, and do not require knot specification (Wood 2006), The shrinkage allows the smoothness selection to shrink to 0, which minimizes the importance of covariates that should not be included in the model (Zuur et al. 2009). If the

estimated degree of freedom of smoother terms was near 1 then the smooth was replaced with a parametric term and the results from the 10-fold CV model selection process were re-calculated. Any interaction effects included in the GAM(M)s were specified using the "ti" tensor product interaction term, which should be used when main effects are already included in the model structure (Wood 2015. Both the GAM and GAMM used Restricted Maximum Likelihood (REML) for estimation of the smoother functions for the normal data. The GAMM used Penalized Quasi-Likelihood (PQL) by default for non-normal smoother estimation. The GAM(M) analysis was conducted using the statistical package "mgcv" (version 1.8-12).

The "mgcv" package does not require the specification of a random effect in the GAMM. The wiggly (i.e., penalized) components of the spline functions are automatically treated as random effects within the GAMM, and no additional random effect is required for the estimation process (Wood 2015). If no additional random effect is specified, then the response will not be grouped, and the residual correlation structure will measure autocorrelation across the entire time series rather than only measuring 24 hour autocorrelation (Wood 2015). No additional random effect was chosen for the GAMM as it was unnecessary to add more complexity to the data structure to capture the temporal correlation in the time series. This was the purpose of extending the GAM to a GAMM.

Time Series Regression Models

Time series data are thought to be composed of a trend, seasonality, and stationary residuals (Chatfield 1989). Time series models are structured to estimate the main components of environmental monitoring time series data, including: autocorrelation, seasonality, stationary properties, process error (i.e., natural variability in the true state of the population) and/or observation error (Pattengill-Semmens et al. 2011). Evaluated time series models included:

Regression Autoregressive Integrated Moving Average model (Reg-Arima), Regression Autoregressive Integrated Moving Average Generalized Autoregressive Conditional Heteroskedasticity model (Reg-ARIMA-GARCH), and a Multivariate Autoregressive State-Space model (MARSS).

Regression Auto-Regressive Moving Average Model

Autoregressive-Moving-Average (ARMA) model is a traditional time series model that is commonly used for modeling stochastic trends (Chatfield 1989, Chandler and Scott 2011). To date, an ARMA model has not been used in MRE biological monitoring studies. ARMA models assume stationarity and require a large number of observations for accurate estimation (Zuur and Pierce 2004, Hyndman and Kostenko 2007). The autoregressive (AR) component regresses a process on past values, and the moving average (MA) component models the error using previous values of that error. The ARMA model was formatted as a Regression-ARMA (Reg-ARMA) to model dependent data using environmental predictors in addition to lagged dependent values. A reg-ARMA is structured as a linear regression, but the error term in the regression is modeled using the ARMA structure (Hyndman 2015).

The ARMA analysis was conducted using the statistical package "forecast" (version 6.2). Maximum Likelihood (ML) estimation was used for analysis. This is not the default estimation process for the "forecast" package, but it ensures convergence of ARMA parameters if the data are non-stationary. Although the baseline data are trend stationary, these models are intended to also be used for operational monitoring, and data may become non-stationary if change occurs due to MRE development.

Regression ARMA-Generalized-Autoregressive-Conditional-Heteroskedasticity Model

Generalized-Autoregressive-Conditional-Heteroskedasticity (GARCH) models have been previously identified as a possible analytic tool for tracking MRE monitoring data over time (Horne et al. 2013). A GARCH model is an econometrics model originally intended to fit and forecast time varying stock return volatility (i.e., variability, measured as standard deviation of returns over time) in financial markets (Engle 2001). A GARCH model is applicable for modeling time series with heteroskedastic variance, such as the Aggregation Index data. An ARMA-GARCH model uses the ARMA model to estimate autoregressive conditional mean values and the GARCH model to estimate autoregressive conditional variance values of the ARMA residuals (Engle 2001, Zivot 2009, Ruppert 2011). When independent predictor variables are included in an ARMA-GARCH model, then the model becomes a linear regression, with errors modeled using an ARMA model, and the variance of the residuals modeled using the GARCH model (Ruppert 2011).

The reg-ARMA-GARCH model assumes a normal distribution for the Sv dataset. A skewed-student-t-distribution (sstd) was selected for the AI dataset. GARCH R packages only permit the use of skewed and heavy-tailed versions of the normal distribution. The skewed-student-t distribution (sstd) can be represented as normal mean-variance mixture model with an Inverse Gamma mixture distribution (Hu and Kercheval 2008). This distribution aligns most closely with the structure of the AI data and the Gamma distribution chosen for the non-normal generalized regression models. The GARCH analysis was conducted using the statistical package "rugarch" (version 1.3-6). The "hybrid" solver was selected for parameter estimation, which systematically runs through all available estimation algorithms in the R package until a successful convergence is produced.

Multivariate-Autoregressive-State-Space model

A Multivariate-Autoregressive-State-Space (MARSS) model is a dynamic time series model that has been used in a wide range of fields (e.g., economics, engineering, and ecology; Holmes et al. 2012), and has been used to characterize the acoustic baseline data from Admiralty Inlet (Jacques 2014). The MARSS model includes an AR-1 lagged dependent process variable to model mean-reversion (i.e., how strongly a stationary time-series data reverts to its mean value): of the dependent variable over time. The state-space model assumes that the underlying process of the data is stochastic and explicitly partitions the total variance into process and observation errors (i.e., stochastic environmental variability and measurement variability).

Two forms of the MARSS model, one with fixed low measurement error and estimated high process error (MARSS-P) and the other with fixed low process error and estimated high measurement error (MARSS-M), were used as candidate models. To produce the most precise estimates and ensure convergence, the mean-reversion (or density-dependence), process, and/or observation error parameters may be fixed in a MARSS model depending on the structure of the data and *a priori* knowledge of the data. Prior knowledge about the measurement error of stationary acoustic data collection permitted the fixed estimate of low measurement error in the MARSS model. Indicator metric values are calculated directly from linear backscatter values collected from the stationary EK-60 echosounder. We assumed that the only potential sources of observation error are due to calibration and hydrographic conditions. These sources of error have been suggested to equate to a maximum of 5% each of the total error (Simmonds and MacLennan 2005). Therefore, a MARSS model (MARSS-P) was structured with fixed measurement error that equated to 10% of the total error for the normal and non-normal data set, and process error was estimated as a parameter. A second MARSS model (MARSS-M) with

fixed process error that equated to 10% of the total error, and observation error estimated as a parameter, was also used as a candidate model. As an aside, if either of these models were estimated in a Bayesian framework, priors could be assigned to these values. Within a MARSS model the B parameter estimates mean-reversion: B=1 indicates a non-stationary random walk, and $B \leq 1$ indicates mean-reversion in the data. B was estimated as a parameter to reflect the stationarity exhibited in the baseline data. The R statistical package "MARSS" (version 3.9) was used to conduct all analyses. The default estimation method of Maximum Likelihood via an Expectation-Maximization algorithm was used for the evaluation process (Holmes et al. 2015).

Nonparametric Regression Models

Nonparametric models are recommended for use when little is known *a priori* about the data, and when accurate predictions are needed (Gitzen 2012). These models do not require a specified distribution of the data nor do they assume linearity. These models have previously been used in renewable energy power prediction studies (e.g., Foley et al. 2012, Perera et al. 2014). They do not allow for parametric estimates of predictor variables or measurements of statistical significance. Lagged variables can be included in nonparametric models to account for autocorrelation in a time series. A Random Forest Regression and Support Vector Regression were included in the evaluation.

Random Forest Regression Model

Random Forest (RF) regression is a nonparametric machine learning regression model that stems from the original Random Forest classification model developed by Brieman (2001). with the strength of being accurate predictive models (Cutler et al. 2007, Kane et al. 2014). Although RF does not include parametric measures of predictor variables, or the significance of these values, it does include a measure of variable importance that can be used to evaluate

predictor variables. RF models have previously been used to characterize the importance of environmental factors and forecast species distribution in wind renewable energy biological monitoring studies (e.g., Belaire et al. 2014, Hayes et al. 2015). A Random Forest is a collection of statistical decision trees applied to random bootstrap samples of data that are averaged to produce predicted values (Liaw and Wiener 2002). It is considered a nonlinear local model, because it is an ensemble of piece-wise, constant functions applied to local subsets in the data (Strobl et al. 2009).

The R statistical package "randomForest" (version 4.6-12) was used to carry out all Random Forest analyses. This package implements the traditional Random Forest methodology based on original Fortran code by Breiman and Cutler (2004).

Support Vector Regression Model

Support Vector Regression (SVR) is a nonparametric machine learning regression model that stems from Support Vector Machine models that were originally used for classifications. SVR models have previously been cited for their time series forecasting ability (Tay and Cao 2001, Hansen et al. 2006), and have been used for species distribution modeling (e.g., Drake et al. 2006, Lorena et al. 2011). Unlike Random Forest models, SVR models do not provide a direct method for measuring variable importance and parameterized SVR models are more difficult to interpret than RF models (Cutler et al. 2007, Lorena et al. 2011). SVR uses a specified kernel to map data into a higher dimensional space to produce a linearly separable regression (Cortes and Vapnik 1995, Hsu et al. 2010). In the common specification of SVR (cf. Vapnik 1995) an ε-insensitive loss function is used for model estimation, which is similar to an absolute loss function but assumes any error < ε, a predefined value of residual error, is equal to zero (Smola

and Schölkopf 2004). Detailed information on model parameters and structure can be found in: Vapnik (1995), Smola and Schölkopf (2004), and Hsu et al. (2010).

A Linear kernel SVR (SVR-L) and Radial-Basis-Function kernel (SVR-RBF) were used as candidate models in the evaluation. The correct kernel choice may not be known *a priori* and the evaluation of both kernels provides a robust evaluation of the SVR model, as the two forms contrast in their flexibility and complexity. The RBF kernel is nonlinear and is the most common initial kernel choice because of its ability to flexibly fit nonlinear data and its generally accurate predictive performance (Berk 2008, Clarke et al. 2009). The linear kernel is the simplest and least flexible kernel, but it has been shown to predict trended time series more accurately than SVR-RBF (e.g., Crone et al. 2006). The R statistical package "e1071" (version 1.6-7) was used to carry out all Support Vector analyses.

2.2.4 Model Selection

To identify a statistical model that is capable of characterizing baseline MRE conditions, the evaluation must be consistent across models, and at the same time, allow each model to have the highest probability of success. Every candidate model underwent the same model selection process to ensure an equal opportunity for each model to be optimally structured. Cross-validation (CV) model selection was used to select the structure of each model and to compare the accuracy of the parameterized candidate models. Cross-validation uses the difference between observed and predicted values to calculate model accuracy. This method can be applied to all models, and is a standard model selection method (Burnham and Anderson 2002). Cross-validation has been recommended when there is not high *a priori* knowledge of model structure, and the goal is to interpolate, or predict within the data range (Gitzen et al. 2012). A measure of

predictive accuracy, rather than fit to the training data, is most appropriate to quantify model accuracy, because the goals of characterizing baseline conditions are to enable future measurement of change, and prediction of potential effects of MRE development.

In cross-validation a portion of the data is used as a training-set to parameterize the model. The parameterized model is then fit to the remainder of the dataset, the test-set. Predicted values from the model applied to the test-set are used to calculate Root-Mean-Squared-Error (RMSE). RMSE is a measure of model accuracy based on the average deviance of model predicted values ($\hat{y}_i$) from observed values (y_i):

$$\sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2.1}$$

Where i is the observed i^{th} value, and n is the sample size. A RMSE value close to 0 indicates a better forecasting ability. Averaging more than a single training/test set allows for a more robust model selection process. A 10-fold CV has been suggested as the most effective number of training/test sets for model selection (Hastie et al. 2009, Arlot and Celisse 2010), and provides a balance between a highly biased model (i.e., one that underfits data), and a highly variable model (i.e., one that overfits data) (James et al. 2015). In 10-fold cross-validation 10 equally sized random subsets of the data are used repeatedly such that 9 subsets compose a training-set and a single subset is the test-set to produce a total of 10 training and test datasets. The 10-fold cross validation model selection allows for the best possible structure of each model, while keeping the selection method equal across all models.

Model selection was performed on 24 versions of each candidate model. Rather than choosing all possible interactions of model covariates (i.e., predictor variables), it is recommended to select a set of model versions based on *a priori* knowledge of the biological

system (Burnham and Anderson 2002). Covariates used in each model included Julian day, tidal range, tidal speed, and a Fourier series defined by a 24 hour period. All covariates were demeaned (i.e., the sample mean was subtracted from observed covariate values to produce a zero mean) before analysis to remove estimate error due to multi-collinearity. No three-way interactions were included in the analysis. Final model versions for selection included two-way interactions: Julian day-tidal speed, Julian day-tidal range, tidal speed-tidal range, and tidal speed- 24 hour period. An interaction between tidal range and the 24 hour period was not included in analysis, because it is an integration of tidal speed throughout the 24 hour period (Jacques 2014). Julian day and the 24 hour period did not interact, because 24 hour Fourier series creates a cyclic relationship within the nekton data that is consistent across each Julian day. All possible combinations of the covariates and these two-way interactions produced 24 versions of each candidate model. Two special cases increased the number of model versions. The MARSS model allows for covariates in both the process and observation equation (i.e., the model for the underlying state of a population process, and the model of observations of that process). Combinations of covariates may equally affect both the process and observation of the response. Similarly, the reg-ARMA-GARCH model allows for covariates in both the conditional mean and variance equation. Therefore, every combination of the 24 models versions in these two-part models were also evaluated in the model selection process to produce a total of 576 versions of the MARSS and reg-ARMA-GARCH models.

A previously established model selection protocol was used in this study to structure model selection. The model selection protocol established by Diggle et al. (1994) and Wolfinger (1993) and used in Zuur et al. (2009) suggests selecting the residual variance structure prior to the predictor variables. The residual variance structure is accounted for first, because it affects

the significance of the covariates (Hedeker and Gibbons 2006). All candidate models assume independent error, and any autocorrelation must be included in the model selection. Autocorrelated error can bias variance estimates of parameters and invalidate hypothesis testing (Stewart-Oaten and Bence 2001). Initially, the optimal autocorrelation structure was deduced using the 10-fold CV method and the full version of each candidate model (i.e., all main effects and interactions), and then the 10-fold CV method was used to select the optimal structure of predictor variables. For models that included an Autoregressive Moving Average (ARMA) correlation structure, a range of AR and MA lagged variable values were used in the autocorrelation selection process. ARMA error structures with an AR or MA lag value larger than 3 tend to generate convergence problems and may not be necessary to model autocorrelation (Schabenberger and Pierce 2002, Zuur et al. 2009). Autocorrelation structure selection for models that included an ARMA error structure included all combinations of AR and MA lagged variable values ranging from 0 to 3. The number of lagged variables in the nonparametric models was also initially selected prior to the covariate structure. Autocorrelation Function (ACF) plots of the detrended data indicated that the minimum number of data points at which autocorrelation was no longer present was 62 for the non-normal data and 50 for the normal data. To be consistent in the selection process, 0 to a maximum of 62 lags were included in the non-parametric model selection for both non-normal and normal datasets.

The nonparametric models included an additional model selection for tuning parameters. The Random Forest regression has three tuning parameters: node size, number of trees, and mtry (Breiman 2001). The node size is the minimum number of data points in the terminal node of a tree. Mtry is the number of randomly chosen subsample of predictor variables used to fit data within each tree. An increase in node size has previously been shown to have little effect on the

fit of a Random Forest model (e.g., Gutierrez 2011, Ishwaran and Malley 2014). The nodesize was set at the default value of 5 data points. The number of trees and value of mtry were selected using 10-fold cross validation. The default value of number of trees, 500, was increased by steps of 500 until the RMSE value stabilized. As previously suggested for mtry tuning selection, the mtry parameter was varied from the default using a step factor of +/- 2 until the RMSE value no longer decreased (Liaw and Wiener 2002).

The Support Vector Regression included kernel tuning parameters. The SVR-L and SVR-RBF models both include the cost tuning parameter, and SVR-RBF additionally includes the gamma tuning parameter. Both parameters affect bias-variance tradeoffs within the SVR (Hastie et al. 2009). Cost affects the penalization of estimation error and the smoothness of the estimated regression. A high cost indicates a low tolerance for error and can potentially overfit the data. The gamma value controls the width of the kernel (Thissen et al. 2003). A high gamma indicates a smaller width that increases the localized fit of the model and potentially overfits the data. An iterative 10-fold cross validation grid search of cost values ranging from 2^{-5} to 2^{15} at intervals of 2^2 and gamma values ranging from 2^{-15} to 2^3 by 2^2 was used following the recommendation of Hsu et al. (2010) and Berk (2008). At each iteration a finer grid search was conducted until the RMSE value could not be further reduced.

2.2.5 Residual Diagnostics

Residuals of each parameterized candidate model must be evaluated to determine if there is any remaining pattern in the data, which may indicate model assumptions are invalid and/or the structure of the data is not being accurately portrayed by the parameterized model. Together, results from the 10-fold CV and the residual diagnostics enable an evaluation of a model's ability to characterize baseline data. The Zuur et al. (2009) protocol used to structure the previous steps

of model selection was used for residual diagnostics assessment. The form of each model that produces the lowest average RMSE value from the 10-fold CV process was refit to the entire dataset. Residual diagnostics included an inspection for homogeneity and independence using residual plots and ACF plots.

2.3 Results

2.3.1 Nekton Density (Sv) Data

Model Selection Results

The most flexible models produced the most accurate interpolation of the data (i.e., prediction of the test data in 10-fold cross-validation model selection) based on average Root-Mean-Squared-Error (Table 2.2). The two nonlinear, non-parametric models, SVR-RBF and RF, produced the best interpolation of the data based on their average RMSE values of 3.05 and 3.16, followed by SVR-L (average RMSE of 3.22). The most flexible linear parametric models, the MARSS-P and MARSS-M models, were the next most accurate candidate models (Table 2.2). The GAM(M)s produced better interpolation of the data than all other parametric regression models, further highlighting the order of model performance from most to least flexible candidate models. The simplest candidate models, GLS and linear regression, had the worst interpolative accuracy (average RMSE value of ~3.54 each).

Table 2.2. Parameterized candidate models from the 10-fold Cross-Validation model selection for the Sv data, including environmental predictors, autocorrelation structure, associated average Root-Mean-Squared-Error (RMSE), and presence of autocorrelation in the residual ACF plots.

Model	**Environmental Predictors**	**Auto-correlation Structure (AR,MA)**	**Average RMSE**	**Residual Auto-correlation**
SVR-RBF	Day, Fourier Series, Tidal Range, Tidal Speed	(1,0)	3.05 (0.147)	Yes
RF	All Environmental Predictors	(14,0)	3.16 (0.274)	No
SVR-L	Fourier Series	(13,0)	3.22 (0.313)	No
MARSS-P	Process Eq: Day, Fourier Series, Tidal Range Observation Eq: Day, Tidal Range, Day-Tidal Range	(1,0)	3.30 (0.238)	Yes
MARSS-M	Process Eq: Tidal Range, Tidal Speed, Tidal Speed-Tidal Range Observation Eq: Day, Fourier Series, Tidal Speed	(1,0)	3.34 (0.239)	Yes
GAM	Day, Fourier Series, Tidal Range†, Day:Tidal Range†	NA	3.43 (0.190)	Yes
GAMM	Day, Fourier Series, Tidal Range†	(1,0)	3.45 (0.190)	Yes
Reg-ARMA-GARCH	Mean Eq: Day, Fourier Series, Tidal Range, Day: Tidal Range Variance Eq: Day, Fourier Series, Tidal Range,	ARMA:(1,0) GARCH:(2,3)	3.53 (0.169)	Yes
Reg-ARIMA	Day, Fourier Series, Tidal Range, Day: Tidal Range	(1,0)	3.54 (0.168)	Yes
GLS	Day, Fourier Series, Tidal Range, Day: Tidal Range	(1,0)	3.54 (0.161)	Yes
LM	Day, Fourier Series, Tidal Range, Day: Tidal Range	NA	3.54 (0.163)	Yes

Notes: Models are ranked in descending order of average RMSE and associated variances of average RMSE are shown in parenthesis. The number of autoregressive (AR) and moving-average (MA) variables in model autocorrelation structures is shown in parenthesis as (AR, MA). The specified (AR, MA) structure of the nonparametric models indicates the number of lagged dependent variables included in the parameterized models. The environmental predictors are listed in alphabetical order of main effects followed by interactions. The Tidal Range* predictor is parametric in the GAM(M)s.

Predictor Variables

All candidate models included the 24 hour Fourier series as an environmental predictor of the Sv data (Table 2.2). SVR-L is the only candidate model that did not also include day and tidal range as environmental predictors (Table 2.2). All parametric regression models also contained the day-tidal range interaction predictor variable, except for MARSS-M and GAMM (Table 2.2). MARSS-M and SVR-RBF are the only candidate models that included tidal speed as a relevant environmental predictor, except for RF, which included all environmental predictors and their subsequent interactions in the model (Table 2.2). The SVR-L and RF models included up to 26 and 28 hour lagged dependent variables as model covariates, whereas SVR-RBF only included a 2 hour lagged dependent variable.

Residual Diagnostics

All models displayed autocorrelated residuals except for SVR-L and RF (Table 2.2). All candidate models that were not structured to account for autocorrelation (i.e., linear regression and GAM) displayed a 2 hour correlation in addition to a 24 hour correlation in the residuals. The candidate models that did account for autocorrelation all modeled a lag-1 correlation in their respective model structures, and did not exhibit the 2 hour residual correlation. SVR-RBF had autocorrelated residuals at an 18 hour lag, which is a different pattern than all other candidate models. All residuals were homogenous for all candidate models, indicating that the assumption of normality was appropriate for modeling the distribution of the nekton density data.

2.3.2 Nekton Aggregation (AI) Data

Model Selection Results

Unlike the model selection results for nekton density, model ability to accurately interpolate the data did not appear to be ranked by the most to least flexible model. The SVR-RBF model produced the lowest average RMSE value (0.0667), while the other two non-parametric models, RF (0.0681) and SVR-L (0.0689), produced the highest average RMSE values (Table 2.3). The MARSS-P model was the second best interpolative model (RMSE = 0.0673) while MARSS-M was the third worst interpolative model (RMSE = 0.0678). The GLM(M)s were the 3rd and 4th best interpolative models, and displayed greater average predictive accuracy than all other parametric regression models, except for MARSS-P. The most simplistic and inflexible models, GLS and linear regression, ranked directly below the GLM(M)s in interpolation performance (Table 2.3).

Predictor Variables

All models included the 24 hour Fourier series as an environmental predictor, except for GLMM, which only included day and tidal range as predictor variables (Table 2.3). The majority of candidate models included few environmental predictors (≤ 3 per model equation), and no interaction effects. RF and GAM included tidal speed as a predictor variable in addition to the 24 hour Fourier series, whereas GLM and MARSS-M included the day predictor variable in addition to the 24 hour Fourier series (Table 2.3). The Reg-ARMA-GARCH model was the parametric model with the most extensive set of environmental predictors, including day and tidal-range in addition to the 24 hour Fourier series (Table 2.3). The SVR-RBF and SVR-L models included the highest number of environmental predictors (7 variables), and are the only

Table 2.3. Parameterized candidate models from the 10-fold Cross-Validation model selection for the AI data, including environmental predictors, autocorrelation structure, error distribution, and associated average Root-Mean-Squared-Error (RMSE).

Model	**Environmental Predictors**	**Auto-correlation Structure (AR,MA)**	**Error Distribution**	**Average RMSE**
SVR- RBF	Day, Fourier Series, Tidal Speed, Day: Tidal Speed, Tidal Speed: Fourier Series	(13,0)	NA	0.0667 (0.000920)
MARSS- P	Observation Eq: Fourier Series	(1,0)	Normal	0.0673 (0.000869)
GLM	Day, Fourier Series	NA	Gamma (identity)	0.0674 (0.000847)
GLMM	Day, Tidal Range	(1,0)	Gamma (identity)	0.0674 (0.000855)
GLS	Fourier Series	(1,0)	Normal	0.0675 (0.000859)
LM	Fourier Series	NA	Normal	0.0675 (0.000858)
Reg-ARIMA	Fourier Series	(1,2)	Normal	0.0675 (0.000865)
Reg-ARMA-GARCH	Mean Eq: Day, Fourier Series, Tidal Range Variance Eq: Fourier Series	ARMA: (1,0); GARCH: (2,0)	Skewed-student-t	0.0675 (0.000953)
GAM	Fourier Series, Tidal Speed	NA	Gamma (identity)	0.0677 (0.000874)
GAMM	Fourier Series	(2,0)	Gamma (identity)	0.0677 (0.000877)
MARSS-M	Process Eq: Day Observation Eq: Day, Fourier Series	(1,0)	Normal	0.0677 (0.000880)
RF	Fourier Series, Tidal Speed	(2,0)	NA	0.0681 (0.000888)
SVR-L	Day, Fourier Series, Tidal Range, Day: Tidal Range, Tidal Range: Fourier Series	(1,0)	NA	0.0689 (0.000858)

Notes: Models are ranked in descending order of average RMSE and associated variances of average RMSE are shown in parenthesis. The number of autoregressive (AR) and moving-average (MA) variables in model autocorrelation structures is shown in parenthesis as (AR, MA). The specified (AR, MA) structure of the nonparametric models indicates the number of lagged dependent variables included in the parameterized models. The environmental predictors are listed in alphabetical order of main effects followed by interactions.

candidate models that included interaction effects between the predictors. Both these models included tidal speed and the 24 hour Fourier series. The SVR-RBF identified the day predictor as a relevant predictor, whereas the SVR-L model replaced the day predictor with tidal-range within the model formula (Table 2.3). The SVR-L and RF models included fewer lagged dependent variables than SVR-RBF. SVR-L and RF included up to 2 and 4 hour lagged dependent variables, whereas SVR-RBF included up to 26 hour lagged dependent variables in the model.

Residual Diagnostics

No candidate model had autocorrelated residuals. All model residuals were heteroskedastic, indicating that no model was able to capture the highly right-skewed distribution of AI data.

2.3.3 Data Characteristics- Nekton Density (Sv)

Variability

Ranking of models in the 10-fold CV selection is ordered most to least flexible model. This result suggests that the data have a highly variable range around the mean, because flexible models were required to most accurately predict the structure of the data. The high variability of nekton density is validated by the ranking of the MARSS models. The MARSS models differed in their partitioning of the total error into process and observation errors (MARSS-P modeled high process error, and MARSS-M modeled low process error). The difference in error structure between these two models affected parameterization of the mean-reversion (parameter B) in the data. MARSS-P estimated higher mean-reversion $B<0.6$, whereas MARSS-M estimated low mean-reversion $B>0.9$. Results of the 10-fold CV model selection suggest that estimating high natural variability and a more mean-reverting population produces a more accurate interpolation

of the data compared to attributing variability as low density-dependence. The nonlinear GAM(M)s produced the best interpolation of the data after the MARSS models. The smoothing parameters in the GAM(M)s do not exhibit strongly nonlinear patterns in the data not modeled by the other candidate models. The only nonlinear smoothing parameters in the GAM(M)s, is the day covariate (i.e., GAM EDF = 2.74, GAMM EDF = 2.69) (Fig. 2.2). The ability of the GAM(M)s to capture the slightly concave trend in nekton density across days enables these models to produce a better fit to the data than the parametric models. Results of the 10-fold CV model selection indicate that the range of nekton density is highly variable, but the data are not strongly nonlinear.

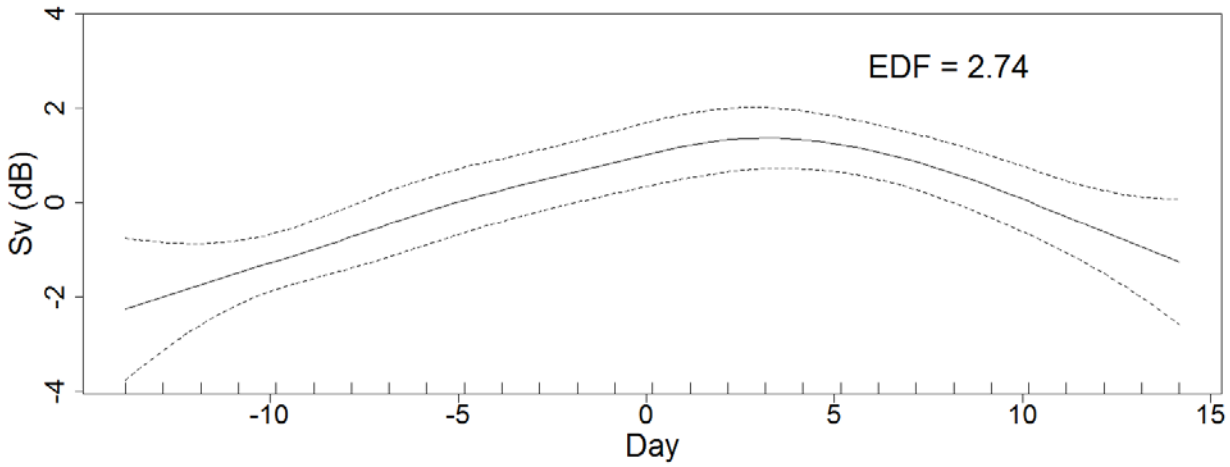

Figure 2.2. Parameterized nonlinear relationship between the demeaned day covariate and nekton density (Sv), and the corresponding effective degrees of freedom (EDF) estimated from the GAM. The dashed line represents 2 standard error bounds on the estimated values. Note: Tidal range is a parametric variable in the parameterized GAM. The day covariate smoother spline in the GAMM is similarly parameterized (not shown).

Environmental Predictors

The pattern of environmental variables included in the final version of each candidate model provides insight into important predictors of nekton density. The most accurate interpolative models included at least one lagged dependent variable, indicating the autoregressive nature of the nekton density, which has previously been documented (Jacques 2014). There is also an indication of strong periodicity in the data, as either tidal range and/or 24 hour Fourier series were included in all final models. The inclusion of the day environmental variable in most candidate models suggests a trend in nekton density across the one month data collection period.

Distribution and Autocorrelation

Residual diagnostics from candidate models supports the assumption of a normal distribution, and reflects the autoregressive structure of the data. All models had homoscedastic residuals, confirming the assumption that the data are normally distributed. Nonparametric models that included 13-14 lagged variables were the only models that did not contain a 24 hour periodic autocorrelation in the residuals, demonstrating a strong periodic daily cycle in nekton density.

2.3.4 Data Characteristics- Nekton Aggregation (AI)

Variability

Model selection results highlight the low variability of the low state AI data, and the inability of the candidate models to characterize variability of the spikes in aggregation index values. Differences in performance of the MARSS models provide insight to the structure of

variability in the AI data. The parameterized MARSS-P model estimated high mean reversion (i.e., B>0.1), compared to the MARSS-M (B>0.7). The MARSS-M model assumes that the spikes are process error, whereas MARSS-M accounts for spikes in aggregation index values as observation error. The 10-fold CV results indicate that the MARSS-P model is a more appropriate model for the data, suggesting that spikes in nekton aggregation are natural stochasticity, not observation error, and that the AI data is strongly mean-reverting. The rank of the most simplistic and least flexible candidate models (i.e., GLM(M), GLS, and linear regression) as the next best interpolation models after MARSS-P validates the characterization of the low state AI data as stationary and low variability as parameterized by the MARSS-P model. The greater interpolation accuracy of the linear regression models over the nonlinear GAM(M)s, remaining time series models, and the nonparametric RF and SVR-L models highlights the inability of the more flexible models to better capture variability of nekton aggregation or accurately predict spikes in nekton aggregation.

Environmental Predictors

The range of environmental predictor variables included in all candidate models provides insight to the dominant processes influencing nekton aggregation. The 24 hour Fourier series is the most consistent environmental predictor among all candidate models, appearing in all but the GLMM, and the only predictor in 5 of the 13 candidate models. The ubiquity of the 24 hour Fourier series highlights the diel periodicity of nekton aggregation (Urmy et al. 2012, Wiesebron et al. 2016). The lack of consistent environmental predictors besides the Fourier series may indicate that the AI data are not predictable using the chosen environmental predictors. By definition nekton movements are independent of fluid motions, and therefore tidal variables may not be primary drivers of nekton aggregation patterns over time (Urmy et al. 2012).

Distribution and Autocorrelation

All candidate models display heteroskedastic errors, illustrating the inability of any candidate model to model the spiked structure of the AI data, regardless of distributional assumptions. Although the SVR-RBF model interpolated the data more accurately relative to all other candidate models, it was still not able to consistently estimate the occurrence and magnitude of spikes in nekton aggregation index values, as reflected by its heteroskedastic error. The MARSS-P model had a much smaller range of residual values compared to other models (0.006 vs ~0.5). The smaller range of the MARSS-P residuals highlights the ability of the MARSS-P model to better fit the structure of the AI data relative to the other candidate models by allocating the spikes in aggregation as process variation. The MARSS-P residuals are heteroskedastic, which highlights the inability of the MARSS models to correctly fit the non-normal distribution of the AI data. All models, including those that did not account for autocorrelation, do not have autocorrelated residuals, suggesting that temporal correlation in the data is weak, and can be modeled by including diel environmental predictors.

2.4 Discussion

2.4.1 Model Efficacy

Overall, the evaluation illustrates the strengths and weaknesses of models used to characterize acoustic-based, empirical MRE baseline monitoring data. State-space models are identified as a suitable model to characterize baseline MRE monitoring data. MARSS state-space models provide the most thorough description of baseline data by accurately interpolating nekton density and aggregation, and quantifying parametric estimates of environmental predictors, and process and observation error. Nonparametric models also excel in interpolating data, but their

predictor variables are not as interpretable or consistent as state-space models, making them unsuitable for data characterization. Deterministic parametric and semi-parametric models (i.e., Reg-ARMA, reg-GARCH, GLS, Lin, GLM(M), and GAM(M)) were less accurate in interpolating data than MARSS models. MARSS state-space models have previously been recommended to characterize renewable energy monitoring data due to their incorporation of process and observation error, inclusion of environmental predictors in the model structure, and predictive abilities (Jacques 2014, Diffendorfer et al. 2015).

Interpolation Accuracy

All nonparametric models excel in interpolating nekton density data, but SVM-RBF is the only nonparametric model that also accurately interpolates nekton aggregation data relative to all other candidate models. RF and SVR models are known for their predictive accuracy due to their lack of structural assumptions (James et al. 2015). RF and SVR-RBF models are also highly flexible due to their inclusion of lagged, dependent variables and nonlinear, localized structure (i.e., fit to local subsets of data rather than fitting a global trend to the data). Localized models are able to interpolate highly variable data because they do not assume a constant global variance (Taddy et al. 2011). SVR-RBF models tend to have greater interpolation accuracy than SVR-L models because of their more flexible kernel, but both have been shown to excel in data prediction (e.g., Crone et al. 2006, Kordon 2009).

In addition to nonparametric models, state-space models also accurately interpolate nekton density and aggregation data relative to all other candidate models. The MARSS estimate of both process and observation error, and the lag-1 structure of the process equation provide the flexibility to accurately fit the stochastic nature of time series data (Dornelas et al. 2012,

Hampton et al. 2013). The MARSS-P model produced a more accurate interpolation of both datasets compared to the MARSS-M model, because the fixed low measurement error was a more appropriate assumption for linear backscatter values by a stationary echosounder.

Deterministic parametric and semi-parametric models (i.e., Reg-ARMA, reg-GARCH, GLS, Lin, GLM(M), and GAM(M)) generally produced less accurate interpolations of the nekton density and aggregation data than nonparametric and state-space models. The time-series models, Reg-ARMA and Reg-ARMA-GARCH, predict both nekton datasets similarly to the linear regression, GLS, GLM, and GLMMs, which is not a surprising result as these time-series models are also linear, parametric, and their predictions are deterministic, regardless of their inclusion of autocorrelated error (Hyndman and Athanasopoulos 2014). Due to the similarity in model structure and predictive abilities, time series (i.e., Reg-ARMA and Reg-ARMA-GARCH) and parametric (i.e., linear regression, GLS, GLM, and GLMM) models are grouped together as deterministic parametric models. Parametric regression models are known for their relative inability to accurately estimate complex, variable data patterns (compared to more flexible models; see Barry and Elith 2006, Shmueli 2010). This inability of parametric models to accurately estimate complex data patterns was validated in the current study. The nekton density data have a highly variable range and an autoregressive structure, as identified in the model evaluation. The nekton aggregation index data includes highly variable spikes and are non-normally distributed. Deterministic parametric and semi-parametric models could not accurately interpolate these complex datasets due to their inflexible properties, the lack of lagged dependent variables, not including stochastic variability as process error, and the assumption of pre-specified error distributions.

Partitioning of Residual Error

The partitioning of total error into process and observation components in a state-space model has been shown to reduce bias and improve accuracy in estimating population abundances (e.g., De Valpine and Hastings 2002, Lindley 2003, Ward et al. 2010). The MARSS models are the only evaluated models that include both process and observation error parameters, which enables a more thorough and accurate characterization of baseline nekton variability. Given that nekton at MRE monitoring sites are known to vary in both in density and behavior (e.g. Jacques 2014, Wiesebron et al. 2016), the MARSS-P model is an appropriate choice as it can quantify variability when using high process error estimates. The MARSS-P model characterized spikes in aggregation index values as process error, which aligns with results of a previous study attributing spikes in AI data to natural variability in nekton aggregations (Urmy et al. 2012). The remaining parametric and semi-parametric regression models did not perform well as they do not explicitly include process error in model structure, and implicitly assume that spikes in aggregation data are observation error. Similarly, nonparametric models did not partition residual model error, because they do not provide explicit estimates of process or observation error.

Environmental Predictors

Even though nonparametric models are flexible and excellent predictive models, they do not consistently identify the same environmental predictors, and are difficult to interpret relative to all other evaluated models. Consistent selection of variables by different statistical models can reduce uncertainty in the choice of relevant environmental predictors (Burnham and Anderson 2002, Jones-Farrand et al. 2011). Covariates identified by nonparametric models generally differed from covariates selected in other models, which reduces confidence in the selected

predictor variables. It is also difficult to assess if the selected predictor variables in nonparametric models were misspecified due to their lack of measureable parameters. SVR-L and SVR-RBF models differed in their environmental predictors, number of lagged dependent variables, and interpolation accuracy for both the nekton density and aggregation data; illustrating the influence of kernel choice and tuning on model characterization of data. The SVR model is known to be highly sensitive to choice of kernel and tuning parameters, and difficult to interpret due to its nonparametric structure (Berk 2008, Lorena et al. 2011). These attributes reduce the ability of SVR models to provide reliable and interpretable characterizations of baseline MRE monitoring data. Unlike the SVR models, the RF model estimates variable importance, which enables the RF model to be used for exploratory analysis of relevant predictors of a dataset (Strobl et al. 2008, Gitzen et al. 2012).

The MARSS state-space models provide parametric estimates of predictor variables, which is an advantage over nonparametric models for making inferences (James et al. 2015). Selection of predictor variables in the MARSS models was influenced by the parameterization of process or observation error in the MARSS models. Fixing the amount of process or observation error in a state-space model as low or high relative to the total variability produces an inverse estimate of the other error parameter (Dennis et al. 2006, Hampton et al. 2013). For example, fixed low observation error produces an estimate of higher process error, while fixed high observation error produces an estimate of lower process error. Proportion choice of observation and process error alters the predictor variables included in MARSS models as illustrated by the difference in the MARSS-P and MARSS-M 10-fold CV model selection results. The MARSS-P models' choice of covariates generally align more with the covariates included in other models, compared to those included in the MARSS-M model. This consistency validates the MARSS-P

model's specification of environmental predictors. In addition, the MARSS-P model more accurately interpolated the nekton density and aggregation data compared to the MARSS-M model, and the low measurement error in the MARSS-P model aligns with *a priori* knowledge of error associated with stationary acoustic data collection. Predictor variables in the MARSS-P model are deemed a more reliable characterization of both baseline nekton datasets compared to the MARSS-M model.

The deterministic parametric and semi-parametric models are not advantageous for interpolating nekton baseline data, and differences in the selection of predictor variables further illustrate effects of model assumptions and structure on baseline characterization. For instance, GAM(M)s varied in their environmental predictors of both nekton density and aggregation due to the structure of smoother splines. GAMs are known to over-fit nonlinear smoother splines in the presence of autocorrelation, while GAMMs are known to have difficulty converging while estimating both autocorrelation and smoother splines (Wood 2006, 2015). Therefore, the GAM may over-estimate variable parameters, or include irrelevant predictors, while the GAMM may converge on a simpler model with few predictors. Distributional assumptions will also impact environmental predictors included in parameterized models. Gamma distributed GLM(M)s both include day as a predictor of nekton aggregation, whereas linear regression and GLS only include the 24 hour Fourier series as a predictor variable. The difference in distributional assumptions between these models consistently impacted the inclusion of day as a predictor of nekton aggregation, regardless of other differences in model structure. Effects of model structure and assumptions illustrate the importance of using a model that is consistent with properties of the data to ensure accurate data characterization (Barry and Elith 2006).

2.4.2 Implications for Operational Monitoring Programs

The model evaluation demonstrated the effect of model choice and parameterization on the characterization of baseline data. Comparisons of models used to monitor change in biological indicators have been repeatedly conducted in ecology (e.g., Thomas 1996, Ward et al. 2014, Bell and Schlaepfer 2016), but have not been used to recommend models for environmental baseline characterization in Marine Renewable Energy monitoring programs. The choice of model has been shown to alter data characterization (e.g., Thomas 1996, Jones-Farrand et al. 2011).This study reinforces these findings, and emphasizes that the choice of a state-space model to characterize MRE baseline data will contribute to an effective design of MRE monitoring programs.

Accurate estimation of baseline variability is critical to designing a monitoring program that has the power to detect change outside the natural range of variability (Klure et al. 2012, McCann 2012). The baseline estimate of variability can be used in a power analysis to calculate the sample size needed detect a predetermined size of change in operational monitoring (Carey and Keough 2002). The accurate interpolation of nekton density and aggregation index data by the nonparametric and state-space models emphasizes the need to use flexible models to accurately predict nekton data variability. Partitioning of residual variability as process and observation error in a state-space model provides an additional assessment of variability that can be used to formulate sampling designs for operational MRE monitoring programs. To illustrate by example, if estimated process error in a state-space model is greater than observation error, then this indicates a need to increase the number of samples, relative to baseline sampling, to improve precision of model fit (e.g., See and Holmes 2015). If the estimate of process error is less than observation error, fewer samples are needed to reach the same target precision of model

fit, which reduces monitoring costs (e.g., See and Holmes 2015). Identifying the need for fewer samples is especially valuable when using traditional sampling techniques (e.g., trawl or trapping surveys) that are associated with greater monitoring costs than remotely sensed, acoustic technology used to collect the case study data. Using a state-space model to characterize baseline data also provides an additional tool when designing operational monitoring programs to measure environmental change caused by MRE development.

Model structure and assumptions influenced the selection of environmental predictors of nekton density and aggregation index data, which also influence the design of MRE operational monitoring programs. Environmental predictors in baseline models and their relationship to the dependent variable are used to understand and identify potential effects of development on the ecosystem (Treweek 1996). Perceived effects of MRE development are then used to design monitoring programs for detecting change from baseline conditions, including sampling resolutions and data collection methods (Klure et al. 2012, McCann 2012, Boehlert et al. 2013). Subsequently, these monitoring programs are used to detect change and inform managers and regulators how to minimize and mitigate harmful effects. As an example, the inclusion of day as a predictor of nekton density and/or aggregation infers a trend across time. By not including day in the baseline model, change may be falsely detected (positive trend) after MRE development. Conversely, if day is included in a baseline model, but its importance is overestimated, then change may be missed (negative trend) after MRE development. Given that the form of predictor variables can influence management decisions (Nuno et al. 2014), a linear trend may infer a constant increase or decrease in the measured indicator, compared to that from an asymptotic, nonlinear trend that infers a possible stabilization over time (Nuno et al. 2014). These two forms

of trend in MRE baseline data will generate different conclusions on the effects of MRE development.

The effect of predictor variables on the design of monitoring programs and subsequent application of monitoring data for management decisions emphasizes the need for a baseline model with predictor variables that provide an interpretable characterization of baseline data. Parametric models are often used in MRE monitoring to provide quantitative measures of the size and shape of predictor variables, and uncertainty around those estimates (Maclean et al. 2014). State-space models consistently produced the most accurate parametric interpolation of nekton data. Therefore, the use of state-space models to characterize nekton data would result in the most accurate and comprehensible assessment of environmental predictors of baseline data, which should lead to the most effective design of operational monitoring programs.

2.4.3 MRE Monitoring Model Recommendations

Recommendations for characterizing baseline MRE monitoring data (Fig. 2.3) are derived from the model evaluation. Criteria used to recommend models include results from the 10-fold CV and residual diagnostics. Advantages or disadvantages of models to provide interpretable and reliable measures of baseline predictor variables were also used to develop model recommendations. As nekton density data was chosen as a representative metric of normally distributed data, the recommended models for characterizing baseline nekton density (i.e., Sv) are identical to those used to characterize Center of Mass and Inertia baseline data.

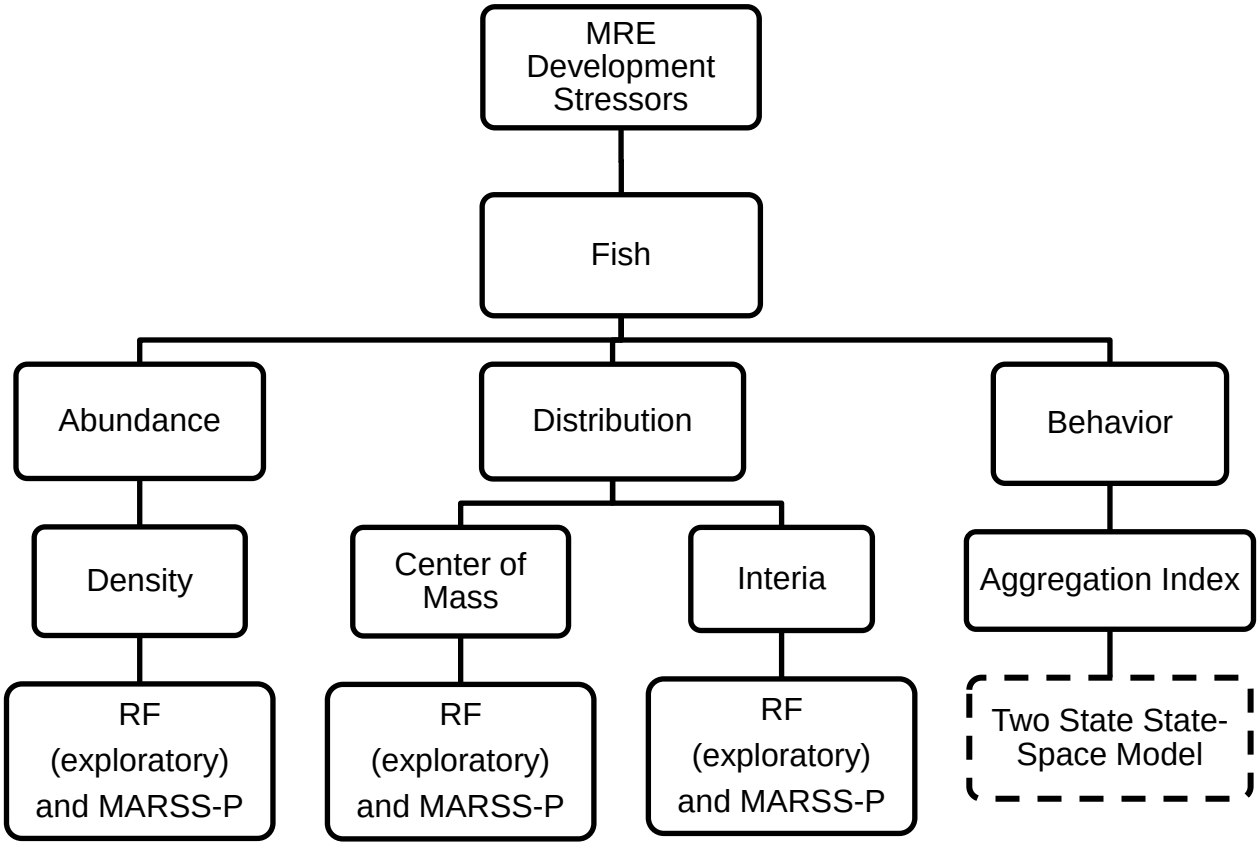

Figure 2.3. Schematic of recommended models to characterize environmental receptor indicators. An example framework for MRE monitoring consists of a stressor (i.e., MRE development), a receptor (e.g., fish), indicators of the state of the receptor (i.e., abundance, distribution, behavior), and metrics representative of the indicator (i.e., Sv (unit: dB re 1 m-1), center of mass (units: M), inertia (units: m^2), and aggregation index (units: m^{-1})). The dashed box indicates the need for further evaluation to identify a two state, state-space model for characterization of aggregation index data.

Normally Distributed Data

The Random Forest regression is recommended as a preliminary model to investigate potential autocorrelation in normally distributed data, and to identify predictor variables for normally distributed, baseline data characterization (Fig. 2.3). The RF model was the second best

interpolator of normally distributed nekton data after SVR-RBF. The inclusion of 14 lagged dependent variables in the RF model captured the 24 hour autocorrelation in the data. The RF model is also able to quantify the importance of all environmental predictors used to characterize the data. The RF model provides the most accurate and complete initial assessment of autocorrelation and relevant environmental predictors of normally distributed metrics. It does not provide explicit estimates of model parameters, or partition observation and process error in the data.

The parametric, state-space MARSS-P model is recommended for baseline characterization of normally distributed nekton data after the application of the RF model (Fig. 2.3). The MARSS-P model was the best interpolator of the data among parametric/semi-parametric models. The parametric estimates of process error, measurement error, density-dependence, and predictor variables provide an interpretable assessment of all primary components of baseline nekton characterization. Environmental predictors in the MARSS-P model were consistent with predictors included in the majority of evaluated models, and have been shown to impact nekton densities and distributions (e.g. Urmy et al. 2012, Wiesebron et al. 2016). Even though there was a 24 hour autocorrelation in the residuals of the MARSS-P model, it did not affect the 10-fold cross-validation RMSE measurement or the resulting recommendation. The MARSS model structure is flexible and adjustable, and may be altered to include a 24 hour lag in the data to remove the observed autocorrelation (Hampton et al. 2013).

Non-normally Distributed Data

No model fully characterized the nekton aggregation data, based on residual diagnostics and 10-fold CV results. Of all candidates, the MARSS-P model best characterized non-normal index data. The MARSS-P model was the second best interpolator of aggregation

index data, after SVR-RBF. The MARSS-P model was also able to characterize spikes in the aggregation data as process error, which is more biologically accurate than the assumption of spikes as observation error as suggested in all other parametric/semi-parametric models. It is important to note that spikes in nekton aggregation may not be random variation, but a periodic pattern driven by diel vertical migration (Urmy et al. 2012). A state-space model that includes diel vertical migration as a covariate may estimate nekton aggregation index data more accurately than a model without this variable, as omission of relevant predictor variables will bias estimated values (Barry and Elith 2006). The MARSS model did have heteroskedastic errors, because it assumes a normal distribution of the data, and as a result, did not fully characterize the distribution of empirical data. Residual diagnostics suggested that a Box-Cox power transformation (Box and Cox 1964) may be necessary to better fit the assumptions of a normal, state-space model.

The model evaluation may not have identified a candidate model that accurately captured all properties of the non-normal baseline data, but it did highlight advantages of using a state-space model to characterize the data. A non-normal, state-space model may provide a more accurate interpolation of data and reduce heteroskedasticity in the residuals. It may also be necessary to fit a state-space model that characterizes spikes in aggregation index data as more than process error. As an example, a Markovian switching, state-space model is capable of modeling AI data as a two-state structure, with the probability of being in the low or high state dependent on the state at the previous time-step (Ghahramani and Hinton 2000). This two-state approach may be a more appropriate way to characterize short period, variable data.

Note on SVR models

The SVR models were not recommended for baseline characterization of either the normal or non-normal data regardless of their performance in the model selection process. Although SVR models may be able to accurately interpolate baseline data, the lack of methods to identify predictor variable importance restricts their use as descriptive models for MRE baseline characterization. Choice of predictor variables in SVR models was also dependent on kernel choice and parameter tuning. A SVR model cannot be used as an exploratory or descriptive model, but would be effective as a predictive model of nekton density and possibly nekton aggregation after a Box-Cox data transformation.

2.5 Conclusion

This study was motivated by the absence of an evaluation of models capable of characterizing Marine Renewable Energy baseline environmental monitoring data. Including all primary classes of regression models provided a palette of candidate models that could be used to characterize normal and non-normal baseline data. The recommended Random Forest and MARRS state-space models have not been commonly used in MRE monitoring studies. Recommended models and the approach used to identify them may be combined as a best practice for the analysis of monitoring data with similar properties as those from the case study. For example, the RF and MARSS model combination recommended for characterizing nekton density is applicable for any temporally continuous, normally distributed data. The combination of RF and MARSS models can also be used to standardize monitoring protocols (Fig. 2.3). The standardization of MRE monitoring, including the choice of analytic model, will reduce cost and uncertainty in MRE permitting in the United States (Dubbs et al. 2013) or consenting in the United Kingdom. The use of recommended models also ensures accurate characterization of

MRE baseline environmental monitoring data. Accurate baseline assessments will lead to the development of effective operational MRE monitoring programs.

The data used as a case study was assumed representative of MRE site baseline data. Models used in the evaluation were assumed to be representative of regression models commonly used in ecological monitoring studies. Mechanistic or Bayesian models were not used to characterize baseline data, but could be evaluated using the same approach. Evaluation of candidate models for characterizing non-normal AI data led to a better understanding of data structure and highlighted the inability of models to characterize short period, large amplitude fluctuations. This insight may be extended by adding diel vertical migration predictors as potential model covariates, using a transformation of the AI data to a normal distribution, or additionally evaluating state-space models that better match the structure of spiky aggregation index data.

This study explicitly developed and applied an evaluation to recommend statistical models capable of characterizing baseline monitoring data. There is an additional need to accurately detect change relative to baseline conditions during MRE operations. To guarantee consistent and comparable results in baseline and operational monitoring, the same techniques should be used during both phases of MRE site development. Therefore, models used for baseline characterization must also be able to detect and forecast change in monitoring variables. An additional evaluation of model ability to detect change is needed to complete a best practices procedure for analyzing Marine Renewable Energy environmental monitoring data.

Chapter 3:

Evaluating Statistical Models to Measure Environmental Change

3.1 Introduction

The need to detect and measure change is ubiquitous in all environmental monitoring to prevent or mitigate detrimental effects on natural resources (e.g., Ferretti 1997, Lovett et al. 2007, Ingersoll et al. 2013). Monitoring programs are used to measure change in the environment, either natural or anthropogenic, and to inform environmental management decisions (Legg and Nagy 2006). Criteria specified to define change and to estimate the size and shape of change affect inferences made about the observed ecosystem (Underwood 1992, Nuno et al. 2014). Monitoring programs often fail to provide accurate and definitive measures of change due to high costs of sampling, analyzing data, time-constraints, and unclear objectives (Busch and Trexler 2003, Gitzen 2012, Lindenmayer et al. 2012). It is critical to define the goals of environmental monitoring and to develop best practice methods for achieving those goals to ensure efficient and accurate measures of change.

Primary objectives of environmental monitoring often include detecting, quantifying, and forecasting change. These objectives are developed from a hypothesis testing framework that is commonly used to structure environmental monitoring programs (Vos et al. 2000, Yoccoz et al. 2001, Nichols and Williams 2006). The hypothesis testing framework uses baseline (i.e., pre-disturbance) environmental studies to identify and assess perceived effects of future natural or anthropogenic disturbance on baseline conditions. Operational monitoring (i.e., post-disturbance

monitoring) is then used to evaluate hypothesized environmental change against the null hypothesis of no change (Fairweather 1991). Quantifying change is used to assess perceived effects of disturbance from baseline assessment to inform environmental management decisions about future monitoring and mitigation plans (Schmitt and Osenberg 1996, Hewitt et al. 2001, Shumchenia et al. 2012). A third objective of monitoring, forecasting, is used to develop future monitoring and mitigation efforts and to provide early warning detection of data reaching a threshold, or predefined level, of change (Magurran et al. 2010, Dornelas et al. 2012, Lindenmayer et al. 2012).

Marine Renewable Energy (MRE) development provides a timely case study of anthropogenic disturbance and exemplifies the need for efficient and effective environmental monitoring methods. Although MRE technology is rapidly expanding, projects within the United States are primarily in demonstration stages rather than full-scale commercial operations. This lag is attributed in part, to the sampling requirements for both baseline and operational monitoring. To comply with the National Environmental Policy Act (NEPA), an environmental assessment prior to project installation is mandatory to obtain permits for any MRE project (FERC 2008, DOE 2009, Portman 2010). Plans for post-installation, compliance monitoring in a Site Assessment Plan (SAP) or Construction and Operations Plan (COP) are commonly required for commercial leases of MRE sites (BOEM 2016). At this time, there are no specific monitoring requirements under these regulations, but the standardization of environmental monitoring protocols has been cited as critical for expediting the MRE permitting process and advancing MRE development (Dubbs et al. 2013, Copping et al. 2016).

To ensure comparable and accurate evaluation of environmental change, limited efforts attempted to develop standard monitoring guidelines for MRE monitoring design and data

collection (e.g., McCann 2012, Copping et al. 2014). These guidelines use the stressor-receptor construct previously defined in Chapter 2 to focus monitoring efforts on effects of concern (Boehlert and Gill 2010, Klure et al. 2012, Boehlert et al. 2013). Predominant MRE environmental stressors include: device presence (i.e., static effects), operational device movement (i.e., dynamic effects), chemical, noise, energy removal, and electromagnetic fields (EMF) (cf., McCann 2012, Klure et al. 2012, Boehlert et al. 2013). Hypothesized effects of these MRE stressors on biological receptors (e.g., fish, marine mammals, birds) are used to structure the design of monitoring programs. The Before-After-Control-Impact (BACI) design is commonly suggested for MRE monitoring studies (McCann 2012, Copping et al. 2014). A BACI design includes sampling a site before and after installation and/or using separate control and impact sites to identify change in the environment. It may be difficult to find control sites that match the characteristics of impact sites due to high environmental variability at MRE device sites (Polagye et al. 2011, Copping et al. 2014). An alternate approach is to use a Before-After monitoring design (McCann 2012). McCann (2012) is the only monitoring guideline document to provide examples of possible statistical methods to measure change caused by stressors, but these recommendations are based on previous applications and do not explicitly compare or evaluate statistical methods.

Previous statistical models used to measure change in MRE Before-After studies vary based on objectives and data properties. To detect change, hypothesis tests (i.e., ANOVA, Mann-Whitney U test) (e.g., Hammar et al. 2013) or significance of change indicator variables in parametric (e.g., Bergström et al. 2013) or semi-parametric regression models (e.g., Peterson et al. 2013) have been used. Parametric and semi-parametric regression models used to characterize trends and patterns in MRE monitoring studies include: linear regression (e.g.,

Hammar et al. 2013, ORPC 2014), Generalized Linear (Mixed) Models (GLM(M)s) (e.g., Embling et al. 2013, Stenberg et al. 2015), and Generalized Additive (Mixed) Models (GAM(M)s) (e.g., Tollit et al. 2013). These models are also used to predict data patterns in MRE Before-After monitoring studies (e.g., Mackenzie et al. 2013, Waggitt et al. 2014, Warwick-Evans et al. 2016). The limited list of statistical models highlights the low diversity of parametric or semi-parametric regression models currently used in MRE monitoring studies. Mackenzie et al. (2013) is the only previous study that evaluated statistical models used to detect environmental change associated with MRE development. This study evaluated three semi-parametric regression models in three scenarios of Before-After step shifts in abundance and distribution of bird and cetacean data at a MRE site.

An evaluation of statistical model classes is needed to identify the most appropriate models capable of measuring change in MRE monitoring data. Objectives of this study are to develop an approach that can be used to evaluate candidate statistical models and to recommend best practices to accurately detect, quantify, and forecast change in environmental monitoring data. Empirical active acoustic data that is representative of common MRE monitoring data are used to develop a scenario analysis evaluation. This approach encompasses a wide range of forms and amplitudes of change in monitoring data caused by hypothesized effects of MRE development. Three classes of statistical models used to analyze monitoring data are evaluated. Results of the evaluation are used to recommend statistical model(s) that are able to detect, quantify, and forecast change in monitoring data. Recommendations from the evaluation can be used to develop best practices for measuring change in monitoring data. Standardizing best practice models shrinks monitoring costs, enables cross-site comparisons among MRE sites, and reduces uncertainty in effects of MRE development.

3.2 Methods

3.2.1 Approach

Models were evaluated using a series of change scenarios in simulated Before-After monitoring studies generated from empirical MRE baseline case study data. A range of amplitudes and relevant scenarios of lagged change were used to assess sensitivity of model performance. Models were parameterized using baseline data (cf., Chapter 2), and indicator variable(s) were added to the parameterized models to measure change in baseline conditions. The evaluation was developed to allow for an equal comparison across classes of parametric and nonparametric models, while allowing each model to have the greatest probability of success.

Candidate statistical models were evaluated on their ability to detect, quantify, and forecast change. A power analysis was used to estimate model ability to detect change. Model power to detect change was defined as inclusion of an indicator variable(s) producing a more accurate interpolation (i.e., prediction of data within the range of the empirical data) of change scenarios than the model without indicator variable(s). This binary definition of power to detect change aligns with traditional hypothesis testing used to detect change in monitoring data at a predetermined level of significance (Morrison 2007). Model ability to quantify change is defined as the accurate measure of size and shape of change, and was evaluated using both fit and forecast accuracy metrics. The use of fit and forecast accuracy metrics provides a comprehensive assessment of model behavior (Shmueli 2010) and ability to quantify change. A model fit to time series data may have high fit accuracy while misspecifying the deterministic and stochastic components of the time series data (Allen and Fildes 2001). Forecast estimates are based solely on parameterized mean model estimates, and consequently any change attributed to stochastic error can be evaluated by comparing forecast accuracy to fit accuracy. The model forecast

accuracy metric can also be used to assess a model's ability to forecast change. Results from the evaluation were then used to recommend model(s) capable of accurately detecting, quantifying, and forecasting change. All analysis was conducted in the R statistical software environment (R Core Development Team 2016).

3.2.2 Case Study Data

Empirical data from Admiralty Inlet, WA (see Chapter 2 for details) was used as case study data. The data was collected from a tidal turbine pilot project site considered by the Snohomish County Public Utility District No. 1, located ~1 kilometer off Admiralty head shoreline, Puget Sound Washington (48.18˚ N, -122.73˚ W), at a depth of ~60 meters (Public Utility District No. 1 of Snohomish County, 2012). This study uses acoustic backscatter data from a Simrad EK60 echosounder, which is are assumed representative of monitoring data throughout the life of an MRE project. Acoustic backscatter data serve as a proxy of nekton (i.e., macro-invertebrates and fish that move independently of fluid motion) within the water column (MacLennan et al. 2002).

A suite of metrics derived from the data were used to quantify nekton density and vertical distribution (Burgos and Horne 2007, Urmy et al. 2012). Two acoustic backscatter metrics were used to represent MRE monitoring data: mean volume backscattering strength (Sv) (dB re 1 m^{-1}) and aggregation index (m^{-1}). Sv is a proxy for nekton density and the aggregation index (AI) measures patchiness of nekton. Both metrics are continuous, display periodic autocorrelation (Jacques 2014), and are trend-stationary (i.e., statistical data properties are constant over time, assuming that the periodicity and trend in the data are associated with deterministic environmental variables). Sv data are considered representative of all normally distributed metrics (cf., Chapter 2). Aggregation index (AI) values are right-skewed, non-normally distributed data composed primarily of low values with high amplitude spikes. The terms *low*

state and *high state* will be used to refer to the two amplitudes of aggregation index values. Characteristics of these two metrics represent data properties common to all monitoring data (i.e., temporally continuous, autoregressive, normal and non-normal data) (Chandler and Scott 2011, Gitzen 2012).

3.2.3 Baseline Simulation Models

A Multivariate-Autoregressive-State-Space (MARSS) model was used to simulate baseline nekton density data, representative of normally distributed MRE monitoring data. The MARSS model adheres to the characteristics of nekton density data because it assumes a normally distributed error structure, and can include an autoregressive lagged dependent process variable to estimate mean-reversion (i.e., how strongly a stationary time-series data reverts to its mean value). The model is also able to estimate both process and observation error that are components of time series data (Pattengill-Semmens et al. 2011, Dornelas et al. 2012).The MARSS model with fixed low measurement error (i.e., MARSS-P) recommended for characterizing normal MRE baseline monitoring data in Chapter 2 was used to simulate the normal data in the current study.

A Hidden-Markov Model (HMM) was used to simulate baseline nekton aggregation index data, representative of non-normally distributed MRE monitoring data. An HMM assumes that there is an underlying process that governs the observed state, where the state at time *t* depends on the state at time *t-1* (Zucchini and MacDonald 2009). The model estimates the probability of the data being within any given number of states at a specified time. This model is applicable to the aggregation data, which contains aperiodic low and high states of aggregation index data.

Normal Data Baseline Simulation Model Specification

MARSS-P was structured using a fixed measurement error of 10% of the total error, and process error was estimated as a parameter. We assumed that observation error in linear backscatter data collected from a stationary echosounder results from calibration and hydrographic conditions. These sources of error have been suggested to each equate to a maximum of 5% of the total error (Simmonds and MacLennan 2005). Within a MARSS model the B parameter estimates mean-reversion: B=1 indicates a non-stationary random walk and B <1 indicates mean-reversion in the data (Holmes et al. 2014). No MARSS model parameters, including B, were fixed prior to parameterizing the model. The MARSS simulation model was fit to the empirical Sv data using the R statistical package "MARSS" (version 3.9).

Non-Normal Baseline Data Simulation Model Specification

The HMM model assumed a log-link Gamma distribution for the nekton aggregation index (AI) data. The HMM statistical package used to fit the AI data requires the input of observed values, Gamma distribution parameters for each data state, chosen covariates within each state, and a transition intensity matrix (Q) to estimate the expected AI data values. Q estimates the rate of transitioning from states *i* to *j*, and is used to calculate the transition probability matrix that is needed to estimate the expected data state at time t:

$$P(t) = \exp(tQ) \tag{3.1}$$

where Q is the transition intensity matrix at time t, and P is the probability transition matrix.

The expected state of data at time t, either low-state or high-state aggregation, was estimated by applying the Viterbi algorithm (Viterbi 1967) to the parameterized model. The expected states

and the linear, parametric, Gamma distributed parameters and covariates within each state were used to simulate the expected AI values. Detailed information on model parameters and structure can be found in Cappé et al. (2005), Zucchini and MacDonald (2009), and Jackson (2016). The HMM simulation model was fit to the empirical AI data using the R statistical package "msm" (version 1.6).

The HMM R statistical package requires initial estimates of the low and high state Gamma distribution parameters (i.e., rate and shape) to estimate expected aggregation index values. A visual inspection was used to initially separate low and high state values using the maximum aggregation index value within the low state of the data. The R statistical package "fitdistr" (version 7.3-45) was used to estimate initial Gamma parameters of identified low and high state AI data. The "msm" package also requires a preliminary estimate of the Q matrix. Random numbers were generated to fit the requirements of the matrix (i.e., the rows must sum to zero, and the diagonal elements are the negative sum of the other elements in a row) (Jackson 2016). An HMM model including all environmental predictors of the AI data within each aggregation state, the randomly generated q matrix, and estimated Gamma distribution parameters was fit to the empirical data to produce an initial estimate of model parameters and the two states of AI data. Results from the initial HMM model indicated that the maximum AI value within the lower state of aggregation was ~0.06. This value was used to separate the high and low state of AI data and re-estimate the Gamma parameters of both states. The re-estimated Gamma parameters, and estimated q matrix, were pre-specified in the HMM model selection to produce the most accurate parameterized simulation model.

3.2.4. Simulation Model Selection

The same ancillary environmental covariates (i.e., daily tidal range (m), tidal speed (m/s), Julian day of year, and a 24 hour Fourier series) used to structure models in Chapter 2 were used to develop 24 versions of each simulation model that were evaluated to select the optimal parameterized structure of the simulation models. Akaike Information Criterion (AIC) was used to select the most accurate version of the simulation models. AIC is a commonly recommended model selection tool that can efficiently select the most parsimonious MARSS and HMM model (Anderson and Burnham 2002, Johnson and Omland 2004). AIC is approximately equivalent to leave-one-out cross-validation model selection (i.e., a single data point is used as the test set repeatedly until all data points have been used). The MARSS model allows for covariates in both the process and observation equation. Combinations of covariates may equally affect both the process and observation of the response. Similarly, the HMM model contains separate equations to estimate expected values in the two states of aggregation. Every combination of the 24 model versions in these two-part models was evaluated in the model selection process to produce a total of 576 model versions. The model version with the lowest AIC and complete convergence was chosen as the simulation model (Table 3.1).

Table 3.1. Parameterized normal data simulation model (MARSS) and non-normal data simulation model (HMM), including selected environmental predictors and error distributions.

Model	**Environmental Predictors**	**Error Distribution**
MARSS-P	Process Eq: Day, Tidal Range, Day:Tidal Range Observation Eq: Fourier Series	Normal
HMM	Low State: Day, Fourier Series, Tidal Range, Tidal Speed High State: Day, Fourier Series	Gamma (log)

3.2.5 Change Scenarios

Change scenarios were developed based on the perceived effects of MRE development, but the scenarios are representative of change that may occur in monitoring data from any natural or anthropogenic disturbance. A literature review resulted in 5 change scenarios in monitoring data hypothesized from effects of primary MRE stressors on nekton indicators: a step shift in mean from device noise (e.g., Inger et al. 2009, Boehlert and Gill 2010), a change in variance resulting from a shift in data caused by dynamic device movement (e.g., Polagye 2011, Klure et al. 2012, McCann 2012), a linear or nonlinear gradual change in mean from the presence of a device acting as a Fish Aggregation Device (FAD) (e.g., Klure et al. 2012, Shields and Payne 2014), and a step reduction in mean plus nonlinear change back to baseline conditions from a chemical spill (e.g., Polagye et al. 2011, Boehlert et al. 2013, Fodrie et al. 2014) (Table 3.2). The specific forms of change hypothesized from effects of MRE stressors were structured using BACI and Intervention Analysis (Box and Tiao 1975) literature that indicate change may occur in the mean and/or variance of a metric, and may be a step-change, linear, or nonlinear (Box and Tiao 1975, Underwood 1994, and Scheiner and Gurevitch 2001).

Table 3.2. Change scenarios modeled in the "After" portion of "Before-After" simulated nekton density and aggregation index data. Change scenarios were developed from hypothesized effects of MRE development stressors on the fish receptor. The predicted effect on fish density and aggregation data, shape of change, the changed statistical property of the data, and change function are identified for each change scenario.

Stressor	**Predicted Effect on Fish**	**Shape of Change**	**Changed Statistical Property**	**Change Function**
Noise	Decrease	Step shift	Mean	$-\mu\delta$
Device Operation Movement	Variable increase and decrease	Periodic shift	Variance	$+/-\ \mu X$
Device Presence (FAD)	Increase	Linear trend	Mean	$\left(\frac{2\mu\delta}{n-1}\right) t_{i-1}$
Device Presence (FAD)	Increase	Nonlinear trend	Mean	$\frac{2\mu\delta}{1+4450e^{-0.05t_{i-1}}}$
Chemical Spill	Decrease	Step-change + nonlinear return to baseline (i.e., Abrupt trend)	Mean	$\left(\frac{1}{1+4450e^{-0.05t_{i-1}}}-1\right) 2\mu\delta$

Notes: μ is the mean of simulated "before" data, δ is the amplitude of change in the mean, t is time-point in the data series, X is a proportional value of μ. The logistic function used to simulate nonlinear FAD and chemical spill scenarios was structured to have an inflection point at 50% of the total mean change at t =168. To simulate the MRE device movement change, the tidal speed was found such that there was a 50:50 divide of data values corresponding to a speed less than or more than 1.07 m/s. The mean (μ) multiplied by an arbitrarily chosen proportion of the mean (X) was added to the data values corresponding to a lower tidal speed, and subtracted from the data values corresponding to a higher tidal speed to produce a change in variance rather than mean.

Change Attributes- Amplitude and Lag of Change

Attributes of change (i.e., shape, size, and lag) were varied in the change scenarios to evaluate the sensitivity of a model's ability to measure change. Previous MRE environmental monitoring studies have simulated varying levels of data attributes (e.g., sample size, effect size, and study duration) to quantify the sensitivity in the power of a study design to detect change in data (e.g., Terrill et al. 2009, Maclean et al. 2013, Vanermen et al. 2015). This approach was incorporated into the current study by altering the amplitude (i.e., size) of change within each of the 5 change scenarios, and including a lag in the onset of change in the FAD change scenarios.

Current MRE monitoring guidelines do not specify sizes of change that are significant, because there are no vetted biological thresholds available. Consequently, a range of amplitudes was chosen that included: 10% of the mean, 25% of the mean, and a mean outside the 5-95th percentile (Munkittrick et al. 2009). For the change in variance scenario, a change of 10%, 25% and 2 times the standard deviation were used.

A time lag in the onset of change was incorporated into the evaluation, because it is critical to be able to detect change that may not be immediately apparent during post-development MRE monitoring (McCann 2012). The potential for a lag in the onset of change is most probable in the FAD scenario, which may not cause an immediate change in fish presence post- MRE installation. The increasing presence of fish around a FAD is not predicted to be immediate as it may take several weeks for fish to appear around a FAD after placement (Matsumoto 1981, FAO 2016). New habitat and algae growth created by a FAD attract smaller fish which in turn attract predators (Barnabe and Barnabe-Quet 2000, Spellman 2015). A lag of 2 weeks in the onset of change was applied to the Fish Aggregation Device change scenarios. To

explore sensitivity to a lagged change, scenarios were only simulated in the 10% and 95% percentile linear and nonlinear scenarios of change.

Simulating Change Data

A function was added to simulated baseline data to create the desired shape and size of change in the "after" (i.e., post-installation) data (Table 3.2). Initially, baseline simulation models were used to generate the same baseline data in the "before" and "after" datasets. Change was then induced by adding an external function to the "after" dataset to simulate an appropriate size and shape of change in the baseline data (Table 3.2). Parameterized covariates and error distributions of each baseline simulation model were kept constant while only the added function produced the change in baseline conditions (e.g., Benedetti-Cecchi 2001, Mackenzie et al. 2013, Vanermen et al. 2015). Changes to the simulated baseline nekton density data were applied to the linear units of acoustic backscatter data and then the data was retransformed to logarithmic (dB) units. "After" data was generated using a different random seed for error distribution than the "before" data to simulate natural variability between the two monitoring periods. Change in lag scenarios was simulated in the same manner, except that the change was only added to the second half of the "after" data. "Before" and "after" simulated datasets each consisted of 336 data points, matching the empirical baseline dataset.

3.2.6 Candidate Models

The same suite of candidate models used in the Chapter 2 baseline model evaluation study were evaluated on their ability to measure change in nekton density and aggregation index values using simulated Before-After datasets (Table 3.3). Since the primary goal of environmental monitoring is to track and measure change relative to baseline conditions (Shumchenia et al.

2012), it is important to use the same methods in baseline and operational monitoring to quantify change (McCann 2012). Evaluation of candidate model abilities to measure change will be used to construct a best practice approach for analyzing data for any and all stages of a monitoring project. All parameterized models from the baseline evaluation study (Tables 3.4 and 3.5) were evaluated on their ability to detect, quantify, and forecast change regardless of model caveats discovered in the previous study, to produce a complete understanding of model behavior.

The MARSS-P candidate model is also used as the normal data baseline simulation model (cf., Chapter 2). Any bias in the MARSS evaluation was minimized by using AIC, because it is a different model selection metric than the metric used to parameterize the candidate models (cf., Chapter 2). An HMM model was not included as a candidate model in the current evaluation to reduce any bias produced from using a HMM as a simulation model. The HMM and MARSS models are both members of the state-space model class, and therefore the MARSS evaluation results indicate possible advantages or disadvantages of using a HMM for measuring change in MRE monitoring data.

Table 3.3. Description of evaluated candidate regression models including model class, linear or nonlinear form, parametric or nonparametric structure, error components (observation and/or process error) and distribution, and autocorrelation structure.

Model	**Class**	**Form**	**Parametric/ Nonparametric**	**Error Components**	**Error Distribution**	**Auto-correlation Structure**
Linear	GR	Linear	Parametric	Observation error	Normal	None
Generalized least squares (GLS)	GR	Linear	Parametric	Observation error	Normal	Residual correlation
Generalized linear model (GLM)	GR	Linear	Parametric	Observation error	Gamma (identity)	None
Generalized linear mixed model (GLMM)	GR	Linear	Parametric	Observation error	Gamma (identity)	Residual correlation
Generalized additive model (GAM)	GR	Nonlinear	Semi-parametric	Observation error	Gamma (identity)	None
Generalized additive mixed model (GAMM)	GR	Nonlinear	Semi-parametric	Observation error	Gamma (identity)	Residual correlation
Multivariate autoregressive state-space model (MARSS)	Time series	Linear	Parametric	Process and Observation error	Normal	AR-1 lagged variable
Regression - autoregressive moving average model (Reg-ARMA)	Time series	Linear	Parametric	Observation error	Normal	ARMA error
Regression - autoregressive moving average – generalized autoregressive conditional heteroscedasticity model (Reg-ARMA-GARCH)	Time series	Linear	Parametric	Observation error	Skewed-student-t	ARMA error; GARCH residual variance
Random forest (RF)	NP	Nonlinear	Nonparametric	N/A	None	Lagged variables
Support vector regression (SVR)	NP	Nonlinear	Nonparametric	N/A	None	Lagged variables

Note: The evaluated model classes are generalized regression (GR), time series, or nonparametric (NP) models.

Table 3.4. Candidate models parameterized using the empirical baseline nekton density (Sv) data with selected environmental predictors and autocorrelation structure for each candidate model. All models assumed a normal distribution.

Model	**Environmental Predictors**	**Autocorrelation Structure (AR,MA)**
LM	Day, Fourier Series, Tidal Range, Day: Tidal Range	NA
GLS	Day, Fourier Series, Tidal Range, Day: Tidal Range	(1,0)
GAM	Day, Fourier Series, Tidal Range†, Day:Tidal Range†	NA
GAMM	Day, Fourier Series, Tidal Range†	(1,0)
Reg-ARIMA	Day, Fourier Series, Tidal Range, Day: Tidal Range	(1,0)
Reg-ARMA-GARCH	Mean Eq: Day, Fourier Series, Tidal Range, Day: Tidal Range Variance Eq: Day, Fourier Series, Tidal Range,	ARMA:(1,0) GARCH:(2,3)
MARSS-M	Process Eq: Tidal Range, Tidal Speed, Tidal Speed-Tidal Range Observation Eq: Day, Fourier Series, Tidal Speed	(1,0)
MARSS-P	Process Eq: Day, Fourier Series, Tidal Range Observation Eq: Day, Tidal Range, Day-Tidal Range	(1,0)
RF	All Environmental Predictors	(14,0)
SVM-L	Fourier Series	(13,0)
SVM-RBF	Day, Fourier Series, Tidal Range, Tidal Speed	(1,0)

Note: The evaluation includes a MARSS model with fixed low observation error (MARSS-P) and fixed low process error (MARSS-M). A SVR model with a linear kernel (SVR-L) and a nonlinear Radial Basis Function kernel (SVR-RBF) were evaluated. The number of autoregressive (AR) and moving-average (MA) variables in model autocorrelation structures is shown in parenthesis as (AR, MA). The specified (AR, MA) structure of the nonparametric models indicates the number of lagged dependent variables included in the parameterized models. Environmental predictors are listed in alphabetical order of main effects followed by interactions. The Tidal Range† predictor is parametric in the GAM(M)s.

Table 3.5.Baseline candidate models parameterized using the empirical nekton aggregation (AI) data with selected environmental predictors, autocorrelation structure, and error distribution for each candidate model.

Model	**Environmental Predictors**	**Autocorrelation Structure (AR,MA)**	**Error Distribution**
LM	Fourier Series	NA	Normal
GLS	Fourier Series	(1,0)	Normal
GLM	Day, Fourier Series	NA	Gamma (identity)
GLM	Day, Tidal Range	(1,0)	Gamma (identity)
GAM	Fourier Series, Tidal Speed	NA	Gamma (identity)
GAMM	Fourier Series	(2,0)	Gamma (identity)
Reg-ARIMA	Fourier Series	(1,2)	Normal
Reg-ARMA-GARCH	Mean Eq: Day, Fourier Series, Tidal Range Variance Eq: Fourier Series	ARMA: (1,0) GARCH: (2,0)	Skewed-student-t
MARSS-M	Process Eq: Day Observation Eq: Day, Fourier Series	(1,0)	Normal
MARSS-P	Observation Eq: Fourier Series	(1,0)	Normal
RF	Fourier Series, Tidal Speed	(2,0)	NA
SVM-L	Day, Fourier Series, Tidal Range, Day: Tidal Range, Tidal Range: Fourier Series	(1,0)	NA
SVM-RBF	Day, Fourier Series, Tidal Speed, Day: Tidal Speed, Tidal Speed: Fourier Series	(13,0)	NA

Note: The evaluation includes a MARSS model with fixed low observation error (MARSS-P) and fixed low process error (MARSS-M). A SVR model with a linear kernel (SVR-L) and a nonlinear Radial Basis Function kernel (SVR-RBF) were evaluated. The number of autoregressive (AR) and moving-average (MA) variables in model autocorrelation structures is shown in parenthesis as (AR, MA). The specified (AR, MA) structure of the nonparametric models indicates the number of lagged dependent variables included in the parameterized models. The environmental predictors are listed in alphabetical order of main effects followed by interactions.

3.2.7 Quantifying Model Ability to Measure Change

Intervention Analysis is a standard method used to detect change post-intervention in time series data with known serial correlation (Stewart-Oaten and Bence 2001, Gilmour et al. 2006). The approach used in Intervention Analysis (IA) was used to structure each model to measure change in baseline conditions. In the original IA literature an ARIMA model is fit to baseline data, and then that same ARIMA model is fit to the entire dataset (pre- and post- intervention) with the addition of an indicator, or intervention function to model change in baseline conditions post-intervention (Biglan et al. 2000, Lagarde 2011). In the current study, an intervention function was added to the parameterized candidate models from Chapter 2 to measure change in the simulated Before-After datasets.

Intervention Model Specification

Candidate models were structured to model change using two intervention-functions: step and ramp (Lewis-Beck et al. 2004, Cryer and Chan 2008). The step function is an indicator variable consisting of 0's before an intervention and 1's after the intervention. The step function was added to all models to measure a step change in mean or variance (i.e., Step and Periodic change scenarios) in the simulated post-installation MRE monitoring data. A ramp function consists of 0's pre-installation, and values equivalent to an indexed data count of time post-installation of MRE devices: $t - n_{before}$; where n is equal to the number of data points before the MRE development. This intervention-function models a linear change in slope. A step-plus-ramp function was added to all models to measure linear, nonlinear, or abrupt change in the simulated post-installation MRE monitoring data. Given that the exact shape of the change would not be known *a priori*, it is most realistic to use the basic step and ramp intervention

functions to model change. The two specified intervention functions were added to each baseline parameterized candidate model to measure change:

$$y = a + BX + b_2 I_2 + b_3 I_3 \quad (3.2)$$

where BX is a matrix of all parameterized covariates used in the baseline model, a is an intercept, I_2 is a step function (an indicator variable equal to 0 before the MRE development, 1 after), and I_3 is a ramp function (an indicator variable equal to 0 before the MRE development, and an indexed data count $t - n_{before}$ after the intervention). The ramp function was only included in scenarios with a change in trend (i.e., FAD and chemical spill). The ramp function was structured as a smoother spline within the GAM(M)s.

The MARSS models required additional specification of the intervention variables within the process and observation equations. The step intervention was only included in the observation equation of the MARSS models to align with the structure of the observation-only candidate models. The ramp intervention was included in the MARSS models by altering the trend parameter (u) in the process equation. The u parameter was estimated before and after the intervention in the simulated monitoring data to allow for a change in the trend parameter.

3.2.8 Quasi-Power Analysis

A quasi-power analysis was used to evaluate model ability to detect change. Statistical power is defined as the probability of rejecting the null hypothesis when it is false. Power can be interpreted as the ability to accurately detect a difference between the null and alternative hypotheses (i.e., change in the data). Therefore, the idea of statistical power is vital to planning and interpreting environmental assessment studies (Osenberg et al. 1994). Power analysis can be

used to compare the probability of different models to detect change. Mackenzie et al. (2013) suggested the use of a power analysis to quantify the ability of statistical models to detect change in MRE monitoring data.

A 10-fold cross-validation (10-fold CV) model selection was used to conduct the power analysis. This method is equally applicable across all candidate models, and was used for model selection in Chapter 2. Model detection of change is commonly identified by using the statistical significance or confidence intervals of indicator variable(s) in the model (Schmitt and Osenberg 1996, Morrison 2007), but these methods were not used because the estimate of statistically significant change may be biased in models that do not account for autocorrelation (Schmitt and Osenberg 1996, Stewart-Oaten and Bence 2001), and evaluated nonparametric models do not include interpretable finite dimensional parameter estimates and confidence intervals equivalent to parametric models. In 10-fold cross-validation 10 equally sized random subsets of the data are used repeatedly such that 9 subsets compose a "training-set" and a single subset is the "test-set" to produce a total of 10 training and test datasets. The predicted values from the model applied to the test-set are used to calculate Root-Mean-Squared-Error (RMSE). RMSE is a measure of model accuracy based on the average deviance of predicted from observed values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

where y_i is the observed i^{th} value, $\hat{y}_i$ is the model predicted value, and n is the sample size.

A RMSE value closer to 0 indicates a more accurate interpolation of the data. The 10-fold CV model selection was applied to each candidate model with and without the indicator variable(s). If the IA version of a model produced a lower average RMSE than the baseline version of that model, then the inclusion of indicator variable(s) produced a more accurate interpolation of the

data. The form of the candidate model that accounted for change in post-installation monitoring data was selected as the more accurate model of the data. The selection of the IA form of a candidate model was defined as the model "detecting" change.

The 10-fold CV model selection process was applied to 1000 simulated Before-After datasets for every change scenario, and the proportion of the 1000 simulated datasets that each candidate model detected change was used to quantify the power of each model. The use of 1000 simulated datasets is validated by previous MRE environmental monitoring studies that used 1000 simulations or less to quantify study design statistical power (Benedetti-Cecchi 2001, Maclean 2013, Mackenzie et al. 2013, Vanermen et al. 2015). To ensure datasets were not identical, different random error seeds were used to create each dataset's error distribution.

3.2.9 Model Fit and Forecast Accuracy Metrics

Root-Mean-Squared Error (RMSE) and Mean-Absolute-Scaled Error (MASE) (Chai and Draxler 2014) were used to evaluate model ability to quantify and forecast change. The RMSE metric was used to quantify model fit to the Before-After monitoring data. The RMSE fit metric was calculated using the first 90% of the data. The remaining 10% of the data (72 data points) was used to estimate forecasting ability. Both RMSE and MASE were used to quantify model forecasting ability. The MASE metric is the average absolute difference between observed and model forecasted value at each time point scaled by the in-sample naïve forecast, which uses the observed value from point $t-1$ as the forecast of point t (Hyndman and Koehler 2006):

$$\text{MASE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\frac{1}{n-1}\sum_{i=2}^{n} |y_i - y_{i-1}|} \tag{3.4}$$

where y_i is the observed i^{th} value, $\hat{y}_i$ is the model predicted value, and n is the total sample size.

A MASE value <1 indicates a better average performance than the naïve in-sample forecast. The MASE metric tends to be less sensitive to outliers compared to RMSE (Ward et al. 2014). Using both metrics provides a more representative description of model performance (Chai and Draxler 2014). RMSE and MASE evaluation metrics were averaged from 30 simulated datasets of every change scenario, each using a different random error seed, to ensure robust evaluation metrics.

3.2.10 Ranking Methods

To identify models that most accurately quantify and forecast change, the models were ranked from most to least accurate using average fit and forecast metrics for each scenario. Power estimates were not explicitly ranked due to the patterns of similarity among models and across change scenarios. It may be misleading to suggest one model outperformed another based on rank position if their average performances are similar. Traditional statistical testing used to quantify performance similarity is not appropriate for this simulation study because p-values are affected by the number of simulations run (White et al. 2014). After an initial ranking, the Empirical Cumulative Distribution Function (ECDF) and associated Kolmogrov-Smirnov D-statistic of the 30 metric values produced for each scenario were used to group models that had similar accuracy. ECDF plots and D-statistics are common nonparametric tools used to assess the similarity of two probability distributions (Bennett et al. 2013).

D-statistic values equaled the maximum distance between each model's individual Empirical Cumulative Distribution Function (ECDF) and the cumulative ECDF of all other models. To illustrate by example, if the D-statistic of a model is +0.5 or -0.5, then the maximum distance between the ECDF of the given model and all other models is 50%, and the given model either has a greater (+) or lesser (-) cumulative probability of producing the estimated RMSE or

MASE value at the point of observed maximum distance. In the context of this study, a more positive D-statistic generally indicates a more accurate fit or forecast of a model relative to all other candidate models. D-statistic values were first inspected to identify similar performing models, and then followed by a visual inspection of ECDF plots to ensure robust and accurate grouping of similarly performing models. If the plot of a model's ECDF displayed greater or less probability of estimating a specified RMSE or MASE value than other models for >50% of 30 metric samples, then the model was ranked separately. After similarly performing models were grouped, all candidate models were re-ranked. Models were re-ranked using a dense ranking method (cf., Kellenberger and Groom 2015), which is defined as ranking grouped models equally using the value of the originally minimum ranked model in the group, and ranking the next models using the immediately following ranking number. The dense ranking method provides the most transparent quantitative values of relative accuracy of model fit and forecast of change in monitoring data.

3.3 Results

To standardize use of labels and to avoid ambiguity in subsequent text sections several terms are explicitly defined and used to group result patterns. The Linear, Nonlinear, and Abrupt change scenarios are referred to as "trended" change scenarios. The 95% amplitude of change is used to refer to both the 95% mean-change in all Step and trended scenarios, as well as the 2x standard deviation change (i.e., largest amplitude of variance-change) in all Periodic scenarios. A summary of the results is provided at the end of the *Results* section (Table 3.6). The ranked fit and forecast model results, and associated average evaluation metrics and D-statistic values, for all change scenarios of the normal and non-normal data are provided in Appendix A and B.

3.3.1 Power

Mean-Change

Normal Data (Sv)

Overall, all models reached ~100% power at the 95% amplitude of change for all mean-change scenarios in the normally distributed monitoring data.

Non-Lagged Dependent Variable (Non-LDV) Models

Models that did not include lagged dependent variables as mean predictors (i.e., Linear regression, GLS, ARIMA, GARCH, and GAM(M)) all produced a similar range of power values for all scenarios of change. All Non-LDV models had zero power for the Step no-change scenario, and low power for trended no-change scenarios (11.8% - 14.6%). The GAMM displayed the largest difference in results relative to all other Non-LDV models at the 10% amplitude of all mean-change scenarios (Fig. 3.1). This pattern continued for these scenarios in the 25% amplitude of change, but the difference in power was smaller. All of these models reached >96% power by the 25% amplitude of change for all mean change scenarios.

Lagged Dependent Variable (LDV) Models

Overall, models with lagged dependent variables (i.e., MARSS-M, MARSS-P, RF, SVM-L, and SVM-RBF) exhibited lower power than non-LDV models for the 10-25% mean-change scenarios (Fig. 3.1). The RF model had the greatest power of all LDV models. RF also had the greatest power of all candidate models in the 10% trended (≥ 70%) and no-change (63.6% for trended no-change and 52.7% for the Step no-change) scenarios. SVM-RBF and SVM-L had distinctly different power in the 10-25% scenarios of change, in which SVM-RBF had the lowest

power of all candidate models, and SVM-L had relatively high power compared to all other LDV models (Fig. 3.1). MARSS-P consistently had lower power in the 10% amplitude mean-change scenarios, but had higher power than MARSS-M in the 25% amplitude of mean-change scenarios. The LDV models generally all detected change in the Step no-change scenario, and displayed greater power in the trended no-change scenarios compared to the non-LDV models. MARSS-P has zero power for all no-change scenarios.

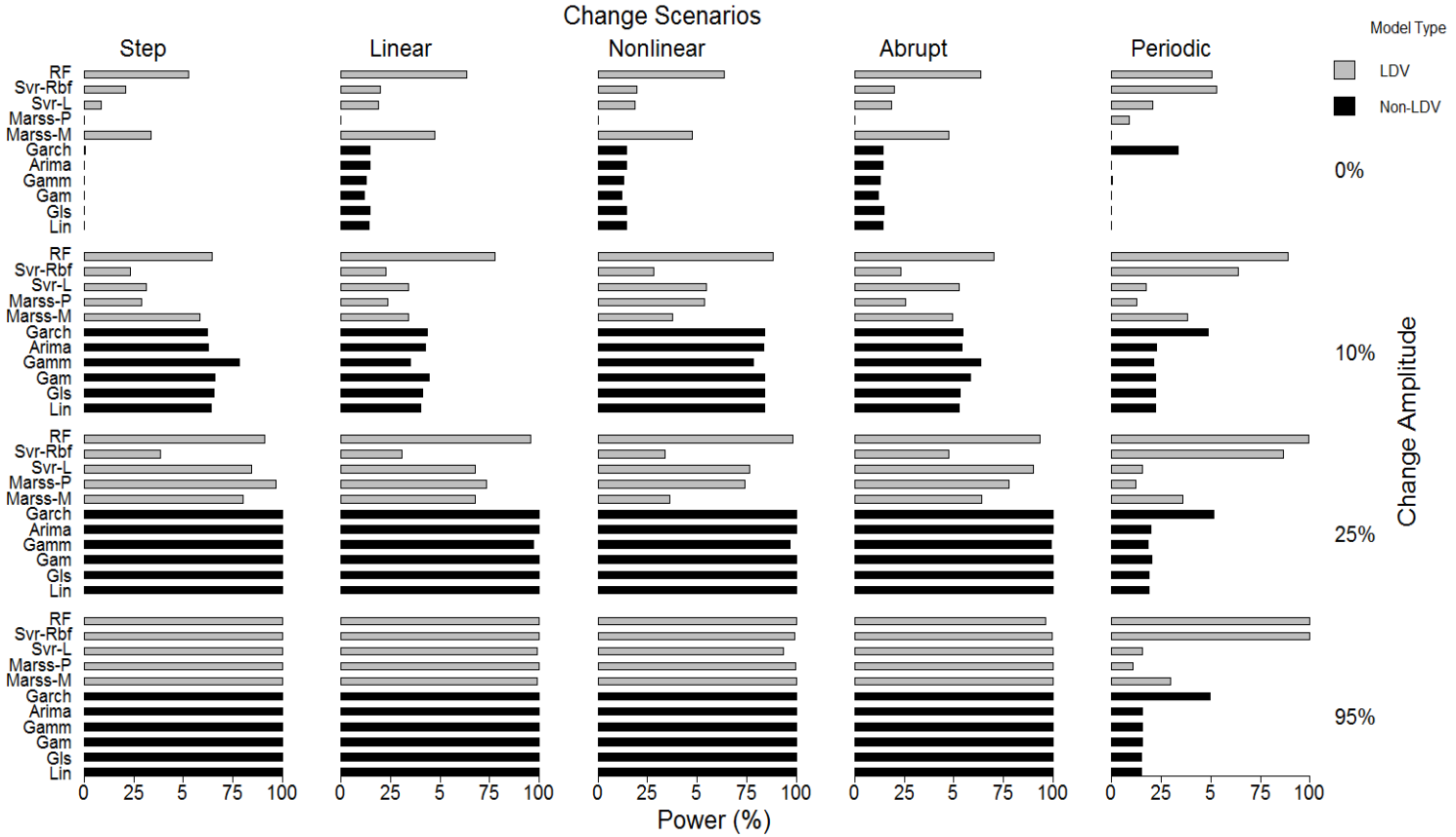

Figure 3.1. Power of candidate models to detect change in simulated normally distributed Before-After MRE monitoring data. Power is calculated across varying forms: Step, Linear, Nonlinear, Abrupt, and Periodic; and amplitudes 0%, 10%, 25%, 95%. Models that included lagged dependent variables (i.e., LDV models) are shown in grey, and the models without lagged dependent variables (i.e., non-LDV models) are shown in black.

Non-Normal Data (AI)

Overall, model power was higher for detecting change in non-normal data relative to normal data. Non-LDV, MARSS-P, and SVM-L models generally reached ~100% power at the 25% amplitude of change for all mean-change scenarios. The models did not display consistent patterns in power to detect change in non-normal data compared to patterns observed in normal data.

Non-LDV Models

All non-LDV models displayed ~100% power for all 25-95% mean-change scenarios, and had ~0% power in the Step no-change scenario, except for the reg-ARMA-GARCH model. The varying power results in the 0% and 10% scenarios of change were primarily influenced by the inclusion of the day predictor variable. The GLM(M)s, which included day as a covariate, had lower power in the 0% and 10% trended change scenarios, and greater power in the 10% Step change scenario relative to the other non-LDV models (Fig. 3.2). The GLM(M)s had ~2% power in the trended no-change scenarios, whereas the power of the other non-LDV models ranged from 13.7% (GAMM) to 40.9% (GAM and reg-ARMA-GARCH model). The linear regression, GLS, and reg-ARMA models displayed consistently similar power for all no-change scenarios. This group of models and the GAM(M)s all had similar power for all trended change scenarios (Fig. 3.2). The reg-ARMA-GARCH power results were not consistent with the other non-LDV models (Fig. 3.2). Similar to the GLM(M)s, the reg-ARMA-GARCH included the day covariate in both its mean and variance equation, but unlike the GLM(M)s reg-ARMA-GARCH and high power in the trended no-change scenarios (40.9%), and also had high power in the Step

no-change scenario (49.1%). The reg-ARMA-GARCH model power did not reach 100% for the Step 25-95% scenarios or Abrupt 25% scenario.

LDV Models

LDV models exhibited a wider range of power to detect change in non-normal data relative to the change observed in the normal data. SVM-L generally displayed ~100% power for all mean-change scenarios, whereas SVM-RBF had almost no power for all mean-change scenarios. The greatest SVM-RBF power was 33% in the Linear 95% scenario followed by 19% power in the Step 95% scenario (Fig. 3.2). RF also had little power for all mean-change scenarios, but had higher power in the Linear and Nonlinear 95% scenarios (94.6% and 96.7%) (Fig. 3.2). The MARSS models had similar power as all non-LDV models in the 25-95% amplitudes of mean-change, and power similar to the non-LDV models with the lowest power in the 10% mean-change scenarios. The MARSS-P model had ~100% power in all 25% scenarios, whereas MARSS-M did not reach 100% for any mean-change scenario until 95% amplitude of change.

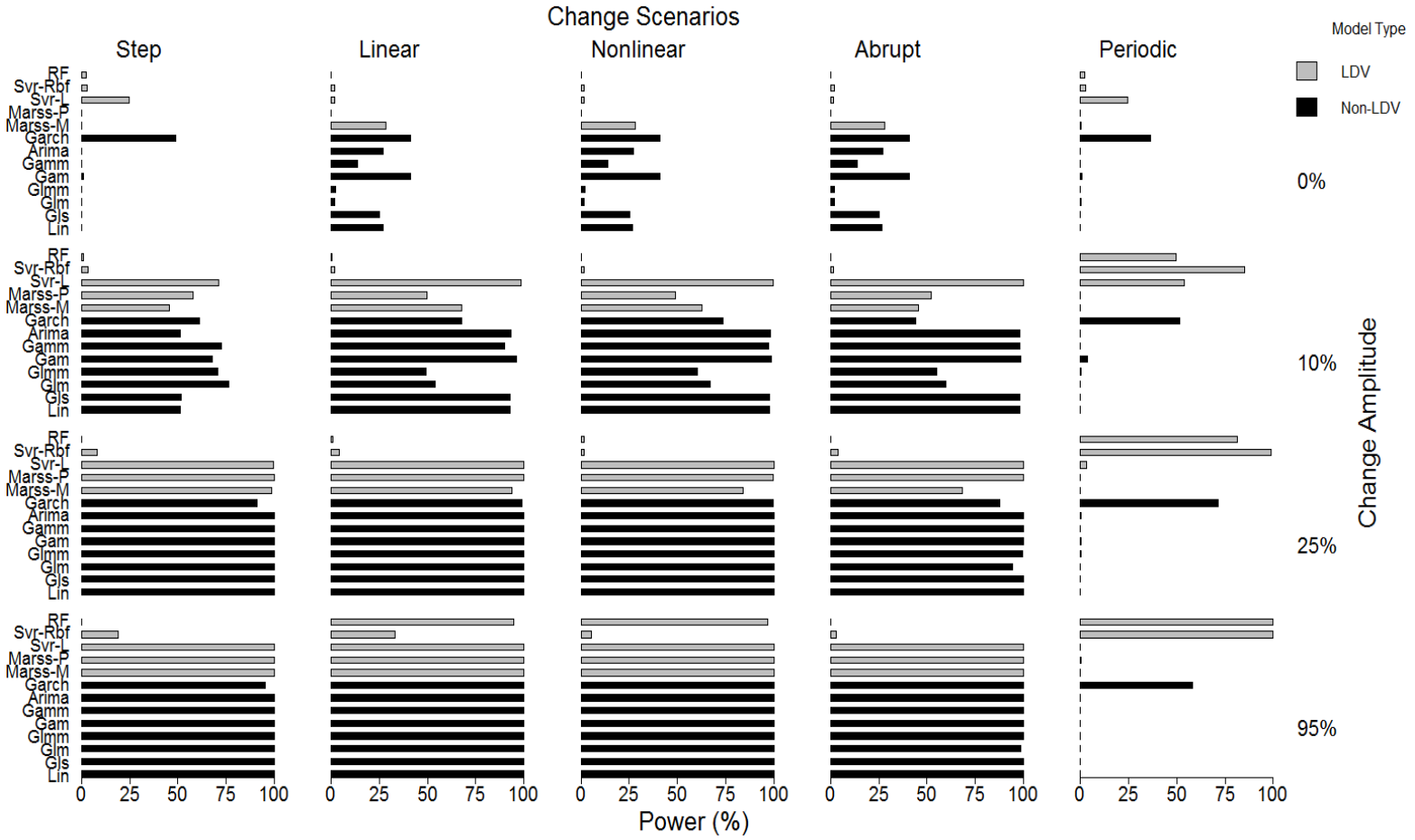

Figure 3.2. Power of candidate models to detect change in simulated non-normally distributed Before-After MRE monitoring data. Power is calculated across varying forms: Step, Linear, Nonlinear, Abrupt, and Periodic; and amplitudes 0%, 10%, 25%, 95%. Models that included lagged dependent variables (i.e., LDV models) are shown in grey, and the models without lagged dependent variables (i.e., non-LDV models) are shown in black.

Variance-Change

Normal Data (Sv)

The Periodic no-change power results are identical to Step no-change scenario results, with the exception of reg-ARMA-GARCH, which had 49.9% power in the Periodic scenario. RF had the greatest power for all Periodic scenarios, followed by SVM-RBF, and reg-ARMA-GARCH. All other models exhibited less than 25% power for all Periodic scenarios. As the amplitude of change increased in the Periodic scenario, the only consistent power increase occurred in the RF and SVM-RBF models; their power reached 100% by the 95% amplitude of change.

Non-Normal Data (AI)

Patterns in model power to detect change in variance within non-normal data are very similar to normal data. The main difference is that the SVM-RBF model displayed higher power for the Periodic 10-25% scenarios than RF. Both models reached 100% power by the 95% amplitude of change.

Lagged Change Scenarios

Ability to detect change was not affected by the introduction of a lag in the onset of change. In general, all candidate models had lower power in the 10% lagged scenarios of change, and consistent or only slightly lower power in the 95% lagged scenarios of change. The relative pattern in the power results among models was consistent with scenarios that did not include a lag in the onset of change.

3.3.2 Fit

Normal Data (Sv)

Model fit performance was consistent across all scenarios and amplitudes of change.

Top Performers

Across all scenarios MARSS-P, followed by SVM-RBF, had the most accurate fit to change scenarios based on RMSE (Fig. 3.3). MARSS-M displayed the 3rd best model fit performance for all mean-change scenarios (Fig. 3.3). The other time series models, reg-ARMA and reg-ARMA-GARCH, were grouped together as the 4th most accurate models for all mean-change scenarios (Fig. 3.3).

Weak Performers

Overall, the GAMM was least able to accurately fit mean-change scenarios (Fig. 3.3). RF was generally the second weakest performer for all mean-change scenarios (Fig. 3.3). The linear regression, GLS, and GAM all performed similarly to RF. GAM generally ranked one position higher in performance than the linear regression and GLS models, but had a relatively similar performance as indicated by its D-statistic values (Fig. 3.3).

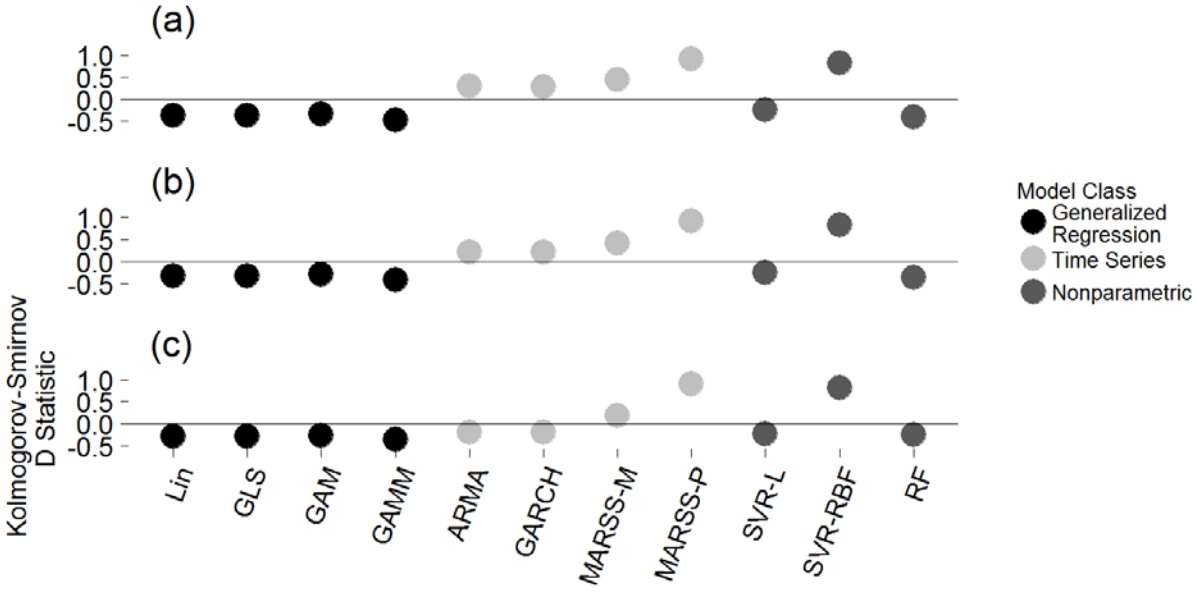

Fig 3.3. Kolmogorov-Smirnov D statistic values of model fit for all change scenarios in normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the root-mean-squared-error (RMSE) values of model fit across all amplitudes of: step mean-change (a), trended mean-change (b), and variance-change (c).

Exceptions to General Fit Performance Patterns

Exceptions to observed performance patterns illustrate the relative improvement of GAM(M)s in nonlinear trend scenarios, and the RF regression in the variance-change scenarios. In the Nonlinear 95% scenario GAM ranked 5th in fit, outperforming the generally 5th ranked model, SVM-L (Table 3.A.1). In the Abrupt 95% scenario, the GAM and GAMM ranked 5th and 6th, outperforming all other candidate models generally considered weak performers (Table 3.A.1). The linear regression and GLS models were grouped as the worst trend fit models for the Nonlinear and Abrupt 95% scenarios rather than GAMM. In the variance-change scenarios RF outperformed all other "weak performers", and was ranked 3rd in fit accuracy in the 95% scenario (Table 3.A.1).

Lagged Change Scenario Highlights

The lagged scenario results were generally consistent with fit performance results from the corresponding non-lagged scenarios. Both 95% lagged scenarios displayed the same relative rank of model fit performance as shown in the non-lagged Nonlinear 95% scenario, with the exception that the GAM model did not outperform SVM-L (Table 3.A.2).

Non-Normal Data (AI)

There appeared to be a generally consistent pattern of model fit performance across all change scenarios in non-normal data.

Top Performers

Similar to the normal data results, the top 3 ranked fit models were generally MARSS-P, SVM-RBF, and MARSS-M (Fig. 3.4). There was less of a difference between the fit performance of these top models and all other candidate models relative to the normal data results. The MARSS-P and SVM-RBF D-statistics are > 0.5 for all scenarios, compared to >0.75 for the normal data (Table 3.B.1). The MARSS-M D-statistic is <0.25 for all scenarios, compared to ~0.5 for the normal data (Table 3.B.1). RF was the 3rd most accurate model for fitting all variance-change scenarios (Fig. 3.4).

Weak Performers

SVM-L, RF, and reg-ARMA-GARCH were generally the worst fit models for the mean-change scenarios (Fig. 3.4). These models were ranked as the bottom 3 worst fit models for all mean-change scenarios except Nonlinear 95%. SVM-L was the least accurate fit model for all 10-25% mean-change scenarios. The SVM-L and reg-ARMA-GARCH models also

generally ranked among the bottom 3 worst fit models for the variance-change scenarios; although their fit performances were relatively similar to the majority of other candidate models (Fig. 3.4).

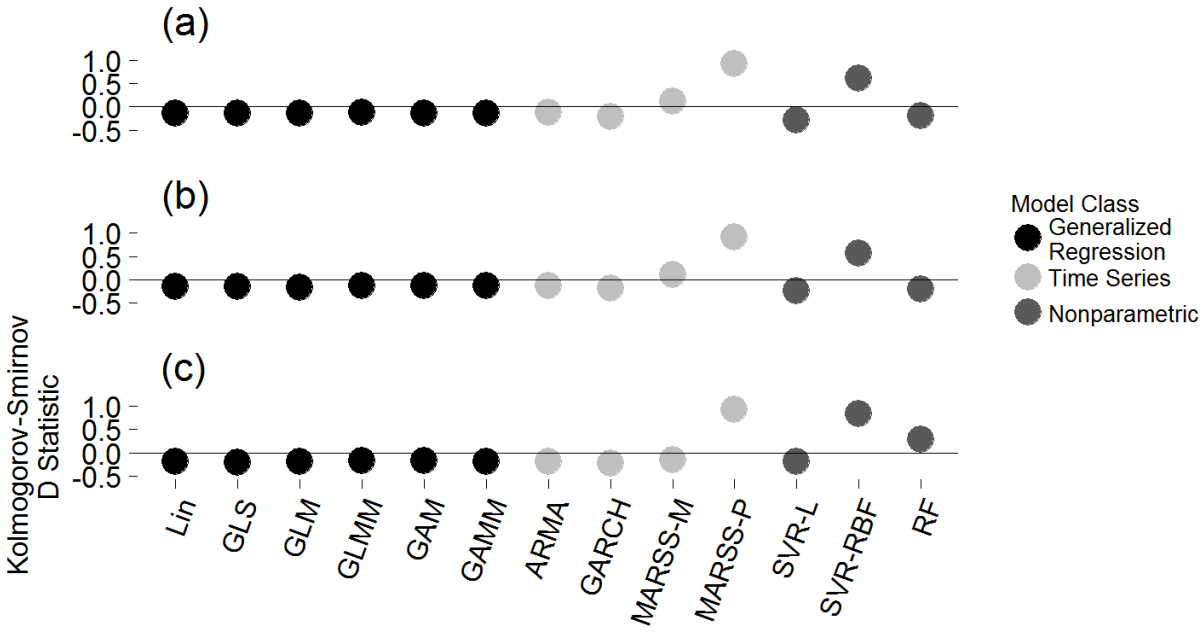

Fig 3.4. Kolmogorov-Smirnov D statistic values of model fit for all change scenarios in non-normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the root-mean-squared-error (RMSE) values of model fit across all amplitudes of: step mean-change (a), trended mean-change (b), and variance-change (c).

Exceptions to General Fit Performance Patterns

The majority of candidate models performed similarly in all change scenarios except in the Nonlinear 95% scenario (Table 3.B.1). Following the general top 3 performers, the GAM(M)s were the best performers, ranking 4th and 5th in fit accuracy. Linear regression and GLS were grouped with reg-ARMA for all other mean-change scenarios, with the exception of

reg-ARMA followed by RF, out-ranked these models in the Nonlinear 95% scenario. Reg-ARMA-GARCH, GLMM, and GLM were ranked as the 3 worst fit models for this scenario.

Lagged Change Scenario Highlights

The lagged scenario results were generally consistent with fit performance results from the corresponding non-lagged scenarios. Both 95% lagged scenarios displayed the same relative rank of model fit performance as shown in the non-lagged Nonlinear 95% scenario. In one exception, the GLM(M)s out-performed the SVM-L and reg-ARMA-GARCH models, which were the worst performing models for both lagged 95% scenarios (Table 3.B.2).

3.3.3 Forecast

Normal Data (Sv)

MASE and RMSE forecast results displayed the same general trends for the normal data and are grouped unless results differed. As the amplitude of change increased the range in model forecasting performance became wider. Differences in performance between ranked models were greatest at the 95% amplitude of change. The relative performance of candidate models was most consistent in the Step scenarios.

Optimal Performers

The SVM-L model was generally the most accurate forecast model for all mean-change scenarios (Fig. 3.5). RF was the best forecast model for all variance-change scenarios and the Abrupt 95% mean-change scenario. RF was the second best forecast model for almost all other mean-change scenarios. SVM-RBF out-ranked SVM-L in the Periodic 95% scenario as the 2nd best forecast model.

Weak Performers

SVM-RBF and GAMM were the worst forecasting models (Fig. 3.5). Although, SVM-RBF performed well in variance-change scenarios (Fig. 3.5). GAM generally had poor forecast accuracy relative to the majority of models, but had a better performance than the GAMM (Fig. 3.5). The MARSS-P and MARSS-M model generally had poor forecast accuracy in the 95% trended mean-change scenarios (Table 3.A.3-4). These models, along with SVM-RBF, were ranked as the 3 worst forecast models for the Linear and Nonlinear 95% scenario (Table 3.A.3-4). In both cases, MARSS-P had less accurate forecasts than MARSS-M. MARSS-P was the worst forecasting model for the Abrupt 95% scenario, whereas MARSS-M was ranked 4th for this scenario (Table 3.A.3-4).

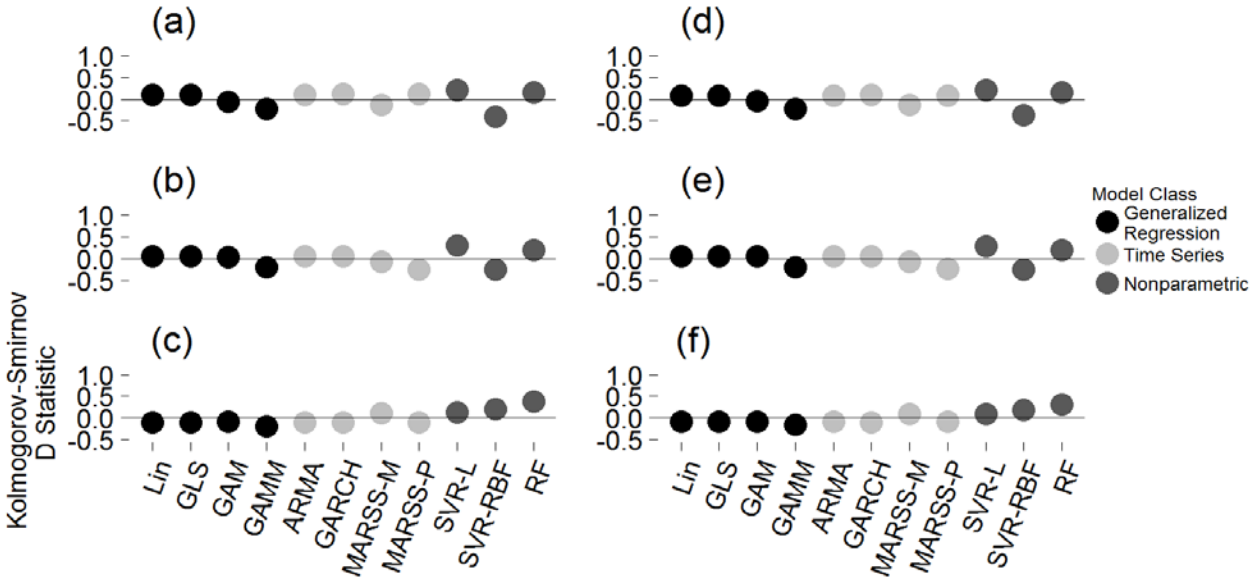

Fig 3.5. Kolmogorov-Smirnov D statistics of model forecast for all change scenarios in normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the mean-absolute-scaled-error (MASE) values of model forecast across all amplitudes of: step mean-change (a), trended mean-change (b), and variance-change (c), and the

root-mean-squared-error (RMSE) values of model forecast across all amplitudes of: step mean-change (d), trended mean-change (e), and variance-change (f).

Exceptions to General Fit Performance Patterns

The only scenario in which SVM-L and RF were not the top two performers was the Linear 95% scenario, in which the linear regression, GLS, reg-ARMA, and reg-ARMA-GARCH were all grouped as the top performing forecast models (Table 3.A.3-4). The relative performance of the GAM(M)s improved in this scenario. GAM and GAMM were ranked 3rd and 5th best forecasting models (Table 3.A.3-4). The GAM(M)s demonstrated a wide ECDF of both RMSE and MASE values in the Abrupt 95% and Nonlinear 95% change scenarios, spanning almost the entire range of metric values from the other candidate models (Fig. 3.6). This wide distribution pattern indicates the presence of outliers in model predictions.

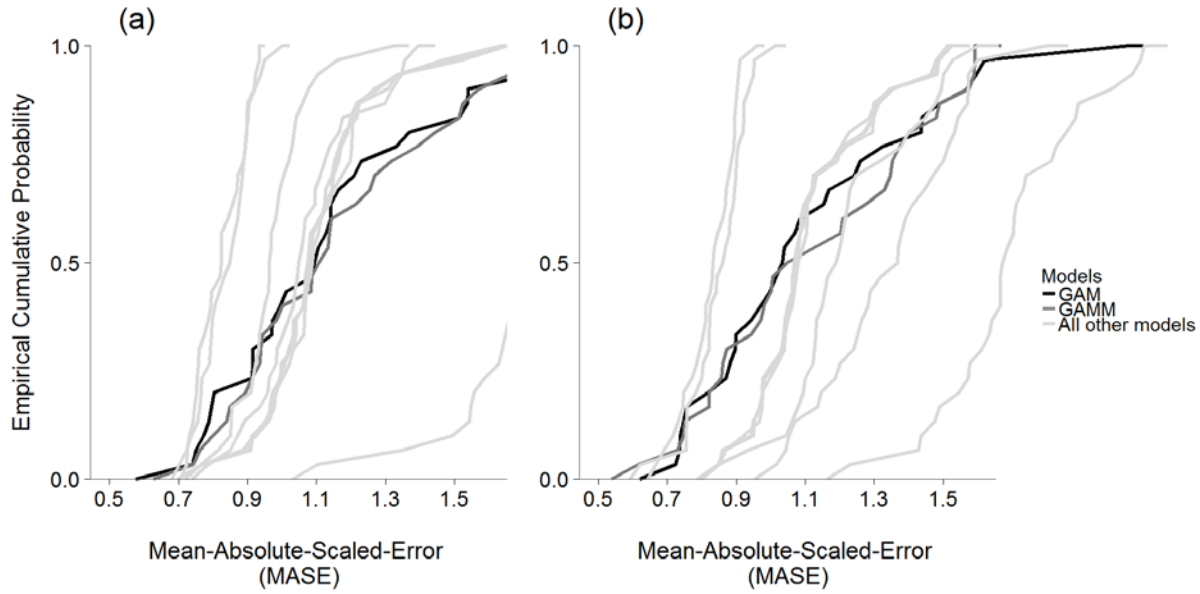

Fig. 3.6. Empirical cumulative probability of the mean-absolute-scaled-error (MASE) values of model forecasts, with the GAM(M) forecasts highlighted for the Abrupt (a) and Nonlinear (b) 95% change scenarios in normally distributed data. MASE plots are representative of the pattern

also displayed in the corresponding root-mean-squared-error (RMSE) cumulative probability plots (not shown).

Lagged Change Scenario Highlights

Results from the lagged change scenarios were similar to those from non-lagged scenarios. The main contrast was between the top forecast performers for the 95% change scenarios. Unlike the 95% non-lagged scenarios, GAM was the top forecast model for the 95% lagged scenarios (Fig. 3.7). GAMM was also a top forecast model along with SVM-L and RF (Fig. 3.7). GAMM and SVM-L were the 2nd and 3rd most accurate forecast models for the Linear 95% scenario, depending on the use of the RMSE or MASE metric, with RF being the 4th best forecaster (Fig. 3.7). The GAMM performance worsened in the Nonlinear 95% scenario, with a ranking that dropped to 4th best forecast model. The poor performance of MARSS-M, MARSS-P, and SVM-RBF in the 95% non-lagged scenarios was generally consistent in the corresponding lagged scenarios (Fig. 3.7). In the Nonlinear 95% scenario MARSS-M was the 5thranked model, out-ranking the grouped Linear regression, GLS, reg-ARMA-GARCH, and reg-ARMA models.

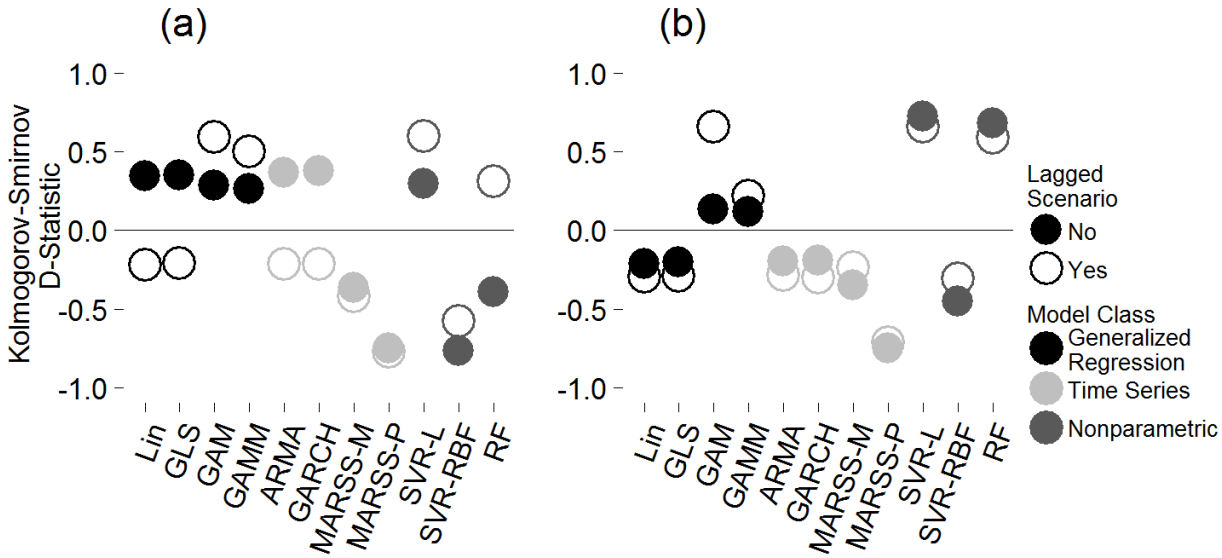

Fig. 3.7. Kolmogorov-Smirnov D statistic values of model forecast for the Linear (a) and Nonlinear (b) 95% change scenarios in normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the mean-absolute-scaled-error (MASE) values of model forecast in both the non-lagged scenarios (solid circles) and corresponding lagged scenarios (open circles). MASE plots are representative of the similar pattern displayed in the corresponding root-mean-squared-error (RMSE) cumulative probability plots (not shown).

Non-normal data (AI)

Interpretation of the non-normal data forecast results differed depending on use of the RMSE or MASE metric. Similar to the normal data results, increased range of relative model performance as the amplitude of change increased was apparent in non-normal data results.

MASE

Optimal Performers

SVM-L was generally the most accurate forecast model among all mean-change scenarios, followed by reg-ARMA-GARCH (Fig. 3.8). RF, GAM, SVM-L, and SVM-RBF were the most accurate forecast models in all Periodic scenarios (Fig. 3.8). SVM-L was ranked above SVM-RBF until the 95% scenario, where SVM-RBF out-ranked SVM-L (Table 3.B.3). SVM-RBF and MARSS-P ranked among the top 4 forecast models in the 10-25% amplitudes of trended mean-change scenarios (Table 3.B.3). The GLM(M)s generally performed similarly and had better forecast accuracy than the other parametric/semi-parametric regression models for all trended mean-change scenarios (Fig. 3.8).

Weak Performers

In general the GAM was a poor forecast model for the trended mean-change scenarios (Fig. 3.8). RF was also generally poor at forecasting the trended mean-change scenarios at the 10-25% amplitudes (Fig. 3.8). The GAMM generally had similar or better forecast accuracy than all other candidate models in its class, but it was ranked among the 3 worst forecast models for Abrupt and Nonlinear 25-95% scenarios (Table 3.B.3). Similar to the normal data results, MARSS-P was a poor forecast model for 95% trended mean-change scenarios. MARSS-P was the worst forecast model for the Linear and Nonlinear 95% scenarios (Table 3.B.3). MARSS-M and SVM-RBF were respectively the 3rd and 2nd worst forecast models for the Linear 95% scenario (Table 3.B.3). MARSS-M was the worst forecast model for all Step scenarios (Fig. 3.8). The GLM(M)s were also weak performers for all Step scenarios, along with SVM-RBF and RF in the 25-95% scenario (Fig. 3.8).

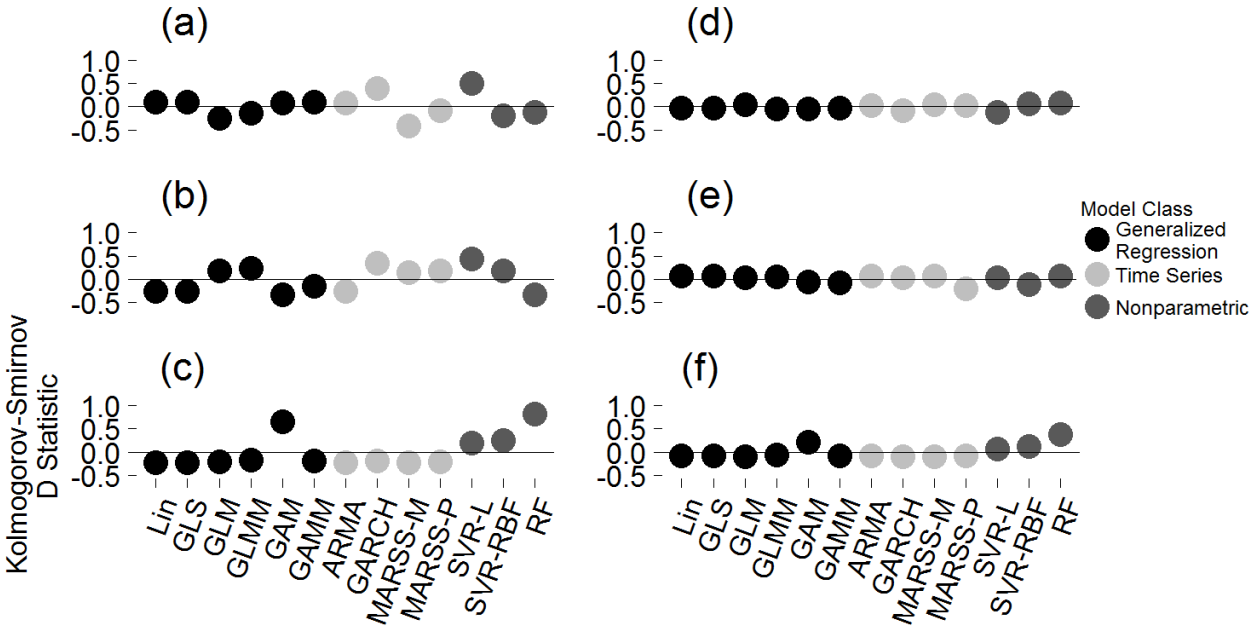

Fig. 3.8. Kolmogorov-Smirnov D statistic values of model forecast for all change scenarios in non-normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the mean-absolute-scaled-error (MASE) values of model

forecast across all amplitudes of: step mean-change (a), trended mean-change (b), and variance-change (c), and the root-mean-squared-error (RMSE) values of model forecast across all amplitudes of: step mean-change (d), trended mean-change (e), and variance-change (f).

Exceptions to General Fit Performance Patterns

Top performing forecast models in the Nonlinear and Abrupt 95% scenarios were inconsistent with the general pattern of top performing models in all other change scenarios. GLM, RF, and GLMM were the top 3 forecast models for the Nonlinear 95% scenario (Table 3.B.3). Reg-ARMA was grouped in ranked performance with the linear regression and GLS model for all scenarios. In the Nonlinear 95% scenario it was ranked as a more accurate forecast model (Table 3.B.3). The majority of candidate models performed similarly in the Abrupt 95% scenario, but SVM-RBF, GLM(M), and MARSS-M all performed slightly better than reg-ARMA-GARCH, resulting in a grouping as the 2nd and 3rd most accurate forecast models (Table 3.B.3).

RMSE

The model forecast performance based on the RMSE metric was similar for all models. An inspection of the D-statistic values and corresponding ECDFs of the forecasting performance showed that differences among models were only distinct in the Periodic and the 95% trended mean-change scenarios.

Optimal Performers

Optimal performers for forecasting the Periodic scenario based on the RMSE metric were the same as those based on the MASE metric results (Fig. 3.8). Overall, there were no distinct optimal performers for the mean-change scenarios.

Weak Performers

The poorest forecast performances occurred in the 95% trended mean-change scenarios. Weak performers in these scenarios matched those based on the MASE metric results. The only exception occurred in the Abrupt 25-95% scenarios, in which MARSS-P was the worst forecast model based on the RMSE metric (Table 3.B.4). The SVM-L and GARCH were the two worst forecasting models for all Step scenarios (Fig. 3.8). In the remaining change scenarios there were generally no specific weak performers among candidate models.

Exceptions in General Fit Performance Patterns

Although there was generally no single best forecast model for the mean-change scenarios based on the RMSE metric, there were a few exceptions. RF and SVM-RBF were the most accurate models for the Step 10-25% scenarios (Table 3.B.4). RF, SVM-L, and MARSS-M were the top 3 forecast models for the Nonlinear 95% scenario (Table 3.B.4). All parametric/semi-parametric regression models and reg-ARMA model were grouped as the most accurate forecast models for the Linear 95% scenario.

Lag Scenario Highlights

MASE and RMSE forecast results displayed the same general trends for the lagged change scenarios and the lag scenario highlights apply to both MASE and RMSE results. Results from the 10% lagged change scenarios were similar to those from 10% non-lagged scenarios. The main contrast was in the 95% lagged change scenarios. Top performers in the 95% lagged scenarios differed from those in the 95% non-lagged scenarios. GAMM, GAM, RF, and reg-ARMA were the top 4 ranked forecast performers for the Linear and Nonlinear 95% lagged

scenarios (Fig. 3.9). Unlike the non-lagged 95% scenarios, GLM(M), reg-ARMA-GARCH, and SVM-L models were all poor forecast models in the lagged 95% scenarios (Fig. 3.9).

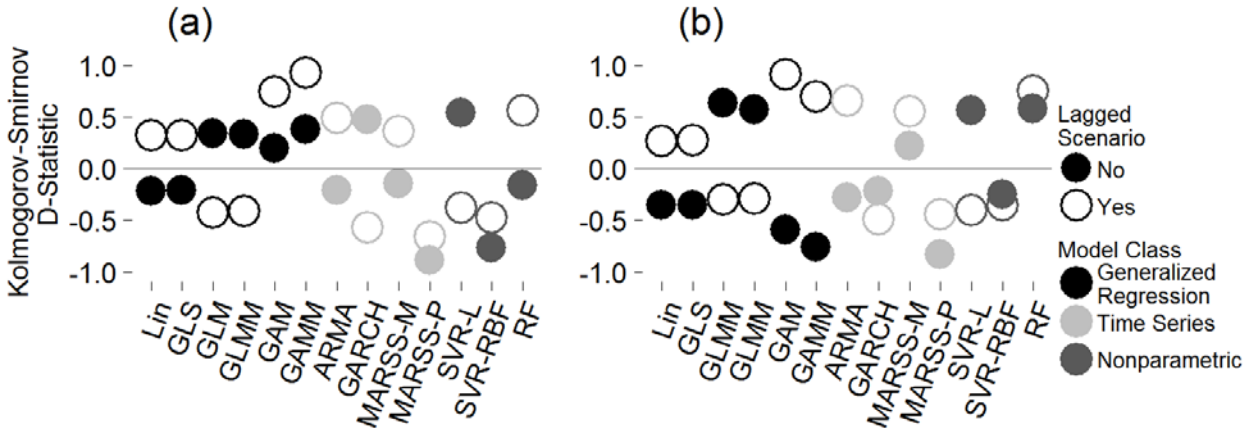

Fig. 3.9. Kolmogorov-Smirnov D statistic values of model forecast for the Linear (a) and Nonlinear (b) 95% change scenarios in non-normally distributed data. The D statistic was calculated from the Empirical Cumulative Distribution Function (ECDF) of the mean-absolute-scaled-error (MASE) values of model forecast in both the non-lagged scenarios (solid circles) and corresponding lagged scenarios (open circles). MASE plots are representative of the similar pattern displayed in the corresponding root-mean-squared-error (RMSE) cumulative probability plots (not shown).

Table 3.6. Summary of model evaluation results. Results are listed based on monitoring objective (i.e., detect, fit, or forecast change) and data (i.e., normal or non-normal data).

Monitoring objective	**Data type**	**Top performing model(s)**	**Weak performing model(s)**	**Exceptions**	**Lagged change highlights**	**Notes**
Detect change	Normal	Non-LDV, SVM-L	LDV	• RF, SVM-RBF are best in variance-change scenarios	None	• Results dependent on baseline covariates
	Non-normal	Non-LDV	LDV	• SVM-RBF, RF are best in variance-change scenarios	None	• Results dependent on baseline covariates
Fit change	Normal	MARSS-P, SVM-RBF, MARSS-M	GAMM	• RF 3rd best for 95% variance-change • GAM(M)s relatively improve in Nonlinear and Abrupt 95% change	None	• ARMA, GARCH grouped as 4th top performing models
	Non-normal	MARSS-P, SVM-RBF, MARSS-M	SVM-L, GARCH, RF	• RF 3rd top performer for all variance-change scenarios	None	• Model results similar except for Nonlinear 95% change
Forecast change	Normal	SVM-L, RF	SVM-RBF, GAMM	• GAMM improves in relative performance in Linear and Nonlinear 95% scenarios • MARSS models poor performers in 95% trended mean-change scenarios	• GAM(M)s top performers for 95% lagged change	• RF best for all variance-change
	Non-normal (MASE only)	SVM-L, GARCH	GAM and RF (trended change); MARSS-M (Step change) (See results for other patterns)	• GLM, RF, GLMM top performers for Nonlinear 95% change • SVM-RBF, GLM(M), MARSS-M outperform GARCH for Abrupt 95% change	• GAM(M), RF, ARMA top performers; GARCH, SVM-L weak performers for lagged 95% change	• RF best for all variance-change • MARSS-P poor performer for 95% trended change

Notes: Forecast results for the normal data were consistent for both RMSE and MASE metrics. Only the MASE forecast results for the non-normal data are listed, because model performances were similar based on the RMSE metric, and any specific patterns in model results did not further contribute to the overall model evaluation findings. The listed lagged change scenario highlights only indicate patterns that differed in the lagged scenarios relative to the corresponding non-lagged scenarios. Reg-ARMA and reg-ARMA-GARCH are listed as ARMA and GARCH.

3.4 Discussion

3.4.1 Model Behavior

A plethora of previous ecological studies have highlighted the strengths and weaknesses of modeling approaches with varying levels of complexity for making data inference and predictions (i.e., Thomas 1996, Elith and Graham 2009, Grilli and Shumchenia 2015). This evaluation is the first to quantify strengths and weaknesses of parametric and nonparametric statistical models that are used to measure change in MRE monitoring data, and to recommend model classes able to detect, quantify, and forecast change. The evaluation also illustrates that the ability of a model to measure change depends on the monitoring objective, which may include detecting, quantifying, and/or forecasting change. Deterministic, parametric models (linear regression, including ANOVA) are most capable of detecting change, while the more flexible time-series models and semi-parametric models are advantageous for quantifying change. To forecast change, the most accurate results were produced by nonparametric models. Findings from this study provide a broad understanding of model behavior in the context of MRE monitoring change scenarios. Recommendations developed here can be used as best practices for measuring change in monitoring data, which will enable effective and efficient environmental monitoring programs that will expedite MRE permitting and development.

Detect Change

Non Lag Dependent Variable (Non-LDV) Models

Non-lagged dependent variable models (i.e., deterministic parametric and semi-parametric models) generally have the greatest power to detect change in the mean-change scenarios, but their ability to detect change is dependent on baseline predictors and model

assumptions (Table 3.6). For example, in the non-normal data change scenarios, GLM(M)s are the only non-LDV models (except reg-ARMA-GARCH) to include day as a covariate. The inclusion of day as a covariate influenced the difference in power results between the GLM(M)s and all other non-LDV models at the 0% and 10% amplitudes of mean-change. Day was included as a covariate in both the high-state and low-state equations of the HMM model used to simulate non-normal data (Table 3.5). Therefore, the simulated non-normal data are known to have a trend across days. By not accounting for baseline trend, any measure of change may be positively biased (Manolov et al. 2010, Parker et al. 2011). If covariates included in the simulation model are assumed to be relevant predictors of the empirical baseline data, then non-LDV models that do not include day as a covariate may be prone to type I error, as illustrated by their higher power in the trended, no-change scenarios relative to the GLM(M)s. Power results for the reg-ARMA-GARCH model and non-normal data highlight the effect of model structure and convergence on model ability to detect change. The reg-ARMA-GARCH model has been shown to have convergence difficulties when including an indicator variable in the mean model equation, especially when outliers are present in the data (e.g., Doornik and Ooms 2005, 2008). The inconsistency of reg-ARMA-GARCH model convergence when including an indicator variable as a covariate may have caused its unique pattern of power results relative to all other models.

Lagged Dependent Variable (LDV) Models

Lagged dependent variable models (i.e., all nonparametric models and MARSS models) generally had lower power to detect mean-change compared to non-LDV models (Table 3.6). The inclusion of lagged predictor variables has been shown to reduce explanatory power of other variables (Achen 2000). Lagged dependent variables account for change in the data as a function

of the dependent variable at previous time steps, suppressing the effect of the indicator variable (Plumper et al. 2005). In this study, lagged dependent variables reduced the explanatory power of the change indicator variable(s). The ability by LDV models to fit change without an indicator variable reduces the power of the model, and indicates that they are less effective than non-LDV models for accurately detecting change.

The effect of including lagged predictor variables as model covariates is best illustrated by the difference in the MARSS-P and MARSS-M power results. The estimate of process and/or observation error in a MARSS model has previously been shown to affect the estimate of mean-reversion; greater process error relative to observation error produces a higher estimate of mean-reversion (Hampton et al. 2013). MARSS-P estimated greater mean-reversion in the normal and non-normal data change scenarios, and generally had greater power than the MARSS-M model. MARSS-M had lower power than MARSS-P, because it estimated less mean-reversion (i.e., a larger effect of the lagged dependent variable), and was better able to flexibly fit change in the data without an indicator variable. Both MARSS-M and MARSS-P models estimate greater mean reversion in the non-normal data relative to the normal data, producing similar power as the non-LDV models for the non-normal data.

The SVM-RBF and RF nonparametric models are the only evaluated models capable of detecting change in variance (Table 3.6). This finding is not surprising as these models have a localized structure, which is better able to detect change in variance because they do not assume a constant global variance (Taddy et al. 2011). However, the ability of the SVM-RBF to measure change in variance may not be as robust or consistent as an RF. The localization of SVM-RBF predicted values is dependent on kernel parameter estimates, which are highly sensitive to their

tuning and prone to overfitting (Eitrich and Lang 2006, Ben-Hur and Weston 2010, Lorena et al. 2011).

Quantifying Change

The MARSS models, especially MARSS-P, excel in accurately fitting change scenarios in normal and non-normal data, but are unable to accurately forecast trended change in monitoring data relative to all other models (Table 3.6). This is due to the baseline parameterization of MARSS models. The baseline structure of evaluated models was not altered, besides inclusion of indicator variable(s), to measure change in baseline conditions. To be consistent with the stationary structure of the baseline data, the B MARSS parameter is estimated as <1 in both MARSS-P and MARSS-M. When the B parameter is <1, the model estimates mean-reversion in the data with the mean level estimated as $\frac{u}{1-B}$, otherwise the u parameter estimates trend in the data (Holmes et al. 2014). Because the B parameter is not fixed as $B = 1$ in baseline models, both baseline MARSS-P and MARSS-M account for trended change as process error. The MARSS models inaccurately forecast predicted values in trended change scenarios, because they do not correctly attribute change to the u trend indicator variable(s).

The reg-ARMA and reg-ARMA-GARCH time series models do not have the same difficulties in attributing change to indicator variable(s), and provide advantages over other parametric models for quantifying change in monitoring data. Generally, the reg-ARMA and reg-ARMA-GARCH predictions are similar to those from the other parametric models, because they are also linear, parametric models, and their predictions are deterministic, regardless of their inclusion of autocorrelated error (Hyndman and Athanasopoulos 2014). However, reg-ARMA and reg-ARMA-GARCH measure stochastic observation error, which provides a more flexible

fit to data than the other parametric models. To illustrate by example, reg-ARMA excels in fit and forecast in the scenarios of lagged 95% change in non-normal data relative to the other parametric models. Reg-ARMA is capable of estimating linear, trend-stationary data. A nonlinear trend may cause non-stationary variable estimates (Hyndman and Athanasopoulos 2014). Reg-ARMA accounts for the non-stationarity by estimating high autocorrelation in the ARMA error structure. The estimate of highly autocorrelated error results in a better fit to nonlinear data (Granger and Newbold 1974, Hyndman and Athanasopoulos 2014). The estimate of autocorrelated error in the deterministic time-series models (i.e., reg-ARMA and reg-ARMA-GARCH) provides greater insight to the form of change (i.e., linear or nonlinear) than other parametric models.

Semi-parametric models (i.e., GAM(M)) excel in fit and forecast in the lagged scenarios of change in normal and non-normal data, indicating their advantage for quantifying nonlinear change (Table 3.6). GAMMs are the only models capable of estimating nonlinear change and providing a quantifiable measure of the nonlinear estimate. However, the inability of GAM(M)s to outperform other models in general fit and forecast accuracy highlight the instability of GAMM estimates. GAMMs are known to have difficulty converging when estimating both autocorrelation and smoother splines (Wood 2006, 2015). GAMs are known to over-fit data by estimating highly nonlinear smoother splines in the presence of autocorrelation (Wood 2006, 2015). As a result, these models are highly sensitive to their parameterization, but are capable of quantifying nonlinear change.

The remaining parametric models do not stand out for either fit or forecast accuracy, and do not provide any advantages over the semi-parametric or deterministic time-series models for quantifying change. These models are unable to fit and forecast nonlinear change, which is

imperative for a model that is to be recommended as a best practice to quantify any change in monitoring data.

Accuracy of nonparametric models to fit change is lower than other models. The SVM-RBF model produced the second most accurate fit for change scenarios in the normal and non-normal data, but its poor forecast of the normal data indicates overfitting. The SVM-RBF does not appear to overfit non-normal data, but the model's forecast accuracy generally decreases as the amplitude of mean-change increases. The SVM-RBF model is more sensitive to model misspecification than SVM-L, which is attributed to its additional tuning parameter (Bahramy and Crone 2013), but both models are known for their sensitivity to tuning parameters (Lorena et al. 2011). As the amplitude of change increases from baseline data, the baseline tuning of the SVM models may not be appropriate for the change scenario datasets. The inability of nonparametric models to provide direct estimates of size and shape of change renders them inappropriate for quantifying change.

Forecasting Change

Nonparametric models, specifically SVM-L and RF, generally excel in forecasting change scenarios relative to all other evaluated models (Table 3.6). These results are not surprising, as these models are regularly cited for their ability to predict both classification (e.g., Cutler et al. 2007, Grilli and Shumchenia 2015) and time-series data (Thissen et al. 2003, Kane et al. 2014). Consistent with this study, SVM-L has been shown to be a more accurate forecast model for trended time-series data than SVM-RBF (e.g., Crone et al. 2006). The relatively high forecast accuracy of the nonparametric models compared to all others decreases by the 95% lagged scenarios. The test data are further outside of the range of the training data in the lagged

scenarios compared to the non-lagged scenarios, as the test data in the lagged scenarios must have a steeper slope to produce the same amplitude of change in the second half of the "after" data. Consequently, model forecast performance in the 95% lagged scenarios illustrates models' abilities to extrapolate outside the range of training data more than the corresponding non-lagged scenarios. Due to the sensitivity of SVM model performance to the model's kernel parameters, baseline SVM models may need to be re-parameterized in these lagged scenarios to accurately forecast outside the data range. The structure of the RF prohibits the model from extrapolating outside the training data range (Kacprzyk and Pedrycz 2015).

The remaining models (i.e., time-series models and parametric regression models) do not generally stand out for their forecasting performance relative to the nonparametric and semi-parametric regression models. The linear regression, GLS, reg-ARIMA, and reg-GARCH are more accurate forecast models than nonparametric models in the normal data Linear 95% change scenario. This result is also not surprising as the data are linear and the complexity of nonparametric models is unnecessary for predicting future change (Merow et al. 2014). Reg-ARMA-GARCH excels in forecasting the non-lagged, non-normal change scenarios based on the MASE metric. Models that had a lower MASE value tended to forecast closer to the low-state non-normal data, because the MASE metric is less sensitive to outliers (i.e., high-state non-normal data) than RMSE (Hyndman and Koehler 2006). The reg-ARMA-GARCH is better able to forecast the low-state non-normal data compared to the other models, except for SVM-L, which may be a result of its measure of heteroskedastic, autoregressive error and its skewed-student-t distribution. As the GARCH model is intended to model stock return volatility (Engle 2001, Ghalanos 2015) (i.e., variability, measured as standard deviation of returns over time), the GARCH R packages only permit the use of skewed and heavy-tailed versions of the normal

distribution to model data. Reg-ARMA-GARCH may be more consistent in its ability to forecast with a transformation of non-normal data to a normal distribution.

3.4.2 Monitoring Application

Power

Non-lagged, dependent variable parametric models are recommended for detecting mean-change in monitoring data. This recommendation is consistent with the previous use of these models to detect change in MRE monitoring studies (e.g., Hammar et al. 2013, Vanermen et al. 2015). Traditionally, linear regression, GLS, or ARMA models are used to measure the effect of a development and/or intervention on baseline conditions (Hewitt et al. 2001, Stewart and Bence 2001). Previous MRE monitoring studies have substituted these traditional methods with GLM(M) or GAM(M)s that use non-normal error distributions (e.g., Poisson or negative binomial distributions) that are more appropriate for the data (e.g., Tollit et al. 2013, Stenberg et al. 2015). This study illustrates the effect of model assumptions and baseline predictor variables on model ability to accurately detect change. For instance, the non-LDV models that did not include day as a covariate had a higher rate of false change detection (i.e., type I error) in non-normal data compared to GLM(M)s. This result suggests that a model that does not account for a baseline trend when it is present may produce a type I error, while a model that accounts for a baseline trend when it is not present may produce a type II error when measuring change. A type II error indicates the failure to detect environmental change and a type I error that falsely estimates environmental change can result in unnecessary costs associated with implementing monitoring and mitigation measures to negate any negative perceived environmental change (Field et al. 2005, Levine et al. 2014). Deterministic parametric and semi-parametric models are

recommended for detecting change in monitoring data, but choice of specific model is dependent on model structure, assumptions, and potential covariates. In the current study the appropriate model covariates are known since the Before-After monitoring data are generated using parameterized simulation models. To identify what may be considered an appropriate baseline characterization, an initial exploration of data characteristics, and a model selection process, as conducted in Chapter 2, using any *a priori* knowledge of the data can be used to make the most valid inferences about the data (Anderson and Burnham 2002).

For monitoring, the inability of the lagged dependent variable models to detect the importance of the indicator variable suggests that they cannot accurately detect change from baseline conditions. Failure to detect change in monitoring data may result in environmental impacts at MRE or other monitoring sites (Fairweather 1991, Legg and Nagy 2006). Nonparametric models do not provide estimates of parameterized variables, so it is not possible to discern how the model is quantifying uncertainty in the estimate of change. Monitoring literature emphasizes the advantage of quantifying the measure and uncertainty of effect size rather than only providing a binary detection of change. The use of a parametric model best aligns with these monitoring goals (Schmitt and Osenberg 1996, Morrison 2007). A quantitative measure of confidence around the detection of change provided by parametric models acts as an objective measure that managers can use to assess risk of alternative management decisions in environmental monitoring (Vos et al. 2000).

Variance

Localized nonparametric models are recommended for detecting, quantifying, and forecasting variance change, because these models were the only candidate models capable of

detecting variance change in the evaluation. Reg-ARMA-GARCH may have also detected variance change through its estimate of heteroskedastic residual error, but this is not reflected as a change in the model's fitted or forecasted predicted values. The Reg-ARMA-GARCH predicted values are based on a mean model equation that does not include residual variance estimates. Model predictions were the only criteria used to evaluate the ability to measure change. This approach facilitated an equal comparison across all parametric and nonparametric models. Previous MRE monitoring studies largely used ANOVA hypothesis testing or indicator variables in parametric/semi-parametric regression models to measure change in mean rather than variance (e.g., Bergstrom et al. 2013, Hammar et al. 2013, Broadhurst and Orme 2014). This evaluation highlights the inability of commonly used parametric/semi-parametric regression models to measure change in variance. It is imperative to measure a change in variance rather than just the mean, because altered variability of populations may indicate perturbations in a population (Underwood 1991, Chapman et al. 1995, Jacques 2014).

Quantifying Change

Deterministic time-series models and semi-parametric regression models are recommended for quantifying change in monitoring data. Deterministic time-series models (i.e., reg-ARMA and reg-ARMA-GARCH) have not previously been used in MRE monitoring studies. GAM(M)s have been used in MRE monitoring studies to measure change in monitoring data that may have nonlinear trends (e.g., Petersen et al. 2011, Tollit et al. 2013). The Mackenzie et al. (2013) MRE model evaluation study limited the candidate models to semi-parametric models (i.e., GAM, GAMM, and CReSS models), because of their ability to quantify nonlinear trend. Mackenzie et al. (2013) recommended CReSS models for measuring change in monitoring data. Unlike GAM(M)s, CReSS models incorporate a Spatially Adaptive Local Smoothing

Algorithm (SALSA) (cf., Walker et al. 2011, Mackenzie et al. 2013). The results of the current study cannot be directly compared to the Mackenzie et al. (2013) study, because the previous evaluation was limited to three scenarios of Before-After step shifts in abundance and distribution of bird and cetacean data, and the study used different evaluation metrics than the current study. The current study evaluated a wider range of model classes, and assumed candidate models are representative of models with similar statistical properties and structures. Therefore, the evaluation results of the GAM(M)s lead to a recommendation of semi-parametric smoother models for quantifying nonlinear trend, which is consistent with the limited evaluation of smoother models in the Mackenzie et al. (2013) study.

The current study illustrates that the reg-ARMA and reg-ARMA-GARCH models generally quantify change similarly to other parametric models, but provide additional information about change in monitoring data by quantifying autocorrelated error. ARMA and ARMA-GARCH models have been used to forecast wind speed because of their ability to measure autocorrelated and heteroskedastic data (e.g., Taylor et al. 2009, Liu et al. 2011). Deterministic time-series models are advantageous for an initial estimate of change in temporally correlated monitoring data, and if the estimate of autocorrelated error is high, then the use of GAM(M)s will directly measure nonlinear change. The use of deterministic time-series and GAM(M)s to quantify change provides robust assessments of both linear and nonlinear change. Accurate assessment of change is imperative for an effective MRE monitoring program because it determines the validity of hypothesized effects from baseline characterization (Treweek 2009), and informs decisions on evolving management practices or mitigation measures (Treweek 2009, Lindenmayer et al. 2012).

MARSS models are not recommended as a primary method for quantifying change in monitoring data, because they are unable to accurately measure trended change using their baseline parameterization. However, an initial estimate of change using the recommended deterministic time-series and semi-parametric models may be used to specify an appropriate parameterization of a state-space model to additionally quantify process error (i.e., stochastic variability) in the data. The estimate of both process and observation error in state-space models provides a more thorough and informative estimate of change by separately quantifying change attributed to natural variability in baseline conditions and systemic change in those conditions (Dornelas et al. 2012). The partitioning of total error into process and observation components may reduce bias and improve accuracy in change estimates (e.g., Lindley 2003, Ward et al. 2010). MARSS models can be parameterized to estimate many different hypothesized structures of population data (Ward et al. 2010, Holmes et al. 2014). Consequently, the initial inspection of change using the primary recommended models may be used to structure state-space models to assess systemic change compared to natural variability in Before-After monitoring data.

Forecasting Change

Nonparametric models have not been used in Before-After biological MRE monitoring studies, but are recommended for forecasting change in MRE monitoring data. Parametric or semi-parametric models in MRE monitoring studies have been used to predict change (e.g., Wade et al. 2014, Warwick-Evans et al. 2016), but MRE monitoring studies do not appear to use models to forecast estimates of a biological variables for early-warning threshold detection. The lack of forecast studies in the MRE biological monitoring literature is attributed to the early stage of development, the retrospective focus of BA(CI) studies to detect if a change occurred, and/or the lack of identified change in existing studies. Selection of accurate forecast models is

imperative for monitoring programs and informing management on timing or conditions when pre-emptive mitigation measures will minimize negative effects of MRE development (Clark et al. 2001, Lindenmayer et al. 2012).

Nonparametric models are recommended for forecasting change because accuracy in prediction rather than estimate of casual relationships is of primary concern. Forecasting change in a monitored response variable does not require hypothesis testing, and does not require parametric measures of casual relationships (Shmueli 2010). Misspecification of structural relationships in parametric models can negatively affect forecasting ability, whereas the flexibility of nonparametric models provides robust predictions against uncertainty (Perretti et al. 2013).

Based on the ability of semi-parametric regression models to provide relatively accurate forecasts in highly nonlinear scenarios, they are recommended as a secondary forecast model for nonparametric data estimates. The measure of change shape may provide further insight into the future trend of the data. For instance, if the change is characterized as highly nonlinear, then the SVM baseline regression model may require a re-tuning of its parameters to provide an accurate forecast. The explicit measure of change can be used to quantify uncertainty around the forecasted change, providing additional information for future monitoring and mitigation efforts (Clark et al. 2001).

3.4.3 MRE Monitoring Model Recommendations

Within the recommended model classes, specific models are suggested for detecting, quantifying, and forecasting change in the MRE case study data (Fig. 3.10). Model recommendations made for the nekton aggregation data are based on the constraints and results

of this study, but may produce biased estimates since they fail to accurately fit the spiked structure of the aggregation data (Barry and Elith 2006). The objective of this study was to evaluate model ability to measure change in monitoring data, not to transform the data to fit requirements of a specific model. Although no models were able to accurately measure spikes in aggregation, the evaluation provided insight into model behavior when fit to highly skewed data. In the future, nekton aggregation data may be altered using a Box-Cox normal data transformation (Box and Cox 1964), to produce unbiased estimates using the recommended models.

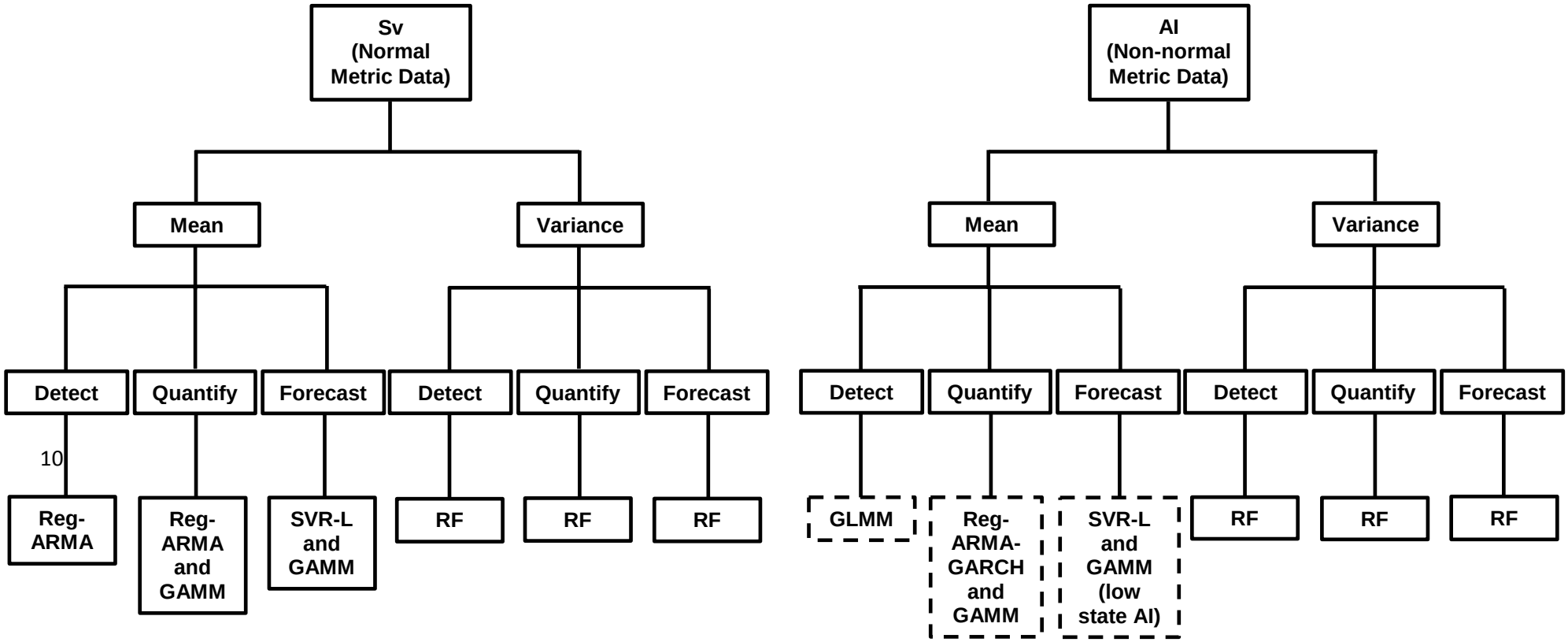

Figure 3.10. Schematic of recommended models to detect, quantify, and forecast change in MRE monitoring data. Models are recommended for detecting, quantifying, or forecasting change in the statistical property of mean or variance of either Sv data (representative of normally distributed data) or AI monitoring data (representative of non-normally distributed data). The dashed boxes indicate potential bias associated with recommended models based on their inability to fit the spikes in aggregation index data.

Detecting Mean Change

Normal Data (Sv)

Reg-ARMA models are recommended for detecting mean change in nekton density data (Fig. 3.10). The reg-ARMA model accurately characterizes data properties, because it assumes a normal distribution, includes environmental covariates that were used as predictor variables in the simulation model, and accounts for autocorrelation. This recommendation is consistent with the recommendation for quantifying change in normal MRE monitoring data.

Non-Normal Data (AI)

The GLMM is recommended for detecting change in nekton aggregation data (Fig. 3.10). Non-LDV models were previously recommended to detect change in MRE monitoring data, and within this group GLMM does not assume normality, appropriately accounts for the day baseline trend in the simulated data, and estimates autocorrelation.

Quantifying Mean Change

Normal Data (Sv)

A reg-ARMA model is recommended for an initial inspection of mean change in the nekton density data, and a GAMM is recommended to provide additional change estimates (Fig. 3.10). A reg-ARMA model is less prone to convergence issues than a GAMM and produces more accurate trend estimations for all scenarios, with the exception of a nonlinear trend. If the reg-ARMA model indicates potential non-stationarity via estimation of highly autocorrelated error, then there may be a nonlinear change. The GAMM may be fit to this data to estimate shape and size of nonlinear trend.

Non-Normal Data (AI)

The reg-ARMA-GARCH model is recommended for an initial inspection of mean change in non-normal data, with the addition of a GAMM to further estimate change (Fig. 3.10). The reg-ARMA-GARCH model is recommended because it generally excels in forecasting the low-state non-normal data, appropriately included the day baseline trend in simulated data, and provides estimates of the spikes in aggregation data through the heteroskedastic residual variance equation. The model may have less convergence difficulties and produce more accurate estimates with a transformation of the data to a normal distribution.

Forecast Mean Change

Normal Data (Sv)

SVM-L is recommended for forecasting change in mean nekton density data (Fig. 3.10). Overall, the SVM-L model was the most accurate forecast model for all mean-change scenarios**.** In the future, re-tuning the baseline SVM-L model to post-construction monitoring data may further improve forecast estimates. A GAMM may be used to provide a semi-parametric assessment of nonlinear forecasts. Although GAMM is sensitive to convergence difficulties in autocorrelated data, the model's ability to fit and extrapolate nonlinear trend allows the GAMM to excel in forecast performance in highly nonlinear data.

Non-Normal Data (AI)

Similar to the normal data recommendations, a SVM-L in conjunction with a GAMM may be used to forecast change in low-state non-normal data (Fig. 3.10). The SVM-L model is most able to accurately forecast mean change in the low-state, non-normal data for almost all

scenarios, with the exception of nonlinear change. No model was able to capture the spiked structure of the non-normal data, and therefore no model can accurately forecast spikes in non-normal data.

Variance Change

The Random Forest regression model is recommended for detecting, quantifying, and forecasting change in variance in both nekton density and aggregation data (Fig. 3.10). The RF and SVM-RBF model are the only two evaluated models that are capable of measuring change in variance. The RF model is recommended over the SVM-RBF model because it had greater forecast accuracy for both datasets. The RF model is more robust in its parameter estimates compared to a SVM-RBF model (Lorena et al. 2011), and the measure of variable importance can be used to interpret the relative importance of an indicator variable in the model.

3.5 Conclusion

Recommended models and the approach used to identify them may be combined as a best practice for the analysis of monitoring data. Standardization of MRE monitoring protocols, including the choice of analytic model, will reduce time and costs of MRE permitting (Dubbs et al. 2013). Standard monitoring methods also enable cross-site comparison, which reduces uncertainty of environmental effects and can streamline monitoring efforts (Froján et al. 2016). The evaluation approach ensures the use of best practices for detecting, quantifying, and forecasting change in monitoring data. The development of best practices for analyzing monitoring data will help lead to the most efficient and effective operational MRE monitoring programs.

The evaluation was applied to MRE monitoring data as a case study of planned anthropogenic disturbance monitoring, but the approach can be applied to assess models and measure change in any monitoring program. Scenarios of change enable a robust estimate of model ability to measure change across the most common range of amplitudes and shapes of change found in ecological literature. The case-study data are assumed representative of monitoring data, and the evaluated statistical models are assumed representative of models used to measure change in monitoring studies. Insight gained on the advantages and disadvantages of the model classes for measuring change are generally applicable.

Chapter 4:

Conclusions and Significance

4.1 Summary of Conclusions

Marine Renewable Energy monitoring data was used as a case study of environmental disturbance monitoring to develop and apply an evaluation of models applicable for analyzing baseline (pre-disturbance) and operational (post-disturbance) monitoring data. Evaluation results enabled specific recommendations of best practices for characterizing baseline data and to detect, quantify, and forecast change in operational monitoring data. Although a single model was not identified for analyzing monitoring data, the evaluation produced consistent patterns in model behavior that highlighted strengths and weaknesses of each evaluated model. The evaluation approach and general model behavior patterns can be applied to identify accurate and robust analytic models for any monitoring program, while specific model recommendations may help standardize MRE monitoring methods and advance MRE development.

The baseline model evaluation resulted in the recommendation of state-space time series models to ensure the most accurate and informative characterization of empirical normal and non-normal baseline monitoring data. Although state-space models have previously been recommended for analyzing renewable energy monitoring data (e.g., Jacques 2014, Diffendorfer et al. 2015), these models are not commonly used in baseline MRE monitoring studies. These models did not only produce an accurate interpolation of the data relative to a wide range of candidate models, but also provided interpretable quantitative estimates, and associated confidence intervals of environmental covariates, data stationarity, and both process and observation error in monitoring data. State-space models are recommended as best practices for

characterizing baseline monitoring data, and can be used to develop effective operational monitoring for measuring change.

Deterministic parametric models were recommended to accurately detect any change in the mean of MRE monitoring data, while localized nonparametric models were identified as the only models capable of detecting and measuring change in the variance of monitoring data. The evaluation results validate the previous use of deterministic parametric models to detect mean-change in MRE monitoring studies (e.g., Hammar et al. 2013, Stenberg et al. 2015, Vanermen et al. 2015), while also highlighting the inability of these models to detect change in variance. The evaluation was critical for identifying statistical models capable of accurately detecting environmental changes caused by a disturbance that only alters the variability of a response variable rather than its mean (Underwood 1991, Chapman et al. 1995). Validation of the previous use of deterministic parametric models to detect mean change in MRE monitoring data reduces uncertainty associated with previous study findings, and the continuing use of these models allows for comparison of future findings with previous studies.

Coupled deterministic time series models and semi-parametric GAM(M)s were recommended to quantify the size and shape of change in monitoring data. Deterministic time series models are not commonly used in MRE monitoring studies, but GAM(M)s have previously been used to quantify change in MRE monitoring studies (e.g., Embling et al. 2013, Tollit et al. 2013). Deterministic time series models provide a robust estimate of linear change, and can be used to identify when GAM(M)s are needed to quantify nonlinear change. The evaluation identified potential convergence problems with GAM(M)s, but also highlighted the ability of GAM(M)s to accurately estimate nonlinear change. The coupling of deterministic time-series models and GAM(M)s provides assessments of both linear and nonlinear change, and will

provide a base for decisions on environmental management practices and/or mitigation measures (Treweek 2009, Lindenmayer et al. 2012).

Nonparametric models were identified as capable of forecasting change in environmental monitoring data. Nonparametric models have not commonly been used in MRE monitoring studies to forecast change, but are generally cited for their predictive accuracy in time-series data (e.g., Thissen et al. 2003, Perretti et al. 2013, Kane et al. 2014). The evaluation highlighted the ability of nonparametric models to accurately forecast a range of change shapes and amplitudes. Selection of accurate forecast models is imperative for informing management on timing or conditions when pre-emptive mitigation measures will minimize negative effects of MRE development (Clark et al. 2001, Lindenmayer et al. 2012).

4.2 Significance

Recommended models and the approach used to identify them may be combined as a best practice for the analysis of monitoring data. Models that can be used for the analysis of both baseline and operational MRE monitoring data are critical for ensuring consistent and accurate estimates of environmental change. Standardization of MRE monitoring methods allows for comparable results of studies across sites and MRE projects (Copping et al. 2016). By enabling precise and comparable analyses of monitoring data, best practice statistical models reduce uncertainty associated with effects of MRE development, and sustain efficient and robust monitoring efforts. In combination, the approach and recommended models provide a pathway to expedite the MRE permitting process (Dubbs et al. 2013).

Evaluations of models that accurately characterize monitoring data are routinely used across a wide range of environmental fields, including fisheries (e.g., Olden and Jackson 2002),

air pollution (e.g., Jerrett et al. 2004), and agriculture (e.g., Michel and Makowski 2013). The evaluation approach developed here can be applied to any monitoring program to provide a comprehensive assessment of model ability to characterize baseline data, and to detect, quantify, and forecast change in operational monitoring data. The case study data are assumed representative of monitoring data, and the evaluated statistical models are assumed representative of models used to measure change in monitoring studies. The scenarios of change simulated as operational monitoring data enable a robust estimate of model ability to measure change across the most common range of amplitudes and shapes of change found in ecological literature. Insight gained on model behavior and advantages and disadvantages of model classes for both characterizing baseline data and measuring change are generally applicable.

4.3 Study Caveats

This approach was specifically structured to allow for an equal comparison across classes of parametric and nonparametric models, while allowing each model to have the greatest probability of success. Consequently, specific parametric model measurements of environmental covariates and indicator variables were not analyzed. An investigation into the estimate and associated uncertainty of environmental covariates included in baseline candidate models may provide a more thorough comparison of parametric models and important environmental predictors of baseline data. Measuring amplitudes and uncertainties of indicator variable(s) in parametric models may also provide additional support for the recommended parametric models to detect and quantify change. Empirical data used in the evaluation was purposely not altered from its raw form, because the purpose of the project was to evaluate model ability to analyze characteristics of the monitoring data. None of the evaluated models, regardless of assumptions, were able to fit the spikes in the non-normal aggregation index data, which may also bias model

estimates of change in the data. To provide more precise model recommendations for the non-normal data, the approach could be repeated on a normalized version of the nekton aggregation index data.

Differences in model recommendations are attributed to differences in the objectives of baseline and operational stages of a monitoring program. Baseline characterization requires statistical models that are able to describe and quantify pattern, including: mean trend, important environmental covariates, and variability within data. Objectives of operational monitoring (i.e., detect, quantify, and forecast change) require an accurate estimate of change from a specified statistical property (i.e., mean and/or variance) of the Before-After dataset. Often the properties of a model that are advantageous for model description may not match other objectives, such as prediction (Shmueli 2010). The model evaluation quantified strengths and weaknesses of candidate models to analyze monitoring data, and resulted in different models that met each monitoring objective. The model recommendation frameworks differ for baseline and operational monitoring but are consistent with the objectives within these stages of a monitoring program.

4.4 Future Work

The next logical step in developing best practice environmental monitoring methods is to extend the results of the evaluation, based on a known point of anthropogenic disturbance, to scenarios of unknown change in real-time monitoring. In the context of MRE, real-time monitoring is becoming more prevalent as the technology for adaptive monitoring is being developed (e.g., Rush et al. 2014, Horne et al. 2016). As real-time operational monitoring continues during an anthropogenic disturbance, there may be unknown points of change (Anderson and Thomas 2004). In future work, the recommended models for detecting,

quantifying, and forecasting change may be evaluated on their ability to measure change in a real-time monitoring context in which change-points may not be known *a priori*.

Appendix 2.A: Candidate model equations and definition of terms

2.A.1 Linear Regression

$y = a + bx + \varepsilon, \varepsilon \sim Normal(0, \sigma)$

The α term is the intercept term, b is the estimated parameter term, x is the predictor variable, ε is the error term, and σ is the standard deviation of the error distribution. This equation also applied to the Generalized Least Squares (GLS) model, with the addition of an autocorrelated residual correlation structure.

2.A.2 Generalized Linear Model (GLM)

$g(\mu) = a + bx + \varepsilon$, $\mu = E(y)$

$g()$ is the link function, which relates the linear predictor to the expected value (μ) of the exponential family distribution function.

2.A.3 Generalized Linear Mixed Model (GLMM)

$g(\mu) = a + X_i B + Z_i b_i + a_i + e_{ij}$, $\mu = E(y_{ij})$

$X_i B$ represents the main effects shown in the GLM equation. X_i is the design matrix for the predictor variables, and B is the matrix of predictor variables. $Z_i b_i$ represents the random effects component of a GLMM. Z_i is the design matrix for the random effects, and b_i is the subject, i, specific effect (or random effect). a_i is the random intercept.

2.A.4 Generalized Additive (Mixed) Model GAM(M)

$g(\mu_i) = a + bx_{1i} + s_1(x_{2i}) + \varepsilon, \mu_i = E(Y_i)$

$s_1()$ is a smoother function, or regression spline. The GAMM can be written similarly to the GLMM, although the X_iB component is a vector of $s(x)$values (Wood 2006).

2.A.5 Regression-Autoregressive-Moving-Average (Reg-ARMA) Model

$y_t = a + \ b_1 x_{1t} + \cdots b_p x_{pt} + n_t$
$n_t = \ b_1 n_{t-1} + \cdots b_p n_{t-p} + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q};\ e_t \sim Normal(0, \sigma)$

n_t is the error remaining from the linear regression model. b_1-b_p represents the parameters multiplied to the autoregressive error terms. $\theta_1 - \theta_q$ represents the parameters multiplied to the moving-average error term in the ARMA component of the model.

2.A.6 Regression-Generalized-Autoregressive-Conditional-Heteroskedasticity-Autoregressive-Moving-Average (Reg-ARMA-GARCH) Model

$\sigma_t^2 = w + \ \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2$

σ_t^2 denotes the conditional variance, w is the intercept, α_1-α_q and β_1-β_p are the Arch and Garch lag parameters

2.A.7 Multivariate-Autoregressive-State-Space (MARSS) Model

Process equation: $x_t = B_t x_{t-1} + u_t + C_t c_t + w_t, w_t \sim MVN(0, Q_t)$

Observation equation: $y_t = x_t + D_t d_t + v_t, v_t \sim MVN(0, R_t)$

The process model contains a density-dependent parameter (B_t), a mean level parameter (u_t), independent parameters (C_t) multiplied to predictor variables (c_t), and error (w_t) that is normally distributed with variance Q_t . The observation model also includes independent covariates (D_t), predictor variables (d_t), and error (v_t) that is normally distributed with variance R_t (Holmes et al. 2014).

2.A.8 Random Forest (RF) Algorithm

The Random Forest is a collection of n_{trees} that are random bootstrap subsamples of the training data. Samples not selected for model training are used as out-of-sample data to calculate error of the model. Within each tree, a randomly chosen subsample of the predictor variables (m_{try}) are used to fit the data. The predictor variable and split of the data based on the predictor variable that produce the best estimate of the dependent variable based on Mean-Squared-Error (MSE) are calculated. This process is repeated until 5 data points remain in each node of the tree. The final predicted values are based on the average of the individual tree predictions.

2.A.9 Support Vector Regression (SVR)

If the linear regression function is denoted as $y_i = wx_i + b$, in which b is the intercept term, w is the estimated parameter term, x_i is the predictor variable, then the minimization attempted by an SVR is denoted as:

$$Q = \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{l}(\varepsilon_i + \varepsilon_i^*)$$

Subject to $\{y_i - wx_i - b \leq \varepsilon + E_i; wx_i + b - y_i \leq \varepsilon + E_i^*; E_i, E_i^* \geq 0\}$ (Vapnik 1995)

In this equation C is a constant that represents the value up to which deviations from ε, a predefined value of residual error, are acceptable. E and E^* are error values above (E) and below (E^*) ε that allow for the optimization problem to be feasible (Smola and Schölkopf 2004, Thissen et al. 2003).

The linear kernel is calculated as $K(x_i, x_j) = (x_i^T x_j)$.

The Radial Basis Function kernel is calculated as $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0$

x_i and x_j are two input vectors, and the gamma value (γ) controls the width of the kernel (Thissen et al. 2003).

Appendix 3.A: Ranked model results for the normal (Sv) metric data.

Table 3.A.1. Ranked in-sample model fit results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Abrupt	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	0.909	1	MARSS-P	0.048	0.909
	2	SVR-RBF	2.844	0.818	2	SVR-RBF	2.815	0.818	2	SVR-RBF	2.805	0.818
	3	MARSS-M	3.620	0.445	3	MARSS-M	3.578	0.458	3	MARSS-M	3.618	0.564
	4	Reg-ARMA	3.711	0.270	4	Reg-ARMA	3.671	0.355	4	Reg-ARMA	3.800	0.297
	4	Reg-ARMA-GARCH	3.711	0.264	4	Reg-ARMA-GARCH	3.671	0.352	4	Reg-ARMA-GARCH	3.801	0.294
	5	SVR-L	3.787	-0.282	5	SVR-L	3.762	-0.385	5	GAM	3.828	0.279
	6	GAM	3.862	-0.300	6	GAM	3.818	-0.361	6	GAMM	3.944	-0.352
	7	Linear reg	3.873	-0.330	6	Linear reg	3.832	-0.352	7	RF	3.998	-0.448
	7	GLS	3.873	-0.333	6	GLS	3.832	-0.355	8	SVR-L	4.037	-0.539
	8	RF	3.876	-0.415	7	RF	3.870	-0.576	9	Linear reg	4.047	-0.497
	9	GAMM	3.951	-0.467	8	GAMM	3.904	-0.482	9	GLS	4.048	-0.500
Linear	1	MARSS-P	0.046	0.909	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	0.909
	2	SVR-RBF	2.868	0.818	2	SVR-RBF	2.825	0.818	2	SVR-RBF	2.840	0.818
	3	MARSS-M	3.641	0.436	3	MARSS-M	3.585	0.512	3	MARSS-M	3.674	0.421
	4	Reg-ARMA	3.737	0.282	4	Reg-ARMA	3.675	0.327	4	Reg-ARMA	3.729	0.321

	4	Reg-ARMA-GARCH	3.737	0.279	4	Reg-ARMA-GARCH	3.675	0.324	4	Reg-ARMA-GARCH	3.729	0.318
	5	SVR-L	3.802	-0.224	5	SVR-L	3.747	-0.267	5	SVR-L	3.852	-0.273
	6	RF	3.880	-0.327	6	GAM	3.820	-0.348	6	GAM	3.885	-0.348
	7	GAM	3.890	-0.336	7	RF	3.828	-0.364	7	Linear reg	3.894	-0.355
	7	Linear reg	3.896	-0.342	7	Linear reg	3.836	-0.391	7	GLS	3.894	-0.358
	7	GLS	3.897	-0.345	7	GLS	3.837	-0.394	8	RF	3.925	-0.488
	8	GAMM	3.955	-0.461	8	GAMM	3.907	-0.582	9	GAMM	3.963	-0.497
Nonlinear	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	0.909	1	MARSS-P	0.049	0.909
	2	SVR-RBF	2.806	0.818	2	SVR-RBF	2.826	0.818	2	SVR-RBF	2.820	0.818
	3	MARSS-M	3.557	0.470	3	MARSS-M	3.573	0.476	3	MARSS-M	3.613	0.558
	4	Reg-ARMA	3.664	0.255	4	Reg-ARMA	3.681	0.252	4	Reg-ARMA	3.788	0.209
	4	Reg-ARMA-GARCH	3.664	0.252	4	Reg-ARMA-GARCH	3.681	0.248	4	Reg-ARMA-GARCH	3.788	-0.209
	5	SVR-L	3.742	-0.264	5	SVR-L	3.754	-0.252	5	GAM	3.829	-0.264
	6	GAM	3.799	-0.345	6	GAM	3.826	-0.364	6	SVR-L	3.858	-0.279
	6	Linear reg	3.809	-0.367	7	RF	3.839	-0.361	7	RF	3.904	-0.370
	6	GLS	3.809	-0.370	7	Linear reg	3.843	-0.400	8	GAMM	3.928	-0.327
	6	RF	3.814	-0.348	7	GLS	3.843	-0.406	9	Linear reg	4.034	-0.527
	7	GAMM	3.869	-0.433	8	GAMM	3.896	-0.530	9	GLS	4.035	-0.530
Periodic	1	MARSS-P	0.044	0.909	1	MARSS-P	0.043	0.909	1	MARSS-P	0.035	0.909
	2	SVR-RBF	2.924	0.818	2	SVR-RBF	2.956	0.818	2	SVR-RBF	3.202	0.818
	3	MARSS-M	3.806	0.348	3	MARSS-M	3.869	0.424	3	RF	4.213	0.715

	4	Reg-ARMA	3.911	0.200	4	Reg-ARMA	4.016	-0.212	4	MARSS-M	4.604	0.458
	4	Reg-ARMA-GARCH	3.913	-0.200	4	Reg-ARMA-GARCH	4.020	-0.215	5	SVR-L	4.764	-0.285
	5	SVR-L	3.979	-0.273	5	SVR-L	4.066	-0.236	6	Reg-ARMA	4.871	-0.348
	6	RF	4.012	-0.294	6	RF	4.096	-0.267	7	Reg-ARMA-GARCH	4.884	-0.358
	7	GAM	4.048	-0.318	7	GAM	4.150	-0.333	8	Linear reg	4.965	-0.421
	7	Linear reg	4.056	-0.312	7	Linear reg	4.155	-0.370	8	GLS	4.965	-0.424
	7	GLS	4.056	-0.315	7	GLS	4.155	-0.373	8	GAM	4.967	-0.415
	8	GAMM	4.112	-0.418	8	GAMM	4.210	-0.503	9	GAMM	5.030	-0.512
Step	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	0.909
	2	SVR-RBF	2.827	0.818	2	SVR-RBF	2.854	0.818	2	SVR-RBF	2.810	0.818
	3	MARSS-M	3.624	0.367	3	MARSS-M	3.600	0.485	3	MARSS-M	3.564	0.542
	4	Reg-ARMA	3.684	0.273	4	Reg-ARMA	3.701	0.364	4	Reg-ARMA	3.664	0.336
	4	Reg-ARMA-GARCH	3.685	0.270	4	Reg-ARMA-GARCH	3.701	0.361	4	Reg-ARMA-GARCH	3.664	0.333
	5	SVR-L	3.750	-0.248	5	SVR-L	3.769	-0.252	5	SVR-L	3.773	-0.318
	6	GAM	3.832	-0.330	6	GAM	3.861	-0.406	6	GAM	3.812	-0.333
	7	RF	3.844	-0.370	7	Linear reg	3.874	-0.418	7	Linear reg	3.827	-0.358
	7	Linear reg	3.850	-0.373	7	GLS	3.875	-0.424	7	GLS	3.828	-0.361
	7	GLS	3.851	-0.376	8	RF	3.878	-0.470	8	RF	3.878	-0.482
	8	GAMM	3.911	-0.476	9	GAMM	3.934	-0.518	8	GAMM	3.891	-0.470

Table 3.A.2. Ranked in-sample model fit results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) for the 95% Linear and Nonlinear change scenarios and corresponding lagged change scenarios.

	Non-Lagged 95% Change Amplitude				**Lagged 95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Linear	1	MARSS-P	0.047	0.909	1	MARSS-P	0.047	1.000
	2	SVR-RBF	2.840	0.818	2	SVR-RBF	2.817	0.900
	3	MARSS-M	3.674	0.421	3	MARSS-M	3.589	0.560
	4	Reg-ARMA	3.729	0.321	4	Reg-ARMA	3.731	0.303
	4	Reg-ARMA-GARCH	3.729	0.318	4	Reg-ARMA-GARCH	3.731	0.293
	5	SVR-L	3.852	-0.273	5	SVR-L	3.789	-0.267
	6	GAM	3.885	-0.348	6	GAM	3.823	-0.307
	7	Linear reg	3.894	-0.355	7	RF	3.841	-0.353
	7	GLS	3.894	-0.358	8	GAMM	3.930	-0.513
	8	RF	3.925	-0.488	9	Linear reg	3.953	-0.533
	9	GAMM	3.963	-0.497	9	GLS	3.953	-0.537
Nonlinear	1	MARSS-P	0.049	0.909	1	MARSS-P	0.047	1.000
	2	SVR-RBF	2.820	0.818	2	SVR-RBF	2.836	0.900
	3	MARSS-M	3.613	0.558	3	MARSS-M	3.612	0.553
	4	Reg-ARMA	3.788	0.209	4	Reg-ARMA	3.774	-0.237
	4	Reg-ARMA-GARCH	3.788	-0.209	4	Reg-ARMA-GARCH	3.775	-0.240
	5	GAM	3.829	-0.264	5	SVR-L	3.821	-0.237
	6	SVR-L	3.858	-0.279	6	GAM	3.831	-0.260
	7	RF	3.904	-0.370	7	RF	3.879	-0.333
	8	GAMM	3.928	-0.327	8	GAMM	3.983	-0.487
	9	Linear reg	4.034	-0.527	9	Linear reg	3.994	-0.530
	9	GLS	4.035	-0.530	9	GLS	3.995	-0.533

Table 3.A.3. Ranked model forecast results and associated average Mean-Absolute-Scale-Error (MASE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**MASE**	**D-stat**	**Rank**	**Model**	**MASE**	**D-stat**	**Rank**	**Model**	**MASE**	**D-stat**
Abrupt	1	SVR-L	0.816	0.358	1	SVR-L	0.834	0.303	1	RF	0.825	0.700
	2	RF	0.828	0.312	1	RF	0.837	0.315	2	SVR-L	0.840	0.645
	3	MARSS-P	0.863	0.136	2	MARSS-M	0.886	0.152	3	SVR-RBF	0.976	0.355
	3	MARSS-M	0.863	0.182	2	Linear reg	0.887	0.091	4	MARSS-M	1.074	-0.170
	4	Reg-ARMA	0.881	-0.091	2	GLS	0.887	0.103	5	Reg-ARMA	1.109	-0.236
	4	Reg-ARMA-GARCH	0.882	-0.097	2	Reg-ARMA	0.888	0.091	5	GLS	1.109	-0.239
	4	GLS	0.882	-0.109	2	Reg-ARMA-GARCH	0.889	0.094	5	Linear reg	1.110	-0.236
	4	Linear reg	0.882	-0.124	3	MARSS-P	0.899	-0.121	5	Reg-ARMA-GARCH	1.117	-0.242
	5	GAM	0.893	-0.158	4	SVR-RBF	0.931	-0.200	6	GAM	1.168	-0.158
	6	SVR-RBF	0.916	-0.270	5	GAM	0.939	-0.167	6	GAMM	1.185	-0.185
	7	GAMM	0.954	-0.391	6	GAMM	0.977	-0.358	7	MARSS-P	1.697	-0.803
Linear	1	SVR-L	0.820	0.324	1	SVR-L	0.820	0.273	1	Reg-ARMA	0.841	0.367
	2	RF	0.831	0.233	2	RF	0.841	0.212	1	Reg-ARMA-GARCH	0.841	0.376
	3	MARSS-P	0.851	0.182	3	Reg-ARMA	0.858	0.079	1	GLS	0.842	0.348
	4	Reg-ARMA-GARCH	0.871	0.094	3	Reg-ARMA-GARCH	0.859	0.085	1	Linear reg	0.843	0.345
	4	GLS	0.872	0.106	3	GLS	0.859	0.088	2	GAM	0.857	0.285

	4	Reg-ARMA	0.872	0.097	3	Linear reg	0.860	0.088	3	SVR-L	0.865	0.297
	4	Linear reg	0.873	0.094	3	MARSS-P	0.861	0.097	4	GAMM	0.919	0.261
	5	MARSS-M	0.877	-0.236	4	MARSS-M	0.869	-0.061	5	RF	1.002	-0.394
	6	GAM	0.887	-0.097	5	GAM	0.881	-0.130	6	MARSS-M	1.156	-0.367
	7	SVR-RBF	0.926	-0.312	6	SVR-RBF	0.926	-0.324	7	MARSS-P	1.569	-0.752
	8	GAMM	0.929	-0.415	7	GAMM	0.939	-0.315	8	SVR-RBF	1.592	-0.770
Nonlinear	1	SVR-L	0.796	0.267	1	SVR-L	0.815	0.300	1	SVR-L	0.829	0.724
	2	RF	0.813	0.176	2	RF	0.839	0.188	2	RF	0.853	0.679
	3	MARSS-P	0.825	0.164	3	MARSS-M	0.852	0.200	3	GAM	1.118	0.133
	4	GAM	0.837	0.088	3	Reg-ARMA-GARCH	0.854	0.124	3	Reg-ARMA	1.120	-0.200
	4	Reg-ARMA	0.840	0.064	3	GLS	0.854	0.121	3	GLS	1.120	-0.203
	4	Reg-ARMA-GARCH	0.840	0.067	3	Linear reg	0.855	0.124	3	Linear reg	1.122	-0.212
	4	GLS	0.840	0.070	3	Reg-ARMA	0.855	0.127	3	Reg-ARMA-GARCH	1.123	-0.194
	4	Linear reg	0.840	0.094	4	MARSS-P	0.861	0.067	3	GAMM	1.132	0.115
	4	MARSS-M	0.840	0.091	5	GAM	0.878	-0.100	4	MARSS-M	1.227	-0.352
	5	SVR-RBF	0.903	-0.312	6	GAMM	0.937	-0.270	5	SVR-RBF	1.351	-0.452
	5	GAMM	0.909	-0.318	7	SVR-RBF	0.940	-0.400	6	MARSS-P	1.702	-0.752
Periodic	1	RF	0.797	0.270	1	RF	0.871	0.230	1	RF	0.765	0.758
	1	SVR-L	0.804	0.291	2	SVR-L	0.886	0.155	2	SVR-RBF	0.873	0.579
	2	Reg-ARMA-GARCH	0.832	0.121	2	MARSS-M	0.893	0.167	3	SVR-L	0.957	0.270
	2	MARSS-P	0.840	-0.106	3	SVR-RBF	0.915	-0.130	4	MARSS-M	0.962	0.333
	2	Reg-ARMA	0.842	-0.115	3	GAM	0.918	-0.082	5	GAM	1.040	-0.221
	2	GAM	0.842	-0.127	3	Reg-ARMA-	0.919	-0.082	6	GAMM	1.050	-0.252

						GARCH						
	2	GLS	0.842	-0.079	3	GLS	0.922	-0.073	6	Linear reg	1.052	-0.258
	2	Linear reg	0.842	-0.073	3	Linear reg	0.922	-0.076	6	GLS	1.052	-0.252
	2	MARSS-M	0.853	-0.155	3	MARSS-P	0.923	-0.085	6	MARSS-P	1.053	-0.255
	3	SVR-RBF	0.876	-0.233	3	Reg-ARMA	0.924	-0.088	6	Reg-ARMA	1.055	-0.258
	4	GAMM	0.877	-0.167	4	GAMM	0.944	-0.209	7	Reg-ARMA-GARCH	1.063	-0.303
Step	1	SVR-L	0.811	0.221	1	SVR-L	0.790	0.233	1	SVR-L	0.806	0.255
	1	RF	0.816	0.179	2	RF	0.801	0.215	2	RF	0.823	0.158
	2	MARSS-P	0.837	0.142	3	MARSS-P	0.811	0.103	2	MARSS-P	0.827	0.170
	2	Reg-ARMA-GARCH	0.839	0.121	3	Reg-ARMA-GARCH	0.811	0.109	2	Reg-ARMA	0.828	0.152
	2	Reg-ARMA	0.839	0.118	3	Reg-ARMA	0.812	0.091	2	Reg-ARMA-GARCH	0.828	0.164
	2	GLS	0.840	0.103	3	Linear reg	0.812	0.091	2	Linear reg	0.829	0.158
	2	Linear reg	0.840	0.106	3	GLS	0.812	0.073	2	GLS	0.829	0.167
	3	GAM	0.856	-0.121	4	GAM	0.827	-0.082	3	GAM	0.835	0.109
	4	MARSS-M	0.860	-0.185	5	MARSS-M	0.838	-0.121	4	MARSS-M	0.857	-0.191
	5	GAMM	0.879	-0.221	6	GAMM	0.862	-0.242	5	GAMM	0.879	-0.297
	6	SVR-RBF	0.897	-0.345	7	SVR-RBF	0.896	-0.312	6	SVR-RBF	0.983	-0.624

Table 3.A.4. Ranked model forecast results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Abrupt	1	SVR-L	3.810	0.361	1	SVR-L	3.963	0.239	1	RF	3.931	0.682
	2	RF	3.880	0.315	1	RF	3.980	0.279	2	SVR-L	4.005	0.642
	3	MARSS-P	4.028	0.155	2	Linear reg	4.154	0.091	3	SVR-RBF	4.635	0.279
	3	MARSS-M	4.046	0.161	2	GLS	4.155	0.112	4	MARSS-M	5.050	-0.218
	4	Reg-ARMA	4.108	-0.109	2	Reg-ARMA	4.156	0.124	5	Reg-ARMA	5.159	-0.242
	4	GLS	4.108	-0.094	2	Reg-ARMA-GARCH	4.160	0.121	5	GLS	5.162	-0.239
	4	Reg-ARMA-GARCH	4.111	-0.106	2	MARSS-M	4.177	0.112	5	Linear reg	5.166	-0.236
	4	Linear reg	4.111	-0.139	3	MARSS-P	4.236	-0.100	5	Reg-ARMA-GARCH	5.193	-0.270
	5	GAM	4.158	-0.148	4	GAM	4.361	-0.136	6	GAM	5.362	-0.167
	6	SVR-RBF	4.304	-0.267	5	SVR-RBF	4.419	-0.221	6	GAMM	5.436	-0.142
	7	GAMM	4.468	-0.436	6	GAMM	4.534	-0.348	7	MARSS-P	7.525	-0.806
Linear	1	SVR-L	3.900	0.300	1	SVR-L	3.838	0.230	1	Reg-ARMA-GARCH	3.986	0.376
	2	RF	3.963	0.245	2	RF	3.941	0.173	1	Reg-ARMA	3.986	0.373
	3	MARSS-P	4.039	0.130	3	Reg-ARMA	3.996	0.091	1	GLS	3.991	0.367
	4	Reg-ARMA-GARCH	4.114	0.088	3	Reg-ARMA-GARCH	3.997	0.094	1	Linear reg	3.996	0.364
	4	Reg-ARMA	4.115	0.085	3	GLS	4.000	0.094	2	GAM	4.055	0.309
	4	GLS	4.116	0.082	3	Linear reg	4.007	0.076	3	SVR-L	4.081	0.294
	4	Linear reg	4.122	-0.061	3	MARSS-P	4.018	0.103	4	GAMM	4.350	0.248

	5	MARSS-M	4.154	-0.161	4	MARSS-M	4.041	0.109	5	RF	4.743	-0.358
	6	GAM	4.165	-0.073	5	GAM	4.105	-0.106	6	MARSS-M	5.331	-0.352
	7	GAMM	4.355	-0.342	6	SVR-RBF	4.320	-0.342	7	MARSS-P	6.877	-0.748
	7	SVR-RBF	4.390	-0.355	7	GAMM	4.348	-0.258	8	SVR-RBF	7.311	-0.779
Nonlinear	1	SVR-L	3.712	0.321	1	SVR-L	3.829	0.252	1	SVR-L	3.885	0.715
	2	RF	3.806	0.203	2	RF	3.916	0.167	2	RF	4.044	0.621
	3	MARSS-P	3.850	0.139	3	Reg-ARMA-GARCH	3.992	0.109	3	GAM	5.114	0.142
	4	GAM	3.911	0.103	3	MARSS-M	3.992	0.158	4	GAMM	5.193	0.118
	4	Linear reg	3.920	0.067	3	GLS	3.994	0.094	5	Reg-ARMA	5.199	-0.164
	4	GLS	3.922	0.064	3	Reg-ARMA	3.997	0.106	5	GLS	5.199	-0.170
	4	Reg-ARMA	3.923	0.061	3	Linear reg	3.998	0.100	5	Linear reg	5.210	-0.176
	4	Reg-ARMA-GARCH	3.923	0.082	4	MARSS-P	4.032	-0.073	5	Reg-ARMA-GARCH	5.212	-0.179
	4	MARSS-M	3.928	0.109	5	GAM	4.085	-0.091	6	MARSS-M	5.573	-0.315
	5	SVR-RBF	4.227	-0.421	6	GAMM	4.340	-0.285	7	SVR-RBF	6.281	-0.464
	5	GAMM	4.235	-0.300	7	SVR-RBF	4.362	-0.373	8	MARSS-P	7.387	-0.712
Periodic	1	RF	3.991	0.288	1	RF	4.438	0.285	1	RF	4.722	0.845
	2	SVR-L	4.026	0.218	2	SVR-L	4.511	0.155	2	SVR-RBF	5.406	0.564
	3	Reg-ARMA-GARCH	4.158	0.103	2	MARSS-M	4.579	0.182	3	MARSS-M	5.782	0.306
	3	MARSS-P	4.192	0.094	3	SVR-RBF	4.669	0.115	4	SVR-L	5.822	0.261
	3	GLS	4.202	-0.103	3	Reg-ARMA-GARCH	4.680	-0.103	5	GAM	6.180	-0.224
	3	Reg-ARMA	4.203	-0.097	3	GAM	4.687	-0.079	6	Linear reg	6.231	-0.242
	3	Linear reg	4.203	-0.106	3	MARSS-P	4.701	-0.118	6	GLS	6.232	-0.245
	3	GAM	4.203	-0.115	3	GLS	4.701	-0.112	6	MARSS-P	6.236	-0.255
	3	MARSS-M	4.286	-0.173	3	Linear reg	4.703	-0.118	6	Reg-ARMA	6.245	-0.261
	4	GAMM	4.359	-0.170	3	Reg-ARMA	4.707	-0.130	7	Reg-ARMA-GARCH	6.268	-0.273
	5	SVR-RBF	4.380	-0.258	4	GAMM	4.822	-0.258	7	GAMM	6.286	-0.279

Step	1	SVR-L	3.744	0.206	1	SVR-L	3.732	0.270	1	SVR-L	3.778	0.273
	2	RF	3.787	0.167	2	RF	3.771	0.161	2	RF	3.858	0.185
	3	MARSS-P	3.885	0.130	3	MARSS-P	3.837	0.094	3	MARSS-P	3.906	0.164
	3	Reg-ARMA-GARCH	3.891	0.088	3	Reg-ARMA	3.839	0.085	3	Reg-ARMA	3.909	0.152
	3	Reg-ARMA	3.892	0.091	3	Reg-ARMA-GARCH	3.839	0.091	3	Reg-ARMA-GARCH	3.909	0.155
	3	GLS	3.894	0.094	3	Linear reg	3.842	0.097	3	Linear reg	3.910	0.152
	3	Linear reg	3.896	0.082	3	GLS	3.843	0.088	3	GLS	3.911	0.142
	4	GAM	3.955	-0.130	4	GAM	3.906	-0.103	4	GAM	3.940	0.118
	5	MARSS-M	3.977	-0.142	5	MARSS-M	3.942	-0.142	5	MARSS-M	4.026	-0.173
	6	GAMM	4.068	-0.185	6	GAMM	4.065	-0.279	6	GAMM	4.127	-0.248
	7	SVR-RBF	4.131	-0.330	7	SVR-RBF	4.174	-0.282	7	SVR-RBF	4.559	-0.679

Appendix 3.B: Ranked model results for the non-normal (AI) metric data.

Table 3.B.1. Ranked in-sample model fit results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Abrupt	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.043	0.641	2	SVR-RBF	0.045	0.723	2	SVR-RBF	0.046	0.577
	3	MARSS-M	0.054	0.136	3	MARSS-M	0.056	0.164	3	MARSS-M	0.055	0.195
	4	GLMM	0.055	-0.126	4	GAM	0.058	-0.151	4	GAM	0.058	-0.144
	4	GAM	0.055	-0.144	4	GLMM	0.058	-0.146	4	GAMM	0.058	-0.146
	4	Reg-ARMA	0.056	-0.141	4	GAMM	0.058	-0.159	5	GLMM	0.059	-0.162
	4	GAMM	0.056	-0.151	4	Reg-ARMA	0.059	-0.149	5	Reg-ARMA	0.059	-0.159
	4	GLM	0.056	-0.144	4	Linear reg	0.059	-0.174	5	Linear reg	0.059	-0.172
	4	Linear reg	0.056	-0.156	4	GLS	0.059	-0.177	5	GLS	0.059	-0.172
	4	GLS	0.056	-0.159	4	GLM	0.059	-0.172	6	GLM	0.059	-0.187
	5	RF	0.057	-0.197	5	Reg-ARMA-GARCH	0.060	-0.259	7	Reg-ARMA-GARCH	0.060	-0.231
	5	Reg-ARMA-GARCH	0.057	-0.233	6	RF	0.061	-0.354	8	RF	0.061	-0.274
	6	SVR-L	0.058	-0.269	7	SVR-L	0.061	-0.328	8	SVR-L	0.061	-0.300
Linear	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.046	0.605	2	SVR-RBF	0.047	0.626	2	SVR-RBF	0.045	0.600
	3	MARSS-M	0.057	0.128	3	MARSS-M	0.058	-0.126	3	MARSS-M	0.055	0.159
	4	GLMM	0.058	-0.131	4	GLMM	0.059	-0.138	4	GLMM	0.058	-0.144
	4	GAM	0.058	-0.138	4	GAM	0.059	-0.133	4	Reg-ARMA	0.058	-0.151
	5	Reg-ARMA	0.058	-0.138	5	Reg-ARMA	0.060	-0.149	4	GAM	0.058	-0.146

	5	GLM	0.058	-0.133	5	GLM	0.060	-0.156	4	GLM	0.058	-0.162
	5	Linear reg	0.058	-0.151	5	GAMM	0.060	-0.162	4	Linear reg	0.058	-0.156
	5	GLS	0.058	-0.154	5	Linear reg	0.060	-0.167	4	GLS	0.058	-0.159
	5	GAMM	0.058	-0.146	5	GLS	0.060	-0.169	4	GAMM	0.058	-0.164
	6	RF	0.060	-0.246	6	RF	0.061	-0.228	5	Reg-ARMA-GARCH	0.059	-0.208
	6	Reg-ARMA-GARCH	0.060	-0.244	6	Reg-ARMA-GARCH	0.061	-0.254	6	SVR-L	0.061	-0.259
	7	SVR-L	0.061	-0.300	7	SVR-L	0.062	-0.308	7	RF	0.061	-0.310
Nonlinear	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.045	0.621	2	SVR-RBF	0.046	0.613	2	SVR-RBF	0.043	0.738
	3	MARSS-M	0.056	0.144	3	MARSS-M	0.056	0.138	3	MARSS-M	0.054	0.331
	4	GLMM	0.057	-0.131	4	GAM	0.057	-0.131	4	GAM	0.057	0.213
	4	GAM	0.057	-0.133	4	GLMM	0.058	-0.128	5	GAMM	0.058	0.177
	5	Reg-ARMA	0.057	-0.144	5	GAMM	0.058	-0.138	6	Reg-ARMA	0.058	0.162
	5	GLM	0.057	-0.156	5	Reg-ARMA	0.058	-0.133	7	RF	0.060	-0.203
	5	GAMM	0.058	-0.159	5	GLM	0.058	-0.144	8	Linear reg	0.063	-0.279
	5	Linear reg	0.058	-0.149	5	Linear reg	0.058	-0.149	8	GLS	0.063	-0.282
	5	GLS	0.058	-0.151	5	GLS	0.058	-0.151	8	SVR-L	0.063	-0.310
	6	RF	0.059	-0.187	6	Reg-ARMA-GARCH	0.060	-0.200	9	Reg-ARMA-GARCH	0.065	-0.336
	7	Reg-ARMA-GARCH	0.059	-0.231	6	RF	0.060	-0.210	10	GLMM	0.066	-0.359
	8	SVR-L	0.060	-0.279	7	SVR-L	0.061	-0.244	11	GLM	0.067	-0.456
Periodic	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.042	0.831	2	SVR-RBF	0.040	0.841	2	SVR-RBF	0.041	0.846
	3	RF	0.059	-0.151	3	RF	0.058	0.344	3	RF	0.062	0.769
	3	MARSS-M	0.059	-0.149	4	MARSS-M	0.062	-0.172	4	MARSS-M	0.091	-0.231
	4	GAM	0.060	-0.177	5	GAM	0.063	-0.179	5	Reg-ARMA	0.093	-0.238

	5	GLMM	0.060	-0.197	6	GLMM	0.064	-0.192	6	GLM	0.093	-0.254
	6	GLM	0.061	-0.215	6	GLM	0.064	-0.192	7	GLMM	0.094	-0.264
	6	Reg-ARMA	0.061	-0.213	6	Reg-ARMA	0.064	-0.203	8	Linear reg	0.094	-0.269
	6	GAMM	0.061	-0.233	6	SVR-L	0.064	-0.190	8	GLS	0.094	-0.272
	6	Linear reg	0.061	-0.236	7	Linear reg	0.064	-0.228	8	GAMM	0.094	-0.285
	6	GLS	0.061	-0.238	7	GLS	0.064	-0.228	9	Reg-ARMA-GARCH	0.094	-0.292
	7	SVR-L	0.062	-0.249	7	GAMM	0.064	-0.223	10	GAM	0.097	-0.400
Step	8	Reg-ARMA-GARCH	0.062	-0.295	8	Reg-ARMA-GARCH	0.065	-0.249	10	SVR-L	0.097	-0.485
	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.044	0.613	2	SVR-RBF	0.044	0.644	2	SVR-RBF	0.046	0.597
	3	MARSS-M	0.055	0.149	3	MARSS-M	0.055	-0.126	3	MARSS-M	0.056	0.174
	4	GLMM	0.056	-0.133	4	GLMM	0.056	-0.138	4	GLMM	0.057	-0.133
	5	GAM	0.056	-0.151	5	Reg-ARMA	0.057	-0.136	5	Reg-ARMA	0.058	-0.131
	5	Reg-ARMA	0.056	-0.154	5	GAM	0.057	-0.144	5	GAM	0.058	-0.144
	5	GLM	0.057	-0.146	5	GLM	0.057	-0.149	5	GLM	0.058	-0.149
	5	GAMM	0.057	-0.169	5	GAMM	0.057	-0.149	5	GAMM	0.058	-0.151
	5	Linear reg	0.057	-0.159	5	Linear reg	0.057	-0.154	5	Linear reg	0.058	-0.154
	5	GLS	0.057	-0.162	5	GLS	0.057	-0.156	5	GLS	0.058	-0.156
	6	RF	0.057	-0.182	6	Reg-ARMA-GARCH	0.058	-0.223	6	Reg-ARMA-GARCH	0.059	-0.244
	7	Reg-ARMA-GARCH	0.058	-0.246	6	RF	0.058	-0.223	7	SVR-L	0.060	-0.336
	8	SVR-L	0.059	-0.356	7	SVR-L	0.059	-0.269	8	RF	0.061	-0.326

Table 3.B.2. Ranked in-sample model fit results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) for the 95% Linear and Nonlinear change scenarios and corresponding lagged change scenarios.

	Non-Lagged 95% Change Amplitude				**Lagged 95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Linear	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.045	0.600	2	SVR-RBF	0.045	0.718
	3	MARSS-M	0.055	0.159	3	MARSS-M	0.057	0.262
	4	GLMM	0.058	-0.144	4	GAM	0.059	0.149
	4	Reg-ARMA	0.058	-0.151	5	GAMM	0.060	-0.146
	4	GAM	0.058	-0.146	6	Reg-ARMA	0.061	-0.159
	4	GLM	0.058	-0.162	7	RF	0.062	-0.203
	4	Linear reg	0.058	-0.156	8	GLMM	0.063	-0.233
	4	GLS	0.058	-0.159	8	Linear reg	0.063	-0.223
	4	GAMM	0.058	-0.164	8	GLS	0.063	-0.226
	5	Reg-ARMA-GARCH	0.059	-0.208	9	GLM	0.063	-0.262
	6	SVR-L	0.061	-0.259	10	SVR-L	0.065	-0.318
	7	RF	0.061	-0.310	11	Reg-ARMA-GARCH	0.066	-0.338
Nonlinear	1	MARSS-P	0.001	0.923	1	MARSS-P	0.001	0.923
	2	SVR-RBF	0.043	0.738	2	SVR-RBF	0.045	0.682
	3	MARSS-M	0.054	0.331	3	MARSS-M	0.056	0.223
	4	GAM	0.057	0.213	4	GAM	0.059	-0.159
	5	GAMM	0.058	0.177	5	GAMM	0.060	-0.167
	6	Reg-ARMA	0.058	0.162	5	Reg-ARMA	0.060	-0.174
	7	RF	0.060	-0.203	5	RF	0.061	-0.154
	8	Linear reg	0.063	-0.279	6	GLMM	0.063	-0.244
	8	GLS	0.063	-0.282	6	Linear reg	0.063	-0.254
	8	SVR-L	0.063	-0.310	6	GLS	0.063	-0.256
	9	Reg-ARMA-GARCH	0.065	-0.336	6	GLM	0.063	-0.277
	10	GLMM	0.066	-0.359	7	Reg-ARMA-GARCH	0.066	-0.372
	11	GLM	0.067	-0.456	8	SVR-L	0.066	-0.382

Table 3.B.3. Ranked model forecast results and associated average Mean-Absolute-Scaled-Error (MASE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**MASE**	**D-stat**	**Rank**	**Model**	**MASE**	**D-stat**	**Rank**	**Model**	**MASE**	**D-stat**
Abrupt	1	SVR-L	0.965	0.536	1	SVR-L	0.957	0.587	1	SVR-L	1.118	0.474
	2	Reg-ARMA-GARCH	1.022	0.451	2	Reg-ARMA-GARCH	1.026	0.462	2	GLM	1.142	0.474
	3	MARSS-P	1.054	0.390	3	MARSS-P	1.046	0.415	2	SVR-RBF	1.147	0.423
	4	SVR-RBF	1.157	0.228	4	GLMM	1.097	0.362	3	MARSS-M	1.183	0.433
	4	GLMM	1.165	0.182	4	SVR-RBF	1.101	0.351	3	GLMM	1.186	0.446
	5	GLM	1.196	0.154	4	MARSS-M	1.106	0.364	4	Reg-ARMA-GARCH	1.227	0.341
	6	MARSS-M	1.228	0.138	4	GLM	1.114	0.326	5	RF	1.510	-0.321
	7	GAMM	1.288	-0.197	5	Linear reg	1.333	-0.390	6	Linear reg	1.552	-0.364
	7	Linear reg	1.315	-0.277	5	Reg-ARMA	1.333	-0.385	6	GLS	1.552	-0.362
	7	GLS	1.316	-0.279	5	GLS	1.333	-0.387	6	MARSS-P	1.553	-0.310
	7	Reg-ARMA	1.319	-0.274	6	RF	1.420	-0.390	6	Reg-ARMA	1.553	-0.367
	8	GAM	1.488	-0.490	7	GAMM	1.538	-0.615	7	GAM	1.837	-0.556
	8	RF	1.524	-0.510	8	GAM	1.564	-0.608	8	GAMM	1.963	-0.685
Linear	1	SVR-L	0.951	0.546	1	SVR-L	0.970	0.551	1	SVR-L	0.953	0.541
	2	Reg-ARMA-GARCH	1.008	0.459	2	Reg-ARMA-GARCH	1.013	0.490	2	Reg-ARMA-GARCH	0.995	0.477
	3	MARSS-P	1.069	0.344	3	SVR-RBF	1.035	0.464	3	GAMM	1.100	0.377
	3	SVR-RBF	1.069	0.336	4	MARSS-P	1.060	0.400	4	GLMM	1.119	0.331
	4	GAMM	1.138	0.210	5	GAMM	1.166	0.197	4	GLM	1.122	0.336
	5	GLMM	1.182	0.133	6	MARSS-M	1.170	0.203	5	GAM	1.247	0.192
	6	GLM	1.212	-0.172	6	GLMM	1.179	0.164	6	RF	1.312	-0.167
	7	MARSS-M	1.249	-0.213	7	GLM	1.207	-0.185	7	Reg-ARMA	1.315	-0.215

	8	Linear reg	1.318	-0.290	8	Linear reg	1.318	-0.336	7	Linear reg	1.316	-0.218
	8	GLS	1.318	-0.285	8	GLS	1.318	-0.341	7	GLS	1.317	-0.215
	8	Reg-ARMA	1.320	-0.292	8	Reg-ARMA	1.319	-0.372	8	MARSS-M	1.347	-0.149
	9	GAM	1.400	-0.354	8	GAM	1.329	-0.382	9	SVR-RBF	2.370	-0.774
	10	RF	1.642	-0.556	9	RF	1.507	-0.590	10	MARSS-P	3.196	-0.890
Nonlinear	1	SVR-L	1.005	0.567	1	SVR-L	0.976	0.462	1	GLM	1.338	0.631
	2	Reg-ARMA-GARCH	1.064	0.451	2	SVR-RBF	1.038	0.413	2	RF	1.468	0.579
	3	MARSS-P	1.098	0.446	2	Reg-ARMA-GARCH	1.050	0.397	3	GLMM	1.468	0.567
	4	SVR-RBF	1.128	0.364	2	MARSS-P	1.053	0.392	4	SVR-L	1.521	0.559
	5	GLMM	1.229	0.154	3	MARSS-M	1.165	0.274	5	MARSS-M	1.873	0.218
	6	GAMM	1.229	0.169	3	GLMM	1.186	0.256	6	Reg-ARMA-GARCH	1.989	-0.223
	7	MARSS-M	1.262	-0.154	4	GLM	1.218	0.221	7	Reg-ARMA	2.036	-0.282
	7	GLM	1.263	-0.190	5	Linear reg	1.402	-0.338	8	Linear reg	2.106	-0.354
	8	Reg-ARMA	1.360	-0.344	5	Reg-ARMA	1.403	-0.331	8	GLS	2.106	-0.356
	8	GLS	1.361	-0.367	5	GLS	1.403	-0.341	9	SVR-RBF	2.148	-0.254
	8	Linear reg	1.361	-0.369	6	GAMM	1.485	-0.333	10	GAM	2.851	-0.595
	9	GAM	1.519	-0.510	7	RF	1.590	-0.482	11	GAMM	3.443	-0.767
	9	RF	1.551	-0.562	8	GAM	1.707	-0.664	12	MARSS-P	4.326	-0.833
Periodic	1	RF	0.934	0.692	1	RF	0.855	0.890	1	RF	0.660	0.923
	2	GAM	1.030	0.467	2	GAM	1.039	0.721	2	GAM	1.012	0.846
	3	SVR-L	1.044	0.474	3	SVR-L	1.162	0.472	3	SVR-RBF	1.358	0.654
	4	SVR-RBF	1.066	0.451	4	SVR-RBF	1.173	0.454	4	SVR-L	1.441	0.526
	5	GLMM	1.187	-0.200	5	GLMM	1.307	-0.187	5	GLMM	1.597	-0.254
	6	Reg-ARMA-GARCH	1.198	-0.205	6	GLM	1.349	-0.249	6	GAMM	1.647	-0.282
	7	GAMM	1.211	-0.213	6	MARSS-M	1.352	-0.246	6	MARSS-M	1.653	-0.326
	8	GLM	1.224	-0.282	6	GAMM	1.362	-0.269	6	GLM	1.654	-0.331

	8	MARSS-P	1.229	-0.244	6	Reg-ARMA-GARCH	1.365	-0.285	7	MARSS-P	1.663	-0.313
	8	Linear reg	1.236	-0.269	7	MARSS-P	1.377	-0.282	7	Reg-ARMA-GARCH	1.664	-0.318
	8	GLS	1.236	-0.267	8	Linear reg	1.385	-0.305	7	Reg-ARMA	1.673	-0.344
	8	Reg-ARMA	1.238	-0.277	8	GLS	1.385	-0.305	7	GLS	1.673	-0.338
	8	MARSS-M	1.241	-0.308	8	Reg-ARMA	1.385	-0.303	7	Linear reg	1.673	-0.341
Step	1	SVR-L	0.957	0.497	1	SVR-L	0.932	0.482	1	SVR-L	0.930	0.572
	2	Reg-ARMA-GARCH	1.012	0.410	2	Reg-ARMA-GARCH	0.968	0.397	2	Reg-ARMA-GARCH	0.979	0.462
	3	RF	1.112	0.128	3	GAMM	1.092	0.103	3	GAMM	1.098	0.144
	3	SVR-RBF	1.116	0.113	3	GLS	1.096	0.087	3	GAM	1.105	0.131
	3	GAMM	1.120	0.126	3	Linear reg	1.096	0.085	3	Linear reg	1.105	0.133
	3	GLS	1.128	-0.103	3	Reg-ARMA	1.098	-0.090	3	GLS	1.105	0.131
	3	Linear reg	1.128	-0.105	3	GAM	1.102	0.090	3	Reg-ARMA	1.107	0.128
	3	GAM	1.128	0.090	3	MARSS-P	1.108	-0.085	4	MARSS-P	1.117	-0.115
	3	Reg-ARMA	1.128	-0.100	4	RF	1.123	-0.144	5	GLMM	1.145	-0.136
	3	MARSS-P	1.136	-0.133	5	SVR-RBF	1.142	-0.177	6	RF	1.171	-0.231
	4	GLMM	1.194	-0.221	6	GLMM	1.155	-0.208	6	GLM	1.175	-0.231
	5	GLM	1.225	-0.300	7	GLM	1.187	-0.313	7	MARSS-M	1.264	-0.428
	6	MARSS-M	1.303	-0.500	8	MARSS-M	1.270	-0.477	8	SVR-RBF	1.272	-0.446

Table 3.B.4. Ranked model forecast results and associated average Root-Mean-Squared-Error (RMSE), and Kolmogorov-Smirnov D-Statistic (D-stat) across 10%, 25%, and 95% amplitudes of change for all change scenarios.

	10% Change Amplitude				**25% Change Amplitude**				**95% Change Amplitude**			
Change Scenario	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**	**Rank**	**Model**	**RMSE**	**D-stat**
Abrupt	1	MARSS-M	0.082	0.095	1	MARSS-M	0.079	0.118	1	MARSS-M	0.087	0.164
	1	GLM	0.082	0.100	1	SVR-RBF	0.079	0.087	1	Reg-ARMA-GARCH	0.088	0.144
	1	SVR-RBF	0.082	0.095	1	Linear reg	0.079	0.121	1	SVR-RBF	0.088	0.141
	1	Linear reg	0.082	0.072	1	GLS	0.079	0.123	1	Linear reg	0.088	0.136
	1	GLS	0.082	0.069	1	Reg-ARMA	0.079	0.118	1	GLS	0.088	0.133
	1	Reg-ARMA	0.082	0.067	1	GLM	0.080	0.090	1	Reg-ARMA	0.089	0.131
	1	GLMM	0.082	-0.056	1	GLMM	0.080	0.064	1	GLM	0.089	0.123
	1	GAMM	0.083	-0.069	1	Reg-ARMA-GARCH	0.080	0.062	1	SVR-L	0.089	0.095
	1	MARSS-P	0.083	-0.082	2	RF	0.081	-0.090	1	RF	0.089	0.118
	1	Reg-ARMA-GARCH	0.084	-0.087	2	SVR-L	0.082	-0.072	1	GLMM	0.090	0.087
	1	GAM	0.084	-0.087	2	GAMM	0.082	-0.097	2	GAM	0.095	-0.231
	1	RF	0.084	-0.108	2	GAM	0.083	-0.123	3	GAMM	0.096	-0.321
	1	SVR-L	0.085	-0.108	3	MARSS-P	0.085	-0.187	4	MARSS-P	0.104	-0.413
Linear	1	GLM	0.081	0.103	1	GLM	0.082	0.105	1	GAM	0.081	0.182
	1	MARSS-M	0.081	0.100	1	Linear reg	0.082	0.105	1	GLM	0.081	0.167
	1	SVR-RBF	0.081	0.121	1	GLS	0.082	0.103	1	GLMM	0.081	0.164
	1	GLMM	0.081	0.067	1	Reg-ARMA	0.082	0.108	1	Reg-ARMA	0.082	0.185
	1	Linear reg	0.081	0.100	1	GLMM	0.082	0.085	1	Linear reg	0.082	0.190
	1	GLS	0.081	0.097	1	MARSS-M	0.083	0.108	1	GLS	0.082	0.187
	1	Reg-ARMA	0.081	0.103	1	GAMM	0.083	0.082	1	GAMM	0.082	0.154
	1	GAMM	0.082	-0.074	1	GAM	0.083	0.097	2	Reg-ARMA-	0.083	0.144

										GARCH		
	1	MARSS-P	0.082	0.062	2	Reg-ARMA-GARCH	0.085	-0.090	3	SVR-L	0.085	0.131
	1	Reg-ARMA-GARCH	0.083	-0.077	2	SVR-RBF	0.085	-0.092	4	RF	0.089	-0.138
	2	GAM	0.083	-0.092	2	MARSS-P	0.086	-0.133	5	MARSS-M	0.092	-0.164
	3	SVR-L	0.085	-0.146	2	SVR-L	0.087	-0.179	6	SVR-RBF	0.125	-0.523
	4	RF	0.087	-0.200	3	RF	0.087	-0.185	7	MARSS-P	0.147	-0.823
Nonlinear	1	GLS	0.085	0.103	1	GLM	0.075	0.110	1	RF	0.086	0.433
	1	Linear reg	0.085	0.100	1	MARSS-M	0.075	0.113	2	SVR-L	0.088	0.395
	1	Reg-ARMA	0.085	0.095	1	GLMM	0.075	0.103	3	MARSS-M	0.092	0.249
	1	GLM	0.085	0.087	1	Reg-ARMA-GARCH	0.076	0.103	4	Reg-ARMA	0.095	0.290
	1	MARSS-M	0.085	0.077	1	Linear reg	0.077	0.105	4	GLM	0.096	0.233
	1	GLMM	0.086	0.069	1	GLS	0.077	0.103	4	Reg-ARMA-GARCH	0.096	0.279
	1	RF	0.087	0.087	1	Reg-ARMA	0.077	0.100	5	Linear reg	0.098	0.249
	1	SVR-RBF	0.087	-0.067	2	SVR-RBF	0.077	0.105	5	GLS	0.098	0.246
	1	GAMM	0.087	-0.077	2	SVR-L	0.078	-0.105	6	GLMM	0.100	0.185
	1	MARSS-P	0.087	-0.105	3	GAMM	0.079	-0.151	7	SVR-RBF	0.119	-0.364
	1	GAM	0.087	-0.105	4	MARSS-P	0.080	-0.149	8	GAM	0.122	-0.367
	1	Reg-ARMA-GARCH	0.088	-0.115	5	GAM	0.082	-0.208	9	GAMM	0.145	-0.646
	2	SVR-L	0.090	-0.146	5	RF	0.083	-0.195	10	MARSS-P	0.187	-0.838
Periodic	1	RF	0.080	0.221	1	RF	0.081	0.513	1	RF	0.087	0.854
	2	SVR-RBF	0.085	0.126	2	GAM	0.089	0.259	2	GAM	0.111	0.628
	2	SVR-L	0.086	0.097	3	SVR-RBF	0.091	0.228	3	SVR-RBF	0.128	0.369
	2	GAM	0.086	0.097	3	SVR-L	0.091	0.269	4	SVR-L	0.141	-0.195
	3	GLMM	0.089	-0.067	4	GLMM	0.096	-0.108	5	GLMM	0.142	-0.156
	3	GLM	0.090	-0.072	5	GAMM	0.097	-0.123	6	GAMM	0.145	-0.169
	3	GAMM	0.090	-0.072	5	GLM	0.097	-0.131	7	MARSS-P	0.145	-0.179

	3	MARSS-P	0.090	-0.059	5	MARSS-P	0.097	-0.136	7	Reg-ARMA-GARCH	0.146	-0.182
	3	Reg-ARMA-GARCH	0.090	-0.082	5	MARSS-M	0.097	-0.149	7	Reg-ARMA	0.146	-0.205
	3	Linear reg	0.090	-0.072	5	Reg-ARMA-GARCH	0.097	-0.149	7	GLS	0.146	-0.203
	3	GLS	0.090	-0.072	5	Linear reg	0.097	-0.144	7	Linear reg	0.146	-0.208
	3	Reg-ARMA	0.090	-0.069	5	GLS	0.097	-0.141	7	GLM	0.146	-0.269
	3	MARSS-M	0.090	-0.126	5	Reg-ARMA	0.097	-0.144	7	MARSS-M	0.147	-0.290
Step	1	RF	0.073	0.162	1	RF	0.076	0.177	1	SVR-RBF	0.076	0.105
	2	SVR-RBF	0.076	0.087	1	SVR-RBF	0.077	0.105	1	GLM	0.076	0.079
	3	GLM	0.078	0.072	2	GLM	0.078	0.067	1	MARSS-M	0.076	0.072
	3	MARSS-M	0.078	-0.062	2	MARSS-M	0.078	0.062	1	MARSS-P	0.076	0.064
	3	GLMM	0.078	0.062	2	MARSS-P	0.078	0.062	1	Reg-ARMA	0.077	0.062
	3	MARSS-P	0.078	-0.033	2	GLS	0.078	0.046	1	Linear reg	0.077	0.059
	3	Linear reg	0.078	-0.041	2	Linear reg	0.078	0.051	1	GLS	0.077	0.056
	3	GLS	0.078	-0.044	2	Reg-ARMA	0.078	0.051	1	GLMM	0.077	-0.059
	3	Reg-ARMA	0.078	-0.044	2	GLMM	0.079	-0.067	1	GAMM	0.077	-0.044
	3	GAMM	0.078	-0.069	2	GAMM	0.079	-0.064	1	GAM	0.077	-0.064
	3	GAM	0.078	-0.067	2	GAM	0.079	-0.105	1	RF	0.078	-0.118
	4	Reg-ARMA-GARCH	0.080	-0.092	3	Reg-ARMA-GARCH	0.080	-0.118	2	Reg-ARMA-GARCH	0.079	-0.100
	5	SVR-L	0.081	-0.149	4	SVR-L	0.081	-0.149	3	SVR-L	0.079	-0.128

References

Achen, C., 2000. Why lagged dependent variables can suppress the explanatory power of other independent variables. Presented at the Annual Meeting of the Society for Political Methodology, UCLA.

Alain F. Zuur, 2009. Mixed effects models and extensions in ecology with R, Statistics for biology and health. Springer, New York, NY.

Allen, P.G., Fildes, R., 2001. Econometric Forecasting, in: Armstrong, J.S. (Ed.), Principles of Forecasting, International Series in Operations Research & Management Science. Springer, pp. 303–362.

Alpaydin, E., 2004. Introduction to machine learning, Adaptive computation and machine learning. MIT Press, Cambridge, Mass.

Anderson, M.J., Thompson, A.A., 2004. Multivariate control charts for ecological and environmental monitoring. Ecol. Appl. 14, 1921–1935.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statist. Surv. 4, 40–79.

Bahramy, F., Crone, S.F., 2013. Forecasting foreign exchange rates using support vector regression, in: 2013 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr). Presented at the 2013 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr), pp. 34–41.

Barnabé, G., 2000. Ecology and management of coastal waters: the aquatic environment, Springer-Praxis books in aquaculture and fisheries. Springer, London ; New York.

Barrio Froján, C.R.S., Cooper, K.M., Bolam, S.G., 2016. Towards an integrated approach to marine benthic monitoring. Mar. Pollut. Bull. 104, 20–28.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423.

Belaire, J.A., Kreakie, B.J., Keitt, T., Minor, E., 2014. Predicting and mapping potential whooping crane stopover habitat to guide site selection for wind energy projects. Conserv. Biol. 28, 541–550.

Bell, D.M., Schlaepfer, D.R., 2016. On the dangers of model complexity without ecological justification in species distribution modeling. Ecol. Model. 330, 50–59.

Benedetti-Cecchi, L., 2001. Beyond Baci: Optimization of environmental sampling designs through monitoring and simulation. Ecol. Appl. 11, 783–799.

Ben-Hur, A., Weston, J., 2010. A user's guide to support vector machines. Methods Mol. Biol. 609, 223–239.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Modell. Softw. 40, 1–20.

Bergström, L., Sundqvist, F., Bergström, U., 2013. Effects of an offshore wind farm on temporal and spatial patterns in the demersal fish community. Mar. Ecol. Prog. Ser. 485, 199–210.

Berk, R.A., 2008. Statistical learning from a regression perspective, Springer series in statistics. Springer Verlag, New York, NY.

Biglan, A., Ary, D., Wagenaar, A.C., 2000. The value of interrupted time-series experiments for community intervention research. Prev. Sci. 1, 31–49.

Boehlert, G.W., Braby, C., Bull, A.S., Helix, M.E., Henkel, S., Klarin, P., Schroeder, D., 2013. Proceedings of the Oregon Marine Renewable Energy Environmental Sciences Conference (Technical Report). U.S. Bureau of Ocean Energy Management.

Boehlert, G.W., Gill, A.B., 2010. Environmental and ecological effects of ocean renewable energy development: a current synthesis. Oceanography 23, 68–81.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. Roy. Stat. Soc. B Met. 26, 211–252.

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc. 70, 70–79.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Breiman, L., Cutler, A., 2004. Random Forests. Version 5.1. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Broadhurst, M., Barr, S., Orme, C.D.L., 2014. In-situ ecological interactions with a deployed tidal energy device; an observational pilot study. Ocean. Coast. Manage., Science in support of governance of wave and tidal energy developments 99, 31–38.

Broadhurst, M., Orme, C.D.L., 2014. Spatial and temporal benthic species assemblage responses with a deployed marine tidal energy device: a small scaled study. Mar. Environ. Res. 99, 76–84.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model Selection: An Integral Part of Inference. Biometrics 53, 603–618.

Bureau of Ocean Energy Management, 2016. Guidelines for information requirements for a renewable energy construction and operations plan (COP). U.S. Department of the Interior, Bureau of Ocean Energy Management, Office of Renewable Energy Programs.

Burgos, J.M., Horne, J.K., 2007. Sensitivity analysis and parameter selection for detecting aggregations in acoustic data. ICES J. Mar. Sci. 64, 160–168.

Burnham, K., Anderson, D., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag New York, Incorporated, Secaucus.

Cappé, O., 2005. Inference in Hidden Markov Models, Springer series in statistics. Springer, New York, NY.

Carey, J.M., Keough, M.J., 2002. The variability of estimates of variance, and its effect on power analysis in monitoring design. Environ. Monit. Assess. 74, 225–241.

Chai, T., Draxler, R.R., Chai, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250.

Chandler, R.E., Scott, M., 2011. Statistical methods for trend detection and analysis in the environmental sciences, Statistics in practice; 90. Wiley, Hoboken, NJ.

Chapman, M.G., Underwood, A.J., Skilleter, G.A., 1995. Variability at different spatial scales between a subtidal assemblage exposed to the discharge of sewage and two control assemblages. J. Exp. Mar. Biol. Ecol. 189, 103–122.

Chatfield, C., 1989. The analysis of time series: an introduction, 4th ed. Chapman and Hall, London ; New York.

Clark, J.S., Carpenter, S.R., Barber, M., Collins, S., Dobson, A., Foley, J.A., Lodge, D.M., Pascual, M., Pielke, R., Pizer, W., Pringle, C., Reid, W.V., Rose, K.A., Sala, O., Schlesinger, W.H., Wall, D.H., Wear, D., 2001. Ecological forecasts: an emerging imperative. Science 293, 657–660.

Clarke, B.S., 2009. Principles and theory for data mining and machine learning, Springer series in statistics. Springer, Dordrecht ; New York.

Copping, A., Hanna, L., Hutchinson, I., 2014. Best practices for monitoring environmental effects of marine energy devices. Aquatera Ltd and Pacific Northwest National Laboratory.

Copping, A., Sather, N., Hanna, L., Whiting, J., Zydlewski, G., Staines, G., Gill, A., Hutchison, I., O'Hagan, A., Simas, T., Bald, J., Sparling, C., Wood, J., Masden, E., 2016. Annex IV 2016 State of the Science Report: Environmental Effects of Marine Renewable Energy Development around the World.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Croll, P., Andina-Pendas, I., 2009. Guidelines for Project Development in the Marine Energy Industry. The European Marine Energy Centre Ltd, London, UK.

Crone, S.F., Guajardo, J., Weber, R., 2006. A study on the ability of support vector regression and neural networks to forecast basic time series patterns, in: Bramer, M. (Ed.), Artificial Intelligence in Theory and Practice, IFIP International Federation for Information Processing. Springer US, pp. 149–158.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.

David E. Busch, Joel C. Trexler, 2003. Monitoring ecosystems: interdisciplinary approaches for evaluating ecoregional initiatives. Island Press, Washington, DC.

De Valpine, P., Hastings, A., 2002. Fitting population models incorporating process noise and observation error. Ecol. Monogr. 72, 57–76.

Dennis, B., Ponciano, J.M., Lele, S.R., Taper, M.L., Staples, D.F., 2006. Estimating density dependence, process noise, and observation error. Ecol. Monogr. 76, 323–341.

Diffendorfer, J.E., Beston, J.A., Merrill, M.D., Stanton, J.C., Corum, M.D., Loss, S.R., Thogmartin, W.E., Johnson, D.H., Erickson, R.A., Heist, K.W., 2015. Preliminary methodology to assess the national and regional impact of U.S. wind energy development on birds and bats (USGS Numbered Series No. 2015–5066), Scientific Investigations Report. U.S. Geological Survey, Reston, VA.

Diggle, P., 1994. Analysis of longitudinal data, Oxford statistical science series ; 13. Clarendon Press ; New York, Oxford.

Doornik, J.A., Ooms, M., 2008. Multimodality in GARCH regression models. Int. J. Forecasting 24, 432–448.

Doornik, J., Ooms, M., 2005. Outlier detection in GARCH Models (Economics Paper No. 2005-W24). Economics Group, Nuffield College, University of Oxford.

Dornelas, M., Magurran, A.E., Buckland, S.T., Chao, A., Chazdon, R.L., Colwell, R.K., Curtis, T., Gaston, K.J., Gotelli, N.J., Kosnik, M.A., McGill, B., McCune, J.L., Morlon, H., Mumby, P.J., Øvreås, L., Studeny, A., Vellend, M., 2013. Quantifying temporal change in biodiversity: challenges and opportunities. Proc. R. Soc. B 280, 20121931.

Drake, J.M., Randin, C., Guisan, A., 2006. Modeling ecological niches with support vector machines. J. Appl. Ecol. 43, 424–432.

Dubbs, L., Keeler, A.G., O'Meara, T., 2013. Permitting, risk and marine hydrokinetic energy development. The Electricity Journal 26, 64–74.

Duck, C., Black, A., Lonergan, M., Mackey, B., 2006. The number and distribution of marine mammals in the Fall of Warness, Orkney July 2005-July 2006.

Eitrich, T., Lang, B., 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. J. Comput. Appl. Math. 196, 425–436.

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77.

Embling, C.B., Sharples, J., Armstrong, E., Palmer, M.R., Scott, B.E., 2013. Fish behaviour in response to tidal variability and internal waves over a shelf sea bank. Prog. Oceanogr., From Physics to Fishing over a Shelf Sea Bank 117, 106–117.

Engle, R., 2001. GARCH 101: The use of ARCH/GARCH models in applied econometrics. J. Econ. Perspect. 15, 157–168.

Fairweather, P., 1991. Statistical power and design requirements for environmental monitoring. Mar. Freshwater Res. 42, 555–567.

Federal Energy Regulatory Commission, 2008. Licensing Hydrokinetic Pilot Projects.

Ferretti, M., 1997. Forest health assessment and monitoring – issues for consideration. Environ. Monit. Assess. 48, 45–72.

Field, S.A., Tyre, A.J., Possingham, H.P., 2005. Optimizing allocation of monitoring effort under economic and observational constraints. J. Wildl. Manage. 69, 473–482.

Fodrie, F.J., Able, K.W., Galvez, F., Heck, K.L., Jensen, O.P., López-Duarte, P.C., Martin, C.W., Turner, R.E., Whitehead, A., 2014. Integrating organismal and population responses of estuarine fishes in Macondo spill research. BioScience 64, 778–788.

Foley, A.M., Leahy, P.G., Marvuglia, A., McKeogh, E.J., 2012. Current methods and advances in forecasting of wind power generation. Renew. Energ. 37, 1–8.

Ghahramani, Z., Hinton, G.E., 2000. Variational learning for switching state-space models. Neural Comput. 12, 831–864.

Ghalanos, A., 2015. Introduction to the rugarch package. R package Version 1.3-1. https://cran.r-project.org/web/packages/rugarch

Gilmour, S., Degenhardt, L., Hall, W., Day, C., 2006. Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: a general method and example. BMC Med. Res. Methodol. 6, 16.

Gitzen, R.A., Millspaugh, J.J., Cooper, A.B., Licht, D.S., 2012. Design and Analysis of Long-term Ecological Monitoring Studies. Cambridge University Press, Cambridge.

Granger, C.W.J., Newbold, P., 1974. Spurious regressions in econometrics. J. Econometrics 2, 111–120.

Grilli, A.R., Shumchenia, E.J., 2014. Toward wind farm monitoring optimization: assessment of ecological zones from marine landscapes using machine learning algorithms. Hydrobiologia 756, 117–137.

Gutiérrez, N.L., Hilborn, R., Defeo, O., 2011. Leadership, social capital and incentives promote successful fisheries. Nature 470, 386–389.

Hammar, L., Andersson, S., Eggertsen, L., Haglund, J., Gullström, M., Ehnberg, J., Molander, S., 2013. Hydrokinetic turbine effects on fish swimming behaviour. PLOS ONE 8, e84141.

Hampton, S.E., Holmes, E.E., Scheef, L.P., Scheuerell, M.D., Katz, S.L., Pendleton, D.E., Ward, E.J., 2013. Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (MAR) models. Ecology 94, 2663–2669.

Hansen, J.V., McDonald, J.B., Nelson, R.D., 2006. Some evidence on forecasting time-series with support vector machines. J. Oper. Res. Soc. 57, 1053–1063.

Hastie, T., 2009. The elements of statistical learning: data mining, inference, and prediction, Second edition, Corrected 7th printing. ed. Springer series in statistics. Springer, New York.

Hayes, M.A., Cryan, P.M., Wunder, M.B., 2015. Seasonally-dynamic presence-only species distribution models for a cryptic migratory bat impacted by wind energy development. PLOS ONE 10, e0132599.

Hedeker, D.R., 2006. Longitudinal data analysis, Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ.

Hewitt, J.E., Thrush, S.E., Cummings, V.J., 2001. Assessing environmental impacts: effects of spatial and temporal variability at likely impact scales. Ecol. Appl. 11, 1502–1516.

Holmes, E., Ward, E., Wills, K., 2015. MARSS: Multivariate Autoregressive Sate-Space Modeling. R Package Version 3.9. https://cran.r-project.org/web/packages/MARSS

Holmes, E.E., Ward, E.J., Scheuerell, M.D., 2014. Analysis of multivariate time-series using the MARSS package Version 3.9. https://cran.r-project.org/web/packages/MARSS

Holmes, E.E., Ward, E.J., Wills, K., 2012. MARSS: Multivariate autoregressive state-space models for analyzing time-series data. R J. 4, 11–19.

Horne, J., Jacques, D.A., Parker-Stetter, S.L., Linder, H.L., Nomura, J.M., 2013. Evaluating Acoustic Technologies to Monitor Aquatic Organisms at Renewable Energy Sites. U.S. Department of the Interior, Bureau of Ocean Energy Management.

Horne, J.K., Hytnen, R., Maxwell, S., Ham, K., Maxwell, A., Condiotty, J., 2016. Fully utilizing the acoustic record for biological monitoring and ecological applications. J. Acoust. Soc. Am. 139, 2173–2173.

Hsu, C., Chang, C., Lin, C., 2010. A practical guide to support vector classification (technical report). National Taiwan University.

Hu, W., Kercheval, A.N., 2008. The skewed t distribution for portfolio credit risk. Adv. Econom. 22, 55–83.

Hyndman, R., 2015. Forecast: forecasting functions for time series and linear models. R package version 6.2. https://cran.r-project.org/web/packages/forecast

Hyndman, R., Athanasopoulos, G., 2014. Forecasting: Principles and Practice. otexts.

Hyndman, R., Kostenko, A.V., 2007. Minimum sample size requirements for seasonal forecasting models. Foresight. 12–15.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Inger, R., Attrill, M.J., Bearhop, S., Broderick, A.C., James Grecian, W., Hodgson, D.J., Mills, C., Sheehan, E., Votier, S.C., Witt, M.J., Godley, B.J., 2009. Marine renewable energy: potential benefits to biodiversity? An urgent call for research. J. Appl. Ecol. 46, 1145–1153.

Ingersoll, T.E., Sewall, B.J., Amelon, S.K., 2013. Improved analysis of long-term monitoring data demonstrates marked regional declines of bat populations in the eastern United States. PLOS ONE 8, e65907.

Ishwaran, H., Malley, J.D., 2014. Synthetic learning machines. BioData Min. 7.

Jackson, C., 2016. Multi-state modeling with R: the msm package. R package version 1.6. https://cran.r-project.org/web/packages/msm

Jackson, L.E., Kurtz, J., Fisher, W.S., 2000. Evaluation Guidelines for Ecol. Indic.. U.S. Environmental Protection Agency, Office of Research and Development.

Jacques, D., 2014. Describing and Comparing Variability of Fish and Macrozooplankton Density at Marine Hydrokinetic Energy Sites. University of Washington.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2015. An introduction to statistical learning: with applications in R, Corrected at 6th printing. ed. Springer texts in statistics. Springer, New York.

Jay, S., Jones, C., Slinn, P., Wood, C., 2007. Environmental impact assessment: Retrospect and prospect. Environ. Impact. Asses. 27, 287–300.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2004. A review and evaluation of intraurban air pollution exposure models. J. Expo. Anal. Environ, Epidemiol. 15, 185–204.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.

Jonathan D. Cryer, 2008. Time series analysis: with applications in R, 2nd ed. Springer texts in statistics. Springer, New York.

Jones-Farrand, D.T., Fearer, T.M., Thogmartin, W.E., Iii, F.R.T., Nelson, M.D., Tirpak, J.M., 2011. Comparison of statistical and theoretical habitat models for conservation planning: the benefit of ensemble prediction. Ecol. Appl. 21, 2269–2282.

Kacprzyk, J., Pedrycz, W., 2015. Springer Handbook of Computational Intelligence. Springer Berlin Heidelberg: Berlin, Heidelberg.

Kane, M.J., Price, N., Scotch, M., Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 15, 276.

Kellenberger, K., Groom, C., 2015. Discovering Rank Functions, in: Expert T-SQL Window Functions in SQL Server. Apress, Berkeley, CA.

Klure, J., Hampton, T., McMurray, G., Boehlert, G., Henkel, S., Copping, A., Kramer, S., Chwaszczewski, R., Fresh, K., 2012. West Coast Environmental Protocols Framework: Baseline and Monitoring Studies. Pacific Energy Ventures LLC.

Kordon, A., 2009. Applying Computational Intelligence: How to Create Value. Springer Science & Business Media.

Kurtz, J.C., Jackson, L.E., Fisher, W.S., 2001. Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development. Ecol. Indic. 1, 49–60.

Lagarde, M., 2011. How to do (or not to do) … Assessing the impact of a policy change with routine longitudinal data. Health Policy Plan. czr004.

Legg, C.J., Nagy, L., 2006. Why most conservation monitoring is, but need not be, a waste of time. J. Environ. Manage. 78, 194–199.

Levine, C.R., Yanai, R.D., Lampman, G.G., Burns, D.A., Driscoll, C.T., Lawrence, G.B., Lynch, J.A., Schoch, N., 2014. Evaluating the efficiency of environmental monitoring programs. Ecol. Indic. 39, 94–101.

Lewis, A., Estefen, S., Huckerby, J., Lee, K.S., Musial, W., Pontes, T., Torres-Martinez, J., Bharathan, D., Hanson, H., Heath, G., Louis, F., Scråmestø, S., Abdulla, A., Moreno, J.M., You, Y., 2011. Ocean Energy, in: Renewable Energy Sources and Climate Change Mitigation. Cambridge University Press.

Lewis-Beck, M.S., Bryman, A., Liao, T.F., 2004. The Sage encyclopedia of social science research methods, SAGE E-Reference Encyclopedias. Sage, Thousand Oaks, Calif.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Lindenmayer, D.B., Gibbons, P., Bourke, M., Burgman, M., Dickman, C.R., Ferrier, S., Fitzsimons, J., Freudenberger, D., Garnett, S.T., Groves, C., Hobbs, R.J., Kingsford, R.T., Krebs, C., Legge, S., Lowe, A.J., Mclean, R., Montambault, J., Possingham, H., Radford, J., Robinson, D., Smallbone, L., Thomas, D., Varcoe, T., Vardon, M., Wardle, G., Woinarski, J., Zerger, A., 2012. Improving biodiversity monitoring. Austral Ecol. 37, 285–294.

Lindley, S.T., 2003. Estimation of population growth and extinction parameters from noisy data. Ecol. Appl. 13, 806–813.

Liu, H., Erdem, E., Shi, J., 2011. Comprehensive evaluation of ARMA–GARCH(-M) approaches for modeling the mean and volatility of wind speed. Appl. Energy 88, 724–732.

Liu, L., Shen, B., Wang, X., 2014. Research on Kernel Function of Support Vector Machine, in: Advanced Technologies, Embedded and Multimedia for Human-Centric Computing. Springer, pp. 827–834.

Lorena, A.C., Jacintho, L.F.O., Siqueira, M.F., Giovanni, R.D., Lohmann, L.G., de Carvalho, A.C.P.L.F., Yamamoto, M., 2011. Comparing machine learning classifiers in potential distribution modeling. Expert Syst. Appl. 38, 5268–5275.

Lovett, G.M., Burns, D.A., Driscoll, C.T., Jenkins, J.C., Mitchell, M.J., Rustad, L., Shanley, J.B., Likens, G.E., Haeuber, R., 2007. Who needs environmental monitoring? Front. Ecol. Environ. 5, 253–260.

Maas, C.J., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1, 86–92.

Mackenzie, M.L., Scott-Hayward, L.A., Oedekoven, C.S., Skov, H., Humphreys, E., Rexstad, E., 2013. Statistical modeling of seabird and cetacean data: guidance document (Guidance document). University of St. Andrews.

Maclean, I.M.D., Inger, R., Benson, D., Booth, C.G., Embling, C.B., Grecian, W.J., Heymans, J.J., Plummer, K.E., Shackshaft, M., Sparling, C.E., Wilson, B., Wright, L.J., Bradbury, G., Christen, N., Godley, B.J., Jackson, A.C., McCluskie, A., Nicholls-Lee, R., Bearhop, S., 2014. Resolving issues with environmental impact assessment of marine renewable energy installations. Front. Mar. Sci 1, 75.

Maclean, I.M.D., Rehfisch, M.M., Skov, H., Thaxter, C.B., 2013. Evaluating the statistical power of detecting changes in the abundance of seabirds at sea. Ibis 155, 113–126.

Maclennan, D.N., Fernandes, P.G., Dalen, J., 2002. A consistent approach to definitions and symbols in fisheries acoustics. ICES J. Mar. Sci. 59, 365–369.

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J., Watt, A.D., 2010. Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. Trends Ecol. Evol. 25, 574–582.

Manolov, R., Arnau, J., Solanas, A., Bono, R., 2010. Regression-based techniques for statistical decision making in single-case designs. Psicothema 22, 1026–1032.

Matsumoto, W.M., Kazama, T.K., Aasted, D.C., 1981. Anchored fish aggregating devices in Hawaiian waters. Mar. Fish. Rev. 43, 1–13.

McCann, J., 2012. Developing Environmental Protocols and Modeling Tools to Support Ocean Renewable Energy and Stewardship. U.S. Department of the Interior, Bureau of Ocean Energy Management, Office of Renewable Energy Programs, Herndon, VA.

Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281.

Michel, L., Makowski, D., 2013. Comparison of statistical models for analyzing wheat yield time series. PLOS ONE 8, e78615.

Morrison, L.W., 2007. Assessing the reliability of ecological monitoring data: power analysis and alternative approaches. Nat. Area. J. 27, 83–91.

Munkittrick, K.R., Arens, C.J., Lowell, R.B., Kaminski, G.P., 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. Environ. Toxicol. Chem. 28, 1361–1371.

Newman, K.B., Buckland, S.T., Lindley, S.T., Thomas, L., Fernández, C., 2006. Hidden process models for animal population dynamics. Ecol. Appl. 16, 74–86.

Nichols, J., Williams, B., 2006. Monitoring for conservation. Trends Ecol. Evol. 21, 668–673.

Niemi, G.J., McDonald, M.E., 2004. Application of ecological indicators. Annu. Rev. Ecol. Evol. Syst. 35, 89–111.

Noss, R.F., 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conserv. Biol. 4, 355–364.

Nuno, A., Milner-Gulland, E.J., Bunnefeld, N., 2015. Detecting abundance trends under uncertainty: the influence of budget, observation error and environmental change. Anim. Conserv. 18, 331–340.

Olden, J.D., Jackson, D.A., 2002. A comparison of statistical approaches for modeling fish species distributions. Freshwater biol. 47, 1976–1995.

ORPC Maine LLC, 2014. Cobscook Bay Tidal Energy Project: 2013 Environmental Monitoring Report. Ocean Renewable Power Company (ORPC).

Osenberg, C.W., Schmitt, R.J., Holbrook, S.J., Abu-Saba, K.E., Flegal, A.R., 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecol. Appl. 4, 16–30.

Parker, R.I., Vannest, K.J., Davis, J.L., Sauber, S.B., 2011. Combining nonoverlap and trend for single-case research: Tau-U. Behav. Ther. 42, 284–299.

Pattengill-Semmens, C.V., Semmens, B.X., Holmes, E.E., Ward, E.J., Ruttenberg, B.I., 2011. Integrating time-series of community monitoring data using multivariate state-space models, in: Proceedings of the Sixty Three Annual Gulf and Caribbean Fisheries Institute. San Juan, Puerto Rico, pp. 214–216.

Perera, K.S., Aung, Z., Woon, W.L., 2014. Machine learning techniques for supporting renewable energy generation and integration: a survey, in: Woon, W.L., Aung, Z., Madnick, S. (Eds.), Data Analytics for Renewable Energy Integration, Lecture Notes in Computer Science. Springer International Publishing, pp. 81–96.

Perretti, C.T., Sugihara, G., Munch, S.B., 2013. Nonparametric forecasting outperforms parametric methods for a simulated multispecies system. Ecology 94, 794–800.

Petersen, I.K., MacKenzie, M.L., Rexstad, E., Wisz, M.S., Fox, A.D., 2011. Comparing pre- and post-construction distributions of long-tailed ducks *Clangula hyemalis* in and around the Nysted offshore wind farm, Denmark : a quasi-designed experiment accounting for imperfect detection, local surface features and autocorrelation. University of St. Andrews.

Pinheiro, J.C., Bates, D.M., 2010. Mixed-Effects Models in S and S-PLUS. Springer Science & Business Media.

Plümper, T., Troeger, V.E., Manow, P., 2005. Panel data analysis in comparative politics: Linking method to theory. Euro. J. Polit. Res. 44, 327–354.

Polagye, B., Copping, A., Suryan, R., Kramer, S., Brown-Saracino, J., Smith, C., 2014. Instrumentation for monitoring around marine renewable energy converters: workshop final report. Pacific Northwest National Laboratory, Seattle, Washington.

Polagye, B., Van Cleve, B., Copping, A., Kirkendall, K., 2011. Environmental effects of tidal energy development: proceedings of a scientific workshop (NOAA technical memorandum). U.S. Department of Commerce, Tidal Energy Workshop, Seattle, Washington.

Polgreen, L.A., Brooks, J.M., 2012. Estimating incremental costs with skew: a cautionary note. Appl. Health Econ. Health Policy 10, 319–329.

Portman, M.E., 2010. Marine renewable energy policy. Oceanography 23, 98.

Prado, J., 2005. Fishing technology equipments: fish aggregating device (FAD). FAO Fisheries and Aquaculture Department. http://www.fao.org/fishery/equipment/fad/en

Public Utility District No. 1 of Snohomish County, 2012. Admiralty Inlet Tidal Project Final Monitoring and Mitigation Plans.

R Development Core Team, 2016. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., 2011. GARCH Models, in: Statistics and Data Analysis for Financial Engineering. Springer New York, New York, NY, pp. 477–504.

Rush, B., Joslin, J., Stewart, A., Polagye, B., 2014. Development of an adaptable monitoring package for marine renewable energy projects part I: conceptual design and operation, in: Proceedings of the 2nd Marine Energy Technology Symposium. Presented at the Marine Energy Technology Symposium.

Schabenberger, O., 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton.

Scheiner, S.M., Gurevitch, J., 2001. Design and analysis of ecological experiments, 2nd ed. Oxford University Press, Oxford ; New York.

Schmitt, R.J., Osenberg, C.W., 1996. Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats. Academic Press.

See, K.E., Holmes, E.E., 2015. Reducing bias and improving precision in species extinction forecasts. Ecol. Appl. 25, 1157–1165.

Shields, M.A., Payne, A.I.L., 2014. Marine renewable energy technology and environmental interactions, Humanity and the sea. Springer, Dordrecht ; New York.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310.

Shumchenia, E.J., Smith, S.L., McCann, J., Carnevale, M., Fugate, G., Kenney, R.D., King, J.W., Paton, P., Schwartz, M., Spaulding, M., Winiarski, K.J., 2012. An adaptive framework for selecting environmental monitoring protocols to support ocean renewable energy development,

an adaptive framework for selecting environmental monitoring protocols to support ocean renewable energy development. Scientific World J., e450685.

Simmonds, E.J., 2005. Fisheries acoustics: theory and practice, 2nd ed. Fish and aquatic resources series; 10. Blackwell Science, Oxford; Ames, Iowa.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.

Spellman, F.R., 2015. Environmental impacts of renewable energy, Energy and the environment. CRC Press/Taylor & Francis, Boca Raton.

Stenberg, C., Støttrup, J., van Deurs, M., Berg, C., Dinesen, G., Mosegaard, H., Grome, T., Leonhard, S., 2015. Long-term effects of an offshore wind farm in the North Sea on fish communities. Mar. Ecol. Prog. Ser. 528, 257–265.

Stewart-Oaten, A., Bence, J.R., 2001. Temporal and spatial variation in environmental impact assessment. Ecol. Monogr. 71, 305–339.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods 14, 323–348.

Taddy, M.A., Gramacy, R.B., Polson, N.G., 2011. Dynamic trees for learning and design. J. Am. Stat. Assoc. 106, 109–123.

Tay, F.E.H., Cao, L., 2001. Application of support vector machines in financial time series forecasting. Omega 29, 309–317.

Taylor, J.W., McSharry, P.E., Buizza, R., 2009. Wind power density forecasting using ensemble predictions and time series models. IEEE T. Energy Conver.24, 775–782.

Terrill, S., Kramer, S., Nelson, P., Zajanc, D., 2009. Baseline data and power analysis for the OWET Dungeness crab and fish baseline study. Oregon Wave Energy Trust.

Thissen, U., van Brakel, R., de Weijer, A.P., Melssen, W.J., Buydens, L.M.C., 2003. Using support vector machines for time series prediction. Chemometr. Intell. Lab. 69, 35–49.

Thomas, L., 1996. Monitoring long-term population change: why are there so many analysis methods? Ecology 77, 49–58.

Thomas, L., Martin, K., 1996. The importance of analysis method for breeding bird survey population trend estimates. Conserv. Biol. 10, 479–490.

Tollit, D., Redden, A., 2013. Passive acoustic monitoring of cetacean activity patterns and movements in Minas passage: pre-turbine baseline conditions. Acadia Centre for Estuarine Research (ACER), Acadia University, Offshore Energy Research Association of Nova Scotia (OERA), and SMRU Consulting.

Treweek, J., 2009. Ecological Impact Assessment. Wiley, Hoboken.

Treweek, J., 1996. Ecology and environmental impact assessment. J. Appl. Ecol. 33, 191–199.

Underwood, A.J., 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecol. Appl. 4, 4–15.

Underwood, A.J., 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. J. Exp. Mar. Biol. Ecol. 161, 145–178.

Underwood, A.J., 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Mar. Freshwater Res. 42, 569–587.

Urmy, S.S., Horne, J.K., Barbee, D.H., 2012. Measuring the vertical distributional variability of pelagic fauna in Monterey Bay. ICES J. Mar. Sci. 69, 184–196.

US Department of Energy, 2009. Report to Congress on the Potential Environmental Effects of Marine and Hydrokinetic Energy Technologies. Federal Energy Regulatory Commission (FERC), National Oceanic and Atmospheric Administration (NOAA), US Department of Energy (DOE), and US Department of the Interior (DOI).

Vanermen, N., Onkelinx, T., Verschelde, P., Courtens, W., Walle, M.V. de, Verstraete, H., Stienen, E.W.M., 2015. Assessing seabird displacement at offshore wind farms: power ranges of a monitoring and data handling protocol. Hydrobiologia 756, 155–167.

Vapnik, V.N., 1995. The nature of statistical learning theory. Springer, New York.

Viehman, H.A., Zydlewski, G.B., McCleave, J.D., Staines, G.J., 2014. Using hydroacoustics to understand fish presence and vertical distribution in a tidally dynamic region targeted for energy extraction. Estuar. Coast. 38, 215–226.

Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE T. Inform. Theory 13, 260–269.

Vos, P., Meelis, E., Ter Keurs, W.J., 2000. A framework for the design of ecological monitoring programs as a tool for environmental and nature management. Environ. Monit. Assess. 61, 317–344.

Wade, H.M., Masden, E.A., Jackson, A.C., Thaxter, C.B., Burton, N.H.K., Bouten, W., Furness, R.W., 2014. Great skua (*Stercorarius skua*) movements at sea in relation to marine renewable energy developments. Mar. Environ. Res. 101, 69–80.

Waggitt, J.J., Bell, P.S., Scott, B.E., 2014. An evaluation of the use of shore-based surveys for estimating spatial overlap between deep-diving seabirds and tidal stream turbines. Int. J. Mar. Energy 8, 36–49.

Wagner, A.K., Soumerai, S.B., Zhang, F., Ross-Degnan, D., 2002. Segmented regression analysis of interrupted time series studies in medication use research. J. Clin. Pharm. Ther. 27, 299–309.

Walker, C.G., Mackenzie, M.L., Donovan, C.R., O'Sullivan, M.J., 2011. SALSA – a spatially adaptive local smoothing algorithm. J. Stat. Comput. Simul. 81, 179–191.

Ward, E.J., Chirakkal, H., González-Suárez, M., Aurioles-Gamboa, D., Holmes, E.E., Gerber, L., 2010. Inferring spatial structure from time-series data: using multivariate state-space models to detect metapopulation structure of California sea lions in the Gulf of California, Mexico. J. Appl. Ecol. 47, 47–56.

Ward, E.J., Holmes, E.E., Thorson, J.T., Collen, B., 2014. Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. Oikos 123, 652–661.

Warwick-Evans, V.C., Atkinson, P.W., Robinson, L.A., Green, J.A., 2016. Predictive Modeling to Identify Near-Shore, Fine-Scale Seabird Distributions during the Breeding Season. PLoS One 11.

White, J.W., Rassweiler, A., Samhouri, J.F., Stier, A.C., White, C., 2014. Ecologists should not use statistical significance tests to interpret simulation model results. Oikos 123, 385–388.

Wiesebron, L.E., Horne, J.K., Hendrix, A.N., 2016. Characterizing biological impacts at marine renewable energy sites. Int. J. Mar. Energy 14, 27–40.

Wiesebron Lauren, E., Horne John, K., Scott Beth, E., Williamson Benjamin, J., 2016. Comparing nekton distributions at two tidal energy sites suggests potential for generic environmental monitoring. Int. J. Mar. Energy. doi:10.1016/j.ijome.2016.07.004

Wolfinger, R., 1993. Covariance structure selection in general mixed models. Commun. Stat. Simulat. 22, 1079–1106.

Wood, S., 2015. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. R Package Version 1.8-12. https://cran.r-project.org/web/packages/mgcv

Wood, S., 2006. Generalized additive models: An introduction with R. CRC press.

Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. 16, 446–453.

Zivot, E., 2009. Practical Issues in the Analysis of Univariate GARCH Models, in: Mikosch, T., Kreiß, J.-P., Davis, R.A., Andersen, T.G. (Eds.), Handbook of Financial Time Series. Springer Berlin Heidelberg, pp. 113–155.

Zucchini, W., 2009. Hidden Markov models for time series: an introduction using R, Monographs on statistics and applied probability (Series); 110. CRC Press/Taylor & Francis Group, Boca Raton.

Zuur, A.F., Pierce, G.J., 2004. Common trends in northeast Atlantic squid time series. J. Sea Res. 52, 57–72.