Uncertainty in Projections of Future Conditions in Marine Ecosystems

Hayley J Bannister

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Animal and Plant Sciences
January 2019

In loving memory of a wonderful grandfather
Hubert R. Hardy

Acknowledgements

I would like to thank the Department of Animal and Plant Sciences and the School of Mathematics and Statistics at the University of Sheffield, as well as the Centre for Environment, Fisheries, and Aquaculture Science (CEFAS), for funding this research. A special thank you goes to the FEMA group at CEFAS for offering me a full-time job to continue my research in marine ecosystem modelling and for providing a great deal of support in the final few months of my PhD.

I am eternally grateful to my PhD supervisors Dr. Tom Webb, Prof. Paul Blackwell, Dr. Kieran Hyder, and Dr. Julia Blanchard for their knowledge, advice, and support over the past four years. Thank you to Tom for guiding me through this entire process and for providing encouragement on an almost daily basis. Thank you to Paul for helping to answer my endless questions about the mathematics and statistics behind this research and for building my confidence in this area. Thank you to Kieran for always checking in to make sure everything was running smoothly, for helping me to look at my research from a different perspective, and for encouraging me to take every opportunity available to me. Lastly, thank you to Julia for the late-night Skype meetings that helped me to better understand the *mizer* model.

A special thank you also goes to everyone involved in the Marine Ecosystems Research Programme and the lab groups of both Tom and Paul for giving me the opportunity to regularly discuss my work and for providing valuable feedback to improve my research. I'd like to thank Dr. Mike Spence for his help throughout my PhD, particularly when running the sensitivity analyses in Chapter 2, as well as all those who distributed and/or completed the online survey described in Chapter 6. Finally, I would like to thank all of my family and friends that have helped me through this process - you know who you are!

I would like to dedicate this work to my parents, who have given me everything and asked for nothing in return, who inspired my love of the ocean, and who have always encouraged me to pursue my passion for marine science. I would also like to dedicate this to my partner, Scott Moakes, for his endless encouragement, patience, love, and support over the past six years; may it continue for many more years to come.

Declaration

The following people were involved in this research project:

Hayley J. Bannister

Prof. Paul G. Blackwell

Dr. Julia L. Blanchard

Dr. Tom J. Webb

Dr. Kieran Hyder

Dr. Michael A. Spence

Chapter 3: Global sensitivity analysis of the *mizer* marine ecosystem model

HJB and MAS conceived the idea for this research;

HJB performed the analyses (with input from MAS and PGB);

HJB wrote the manuscript;

MAS, PGB, TJW, KH, and JLB commented on the drafts of the manuscript.

Chapter 4: Using machine learning to predict the behaviour of the *mizer* marine ecosystem model

HJB conceived the idea for this research (with input from PGB);

HJB performed the analyses (with input from PGB);

HJB wrote the manuscript;

PGB, TJW, and KH commented on the drafts of the manuscript.

Chapter 5: Uncertainty in projections of global and regional sea surface temperature and salinity

HJB conceived the idea for this research;

HJB performed the analyses (with input from PGB);

HJB wrote the manuscript;

PGB, TJW, and KH commented on the drafts of the manuscript.

Chapter 6: Visualising uncertainty in multi-model ensembles

HJB, TJW, PGB, and KH conceived the idea for this research;

HJB performed the analyses (with input from PGB and TJW);

HJB wrote the manuscript;

PGB, TJW, and KH commented on the drafts of the manuscript.

Abstract

The impacts of climate change and anthropogenic pressures on marine ecosystems are becoming of increasing concern, resulting in a growing demand for predictions of future ecosystem conditions to aid the development of robust management solutions. However, the outputs of marine ecosystem models tend to be highly uncertain. Quantifying the impacts of these uncertainties on the model outputs and successfully communicating this information to decision makers and the general public is vital to increasing the credibility of the model outputs and ensuring the projections can make a useful contribution to the decision-making process.

In this thesis, I aim to improve our understanding of the uncertainties associated with the future of marine ecosystems by: (1) identifying the key sources of uncertainty in marine ecosystem models; (2) evaluating the impacts of these uncertainties on projections of key parameters for marine environmental policy; and (3) identifying methods to effectively communicate the outputs of complex marine ecosystem models to a non-specialist audience.

In order to achieve this goal, I have used various methods of sensitivity analysis and machine learning to better understand the behaviour of a widely-used marine ecosystem model known as *mizer* in response to parameter uncertainties (see Chapters 3 and 4). I have also assessed the relative contributions of internal variability, model, and scenario uncertainties to the total variance of the projections of sea surface temperature and salinity from over ten different global climate models (see Chapter 5). Finally, I have conducted an in-depth online survey aimed at identifying the most effective methods for visually communicating the outputs of complex models alongside their associated uncertainties (see Chapter 6). Overall, it is hoped that the information gleaned from this research may be used to help improve model behaviour and increase the confidence that decision-makers have in marine ecosystem models.

Contents

Chapter 1

Introduction

The impacts of climate change and anthropogenic pressures on marine ecosystems are becoming of increasing concern, resulting in a growing demand for predictions of future ecosystem conditions to aid the development of management solutions (Brander et al., 2013). Ecosystem models may be used to predict the likely impacts of both natural and anthropogenic pressures on species biomass and the trophic dynamics between species in an ecosystem over time (Pauly et al., 2000). Forecasting such changes within an ecosystem also enables us to predict the impacts of external pressures on the goods and services provided by marine ecosystems, including fish production, nutrient cycling, and climate regulation (Beaumont et al., 2007).

Ecosystem models vary in structure and complexity from single-species biomass models to multispecies food web models. Single-species models, which typically do not take into account the often complex interactions between species and processes in an ecosystem (although some now include predation mortality rates), are currently widely used in fisheries management, particularly in annual stock assessments (Möllmann et al., 2013). However, recent legislation, such as the EU Marine Strategy Framework Directive (MSFD) (European Commission, 2008b) and Common Fisheries Policy (European Commission, 2013), calls for a more holistic ecosystem-based approach to management that incorporates not only the interactions between species, but also the interactions between different ecosystem processes, services, and pressures (Möllmann et al., 2013). To do this, we need to increase the contribution of more complex marine ecosystem models to the management process. One of the largest barriers to achieving this goal lies in the fact that although complex ecosystem models are designed to depict the interactions between species within an ecosystem in a more realistic manner than simpler single-species models, their performance in producing accurate projections may not always be superior (Morissette, 2005). In some instances, the increase in model complexity results in less informative projections due to an increase in the number of uncertain components included within the model (Morissette, 2005). This uncertainty stems from our fundamental lack of understanding of the important processes and species interactions within an ecosystem, as well as our inability to capture the inherent variability of both human and natural phenomena, thus preventing the accurate parameterisation of all of the components within the model (Morissette, 2005).

Failing to acknowledge the presence of uncertainties in ecosystem models may result in severe ecological and economic consequences due to misguided management decisions (Uusitalo et al., 2015), including species extinctions and unintended fishery collapse (Roughgarden and Smith, 1996). The MSFD attempts to avoid such consequences by requiring all Member States to protect, preserve, and restore the quality of the marine environment wherever possible, whilst also allowing the sustainable use of natural resources based on the 'precautionary principle' (European Commission, 2008b). As the precautionary principle necessitates the prevention of any potential adverse risks to the marine environment, Member States are obliged to explicitly incorporate uncertainty into the decision-making process (Refsgaard et al., 2007). Identifying and understanding the implications of uncertainties in ecosystem models, as well as successfully communicating these uncertainties to decision makers, is therefore vital to ensuring the successful development of robust management solutions for the future (Walker et al., 2003).

More specifically, it is important to ascertain the source, level, and nature of the uncertainties to describe why, how, and to what extent the model projections are uncertain (Walker et al., 2003; Refsgaard et al., 2013). Although the sources of uncertainty within ecosystem models may be well established in the scientific literature, the 'correct' terminology to describe these uncertainties is still an area of discussion. Similarly there is much debate within the literature as to which methods should be used to quantify or qualitatively describe these uncertainties, as well as how to effectively communicate this information to non-specialist audiences, including decision makers and the general public (Wesselink et al., 2015). As a result, few studies in ecosystem modelling have attempted to describe all sources of uncertainty and their impacts on the modelled projections, most likely due to the vast number of uncertainties in ecosystem models and the difficulties associated with disentangling the impacts of multiple sources of uncertainty (Morissette, 2005; Gårdmark et al., 2013).

By synthesising previous research regarding uncertainties in both ecosystem modelling and in other subject areas, such as climate science, we can begin to evaluate and communicate uncertainties more effectively. In Section 1.1, we adopt the terminology of Walker et al. (2003) to describe the numerous sources of uncertainty in ecosystem models. In Section 1.2 we introduce some of the most promising methods for evaluating the impacts of these uncertainties on projections of future conditions in the marine environment, including sensitivity analysis, machine learning, and Multi-Model Ensembles (MMEs). Finally, we discuss methods to visually communicate the uncertainties associated with MMEs to non-specialist audiences in Section 1.3.

1.1 Sources of uncertainty

In the context of ecosystem management, uncertainty may be loosely defined as an incomplete understanding of the system to be managed (Brugnach et al., 2008). Although potential sources of uncertainty within environmental models have been discussed extensively within the literature (see Schneider and Moss (1999); Regan et al. (2002); Van Asselt and Rotmans (2002); Walker et al. (2003) for example), there is little agreement regarding the terminology that should be used to describe these uncertainties (Daish, 2011). Here, we adopt the typology of Walker et al. (2003) to describe three partially overlapping 'dimensions' of uncertainty that are present in a model-based decision support context: the location, level, and nature of uncertainty (Figure 1.1). It is important to note it is not our intention to cover every aspect of the uncertainties associated with the modelling process in detail but simply to provide a brief overview (for a more detailed review see Regan et al. (2002) or Van Asselt and Rotmans (2002) for example).

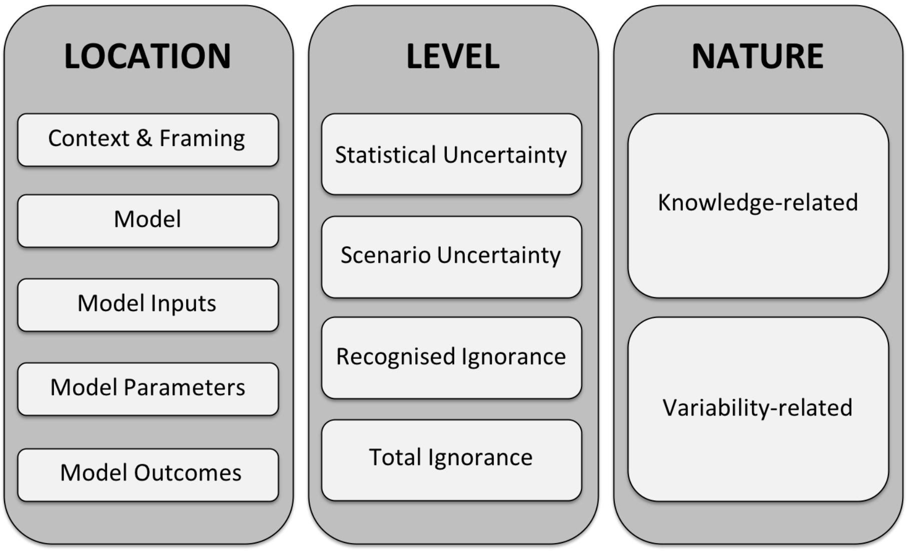


Figure 1.1: The three 'dimensions' of uncertainty: location, level, and nature as described by Walker et al. (2003).

1.1.1 Location of uncertainty

The first dimension of uncertainty identifies the location in which the uncertainty occurs within the model complex, including the model context and framing, structure, inputs, parameters, and outcomes (Figure 1.1 and 1.2). The model context and framing refers to the chosen boundaries of the ecosystem described within the model, as well as the framing of the issues to be addressed within these boundaries (Walker et al., 2003). Context uncertainty may include external variations in environmental, social, economic, technological, and political aspects that may have some impact on the modelled ecosystem (Vasantrao, 2011). For example, socio-economic development and technological advances will affect future greenhouse gas emissions in ways we cannot predict (Lorenz et al., 2015). Consequently, the context uncertainty surrounding predictions relating to the impacts of climate change on a modelled ecosystem will be high. Uncertainty in the framing of a model may arise through the differing perspectives of those involved in the decision-making process, resulting in confusion or disagreements over the issue(s) to be addressed (Van Asselt and Rotmans, 2002).

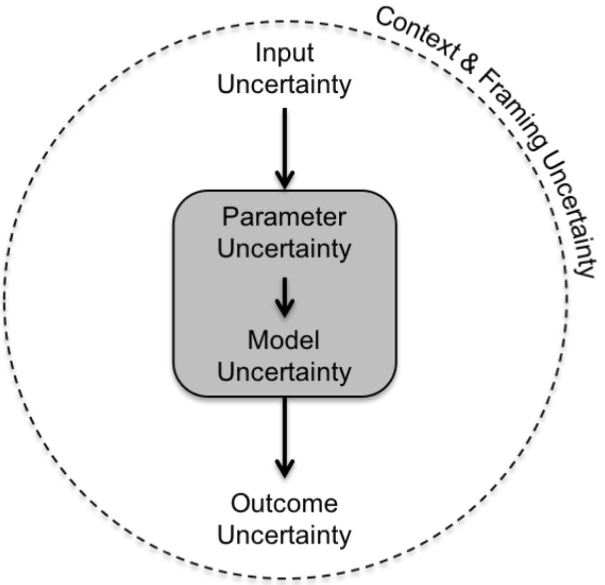


Figure 1.2: A schematic of the five locations of uncertainty in ecosystem models, including context and framing, input, parameter, model, and outcome (or output) uncertainties.

Model uncertainty includes both structural and technical uncertainty. The former arises through the choice of variables or processes deemed necessary to include or exclude in the model, how these components are represented mathematically, and the relationships between these variables and the model inputs and outputs (Walker et al., 2003; Ascough II et al., 2008). Beven (1993, 2006) describes model structure uncertainty as an issue of 'equifinality', whereby a number of equally plausible model structures may generate very different predictions for the future. Consequently, differentiating between acceptable model structures to identify the best possible structure to be used for predictive purposes may be extremely difficult (Beven, 2012). Model technical uncertainty, the latter form of model uncertainty, refers to flaws in the software and hardware used in the modelling process, including errors or bugs (van der Sluijs, 1997).

Model input uncertainty refers to the data used to describe the ecosystem of interest and the external driving forces that act upon this ecosystem (Walker et al., 2003). Much of this uncertainty stems from an inability to accurately capture the inherent complexity and variability of both natural and anthropogenic phenomena within a model (see Section 1.1.3) (Salling and Leleur, 2012). Nonetheless, these uncertainties may also arise as a result of non-representative data collection caused by time constraints, equipment failures, inappropriate methodologies, or financial limitations (Ascough II et al., 2008). In a management context, it is likely that the most important model input uncertainties would be external Forces Driving System Change (FDSCs), particularly those that might result in negative ecosystem responses (Walker et al., 2003). FDSCs are often not well understood both in terms of the magnitude of their impacts (Walker et al., 2003) and the response of the ecosystem to multiple interacting drivers (Nelson et al., 2006), some of which may have cumulative additive, synergistic, or antagonistic effects (Crain et al., 2008). Perhaps the most obvious example of an FDSC in the context of ecosystem management would be climate change, the uncertainties of which are widely discussed (see Pachauri et al. (2014) for example).

Model parameter uncertainty occurs largely as a result of uncertainties in model input data, as parameter estimates are calculated either directly from or via calibration of input data (Briggs et al., 2012). If the quantity and/or quality of the data available for parameter calibration is lacking, the resultant parameter estimates will become uncertain (Zhu and Zhuang, 2013). Parameter uncertainty is therefore of particular concern in the more complex ecosystem models, which contain large numbers of parameters (McElhany et al., 2010).

Finally, model outcome (or output) uncertainty refers to the accumulated uncertainty that is propagated through the model from each of the locations previously discussed. This type of uncertainty represents the difference between the modelled value of an outcome and its true value, with the true value being unknown in all examples in which models are used to predict future impacts on an ecosystem via extrapolation (Walker et al., 2003). Because of this, model outcome uncertainty is extremely difficult, if not impossible, to quantify in its entirety. Uncertainties relating to the interpretation or communication of model outcomes, such as subjective and linguistic uncertainty (arising from ambiguous, vague, or context-dependent scientific vocabulary), may further exacerbate outcome uncertainty (Regan et al., 2002; Daish, 2011).

1.1.2 Level of uncertainty

The second dimension of uncertainty describes the levels of uncertainty in the form of a continuum ranging from statistical uncertainty to total ignorance, i.e. from determinism to indeterminism (Figure 1.1). The closest level of uncertainty to determinism, or precise knowledge of each model component, is statistical uncertainty. This level of uncertainty applies to any of the aforementioned uncertainties where the extent to which a modelled value deviates from its true value may be expressed statistically (Walker et al., 2003). Statistical uncertainty thus allows for the probability distribution of a particular outcome to be calculated (Tyre and Michaels, 2011). Perhaps the most frequently described statistical uncertainty is measurement uncertainty (Walker et al., 2003), which results from incomplete, inaccurate, or biased data collection (Maier et al., 2008).

The second level, known as scenario uncertainty, refers to those uncertainties in which there are multiple possible values for a given model component, resulting in a range of conceivable outcomes (Uusitalo et al., 2015). This level of uncertainty is most often associated with the unknown nature of future conditions, including the degree of socio-economic development, technological advances, and the impacts of climate change. Unlike statistical uncertainty, it is not possible to assign a probability to a given outcome under scenario uncertainty, as each scenario is based on assumptions that we are unlikely to be able to verify in reality (Walker et al., 2003).

Recognised and total ignorance constitute the third and fourth levels of uncertainty and encompass situations in which we do not know enough about some or all of the model components to formulate plausible scenarios (Spangenberg, 2006). Recognised ignorance refers to a situation in which we acknowledge what we do not know, whilst total ignorance (or 'deep uncertainty') refers to a state of indeterminism in which we do not know what we do not know (Spiegelhalter and Riesch, 2011). Walker et al. (2003) subdivide recognised ignorance into two groups based on whether it is possible to improve the level of uncertainty via further scientific research or not, the former being referred to as reducible and the latter as irreducible ignorance. An example of reducible ignorance in climate science might include geophysical feedbacks, whilst irreducible ignorance may include the role of sunspots (Van Asselt and Rotmans, 2002).

1.1.3 Nature of uncertainty

The third dimension of uncertainty divides the nature of a given source of uncertainty into either knowledge- or variability-related categories (Figure 1.1). Knowledge, or epistemic, uncertainty refers to a lack of accurate scientific evidence or understanding of the phenomena described by the model (Ascough II et al., 2008). For instance, the aforementioned incomplete, inaccurate, or biased data may constitute knowledge uncertainty, as might subjective and linguistic uncertainty (Maier et al., 2008). Variability uncertainty, also referred to as ontologic or aleatory uncertainty, arises due to the inherent variability (or randomness) and subsequent unpredictability of both the natural and anthropogenic systems described within the model complex (Ascough II et al., 2008). Variability-related uncertainty encompasses examples such as 'non-rational' human behaviour and the unknown nature of future technological developments (Walker et al., 2003). Though variability-related uncertainty is largely irreducible, knowledge-related uncertainty may often be reduced following further scientific research or model development (Figure 1.1) (Stainforth et al., 2005; Ascough II et al., 2008). Nonetheless, any novel information gleaned from such research could highlight further gaps in our understanding, thereby seemingly increasing uncertainty rather than reducing it (Van Asselt and Rotmans, 2002).

1.2 Methods to quantify uncertainty

Although there is a general consensus within the literature regarding the sources of uncertainty within ecosystem models, there is much debate regarding the methods that should be used to quantify or qualitatively describe these uncertainties. In many circumstances there are multiple methods that may be used to describe uncertainties under each of the aforementioned locations (Table 1.1). However, it is outwith the scope of this research to discuss these methods in detail (see previous reviews e.g. van der Sluijs et al. (2004), Refsgaard et al. (2007), or Uusitalo et al. (2015)) and instead we focus on two of the most promising methods to help quantify uncertainties: sensitivity analysis and Multi-Model Ensembles (MMEs). We also discuss the potential for machine learning algorithms, which seem to have been largely overlooked in this field of research in the past, to play a larger role in uncertainty analyses in the future (Shrestha et al., 2009).

1.2.1 Sensitivity analysis

Sensitivity analysis (SA) are used to quantitatively or qualitatively determine how uncertainties in various aspects of the modelling process affect the outputs of a model (van der Sluijs et al., 2004). The model outputs are deemed to be insensitive to a given model component if variations in the component cause a negligible change in the model outputs. Conversely, the model outputs are deemed to be sensitive to a given model component if variations in the component result in a large change in the model outputs. Identifying the sensitivity of the model outputs to different components within the model may be useful for a wide range of purposes, including model simplification (Rose and Harmsen, 1978; Cariboni et al., 2007; Saltelli et al., 2008), testing the robustness of the model outputs (Cariboni et al., 2007; Saltelli et al., 2008), investigating the behaviour of the model (Rose and Harmsen, 1978; Saltelli et al., 2008), and identifying areas in which to focus future research efforts to reduce the uncertainty

Table 1.1: Summary of frequently used methods to quantify or qualitatively describe each location and level of uncertainty. Methods include: Data Uncertainty Engine (DUE), Data Validation (DV), Expert Elicitation (EE), Error Propagation Equation (EPE), Extended Peer Review (EPR), Inverse Modelling (parameter estimation; IN-PA), Inverse Modelling (predictive uncertainty; IN-UN), Model Comparison (MC), Monte Carlo Analysis (MCA), Multi-Model Ensemble (MME), Model Validation (MV), Numerical Unit Spread Assessment Pedigree (NUSAP), Quality Assurance (QA), Sensitivity Analysis (SA), Scenario Analysis (SC), Stakeholder Involvement (SI), Uncertainty Matrix (UM). Modified from: van der Sluijs et al. (2004) and Vasantrao (2011).

Location of uncertainty		**Level of uncertainty**			
		Statistical	**Scenario**	**Qualitative**	**Recognised ignorance**
Context and framing	Natural, technological economic, social, political	EE, QA, SA	EE, QA, SC, SI	EE, EPR, NUSAP, SI UM	EE, EPR, NUSAP, QA SC, SI, UM
Inputs	System data	DUE, EPE, EE, MCA QA, SA	DUE, EE, SC, QA	DUE, EE	DUE, DV, EE, MV NUSAP, QA, SC
	Driving forces	DUE, EPE, EE, MCA QA, SA	DUE, EE, SC, QA	DUE, EE, EPR	DUE, DV, EE, EPR MV, NUSAP, QA, SC
Model	Model structure	EE, MC, MME, QA, SA	EE, MME, SC, QA	EE, NUSAP, QA	EE, MC, MV, NUSAP QA
	Technical	QA, SA	QA, SA		QA, SA
Parameters		EE, IN-PA, MCA, SA	EE, IN-PA	QA	DV, EE, MV, NUSAP QA, SC
Model outputs		EPE, EE, IN-UN, MC MCA, MME, SA, SC	EE, IN-UN, MME, SA SC	EE, NUSAP	EE, NUSAP

in the model outputs (Saltelli et al., 2008). The model components that are most often included in SA are the inputs and parameters, but SA may also be used to better understand the sensitivity of a model to different model structures or to various scenarios of the future (Table 1.1) (van der Sluijs et al., 2004). For example, in marine ecosystem modelling SA may be used to determine the sensitivity of fish populations to a wide range of management options for the future, such as harvest control rules or area closures, as well as to different climate change scenarios (Serpetti et al., 2017; Fu et al., 2018; Bentley et al., 2019b; Stäbler et al., 2019). Understanding the sensitivity of the model to these scenarios may subsequently aid the development of robust management decisions if the models are used to support policy (Saltelli et al., 2008; Uusitalo et al., 2015).

Various different methods are available for conducting SA, which tend to fall into two broad categories: local and global methods (Saltelli et al., 2008). Local methods of SA typically quantify the sensitivity of the model outputs to small variations in a single uncertain model component, whilst all other components remain fixed (Pianosi et al., 2016). Conversely, global methods of SA typically quantify the sensitivity of the model outputs to much larger variations in the uncertain model components and allow each component to vary at the same time (Pianosi et al., 2016). Global methods of SA may be particularly useful in environmental modelling as they take into account the often complex interactions between model components, whilst local methods do not (Saltelli et al., 2008). However, both local and global methods of SA are often deemed to be too time-consuming and computationally expensive to be included in modelling frameworks, most often due to the large number of model evaluations required to estimate the sensitivity of the model outputs to each of the components (Arhonditsis et al., 2006; Roeder and Hill, 2009). Nevertheless, ignoring the sensitivity of the model outputs risks suboptimal, ineffective, or potentially damaging management decisions if small (but realistic) variations in individual model components result in highly variable model outputs (Uusitalo et al., 2015).

The Sobol' variance-based method, which is often seen as the 'gold standard' of SA, allows us to estimate total-effect indices by decomposing of the model output variance into contributions associated with individual parameters, as well as their interactions with every other parameter in the model (Chen et al., 2004). However, this method is perhaps one of the most computationally expensive forms of sensitivity analysis (Iooss et al., 2012) and may therefore not be a viable option when exploring the sensitivity of a complex marine ecosystem model. Fortunately, a less computationally expensive method, known as derivative-based global sensitivity analysis, has recently been developed to estimate the upper bounds of the Sobol' variance-based sensitivity indices by integrating the squared partial derivatives of the model outputs (Kucherenko et al., 2009; Sobol' and Kucherenko, 2009, 2010; Iooss et al., 2012).

This method of sensitivity analysis has increased in popularity in recent years and has been used in a number of research areas, including biochemical pathway modelling (Rodriguez-Fernandez et al., 2012), reservoir modelling (Touzani and Busby, 2014), and the modelling of predator-prey interactions between grazers and periphyton in aquatic mesocosms (Iooss et al., 2012), among others (Kucherenko and Song, 2016). However, to the best of our knowledge the derivative-based method has not yet been applied to a complex model that includes many hundreds of interacting parameters. Furthermore, few examples exist of a direct comparison between the performance of Sobol' variance-based SA and derivative-based SA when applied to complex models (see Iooss et al. (2012) for example), and none that we are aware of in marine ecosystem modelling. The potential benefits of applying the derivative-based method of sensitivity analysis to marine ecosystem models are therefore largely unexplored but may be vital in helping to better understand and improve these models in the future. Such improvements are of particular importance if marine ecosystem models are to play a larger role in fisheries management in the future.

1.2.2 Machine learning algorithms

Machine Learning (ML) is a branch of artificial intelligence that is used to construct algorithms that learn from and detect patterns in 'big data' (Alpaydin, 2014). Similar to sensitivity analysis, ML algorithms may be used to investigate the behaviour of a model under different parameter combinations and thus help to identify areas in which to focus future research efforts to reduce the uncertainty in the model outputs (Saltelli et al., 2008). Perhaps more importantly, ML algorithms may also be trained to predict (or emulate) the outputs of a model. Being able to accurately predict the outputs of a given model would reduce the need to run the full model and enable us to explore the parameter space more efficiently, thereby helping to lessen the costs associated with marine ecosystem modelling both in terms of human and computational resources; this may be especially important for scientists and decision makers that have limited funding and/or short deadlines.

A wide variety of ML algorithms are currently available, all of which may be grouped into two main categories: supervised and unsupervised learning techniques. Supervised ML techniques, which include decision trees, random forests, support vector machines, and neural networks (Tan and Gilbert, 2003; Mohri et al., 2012), may be particularly useful when attempting to predict the behaviour of an environmental model as they learn classification rules from pre-labelled training data to make predictions about unlabelled testing data (Maglogiannis et al., 2007). A model may therefore be run under multiple parameter combinations and the outputs can be used to train the algorithm to predict the behaviour of the model under a new set of parameter combinations. As supervised methods of ML are often capable of performing classification and regression tasks, they may be used to predict both continuous and discrete model outputs, such as species biomass and species extinctions respectively (Tan and Gilbert, 2003; Mohri et al., 2012). Conversely, unsupervised ML techniques may be used when pre-labelled data is not available. Instead, unsupervised ML algorithms attempt to classify unlabelled data by identifying hidden patterns in the dataset (Maglogiannis et al., 2007). Unsupervised methods include k-means clustering, hierarchical clustering, Principal Components Analysis (PCA), belief networks, and Hidden Markov models (Maglogiannis et al., 2007; Ghahramani, 2004).

Both supervised and supervised ML algorithms have been used in various contexts in ecosystem management in the past. For example, supervised ML algorithms have been used in the prediction of: (1) ocellated turkey (*Meleagris ocellata*) abundance on the Yucatán peninsula (Kampichler et al., 2010); (2) species richness and diversity on two coral reefs located between Tanzania and Zanzibar (Knudby et al., 2010); and (3) bio-indicators of aquatic ecosystems in the Taizi River in northeast China (Fan et al., 2017). However, there are few examples of ML algorithms being used to predict the behaviour of a complex environmental model (see Lucas et al. (2013) for example), and none that we are aware of in marine ecosystem modelling. Perhaps one of the greatest barriers to the widespread use of ML algorithms in marine ecosystem modelling (and elsewhere) is the lack of transparency regarding the internal workings of the algorithms, some of which are often referred to as 'black boxes' (Quetglas et al., 2011). This lack of transparency can make it difficult to implement and interpret ML algorithms without specialist training (Gardner and Dorling, 1998). Nevertheless, there are methods of ML that are more transparent and easy to use than others, with decision trees and random forests being some of the simplest methods of ML (Westreich et al., 2010). By exploring the ability of these simpler methods of ML to predict the outputs of a complex marine ecosystem model, we may not only be able to improve the behaviour of the model and reduce the computational costs associated with running the model, but also ensure that we can effectively communicate the internal workings of the algorithm to non-specialist audiences, particularly decision makers. This research may in turn help to increase the confidence that decision makers have in marine ecosystem models and ensure they can be used more widely in fisheries management in the future.

1.2.3 Multi-model ensembles

Multi-Model Ensembles (MMEs) involve the use of multiple structurally different models to predict future ecosystem responses to natural and anthropogenic pressures (Gårdmark et al., 2013). This method reduces the requirement to identify a single 'best' model and allows a wider range of possible outcomes to be considered (Wang et al., 2011; Beven, 2012). One of the major benefits of MMEs is that we may use the differences in the structural representation of a given ecosystem in each model to disentangle the effects of multiple uncertainties (Tebaldi and Knutti, 2007; Gårdmark et al., 2013). For instance, by comparing the outputs of multiple models run under a single scenario of the future and by comparing the outputs of a single model under various scenarios of the future, it is possible to disentangle the effects of model structure uncertainty and outcome (or 'scenario') uncertainty. This is feasible since any variations in the outputs of multiple models under a single scenario will be caused solely by differences in the structure of the models (Wang et al., 2011; Gårdmark et al., 2013). Conversely, variations in the outputs of a single model under multiple scenarios will represent the extent to which uncertainties in these scenarios are propagated through the model to the model output(s) (Knutti and Sedláček, 2013). Using MMEs to identify events with a high model probability (i.e. results common among many simulations within the ensemble) may also give some indication of the model outputs that are robust to different model formulations (Jones and Cheung, 2014). Robust outputs are of particular interest to decision makers as they provide a strong indication of the likely future conditions on which to base management plans.

Using MMEs has previously been shown to increase the skill and reliability of model predictions in sectors such as public health (Thomson et al., 2006), agriculture (Cantelaube and Terres, 2005), and terrestrial ecosystem modelling (Dormann et al., 2008). MMEs have also been used extensively in climate modelling (see Giorgi and Mearns (2003); Murphy et al. (2004); Tebaldi et al. (2005); Greene et al. (2006); Christensen and Christensen (2007); Furrer et al. (2007); Kjellström et al. (2011); Nikulin et al. (2011); Meier et al. (2012a,b) for example). In particular, Hawkins and Sutton (2009) developed a novel approach to both quantifying and visualising the contributions of internal variability, model, and scenario uncertainty to the total uncertainty of the projections of Surface Air Temperature (SAT) from a world-renowned climate MME (Figure 1.3). To achieve this, the authors used weighted averages of SAT from 15 different global climate models exposed to three climate change scenarios. Uncertainties were estimated by smoothing the model projections and using: (1) the multi-model mean of the variance of the model residuals from the model fits, independent of lead time, to represent internal variability; (2) the multi-scenario mean of the variance of the model fits to represent model uncertainty; and (3) the variance of the multi-model mean of the smooth fits to represent scenario uncertainty (see Hawkins and Sutton (2009) for further details). The relative contributions of each of the three sources of uncertainty to the total variance of the projections was calculated for 180 different regions across the globe and later mapped to indicate areas most affected by each of the uncertainties (Figure 1.4). Not only is this an efficient method to visualise the impacts of multiple sources of uncertainty on the predictions of an MME, such visualisations may also help to identify areas in which investments may help to reduce uncertainty (Hawkins and Sutton, 2009). This type of research has proven to be extremely popular in recent years, with similar methods being applied to a wide range of different climate variables, including precipitation (Hawkins and Sutton, 2011), tropical storm frequency (Villarini and Vecchi, 2012), sea surface temperature (Villarini and Vecchi, 2012; Cheung et al., 2016), sea level (Little et al., 2015), and the Atlantic Meridional Overturning Circulation (Reintges et al., 2017).

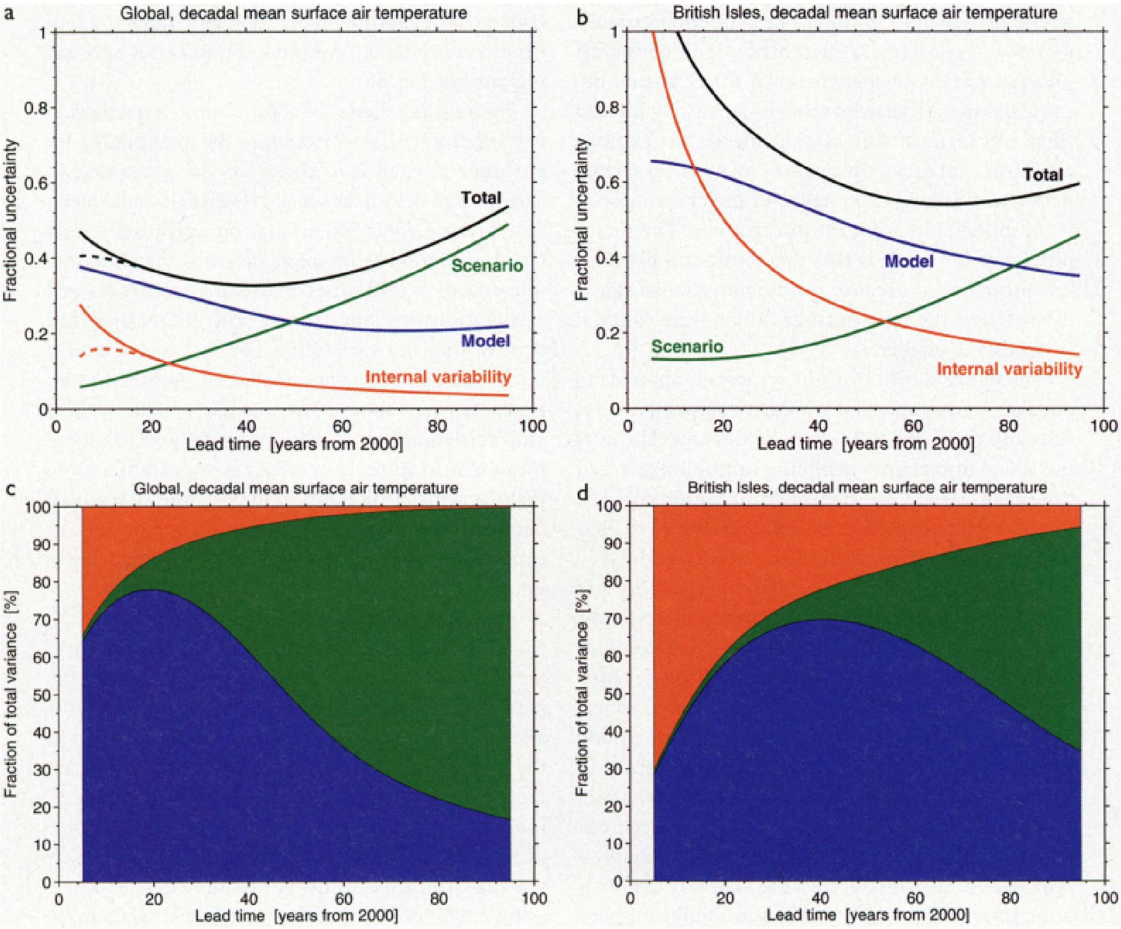


Figure 1.3: The relative contributions of internal variability (orange), model (blue), and scenario (green) uncertainty to the total variance of decadal mean surface air temperature projections for a) the global mean b) the British Isles mean. The proportion of the total variance in decadal mean surface temperature projections explained by each of the three sources of uncertainty for c) the global mean and d) the British Isles mean. Source: Hawkins and Sutton (2009). ©American Meteorological Society. Used with permission.

Although popular in climate science, there are few examples of MMEs in marine science (St-Louis et al., 2012; Jones and Cheung, 2014). The lack of uptake of this methodology in marine ecosystem modelling has occurred largely as a result of inconsistent model outputs and a lack of common parametrisations and scenarios with which to run in each of the models in the ensemble (Spence et al., 2018). However, interdisciplinary research projects such as the

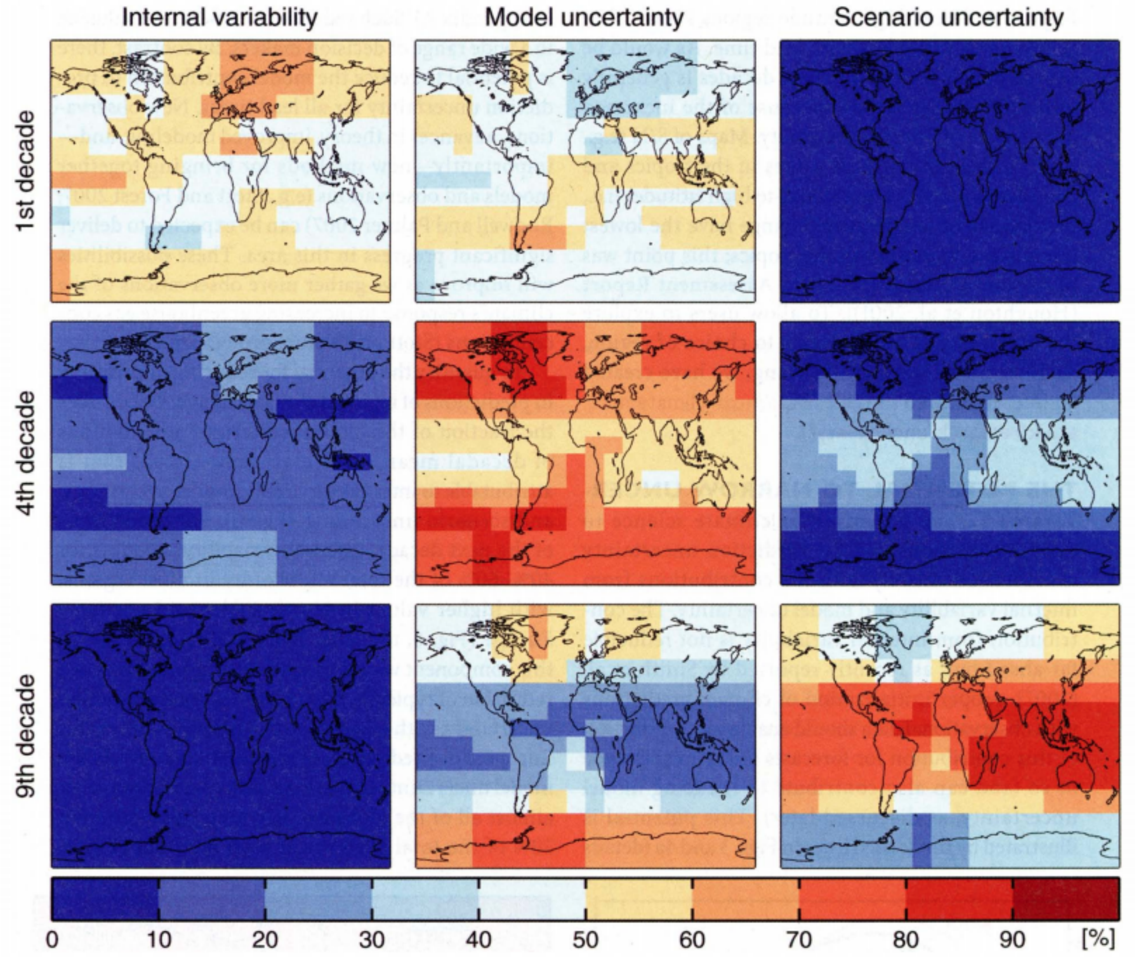


Figure 1.4: The relative contributions of internal variability (left), model (middle), and scenario (right) uncertainty to the total variance of decadal mean surface air temperature projections across the globe under three lead times (years from 2000): one (top), four (middle) and nine (bottom) decades. Source: Hawkins and Sutton (2009). ©American Meteorological Society. Used with permission.

Marine Ecosystems Research Programme (MERP; `marine-ecosystems.org.uk`) and the Fisheries and Marine Ecosystems Model Intercomparison Project (FISH-MIP, `isimip.org/gettingstarted/marine-ecosystems-fisheries/`) have begun to overcome these issues, with a number of recent publications successfully implementing marine ecosystem MMEs (see Spence et al. (2018) and Paine et al. (prep) for example). We may now begin to use these MMEs to assess the impacts of prospective policy options on a given ecosystem by applying management techniques such as fishing quotas under various scenarios of the future (Fu et al., 2018; Shin et al., 2018; Spence et al., 2018). Marine ecosystem MMEs may thus play an important role at the science-policy interface by providing decision support at local, national, multinational, and global scales.

However, the outputs of an MME are often extremely complex and different models within the MME may give very different predictions for the future (Hansen and Hoffman, 2011). In the past, a lack of effective communication of such complex and highly variable model outputs, both to decision makers and the general public, has been blamed for ineffective management decisions (Janssen et al., 2005). This in turn has contributed to public distrust of scientific evidence, particularly in regards to climate science (Frewer, 2004). Improving the communication of the uncertainties associated with MMEs to non-specialist audiences is therefore vital to ensuring these models can continue to make a significant contribution to the decision-making process.

1.3 Communicating uncertainty

It is often suggested that non-specialist audiences, including decision makers and the general public, are unable to understand uncertainty analyses (Morgan, 2009). Nevertheless, the successful communication of uncertainties to decision makers is vital to increasing transparency and reducing misinterpretations between scientists and decision makers, thus helping to ensure that management efforts are not misplaced (Janssen et al., 2005). Scientists therefore need to focus on developing methods that may be used to communicate uncertainties to non-specialist audiences in a simple, understandable, and effective manner.

Whilst there is an abundance of scientific literature regarding how best to overcome the linguistic uncertainties associated with communicating complex model outputs (see Patt and Schrag (2003); Patt and Dessai (2005); Morgan (2009) or Mastrandrea et al. (2010) for example), there is little guidance concerning how best to visualise uncertainties (Spiegelhalter et al., 2011). In the past, many of the techniques used for data visualisation ignored the presence of uncertainties or were only able to depict one source of uncertainty at a time (MacEachren et al., 2005; Brodlie et al., 2012). This is especially problematic when attempting to communicate the outputs of MMEs, which typically require a visual representation of changes in both model and scenario uncertainties over time. Animated and interactive visualisations may be particularly useful when communicating multiple uncertainties or changes in uncertainty over time, but these methods are limited to television, film, and digital media. By first focusing on improving the communication of uncertainty using static visualisations, which can also be used in print, we may be able to access a much broader cross-section of society than would be possible using animated or interactive visualisations; we can then build on the information gleaned from this research to develop more effective interactive and animated visualisations.

Examples of traditional methods of static visualisation that are often used to communicate uncertainty in environmental modelling include line plots with uncertainty bands (also referred to as envelopes) and box plots with error bars. These visualisations include both summaries of the data, such as averages, and the estimated uncertainty surrounding this information. Static visualisations may also be created solely to communicate uncertainties, although they may require a more in-depth description of what the figure depicts to ensure they are understandable (Kloprogge et al., 2007). For example, radar (or spider) plots and pedigree charts, which are based either on the direct quantification of uncertainties or through expert elicitation, may be used to depict the estimated uncertainty of various model components. It may also be possible to re-format these visualisation methods, particularly the radar plots, to display both a summary of the model outputs and the uncertainty surrounding the projections. Although there is evidence to suggest that some of these visualisation methods may be effective at communicating uncertainties to specific groups of people (see Ibrekk and Morgan (1987) for example), relatively little is known about the ability of decision makers and the general public to interpret these types of visualisations, particularly when used to communicate multiple sources of uncertainty.

Even less is known about the effectiveness of more modern methods of visualising uncertainty, such as cascade plots (see Wilby and Dessai (2010) and Hawkins (2014) for example) and infographics. Although cascade plots are not yet widely used in science communication, infographics have become increasingly popular with non-specialist audiences in recent years and are frequently used by the media. Infographics may represent data or knowledge and tend to consist of a combination of symbols, pictures, figures, maps, and/or diagrams (Mol, 2011). In ecosystem management an infographic may be used to communicate aspects such as key ecosystem services, the impacts of human use on the environment, and possible management scenarios (NART, 2013). Infographics are believed to be highly effective as they often provide context to the data presented in the visualisation, unlike many other forms of visual communication (Kosara and Mackinlay, 2013). Because of this, infographics tend to be more memorable to non-specialist audiences, particularly the general public (Bateman et al., 2010). However, there are few examples in which uncertainty is incorporated into an infographic and little is known about their performance relative to more traditional methods of visualisation. Although an infographic that is successfully able to communicate the full complexity of the uncertainties associated with MMEs is difficult to envisage at present, we must begin to explore the potential uses of this type of visualisation, as well as some of the more traditional methods of visualising uncertainty, if we are to improve communication between scientists and non-specialist audiences in the future.

1.4 Thesis summary

The presence of uncertainties within the ecosystem modelling process is unavoidable, yet despite this management decisions must still be made. By improving the identification, evaluation, and communication of uncertainties and their impacts on the projections of a model, scientists can provide decision makers and the general public with a more realistic insight into the likely future conditions of an ecosystem under various management scenarios (Burgman, 2005; Power and McCarty, 2006; Hill et al., 2007; Tyre and Michaels, 2011). I hope to help achieve this goal by applying methods such as global sensitivity analysis, machine learning, and MMEs to better understand the uncertainties regarding the future of marine ecosystems. If such improvements to our understanding of uncertainty are to be useful in a management context, it must be reinforced with improved communication and knowledge-exchange at the science-policy interface (Krupnick et al., 2006). I therefore also aim to help achieve this goal by identifying how best to visually communicate the outputs of MMEs to non-specialist audiences.

The rest of this thesis is thus organised as follows:

Chapter 2: The *mizer* marine ecosystem model

In Chapter 2, I provide an introduction to a marine ecosystem model known as *mizer*. This model is used in Chapters 3 and 4 to demonstrate how global sensitivity analysis and machine learning may be used to analyse parameter uncertainties in marine ecosystem modelling. Although this chapter does not introduce any new model development, it is included to provide a detailed explanation of the inner workings and assumptions of the model and to describe all of the parameters that were included in the novel analyses described in Chapters 3 and 4. The reasons for using *mizer* are also provided at the end of Chapter 2.

Chapter 3: Global sensitivity analysis of the *mizer* marine ecosystem model

In Chapter 3, I apply two different methods of global sensitivity analysis to the *mizer* marine ecosystem model. More specifically, I estimate the Sobol' variance-based and derivative-based sensitivity indices of the trait-based version of the *mizer* model, which includes 24 parameters, to allow for a direct comparison between these two methods of sensitivity analysis. I also apply a derivative-based sensitivity analysis to the North Sea multispecies version of the *mizer* model, which includes over 300 parameters, to demonstrate the ability of this method of sensitivity analysis to handle a highly complex model. I consider the sensitivity of multiple model outputs, including community biomass, population size, spawning stock biomass, fisheries yield, and the Large Fish Indicator (LFI), to small changes of $\pm$ 10% of the nominal values of the parameters. The results of the sensitivity analyses are used to discuss where further research might be focused to help reduce the uncertainty in the model outputs, thus enabling us to produce more accurate model projections.

Chapter 4: Using machine learning to predict the behaviour of the *mizer* marine ecosystem model

In Chapter 4, I explore the ability of a method of machine learning known as the random forest algorithm to predict the behaviour of the North Sea multispecies *mizer* model under different parameter combinations. I also use the random forest algorithm to identify the parameters that drive certain model behaviours, such as species extinctions, thus enabling us to better understand and subsequently improve the behaviour of the model. At the end of this chapter, I discuss the similarities between the parameters that are identified as being important by the machine learning algorithm and the sensitivity analyses in Chapter 3, further supporting the conclusions given in both chapters regarding areas in which to focus future research efforts to reduce the uncertainties in the outputs of the *mizer* model.

Chapter 5: Uncertainty in projections of global and regional sea surface temperature and salinity

In Chapter 5, I use global and regional projections of Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) from a world-renowned climate MME to quantify spatio-temporal changes in the contributions of internal variability, model, and scenario uncertainties to the total variance of the projections. The results are used to highlight areas in which the uncertainties may be reduced via further research, as well as to identify irreducible uncertainties. I also explore the signal-to-noise ratio of the projections to allow us to identify regions and time periods in which the projections are most certain and are thus most useful to decision makers in terms of adaptation planning. Please note that I chose to use a climate MME in this chapter as the marine ecosystem MMEs developed during MERP and FISH-MIP were not complete at the time of writing. Nevertheless, the results are extremely relevant to marine ecosystem modelling and the methods may be easily applied to MMEs from a wide variety of research areas in the future.

Chapter 6: Visualising uncertainty in multi-model ensembles

In Chapter 6, I conduct an in-depth online survey to identify the most effective methods for communicating the outputs of a MME to different audiences using static visualisations. I test the performance of 10 different visualisations, all of which depict exactly the same data but in slightly different ways. The performance of each visualisation is measured based on the accuracy, confidence, and ease with which the participants were able to interpret each visualisation, as well as their preferences for different visualisations across a number of categories. I take into account the education level, background, and expertise of the participants to determine whether different groups of people are better able to interpret certain visualisations, thus enabling us to target visualisations at specific audiences to maximise their impact, whilst also minimising the potential for misinterpretations. Please note that the visualisations were produced using the climate MME described in Chapter 5 for the same reasons as stated previously.

Chapter 7: Discussion

In the final chapter of this thesis, I synthesise the results of the research described in Chapters 3 to 6. I highlight the key findings of this work and identify areas of future research that may be required to further advance our understanding of the uncertainties associated with marine ecosystem models.

Chapter 2

The *mizer* marine ecosystem model

The aim of Chapters 3 and 4 is to better understand the impact of parameter uncertainties on the outputs of a marine ecosystem model using global sensitivity analysis and machine learning techniques. In both of these chapters, the *mizer* model is used as an example of how these methods may be applied in practice. The inner workings and assumptions of the *mizer* model are therefore described in detail below. Please note that I have not contributed to the development of the *mizer* model prior to or during the research presented in this thesis. The model is described in full here only as supporting text for Chapters 3 and 4 and the following sections are based solely on the information given by those involved in the development of the model (see Scott et al. (2014), Andersen et al. (2015), and Scott et al. (2018) for example). A full list of the parameters of the model, their nominal values, and associated references is given in Tables 2.1 to 2.5.

2.1 Introduction to the model

mizer is a dynamic size spectrum ecological model that is used to better understand how the growth of a set of individuals (and their subsequent changes in trophic level) affects fish community dynamics (Scott et al., 2014). All processes included within *mizer* are formulated at the level of the individual and all of the parameters are directly related to body size, thereby allowing the model to be formulated using a relatively small number of parameters (Scott et al., 2014). Individual body size is aggregated to describe the entire fish community using a single size distribution, known as the size spectrum (Andersen et al., 2015).

There are currently three different versions of the *mizer* model, increasing in complexity from the comparatively simple community model to the more complex trait-based and multispecies models (Scott et al., 2014) (see Section 2.4 for more details). All three versions of *mizer* are based on the same two central assumptions (Section 2.2) and a number of ’standard’ assumptions (Section 2.3) that are often used in ecology to describe food consumption, growth $g_i(w)$, recruitment R_i, and mortality $\mu_i(w)$ (Scott et al., 2014).

2.2 Central assumptions

In this section, the two central assumptions of the *mizer* model are described in detail. The first central assumption of the *mizer* model is that individuals may be characterised solely by their weight w and species ID i. The size spectrum $N_i(w)$ of species i represents the density of individuals at a given size at time t (where the default time step is equal to 0.25 years) and is calculated by scaling the individual-level processes of somatic growth $g_i(w)$ and mortality $\mu_i(w)$ using the McKendrick-von Foerster equation:

$$\frac{\partial N_i(w)}{\partial t} + \frac{\partial g_i(w) N_i(w)}{\partial w} = -\mu_i(w) N_i(w) \tag{2.1}$$

Individual growth and mortality are determined by food availability, predation, and fishing mortality. Food may be sourced from other individuals or from a background resource spectrum $N_R(w)$, which represents the planktonic community and any other food sources that are not directly included in the model. Only the smallest individuals feed on the background resource spectrum, which is modelled using a dynamic semi-chemostat growth equation:

$$\frac{\partial N_R(w,t)}{\partial t} = r_0 w^{p-1} \big[\kappa w^{-\lambda} - N_R(w,t)\big] - \mu_p(w) N_R(w,t) \tag{2.2}$$

where $r_0 w^{p-1}$ is the population regeneration rate (Fenchel, 1974; Savage et al., 2004), $\kappa w^{-\lambda}$ is equal to κw^{-2+q+n}, where $\kappa w^{-\lambda}$ represents the carrying capacity of the population, n represents the scaling of food intake, p represents the scaling of standard metabolism, q represents the search volume exponent, and μ_p is the predation mortality rate given in Equation 2.16.

The second central assumption of the *mizer* model is that an individual's food preferences are determined both by species preference and a combination of individual weight and prey weight preference. Prey weight preferences are described using a log-normal selection model (Ursin, 1973) in terms of the ratio between the weight of the predator w and the weight of the prey w_p:

$$\phi(w_p/w) = \exp\Big[\frac{-(\ln(w/(w_p \beta_i)))^2}{2\sigma_i^2}\Big] \tag{2.3}$$

where β_i is the preferred predator-prey mass ratio and σ_i is the width of the prey size selection function. Due to predation being size-based, cannibalism is an inherent part of the model.

2.3 Standard assumptions

In this section the 'standard' assumptions of the *mizer* model that relate to food consumption, growth, reproduction, recruitment, and mortality are described.

2.3.1 Predator-prey encounters

Predator-prey encounters are based on the "Andersen-Ursin" encounter model that was originally developed to represent the North Sea marine ecosystem (see Andersen and Ursin (1977) and Andersen and Beyer (2006) for further details) and within which the general rule that "big fish eat smaller fish" was formalised. The food available (mass per volume) for an individual of weight w, denoted as $E_a, i(w)$, is determined by integrating over the number of individuals in the model and the size of the background resource, weighted by the size selection function in Equation 2.3:

$$E_a, i(w) = \int \left(N_R(w_p) + \sum_j \theta_{ij} N_j(w_p) \right) \phi_i(w_p/w) w_p dw_p \tag{2.4}$$

where θ_{ij} is the preference of species i for species j. The amount of food encountered by a predator $E_{e,i}$ (biomass per time) is dependent on the individual's search rate γ_i (volume per time), which scales with weight such that larger fish are capable of searching greater volumes of water for food than smaller fish:

$$E_e, i(w) = \gamma_i w^q E_{a,i} \tag{2.5}$$

2.3.2 Food consumption

Encountered food is consumed following a standard Holling type II functional response (Holling, 1959) to represent satiation. This functional response is used to determine the feeding level $f_i(w)$ of an individual. The feeding level is a dimensionless number between 0 and 1, which represents a total lack of food and full satiation respectively:

$$f_i(w) = \frac{E_{e,i}}{E_{e,i} + h_i w^n} \tag{2.6}$$

where h_i is the maximum food intake and $h_i w^n$ is the maximum consumption rate. The food consumption rate is then $f_i(w) h_i w^n$. If h_i is not directly specified by the model user, it is calculated as:

$$h_i = \frac{3k_{vb}}{\alpha f_0} W_i^{1/3} \tag{2.7}$$

where k_{vb} represents the von Bertalanffy K parameter, α is the assimilation efficiency, f_0 represents the level at which the smallest individuals in the population feed on the background resource spectrum when it is at carrying capacity, and W_i is the asymptotic weight. f_0 is used to control resource productivity and to calculate the search rate parameter γ_i if it is not directly specified by the model user (see Andersen and Beyer (2006), Appendix B for more details):

$$\gamma_i(f_0) = \frac{f_0 h_i \beta_i^{2-\lambda} \exp{(-\lambda - 2)^2 \sigma_i^2/2)}}{(1-f_0)\sqrt{2\pi}\kappa\sigma_i} \tag{2.8}$$

where κ represents the carrying capacity of the background resource spectrum.

2.3.3 Energy budget

Consumed food is used as energy for standard metabolism, activity, growth, and reproduction. The consumed food is assimilated with efficiency α and the acquired energy is used firstly for standard metabolism at a rate of $k_{s,i}$ (defined as 20% of h_i if not directly specified by the user) and secondly for activity at a rate of $k_i w$. Once these costs have been accounted for, the remaining energy $E_{r,i}(w)$ (if any) is used for growth and reproduction:

$$E_{r,i}(w) = \max(0, \alpha f_i(w) h_i w^n - k_{s,i} w^p - k_i w) \tag{2.9}$$

If the energy acquired for standard metabolism and activity is not sufficient to meet the needs of the individual, growth and reproduction ceases; growth cannot be negative and therefore individuals cannot decrease in size. It is important to note that individuals are not subjected to starvation mortality in the model at present as starvation was not found to be an important process in the model when using a "Beverton-Holt" recruitment function (see Section 2.3.5; Scott et al. (2014)).

The proportion of $E_{r,i}w$ that is used for reproduction $\psi_i(w)$ is defined as:

$$\psi_i(w) = \left[1 + \left(\frac{w}{w_{m,i}}\right)^{-10}\right]^{-1} \left(\frac{w}{W_i}\right)^{1-n} \tag{2.10}$$

where the function in the square bracket varies smoothly from 0 to 1 around the individual's weight at maturation. This means that juveniles use all of their remaining energy solely for growth whereas mature individuals use their remaining energy for both growth and reproduction. The last term in Equation 2.10 describes the relative increase in energy available for reproduction as an individual nears its asymptotic weight.

The proportion of $E_{r,i}w$ that is used for somatic growth is therefore defined as:

$$g_i(w) = E_{r,i}(w)(1 - \psi_i(w)) \tag{2.11}$$

When the feeding level is constant, the growth curve given by Equation 2.11 approximates a von Bertalanffy growth curve (Hartvig et al., 2011). However, the growth curve will depend on the feeding level and growth may therefore be stunted if the feeding level drops below a critical level f_c, after which the amount of food assimilated by an individual is only sufficient to cover the costs of standard metabolism:

$$f_{c,i}(w) = \frac{k_{s,i}w^p + k_i w}{\alpha h_i w^n} \tag{2.12}$$

2.3.4 Reproduction

Both reproduction and recruitment are determined by considering the reproductive contributions of all of the individuals in the population. Egg production $R_{p,i}$ (numbers per time step) is defined as:

$$R_{p,i} = \frac{\epsilon}{2w_0}\int N_i(w)E_{r,i}(w)\psi_i(w)dw \tag{2.13}$$

where w_0 is the egg weight and ϵ represents a penalty on egg production due to egg mortality and the cost of spawning.

2.3.5 Recruitment

Recruits enter the size spectrum at the smallest body size (the egg weight w_0 by default). However, it is widely assumed that juvenile marine fish experience significant density dependence (Ricker, 1954) and this density dependence is incorporated into *mizer* in the form of a "Beverton-Holt" stock-recruitment relationship (SRR) that acts to compensate egg production. Using a SRR to represent density dependence helps to prevent competitive exclusion (Hartvig and Andersen, 2013) and the subsequent extinction of a trait or species group, thus acting to stabilise the model outputs. The "Beverton-Holt" SRR ensures that recruitment R_i (numbers per time step) approaches maximum recruitment $R_{max,i}$ (i.e. the population carrying capacity) with increasing egg production $R_{p,i}$:

$$R_i = R_{\max,i} \frac{R_{p,i}}{R_{p,i} + R_{\max,i}} \tag{2.14}$$

In actuality, $R_{max,i}$ is used as a model tuning parameter that represents any phenomena affecting the population that are not explicitly included in the model.

2.3.6 Mortality

There are three different types of mortality that affect the overall mortality rate of an individual $\mu_i(w)$ in the *mizer* model, including predation mortality $\mu_{p,i}(w)$, background (or natural) mortality $\mu_{b,i}(w)$, and fishing mortality $\mu_{f,i}(w)$. Predation mortality is dependent on the trophic dynamics of the model, with food consumption resulting in a corresponding decline in the population size of prey individuals (see Hartvig et al. (2011), Appendix A for further details):

$$\mu_{p,i}(w_p) = \sum_j \theta_{ji} \int \phi_j(w_p/w)(1 - f_j(w))\gamma_j w^q N_j(w) dw \tag{2.15}$$

The predation mortality of the background resource spectrum $\mu_p(w_p)$ is defined as:

$$\mu_p(w_p) = \sum_j \int \phi_j(w_p/w)(1 - f_j(w))\gamma_j w^q N_j(w) dw \tag{2.16}$$

Background mortality represents death by natural causes and is assumed to be independent of individual body size but dependent on species ID and inversely proportional to generation time (Peters, 1983). If not directly specified by the model user, $\mu_{b,i}$ is calculated as:

$$\mu_{b,i} = \mu_0 W_i^{n-1} \tag{2.17}$$

Fishing mortality $F_{g,i}(w)$ is size- and species-specific and is imposed by fishing gears g:

$$F_{g,i}(w) = S_{g,i}(w) Q_{g,i} E_g \tag{2.18}$$

where S is the selectivity, Q is the catchability, and E is the fishing effort associated with each gear type. Gear selectivity ranges from 0 to 1, with a value of 0 indicating that the gear is not capable of selecting (or catching) the species at size w and a value of 1 indicating the species is fully selected by the fishing gear at size w. By default, a "knife-edge" selectivity function is used in the trait-based model and therefore the selectivity of each gear type instantaneously changes from 0 to 1 at a given size. Conversely, a sigmoid function is used in the multispecies model such that the selectivity changes from 0 to 1 more smoothly. The catchability term Q is an additional scalar that is used to link fishing mortality and population size. Both the selectivity and catchability of the gear remain constant in time, although the fishing effort can be varied through time to allow for the simulation of dynamic fishing patterns.

The total fishing mortality imposed on each species at a particular weight is defined as the sum of the fishing mortalities imposed by all gears:

$$\mu_{f,i}(w) = \sum_g F_{g,i}(w) \tag{2.19}$$

2.4 Model types

As previously mentioned, there are three different versions of the *mizer* model, increasing in complexity from the community model to the trait-based and multispecies models.

2.4.1 The community model

In the community version of the model, individuals are characterised solely by their size and there are no distinctions between different species; the population forms one group that represents an average across all species. In this version of *mizer*, maturation and reproduction are ignored, the recruitment flux is constant and the energy budget is simplified such that growth is equal to the remaining energy available once the individual costs of standard metabolism and activity have been accounted for, multiplied by an "average growth efficiency" (see Andersen et al. (2015) Appendix B and Zhang et al. (2013) for full details).

2.4.2 The trait-based model

The trait-based version of the model can include any number of species. However, the species are distinguishable solely by their asymptotic sizes W_i, which are evenly spaced on a continuum ranging from the smallest possible size to the maximum asymptotic size. The number of species that are included in the trait-based model is unimportant and has little impact on the dynamics of the model when more than ten species are specified in the model. In the trait-based *mizer* model, maximum recruitment R is a function of asymptotic size (see Andersen et al. (2015), Appendix A for full details) and an individual's food supply is determined solely by body size. In the past, the trait-based version of the model has proven to be useful in understanding the community-level impacts of changes in species-specific fishing mortality rates (Andersen and Pedersen, 2010; Jacobsen et al., 2013) without requiring large amounts of species-specific information for model parameterisation.

2.4.3 The multispecies model

The multispecies version of the *mizer* model is the most complex of the three versions, with individual species being resolved in much greater detail than the community and trait-based models. The multispecies model includes up to 21 species-specific parameters, as well as an interaction matrix that represents the spatial co-occurrence of each pair of species in the model (Blanchard et al., 2014). This version of *mizer* therefore requires large amounts of data for parameterisation, but it acts as a more realistic representation of the ecosystem than both the community and trait-based versions of the model.

In the past, the multispecies version of *mizer* has been used to better understand the impacts of fishing (Blanchard et al., 2014), seasonal spawning and plankton blooms (Datta and Blanchard, 2016), and changing environmental conditions (Marshall, 2017) on the community structure of the North Sea marine ecosystem. The North Sea multispecies *mizer* model is focused on 12 common and commercially important North Sea fish species including: sprat (*Sprattus sprattus*), sandeel (*Ammodytes marinus*), Norway pout (*Trisopterus esmarkii*), Atlantic herring (*Clupea harengus*), dab (*Limanda limanda*), whiting (*Merlangius merlangus*), common sole (*Solea solea*), grey gurnard (*Eutrigla gurnardus*), European plaice (*Pleuronectes platessa*), haddock (*Melanogrammus aeglefinus*), Atlantic cod (*Gadus morhua*), and saithe (*Pollachius virens*), whose aggregated size spectra forms the community spectrum of the model. Together, these 12 species account for almost 90% of the total biomass of fish sampled by research trawl surveys within the area (Blanchard et al., 2014).

2.5 Nominal parameter values

The nominal parameter values in the community and trait-based versions of the model (see Tables 2.1 and 2.2) were determined based on meta-analyses of marine fish data, including those obtained from laboratory experiments and in the field (see Hartvig et al. (2011), Appendix E for full details). All nominal parameter values in the North Sea multispecies *mizer* model (see Tables 2.3 to 2.5) were originally estimated and/or calibrated using publicly-available vessel survey data, stock assessment estimates, and fisheries landings data collected between 1985 and 1995 (`ices.dk`). The maximum recruitment R_{max} of each species and the carrying capacity κ of the background resource spectrum, both of which are particularly difficult to measure in the natural environment, were estimated by Spence et al. (2016) using Bayesian statistics.

2.6 Model outputs

The outputs of the *mizer* model include the population size, total biomass, Spawning Stock Biomass (SSB), and fisheries yield of each species or trait group through time, as well as three fish community indicators that are often used to determine the health of a marine ecosystem (Blanchard et al., 2014). These indicators include: (1) the Large Fish Indicator (LFI; the proportion of fish (by weight) of length > 40cm); (2) the mean weight of all of the individuals in the community spectrum; and (3) the slope of the community spectrum (based on a linear regression of log-transformed numbers of individuals against log-transformed body mass) (Blanchard et al., 2014). These indicators are usually quantified based solely on demersal fish with a weight of between 10g and 100kg to maintain consistency with empirical fish community indicators related to the implementation of policy such as the EU Marine Strategy Framework Directive (MSFD; European Commission (2008b)) (Blanchard et al., 2014).

2.7 Why use the *mizer* model?

Size-based models, such as *mizer*, are powerful yet relatively simple tools that allow us to explore the potential impacts of changes in human- and environmentally-induced pressures on marine and freshwater ecosystems (Blanchard et al., 2017). It is this simplicity that makes the size-based approach particularly well-suited to demonstrating the possible applications of computationally-demanding methods, such as global sensitivity analysis (see Chapter 3) and machine learning (see Chapter 4), to marine ecosystem models. This type of modelling is also supported by over 50 years of research into the logarithmic relationship between biomass and body size, as well as the correlations between various physiological and ecological processes, including metabolism, respiration, movement, and trophic interactions, with body size (Giacomini et al., 2016; Blanchard et al., 2017). As a result of this large body of work, size spectrum models have proliferated in recent years (Blanchard et al., 2017), with UK examples (as highlighted by Hyder et al. (2015)) including the Coupled Community Size-Spectrum Model (CCSSM) (Blanchard et al., 2009), the Species Size-Spectrum Model (SSSM) (Rossberg, 2012), FishSUMS (Speirs et al., 2010), the Fish Community Size-Resolved Model (FCSRM) (Hartvig et al., 2011), the Length-based Multispecies Analysis by Numerical Simulation model (LeMANS) (Hall et al., 2006), and *mizer* (Scott et al., 2014). Although largely focused on the North Sea, these models have been used in a wide variety of contexts, including management strategy evaluation, risk assessment, and the testing of various size-based indicators of community or ecosystem health (Hyder et al., 2015). Size-based models have also been applied to various marine ecosystems outside of the North Sea (see Canales et al. (2015) for an example off the coast of Chile and Rowan et al. (2017) for an example in the Southern Ocean), to several freshwater ecosystems (see van Zwieten et al. (2015) and Kolding et al. (2015) for example), as well as at a global scale (see Watson et al. (2015) for example).

Of all of the size spectrum models that are currently available, we chose to use *mizer* for the following reasons: (1) the model is well-developed and has been published in several peer-reviewed journals (see Blanchard et al. (2014), Zhang et al. (2015), Spence et al. (2016) for example); (2) one of the model developers (Dr. Julia Blanchard) was involved in this research project and was therefore able to provide access to the R code used to build and run the model, making it easy to modify the code to run the sensitivity analysis and machine learning algorithm described in Chapters 3 and 4; (3) the model was originally developed (and has been calibrated) to represent the North Sea (Blanchard et al., 2014), an ecosystem that is well-known to all of those involved in this work; (4) the model was the first dynamic size spectrum model to be incorporated into a Bayesian framework that explicitly addresses parameter uncertainties (Spence et al., 2016), which are the focus of Chapters 3 and 4. Furthermore, many of the aforementioned size spectrum models typically rely on a very similar set of assumptions and equations. As such, we expect the conclusions reached in Chapters 3 and 4 to be applicable to many of these models, thus making the choice of size-based model less important.

However, it is important to note that there are some drawbacks associated with size spectrum models and/or *mizer* in particular (see Blanchard et al. (2017) for a review). Of perhaps greatest importance to this research is the fact that size-based models are not end-to-end ecosystem models and therefore they do not cover the entire food web. Most size-based models, including the *mizer* model, have only a crude representation of the planktonic and benthic organisms that typically form the base of the food web, and only a few include top predators, such as sea birds and marine mammals (Blanchard et al., 2017). This means that the results of Chapters 3 and 4 may be focused predominately on the fish part of the community. Nevertheless, work is currently underway to incorporate seals into the North Sea version of the *mizer* model (Spence, pers. comm.), as well as to develop a dynamic coupling between *mizer* and the European Regional Seas Ecosystem Model (ERSEM) (MERP, 2017), which includes a detailed representation of the biogeochemistry and lower trophic levels of the North Sea (Butenschön et al., 2016). It may therefore be possible to extend Chapters 3 and 4 to gain a better understanding of the ecosystem components outside of the fish community in the near future.

Table 2.1: Nominal parameter values of the community version of the *mizer* model. All nominal values were taken from the *mizer* R package (see Scott et al. (2018)).

Parameter	**Description**	**Nominal value**
w_{min}	Minimum size of the community spectrum	0.001
w_{max}	Maximum size of the community spectrum	1 x 10^6
β	Preferred predator-prey mass ratio	100
σ	Width of the prey size preference	2
α	Assimilation efficiency	0.2
h	Maximum food intake rate	10
z_0	Background mortality of the community spectrum	0.1
n	Scaling of the food intake	2/3
q	Search volume exponent	0.8
λ	Exponent of the background resource	2+q-n
κ	Carrying capacity of the background resource	1000
f_0	Average feeding level of individuals feeding mainly on the background resource	0.7
r_{pp}	Growth rate of primary productivity	10
KES	Size at the edge of the knife-selectivity function	1000
rec	Constant recruitment in the smallest size class of the community spectrum	2.51 x 10^9
rec_{mult}	Multiplier for constant recruitment	1
F	Fishing effort	0

Table 2.2: Nominal parameter values of the trait-based version of the *mizer* model. All nominal values were taken from the *mizer* R package (see Scott et al. (2018)).

Parameter	**Description**	**Nominal value**
w_{min}	Minimum size of the community spectrum	0.001
w_{max}	Maximum size of the community spectrum	1.1 x 10^5
$w_{\infty_{min}}$	Asymptotic size of the smallest species in the community spectrum	10
$w_{\infty_{max}}$	Asymptotic size of the largest species in the community spectrum	1 x 10^5
η	Factor to calculate w_{mat} from w_∞	0.25
$w_{pp_{min}}$	Smallest size of the background resource	1 x 10^{-10}
$w_{pp_{cut}}$	Maximum size of the background resource	1
β	Preferred predator-prey mass ratio	100
σ	Width of the prey size preference	1.3
α	Assimilation efficiency	0.6
h	Maximum food intake rate	30
γ	Volumetric search rate	600.44
$z_{0_{pre}}$	Coefficient of the background mortality of the community spectrum	0.6
n	Scaling of the food intake	2/3
p	Scaling of the standard metabolism	0.75
q	Search volume exponent	0.9
λ	Exponent of the background resource	2.23
κ	Carrying capacity of the background resource	0.005
f_0	Average feeding level of individuals feeding mainly on the background resource	0.5
r_{pp}	Growth rate of primary productivity	4
ks	Coefficient of standard metabolism	4
KES	Size at the edge of the knife-selectivity function	1000
k_0	Multiplier for maximum recruitment	50
F	Fishing effort	0

Table 2.3: Nominal values of the species-independent parameters of the multispecies *mizer* model. All nominal values were taken from Blanchard et al. (2014) excluding κ, which was taken from Spence et al. (2016).

Parameter	**Description**	**Nominal value**
w_{max}	Maximum size of the community spectrum	4.40 x 10^4
$w_{pp_{cut}}$	Maximum size of the background resource	10
$z0_{pre}$	Coefficient of the background mortality of the community spectrum	0.6
$z0_{exp}$	Exponent of the background mortality of the community spectrum	-1/3
n	Scaling of the food intake	2/3
p	Scaling of the standard metabolism	0.7
q	Search volume exponent	0.8
λ	Exponent of the background resource	2.13
κ	Carrying capacity of the background resource	8.45 x 10^{10}
$slope_0$	Starting slope of the community spectrum	-1.17
f_0	Average feeding level of individuals feeding on the background resource	0.6
r_{pp}	Growth rate of primary productivity	10

Table 2.4a: Nominal values of the species-specific parameters of the multispecies *mizer* model: sprat, sandeel, Norway pout, and herring. All nominal values were taken from Blanchard et al. (2014) excluding R_{max}, which was taken from Spence et al. (2016).

Parameter	**Description**	**Sprat**	**Sandeel**	**Norway pout**	**Herring**
a	Length-weight converter	0.007	0.001	0.009	0.002
b	Length-weight converter	3.01	3.32	2.94	3.43
W_∞	Asymptotic weight	32.9	35.6	100.4	333.9
W_{mat}	Weight at maturity	12.55	3.56	22.72	98.52
W_{min}	The size class of recruits	0.001	0.001	0.001	0.001
β	Preferred predator prey mass-ratio	51076	398849	21.5	280540
σ	Width of the prey size preference	0.8	1.9	1.5	3.2
L_{25}	Length at which 25% of the stock is selected by fishing gear	7.64	9.83	8.69	10.13
L_{50}	Length at which 50% of the stock is selected by fishing gear	8.14	11.82	12.24	20.79
α	Assimilation efficiency	0.6	0.6	0.6	0.6
h	Maximum food intake rate	18.18	27.41	32.88	35.03
ks	Coefficient of standard metabolism	3.64	5.48	6.58	7.01
k	Activity coefficient	0	0	0	0
γ	Volumetric search rate	3.79 x 10^{-11}	1.83 x 10^{-11}	1.03 x 10^{-11}	1.46 x 10^{-11}
R_{max}	Maximum recruitment	4.22 x 10^{11}	2.03 x 10^{11}	2.63 x 10^{13}	3.55 x 10^{11}
N_0	Initial population size	6.11 x 10^{8}	5.74 x 10^{8}	2.50 x 10^{8}	9.57 x 10^{7}
$eRepro$	Reproductive efficiency	1	1	1	1

Table 2.4b: Nominal values of the species-specific parameters of the multispecies *mizer* model: dab, whiting, common sole, and grey gurnard. All nominal values were taken from Blanchard et al. (2014) excluding R_{max}, which was taken from Spence et al. (2016).

Parameter	**Description**	**Dab**	**Whiting**	**Sole**	**Grey gurnard**
a	Length-weight converter	0.01	0.006	0.008	0.004
b	Length-weight converter	2.99	3.08	3.02	3.20
W_∞	Asymptotic weight	324.4	1192.3	866.1	668.1
W_{mat}	Weight at maturity	21.20	75.14	78.12	39.11
W_{min}	The size class of recruits	0.001	0.001	0.001	0.001
β	Preferred predator prey mass-ratio	191	22	381	283
σ	Width of the prey size preference	1.9	1.5	1.9	1.8
L_{25}	Length at which 25% of the stock is selected by fishing gear	11.52	19.81	16.40	19.81
L_{50}	Length at which 50% of the stock is selected by fishing gear	17.04	29.02	25.80	29.02
α	Assimilation efficiency	0.6	0.6	0.6	0.6
h	Maximum food intake rate	30.69	28.54	22.56	19.38
ks	Coefficient of standard metabolism	6.14	5.71	4.51	3.88
k	Activity coefficient	0	0	0	0
γ	Volumetric search rate	5.68 x 10^{-11}	8.92 x 10^{-11}	3.81 x 10^{-11}	3.59 x 10^{-11}
R_{max}	Maximum recruitment	1.14 x 10^{10}	2.31 x 10^{11}	9.63 x 10^{9}	1.17 x 10^{11}
N_0	Initial population size	9.80 x 10^{7}	3.46 x 10^{7}	4.47 x 10^{7}	5.50 x 10^{7}
$eRepro$	Reproductive efficiency	1	1	1	1

Table 2.4c: Nominal values of the species-specific parameters of the *mizer* model: European plaice, haddock, Atlantic cod, and saithe. All nominal values were taken from Blanchard et al. (2014) excluding R_{max}, which was taken from Spence et al. (2016).

Parameter	**Description**	**Plaice**	**Haddock**	**Cod**	**Saithe**
a	Length-weight converter	0.007	0.005	0.005	0.007
b	Length-weight converter	3.10	3.16	3.17	3.08
W_{∞}	Asymptotic weight	2975.9	3484.7	40044.3	16856.4
W_{mat}	Weight at maturity	104.66	164.91	1606.00	1076.46
W_{min}	The size class of recruits	0.001	0.001	0.001	0.001
β	Preferred predator prey mass-ratio	113	558	66	40
σ	Width of the prey size preference	1.6	2.1	1.3	1.1
L_{25}	Length at which 25% of the stock is selected by fishing gear	11.52	19.09	13.20	35.32
L_{50}	Length at which 50% of the stock is selected by fishing gear	17.04	24.35	22.87	43.55
α	Assimilation efficiency	0.6	0.6	0.6	0.6
h	Maximum food intake rate	14.62	34.24	61.58	37.39
ks	Coefficient of standard metabolism	2.92	6.85	12.32	7.48
k	Activity coefficient	0	0	0	0
γ	Volumetric search rate	3.45 x 10^{-11}	4.97 x 10^{-11}	1.92 x 10^{-10}	1.47 x 10^{-10}
R_{max}	Maximum recruitment	2.39 x 10^{13}	2.38 x 10^{12}	8.36 x 10^{9}	5.63 x 10^{11}
N_0	Initial population size	1.66 x 10^{7}	1.47 x 10^{7}	2.08 x 10^{6}	4.16 x 10^{6}
$eRepro$	Reproductive efficiency	1	1	1	1

Table 2.5: Nominal values of the interaction matrix θ of the multispecies *mizer* model (taken from Blanchard et al. (2014)). This represents the spatial overlap within and between species as a fraction.

	Sprat	Sandeel	Norway Pout	Herring	Dab	Whiting	Sole	Grey Gurnard	Plaice	Haddock	Cod	Saithe
Sprat	0.729	0.034	0.064	0.274	0.362	0.265	0.298	0.175	0.371	0.081	0.338	0.017
Sandeel	0.034	0.681	0.049	0.059	0.097	0.075	0.060	0.060	0.078	0.094	0.099	0.016
Norway Pout	0.064	0.049	0.797	0.298	0.091	0.300	0.017	0.306	0.079	0.549	0.325	0.295
Herring	0.274	0.059	0.298	0.659	0.290	0.374	0.200	0.275	0.278	0.348	0.405	0.126
Dab	0.362	0.097	0.091	0.290	0.808	0.334	0.380	0.220	0.565	0.132	0.416	0.031
Whiting	0.265	0.075	0.300	0.374	0.334	0.709	0.192	0.371	0.295	0.392	0.441	0.102
Sole	0.298	0.060	0.017	0.200	0.380	0.192	0.716	0.107	0.391	0.034	0.258	0.012
Grey Gurnard	0.175	0.060	0.306	0.275	0.220	0.371	0.107	0.880	0.165	0.357	0.352	0.124
Plaice	0.371	0.078	0.079	0.278	0.565	0.295	0.391	0.165	0.719	0.112	0.350	0.033
Haddock	0.081	0.094	0.549	0.348	0.132	0.392	0.034	0.357	0.112	0.858	0.396	0.262
Cod	0.338	0.099	0.325	0.405	0.416	0.441	0.258	0.352	0.350	0.396	0.787	0.209
Saithe	0.017	0.016	0.295	0.126	0.031	0.102	0.012	0.124	0.033	0.262	0.209	0.664

Chapter 3

Global sensitivity analysis of the *mizer* marine ecosystem model

3.1 Abstract

Models often include many highly uncertain parameters, some of which may have a large impact on the model projections. Ignoring the sensitivity of the model projections to uncertain parameter values risks suboptimal, ineffective, or potentially damaging management decisions if the model is used to support policy. Despite this, there are few examples of a sensitivity analysis being applied to a complex model, particularly in marine ecosystem modelling. We aim to fill this research gap by conducting a global sensitivity analysis of a widely-used size spectrum model, known as *mizer*. We apply both Sobol' variance- and derivative-based methods of sensitivity analysis to the trait-based *mizer* model, which includes 24 uncertain parameters, to allow for a direct comparison between the sensitivity indices given by each of these methods. We also apply a derivative-based sensitivity analysis to the North Sea multispecies version of *mizer*, which includes over 300 uncertain parameters. The sensitivity of multiple model outputs, such as community biomass and the Large Fish Indicator (LFI), are considered. We use the results of the sensitivity analyses to discuss: (1) the relationship between the variance- and derivative-based sensitivity indices; (2) areas in which to focus future research to reduce the uncertainty in the parameters associated with the greatest sensitivity indices; and (3) the convergence of the variance- and derivative-based sensitivity indices. Overall, we hope that this research will enable us to produce more accurate model projections and ensure multi-species size spectrum models such as *mizer* are well placed to support ecosystem-based fisheries management in the future.

3.2 Introduction

Marine ecosystem models vary in complexity from single-species models with relatively few parameters to multispecies models and whole ecosystem models that may contain hundreds or thousands of parameters (Plaganyi, 2007). Even marine ecosystem models that are considered to be of intermediate complexity may include many hundreds of parameters. For example, the *mizer* model (described in detail in Chapter 2) is a marine ecosystem model of intermediate complexity that is used to simulate the size dynamics of a fish community (Spence et al., 2016). Size-based models such as *mizer* tend to have relatively few parameters when compared with other types of marine ecosystem models as species are often not resolved in detail (Scott et al., 2014). In particular, the community and trait-based versions of *mizer* both include fewer than 25 parameters. However, species-specific information such as life history data and diet preferences can also be incorporated into the multispecies version of the *mizer* model (Blanchard et al., 2014). The number of parameters in the multispecies model therefore largely depends on the number of species in the modelled community. Previous applications of *mizer* have primarily focused on 12 common and commercially important fish species found in the North Sea (see Blanchard et al. (2014), Datta and Blanchard (2016) and Spence et al. (2016) for example), which resulted in more than 300 parameters being included in the model.

Although many of these parameters can be estimated using empirical data, it is often difficult or impossible to identify the value that some of the parameters should take with a high degree of certainty (Ward, 2009). Because of this, models such as *mizer* often include many highly uncertain parameters, some of which may have a large impact on the model outputs. Ignoring the sensitivity of the model projections to these uncertain parameter values risks suboptimal, ineffective, or potentially damaging management decisions if the model is used to support policy (Uusitalo et al., 2015). Fortunately, various methods of sensitivity analysis may be used to identify which of the parameters a model is most sensitive to, thereby enabling us to identify research areas in which to focus future data collection to reduce the uncertainty in the model outputs and thus help to prevent misguided management decisions (Saltelli et al., 2008).

Although the importance of conducting a sensitivity analysis is widely recognised, the process is often deemed to be too time-consuming and is therefore frequently neglected in modelling frameworks (Arhonditsis et al., 2006; Roeder and Hill, 2009). A sensitivity analysis tends to be particularly computationally expensive due to the large number of required model evaluations, as well as potentially long model run times. The number of model evaluations can vary greatly depending on the method of sensitivity analysis used, the intended purpose of the sensitivity indices, and the number of uncertain model parameters. For example, 'local' sensitivity analysis tends to require fewer model evaluations than 'global' sensitivity analysis (Pianosi et al., 2016). Local methods of sensitivity analysis quantify the sensitivity of the model outputs to small variations in parameters around a nominal value, whilst global methods estimate sensitivity indices by varying parameters across the entire range of possible values (Iooss and Lemaître, 2014). Local methods typically employ a One-At-a-Time (OAT) approach in which one parameter is varied whilst all other parameters are fixed at their nominal value (Pianosi et al., 2016). Conversely, global methods tend to estimate the sensitivity of the model outputs by varying all parameters at the same time (referred to as the All-At-a-Time (AAT) approach), although they may also use the OAT approach (Pianosi et al., 2016). Implementing a global, AAT approach therefore requires a much larger number of model evaluations than a local, OAT approach as the former requires more extensive sampling of the parameter space (Pianosi et al., 2016).

The Sobol' variance-based method is a global AAT approach that is often seen as the 'gold standard' of sensitivity analysis (Donders et al., 2015). This method allows us to estimate total-effect indices, which result from the decomposition of the model output variance into contributions associated with individual parameters, as well as their interactions with every other parameter in the model (Chen et al., 2004). However, the number of required model evaluations may not be computationally feasible with large numbers (e.g. >10) of model parameters (Iooss et al., 2012). Instead, Derivative-based Global Sensitivity Analysis (DGSA) may be used to estimate the upper bound of the total-effect sensitivity indices by integrating the squared partial derivatives of the model outputs (Sobol' and Kucherenko, 2009, 2010; Iooss et al., 2012). Using DGSA instead of the Sobol' variance-based method has been shown to reduce the computational time required to estimate the total-effect sensitivity indices by many orders of magnitude (Kucherenko et al., 2009).

The computational time required to undertake a sensitivity analysis also depends on the intended use of the sensitivity indices. For example, a sensitivity analysis may be used solely to differentiate between parameters with negligible and non-negligible impacts on the model outputs, a process known as screening, or to rank the parameters according to their influence on the model outputs (Saltelli et al., 2008). The number of model evaluations required for the parameter rankings to converge (i.e. remain stable with increasing numbers of model evaluations) is usually much higher than for parameter screenings (Sarrazin et al., 2016). However, it is difficult to identify the exact number of model evaluations required for either ranking or screening the parameters prior to conducting the sensitivity analysis itself (Pianosi et al., 2016). Instead, it is necessary to assess whether the screening, ranking, and/or the exact value of the sensitivity indices has reached convergence once the sensitivity analysis is complete. Methods to assess convergence include the estimation of 95% confidence intervals via bootstrapping of the sensitivity indices and rank correlation amongst bootstrap resamples (Sarrazin et al., 2016).

Despite recent advances in computing power making it more feasible to run large numbers of model evaluations, the concurrent increase in the number of parameters and/or the run times of models such as *mizer* has resulted in relatively few examples of an extensive global sensitivity analysis being applied to marine ecosystem models (Arhonditsis et al., 2006; Morris et al., 2014; Hines et al., 2018). One such example includes the application of two different methods of sensitivity analysis to a marine ecosystem model known as StrathE2E; The Morris method (Morris, 1991), a popular OAT approach often used for parameter screening, was first used to identify parameters with a non-negligible impact on the model outputs (Morris et al., 2014). The Sobol' variance-based method was then used to rank the remaining parameters in order of sensitivity (Morris et al., 2014). However, the analysis required 540,000 model evaluations to reach convergence (Morris et al., 2014), a potentially computationally infeasible number for larger models. Other examples apply only OAT approaches (e.g. Niiranen et al. (2012) and Livingston (2013)), do not run each model evaluation to equilibrium (e.g. Morris et al. (2014) and Zhang et al. (2015)), and/or quantify the sensitivity of the model outputs to groups of parameters (e.g. Zhang et al. (2015)), thereby making it difficult to attribute model output sensitivity to specific parameters. Additionally, there are few examples in the literature of a direct comparison between the performance of Sobol' variance- and derivative-based methods of sensitivity analysis when applied to complex models (see Iooss et al. (2012) for example), and none that we are aware of in marine ecosystem modelling. Please note that although numerous sensitivity or uncertainty analyses have been applied to models of marine ecosystem flow networks (see Borrett et al. (2016); Hines et al. (2018); Bentley et al. (2019a) for example), the outputs of these models and the methods used to quantify sensitivity are typically not directly comparable with those that are of interest here and they are therefore not discussed in detail in this chapter.

In this study we aim to fill these research gaps by conducting a global sensitivity analysis of the trait-based and North Sea multispecies versions of the *mizer* model, considering the sensitivity of multiple model outputs to individual parameters based on equilibrated model evaluations. Both variance- and derivative-based methods will be applied to the trait-based version of the model to allow for a direct comparison between these two methods of sensitivity analysis. A derivative-based sensitivity analysis will also be applied to the North Sea multispecies version of the model. Not only will this research help us to better understand how parameter uncertainties impact the outputs of a size-based marine ecosystem model, it will also highlight where further research should be focused to reduce these uncertainties.

Narrowing down the range of possible values a parameter may realistically take will enable us to produce more accurate model projections, thus helping to ensure the model is well placed to support ecosystem-based fisheries management in the UK.

3.3 Methods

We applied both variance- and derivative-based methods to evaluate the sensitivity of the *mizer* model to small variations in the values of the model parameters. The two different methods of sensitivity analysis are formally introduced in Section 3.3.1, whilst their application to the *mizer* model is described in Section 3.3.2. Finally, the methods used to analyse the convergence of the sensitivity indices are described in Section 3.3.3.

3.3.1 Variance- and derivative-based sensitivity analysis

In this section, the Sobol' variance- and derivative-based methods of sensitivity analysis are formally introduced using the definitions and notation of Touzani and Busby (2014).

Let $Y = f(\boldsymbol{X})$ represent the *mizer* model, where Y represents the model outputs (see Chapter 2, Section 2.6 for further details of the model outputs), $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a d-dimensional input vector that represents the uncertain parameters in the model (which are assumed to be independent and uniformly distributed over a unit hypercube i.e. $\boldsymbol{X} \in [0,1]^d$), and $f : [0,1]^d \to \mathbb{R}$ is a function that maps the model inputs to the model outputs. Please note that although a uniform distribution is potentially unrealistic, we chose to use this distribution due to the lack of available data with which to accurately determine the distributions of all of the parameters in the *mizer* model. A uniform distribution is also thought to provide an unbiased, conservative estimate of the plausible range of model outputs (Hines et al., 2018).

Sobol' variance-based sensitivity analysis

The Sobol' method of sensitivity analysis, also referred to as the variance-based method, involves decomposing the variance of the model outputs Y into contributions (or summands) associated with each uncertain parameter, as well as their interactions with all of the other parameters in the model (Chen et al., 2004):

$$f(\boldsymbol{X}) = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \ldots + f_{1,\ldots,d}(X_1, \ldots, X_d) \tag{3.1}$$

where f_0 is a constant, f_i represents the 'main effect' of a given uncertain parameter, and f_{ij} represents the 'interaction effect' of two uncertain parameters.

Assuming each term in Equation 3.1 is square-integrable with zero average, all of the summands are mutually orthogonal and the decomposition of variance of $f(\boldsymbol{X})$ is unique (Sobol, 1993; Nossent et al., 2011). All terms in Equation 3.1 may therefore be defined as:

$$f_0 = E(Y) \tag{3.2}$$

$$f_i(X_i) = E(Y|X_i) - f_0 \tag{3.3}$$

$$f_{ij}(X_i, X_j) = E(Y|X_i, X_j) - f_i - f_j - f_0 \tag{3.4}$$

where $E(Y)$ represents the expectation of the output Y and $E(Y|X_i)$ represents the conditional expectation of the output Y given the uncertain parameter X_i. Similar formulae can also be derived for higher-order interactions.

If we also assume that $f(\boldsymbol{X})$ is square integrable, the total variance V of the model outputs Y can be defined as:

$$V = E(Y^2) - f_0^2 = \sum_{i=1}^{d} V_i + \sum_{1 \leq i < j \leq d} V_{ij} + \ldots + V_{1,2,\ldots,d} \tag{3.5}$$

where $V_i = V[E(Y|X_i)]$ is the variance of the conditional expectation that measures the main effect of the uncertain parameter X_i on the model output Y and $V_{ij} = V[E(Y|X_i, X_j)] - V_i - V_j$ is the variance of the conditional expectation that measures the joint effect of parameters X_i and X_j on the model output Y minus their first order effects. Decomposing the variance of the model outputs in this way enables the quantification of the Sobol' variance-based sensitivity indices $S_{i_1,\ldots,i_s}$ and total-effect indices S_{T_i}:

$$S_{i_1,\ldots,i_s} = \frac{V_{i_1,\ldots,i_s}}{V} \tag{3.6}$$

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \ldots \tag{3.7}$$

where $1 \leq i_1 < \ldots < i_s \leq d$ and $s = 1, \ldots, d$. $S_i = V_i/V$ is referred to as the first-order sensitivity index, $S_{ij} = V_{ij}/V$ for $i \neq j$ is referred to as the second-order sensitivity index and so on for higher-order effects. The total-effect index S_{T_i} measures the overall contribution of a given uncertain parameter to the total model output variance. Whilst the sensitivity indices defined in Equation 3.6 sum to one, the sum of all of the total-effect indices will be greater than one due to the double counting of interactions. Importantly, the total-effect indices may be estimated without quantifying all higher-order effects and this measure can therefore be used to identify parameters with a non-negligible impact on the model outputs. Often, a threshold of 0.01 is used to determine which parameters have a non-negligible impact on the model outputs (Touzani and Busby, 2014).

Derivative-based sensitivity analysis

A relatively new method of global sensitivity analysis, known as Derivative-based Global Sensitivity Analysis (DGSA), may be used to estimate the upper bound of the Sobol' total-effect indices S_{T_i} using far fewer model evaluations than is required by the Sobol' variance-based method. The derivative-based method involves integrating the squared partial derivatives of the model output Y (Sobol' and Kucherenko, 2009, 2010; Iooss et al., 2012).

If we assume that $\partial f(\boldsymbol{X})/\partial x_i$ for $i = 1, \dots d$ are square-integrable, the derivative-based sensitivity indices are defined as:

$$v_i = \mathbb{E}\left[\left(\frac{\partial f(\boldsymbol{X})}{\partial x_i}\right)^2\right] = \int \left(\frac{\partial f(\boldsymbol{X})}{\partial x_i}\right) d\boldsymbol{x} \tag{3.8}$$

Monte Carlo techniques or Latin Hypercube Sampling can be used to evaluate the integrals in Equation 3.8 and provide an empirical estimation of the derivative-based sensitivity indices:

$$\hat{v_i} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial f(\boldsymbol{X}_j)}{\partial x_i}\right)^2 \tag{3.9}$$

Link between variance- and derivative-based sensitivity indices

Assuming that the uncertain parameter X_i follows a uniform distribution between 0 and 1 for $i = 1, \dots, d$, the link between the variance-based total-effect indices S_{T_i} (referred to simply as the variance-based sensitivity indices from here on) and the derivative-based sensitivity indices v_i is shown by Sobol' and Kucherenko (2009) to be:

$$S_{T_i} \leq \frac{v_i^*}{V} = \Upsilon_i \tag{3.10}$$

where $v_i^* = v_i/\pi^2$ is a version of v_i that is scaled to be directly comparable with the variances used to calculate the variance-based sensitivity indices.

However, it is important to note that if $f(\boldsymbol{X})$ is highly non-linear, there may be differences in the rankings of the most important parameters when using Υ_i instead of S_{T_i}. A non-linear function may also result in Υ_i exceeding one despite S_{T_i} being bounded by zero and one (Lamboni et al., 2013). In this situation, the upper bound Υ_i is deemed to be 'useless' (Lamboni et al., 2013). In cases where Υ_i does not exceed one, Υ_i may still be much greater than S_{T_i} when applying the derivative-based method to a complex model, making it difficult to determine which uncertain parameters have a non-negligible impact on the model outputs (Lamboni et al., 2013). A possible solution to this issue is to use a normalised version of Υ_i, although doing so breaks the link between the variance- and derivative-based sensitivity indices (Touzani and Busby, 2014):

$$\Upsilon_i^* = \frac{v_i^*}{\sum_{j=1}^{d} v_j^*} \tag{3.11}$$

3.3.2 Sensitivity analysis of the *mizer* model

To evaluate the sensitivity of the *mizer* model to small variations in the model parameters, we applied both variance- and derivative-based methods of sensitivity analysis to the trait-based *mizer* model. We also applied a derivative-based sensitivity analysis to the North Sea multispecies *mizer* model. The variance-based method was not applied to the multispecies model due to the large number of model evaluations required to estimate the sensitivity indices.

In order to sample the parameter space of both versions of the *mizer* model, we assigned a uniform distribution to each parameter with an upper and lower limit of ± 10% of their nominal value (see Chapter 2, Tables 2.2 to 2.5 for a list of the parameters included in the sensitivity analysis and their nominal values). If the nominal value of a parameter was set to one and it could not be increased further, the upper and lower limits were set to 1 and 0.9 respectively. Conversely, if the nominal value of a parameter was zero and it could not take a negative value, the upper and lower limits were set to 0.1 and 0 respectively. Although the fishing effort parameter F can be varied at each time step within the model, we chose to maintain effort at a constant level for the duration of each model evaluation. In the trait-based model, the upper and lower limits of F were set to 0.1 and 0 respectively. In the multispecies model, the upper and lower limits of F were set to 1.5 and 0 respectively. A maximum fishing effort of 1.5 was chosen to reflect the mean maximum catch of each of the 12 modelled species (see

Chapter 2, Section 2.4.3) in the North Sea between 1957 and 2011 (`ices.dk`; Blanchard et al. 2014).

It is important to note that the values of some of the parameters included in the sensitivity analyses are typically calculated using other parameters in the *mizer* model. For example, the volumetric search rate γ is estimated using the maximum food intake rate h, the preferred predator-prey mass ratio β, the width of the prey size preference σ, and the carrying capacity of the background resource κ etc. (see Chapter 2, Equation 2.8). However, each parameter is treated as independent throughout the sensitivity analyses and therefore any relationships between the parameters are ignored (as in Borrett et al. (2016) for example). We chose to ignore these relationships so that the sensitivity indices reflect only the effects of the parameter in question, not the cumulative effects of multiple parameters.

Sampling of the parameter space

A stratified sampling technique, known as Latin Hypercube Sampling (LHS) (McKay et al., 1979), was used to generate the parameter sets required for both the variance- and derivative-based methods of sensitivity analysis. LHS was used as it tends to provide a better representation of the parameter space than random sampling, as random sampling may result in a cluster of points within the parameter space (Agarwal et al., 2012; O'Sullivan and Perry, 2013). Such clustering means that certain regions of the parameter space may be undersampled. LHS prevents this clustering by dividing the parameter space into equiprobable subregions and selecting samples such that each subregion is sampled once and only once (Koziel and Leifsson, 2013) (see Figure 3.1 for a simple example of the differences between random sampling and LHS).

For the variance-based sensitivity analysis, we used LHS to generate two matrices (matrix A and matrix B) with dimensions $N \times d$ (where N is equal to the sample size and d represents the number of uncertain parameters included in the sensitivity analysis). Matrices A and B were then used to build d further matrices $A_{B,i}$ for $i = 1, ..., d$ such that the ith column of $A_{B,i}$ was equal to the ith column in matrix B, whilst all other columns originated from matrix A (Merle, 2016). For the trait-based model, we set N equal to 1000 and d equal to 24, resulting in a total of 26,000 parameter combinations (Table 3.1). The trait-based *mizer* model was then evaluated for all parameter sets given in $A_{B,i}$ and the variance-based sensitivity indices were quantified using the `soboljansen()` function from the `sensitivity` R package (Iooss et al., 2018), which is based on the estimator given in Jansen et al. (1994), Jansen (1999) and

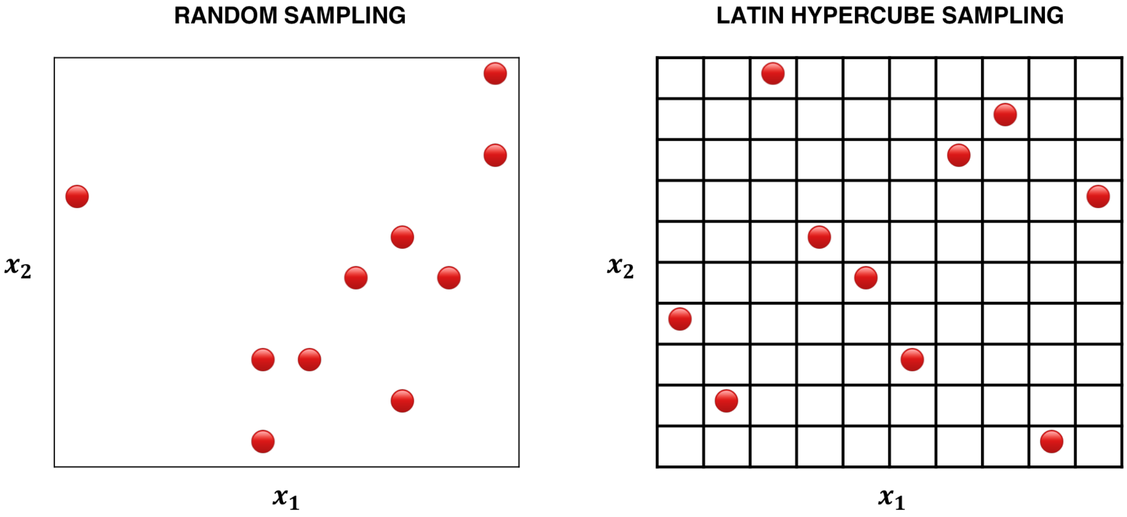


Figure 3.1: An example of random sampling (left) and Latin Hypercube Sampling (LHS; right) of a two-dimensional parameter space. The random sample exhibits clustering in one corner of the parameter space, whilst the sample points selected using LHS are more evenly distributed. LHS avoids clustering by dividing the parameter space into equiprobable subregions (represented as rows and columns) and randomly selecting points within each subregion such that each subregion is sampled once and only once (i.e. none of the sample points share a row or column with any other sample point). Please note that in this example, the parameters are assumed to be uniformly distributed.

Saltelli (2002):

$$S_i = 1 - \frac{\frac{1}{2N}\sum_{j=1}^{N}(Y_{B_j} - Y_{A_{B_{i,j}}})^2}{V} \tag{3.12}$$

$$S_{T_i} = \frac{\frac{1}{2N}\sum_{j=1}^{N}(Y_{A_j} - Y_{A_{B_{i,j}}})^2}{V} \tag{3.13}$$

For the derivative-based sensitivity analysis, we used LHS to generate a single matrix (matrix C) with dimensions $N \times d$. The derivative-based sensitivity indices were estimated by comparing the model outputs obtained by evaluating the model for all parameter sets given in matrix C with those obtained by running the same parameter sets with a small increase (+0.0001 on the scale of the Latin Hypercube) in one parameter value per model evaluation (see Equation 3.9). These methods were applied to both the trait-based and North Sea multispecies *mizer* models with an N of 1000 and d equal to 24 and 306 respectively.

The total number of model evaluations required for both the variance- and derivative-based methods of sensitivity analysis is shown in Table 3.1. Although we used a similar number of model evaluations to conduct the variance- and derivative-based sensitivity analyses of the trait-based *mizer* model, we expect the derivative-based sensitivity indices to converge much more quickly than the variance-based sensitivity indices.

Table 3.1: The number of model evaluations conducted for both the variance- and derivative-based sensitivity analyses of the trait-based and multispecies *mizer* models.

Method	**Model version**	**Number of model evaluations**
Variance-based	Trait-based	26,000
Variance-based	Multispecies	-
Derivative-based	Trait-based	25,000
Derivative-based	Multispecies	307,000

Model equilibrium

All model evaluations were run in *mizer* until the biomass of each species reached equilibrium. In many cases, the chosen parameter sets resulted in a slow decline in the biomass of a given species to infinitesimal values. We therefore assumed a species had reached equilibrium if the total biomass of the species dropped below the egg weight (w_0; 0.001g). The biomass of a species was also deemed to have reached equilibrium if it remained within $\pm$10^{-6}g of the mean throughout the final 400 time steps (100 model years). However, a large number of model evaluations displayed both regular and irregular periodicity (often referred to as quasiperiodicity; Huggett (2003)) in the biomass of a given species. In cases where there was regular periodicity, we used the `ADCF()` function in the `dCovTS` R package (Pitsillou and Fokianos, 2016) to identify the time lag L (or period length) (Zhou, 2012). The time series was then assumed to have reached equilibrium if the ratio between the final L time steps and the penultimate L time steps was within 1$\pm$10^{-6}. Where there was irregular periodicity, we assumed equilibrium had been reached when a linear regression of the biomass of a given species indicated there was no significant change (p > 0.05) in biomass over time. The linear regression was applied over the second half of the time series to avoid the initial 'spin-up' period of the model and was used only when all other checks for equilibrium had failed.

Model outputs

The sensitivity of seven model outputs, including the community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, and the slope of the community spectrum (see Chapter 2, Section 2.6 for further details of these model outputs), were considered in the sensitivity analyses of both the trait-based and multispecies *mizer* models. These model outputs were selected as they have been widely used in the literature as proposed indicators of ecosystem health in the North Sea (see Nicholson and Jennings (2004), Blanchard et al. (2014), Thorpe et al. (2015), and Marshall et al.

(2016) for example). The sensitivity indices were estimated using the mean of each model output in the final 400 time steps (100 model years) of each model evaluation.

3.3.3 Convergence of the sensitivity indices

It is important to ensure that the sensitivity indices have reached convergence as they may change considerably depending on the number of model evaluations included in the analysis. However, there are three different definitions of convergence depending on the intended purposes of the sensitivity analysis, including convergence of the sensitivity indices themselves, the parameter rankings, and the parameter screening (Sarrazin et al., 2016). The sensitivity indices are assumed to have reached convergence when their value remains stable with increasing numbers of model evaluations (Sarrazin et al., 2016). This type of convergence requires the greatest number of model evaluations and is difficult to achieve when applying a sensitivity analysis to a complex model with large numbers of uncertain parameters (Sarrazin et al., 2016). Conversely, parameter rankings and screenings tend to converge with far fewer model evaluations, with screening usually requiring the least number of model evaluations (Sarrazin et al., 2016). The parameter rankings are deemed to have reached convergence when the order of the parameters from high to low sensitivity does not change with added model evaluations (Sarrazin et al., 2016). Finally, the parameter screenings reach convergence when the groups of parameters defined as having a negligible or non-negligible impact on the model outputs do not change with increasing numbers of model evaluations (Sarrazin et al., 2016).

Following the methods of Sarrazin et al. (2016), we assessed the convergence of both the variance- and derivative-based sensitivity indices by estimating the maximum width of the 95% confidence intervals via bootstrapping:

$$Stat_{indices} = \max_{i=1,\ldots,M}(S_i^{ub} - S_i^{lb}) \tag{3.14}$$

where S_i^{ub} and S_i^{lb} are the upper and lower bounds of the 95% confidence interval of the sensitivity indices associated with the i-th parameter. If $Stat_{indices}$ is close to zero, the sensitivity indices are assumed to have reached convergence. An arbitrary threshold of 0.05 is often used to indicate convergence, but this applies only when using normalised sensitivity indices that vary between 0 and 1 (Sarrazin et al., 2016). The variance-based sensitivity indices S_{T_i} are bounded by 0 and 1, whilst the derivative-based sensitivity indices Υ_i are not. Because of this, we estimated the normalised derivative-based sensitivity indices Υ_i^* as defined in Equation 3.11. The normalised sensitivity indices were used throughout the assessment of convergence to maintain consistency.

To identify whether the rankings of the parameters included in the sensitivity analyses had reached convergence, we used an adjusted and weighted rank correlation coefficient using pairs of bootstrap resamples (Sarrazin et al., 2016):

$$\rho_{s,j,k} = \sum_{i=1}^{M} |R_i^j - R_i^k| \frac{\max\limits_{j,k}(S_i^j, S_i^k)^2}{\sum_{i=1}^{M} \max\limits_{j,k}(S_i^j, S_i^k)^2} \tag{3.15}$$

where S_i^j and S_i^k are the sensitivity indices and R_i^j and R_i^k are the ranks of the i-th parameter as estimated using the j-th and k-th bootstrap resamples. The rank correlation coefficient is weighted by the sensitivity indices in an attempt to ensure that changes in the rankings of the parameters associated with very low sensitivities have less of an effect on $\rho_{s,j,k}$ than changes in the rankings of the parameters associated with the greatest sensitivities (Sarrazin et al., 2016). The parameter rankings are assumed to have reached convergence when the value of the 95% quantile of the rank correlation coefficients from all possible pairs of bootstrap resamples falls below one (Sarrazin et al., 2016). This threshold indicates that the average distance between the rankings of the parameters with the greatest sensitivities is less than one rank position across all bootstrap resamples (Sarrazin et al., 2016).

Finally, the convergence of the parameter screenings depends on the chosen threshold between those parameters that are considered to have a negligible and non-negligible impact on the model outputs. Although a threshold of 0.01 is often used to distinguish between parameters with a negligible or non-negligible impact, we used a threshold of 0.05 to identify parameters associated with 'lower sensitivity' as described in Sarrazin et al. (2016). The maximum width of the 95% confidence interval across all parameters deemed to have a negligible impact on the model outputs was then used to assess the convergence of the parameter screenings (Sarrazin et al., 2016):

$$Stat_{screening} = \max_{x_i \in X_0} (S_i^{ub} - S_i^{lb}) \tag{3.16}$$

where x_i is the i-th parameter and X_0 represents the set of parameters with a sensitivity index of less than 0.05. The parameter screenings are assumed to have reached convergence when $Stat_{screening}$ falls below 0.05 (Sarrazin et al., 2016). Please note that in all cases, we used 1000 bootstrap resamples to estimate $Stat_{indices}$, $\rho_{s,j,k}$, and $Stat_{screening}$.

3.4 Results

Both the variance- and derivative-based methods of sensitivity analysis were applied to the trait-based *mizer* model (see Section 3.4.1). The derivative-based method was also applied to the North Sea multispecies *mizer* model (see Section 3.4.2). The sensitivity of seven model outputs, including community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, and the community slope, was determined for both versions of the model. These model outputs may be considered to be sensitive to a given parameter if either the variance-based sensitivity index S_{T_i} and/or the derivative-based sensitivity index Υ_i is greater than 0.01. As previously mentioned, if Υ_i exceeds one it is deemed to be 'useless' as an upper bound of S_{T_i}. However, Υ_i may still be used to better understand the sensitivity of the model outputs to the parameters included in the analyses (see below for further details).

3.4.1 Variance- and derivative-based sensitivity analysis of the trait-based *mizer* model

In this section, we present the results of the variance- and derivative-based sensitivity analyses of the trait-based *mizer* model. We first provide a direct comparison of the variance- and derivative-based sensitivity indices (S_{T_i} and Υ_i respectively), before describing the convergence of S_{T_i} and the normalised derivative-based sensitivity indices Υ_i^* and comparing the rankings of the parameters based on S_{T_i} and Υ_i^*.

Comparison of the variance- and derivative-based sensitivity indices

Both the variance- and derivative-based methods resulted in similar patterns of sensitivity across all of the parameters and model outputs considered in the sensitivity analyses of the trait-based *mizer* model. For example, both methods of sensitivity analysis showed that all of the model outputs were sensitive (S_{T_i} and Υ_i > 0.01) to the exponent of the background resource λ, the scaling of food intake n, the scaling of standard metabolism p, and the search volume exponent q (Figure 3.2 and 3.3). However, Υ_i was often much greater than S_{T_i} for many of these 'high sensitivity' parameters (Figure 3.3), especially for the mean weight and community slope indicators (Figure 3.2). The sensitivity of the community slope to λ differed by the largest amount between the two methods of sensitivity analysis, with an S_{T_i} of 0.52 and an equivalent Υ_i of 23468.6 (Figure 3.2).

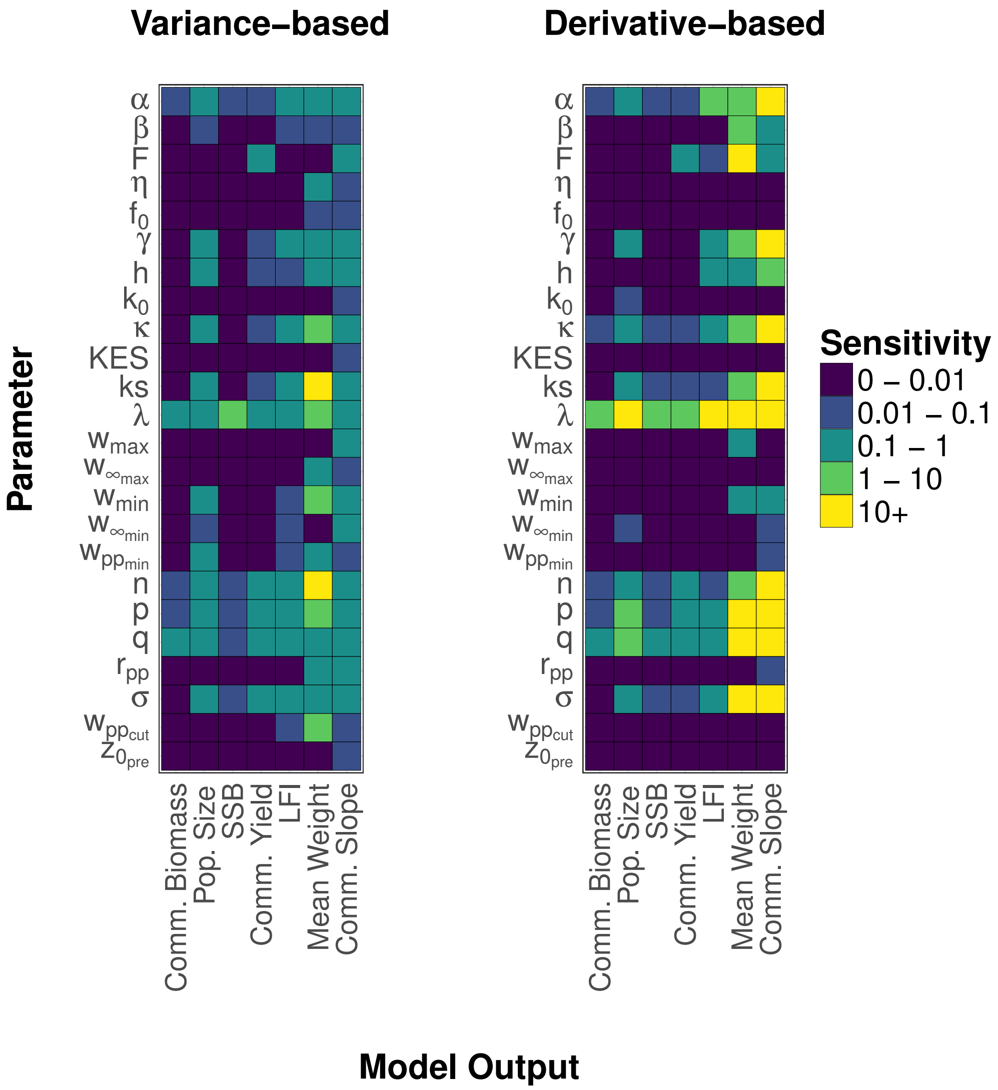


Figure 3.2: The variance- (S_{T_i}; left) and derivative-based (Υ_i; right) sensitivity indices of the model outputs to the parameters of the trait-based *mizer* model. Green and yellow indicate that the sensitivity index exceeds one and is therefore not useful as an upper bound of S_{T_i}, whilst purple indicates the sensitivity index is less than 0.01 and the parameter is thus deemed to have a negligible impact on the model output.

The derivative-based sensitivity indices suggest that the model outputs were insensitive (S_{T_i} and Υ_i < 0.01) to the conversion factor used to calculate weight at maturity from asymptotic weight η, the average feeding level of individuals feeding mainly on the background resource f_0, body size at the edge of the knife-edge selectivity function KES, the asymptotic weight of the largest species in the model $w_{\infty_{max}}$, the maximum weight of the background resource $w_{pp_{cut}}$, and the coefficient of background mortality for the community spectrum $z_{0_{pre}}$ (Figure 3.2 and 3.3). However, the variance-based sensitivity indices suggest the mean weight and community slope indicators may in fact be sensitive to many of these parameters, particularly $w_{pp_{cut}}$ (Figure 3.2). Nevertheless, we must be cautious when interpreting the variance-based sensitivity indices for mean weight as S_{T_i} exceeds one for six of the model parameters, despite the index being bounded by 0 and 1 (Figure 3.2). These results suggest the sample size may not have been large enough for the variance-based sensitivity indices associated with

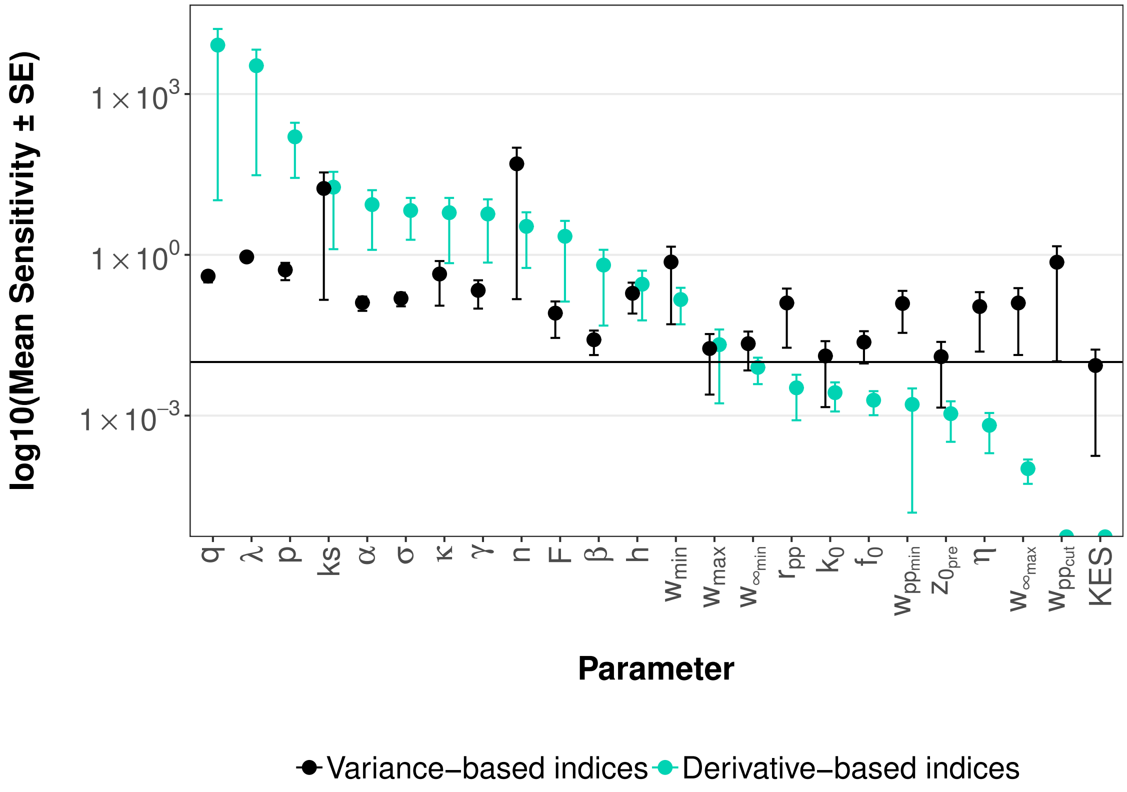


Figure 3.3: The log-transformed mean sensitivity (± standard error) of the trait-based *mizer* model to the parameters included in the sensitivity analysis. The black points represent the variance-based sensitivity indices S_{T_i}, whilst the teal points represent the derivative-based sensitivity indices Υ_i. The black line represents the threshold between those parameters with a negligible and non-negligible impact on the model outputs.

the mean weight indicator to converge (see the next section for further details). The S_{T_i} representing the influence of λ on SSB also exceeded one, whilst all other parameters and model outputs were unaffected by this issue (Figure 3.2).

The derivative-based sensitivity indices Υ_i associated with the mean weight and community slope model outputs exceeded one for many of the parameters (Figure 3.2), indicating the indices would not be useful as an upper bound of S_{T_i}. However, a linear regression of the log-transformed sensitivities showed that there was a significant positive relationship between the variance- and derivative-based sensitivity indices (λ_i = -0.24 + 0.95 $\cdot S_{T_i}$, r^2 = 0.37, p < 0.01; Figure 3.4). The coefficient of determination (r^2) was relatively low at 0.37, but this was largely caused by the derivative-based indices being much greater than the variance-based indices at high sensitivities (Figure 3.3 and 3.4). Such differences between the derivative- and variance-based indices occurred as very few of the variance-based indices exceeded the threshold of one, whilst a number of the derivative-based sensitivity indices exceeded this threshold. As previously suggested, the derivative-based method also underestimated the sensitivity of some of the parameters with low to mid-levels of variance-based sensitivity, although this issue affected fewer than 10% of the sensitivity indices (Figure 3.3 and 3.4).

The largest underestimations occurred when the variance-based sensitivity indices exceeded the threshold of one, whilst the derivative-based sensitivity indices did not (Figure 3.3 and 3.4).

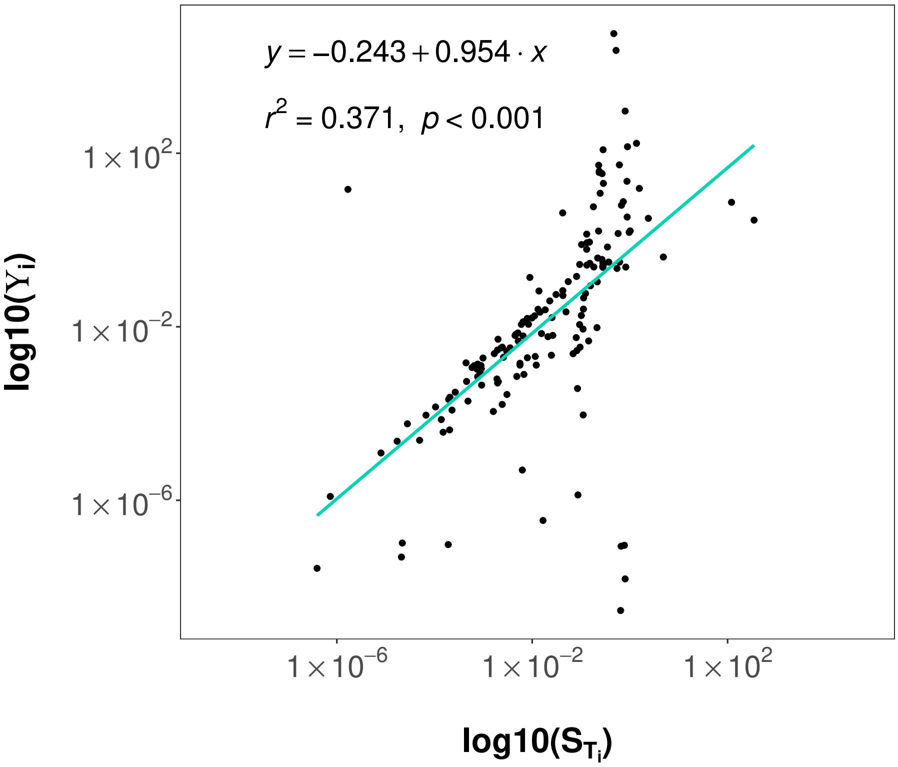


Figure 3.4: A comparison of the log-transformed variance- (S_{T_i}) and derivative-based (Υ_i) sensitivity of the model outputs to the parameters of the trait-based *mizer* model (n = 168). The teal line represents a linear regression of the log-transformed sensitivity indices.

Convergence of the variance- and derivative-based sensitivity indices

The sensitivity indices, parameter rankings, and parameter screenings are assumed to have reached convergence when $Stat_{indices}$ < 0.05 (Equation 3.14), $\rho_{s,j,k}$ < 1 (Equation 3.15), and $Stat_{screening}$ < 0.05 (Equation 3.16). When using the variance-based sensitivity indices S_{T_i}, $Stat_{screening}$ was below the threshold of 0.05 for the community biomass, population size, and SSB (Figure 3.5). A value of less than 0.05 indicates that we are successfully able to differentiate between those parameters with a negligible and non-negligible impact on these model outputs. The screening process identified two parameters that were associated with higher sensitivities (S_{T_i} > 0.05) for the community biomass and SSB: the exponent of the background resource λ and the search volume exponent q. A total of 12 parameters were associated with higher sensitivities for the population size, including λ, q, the scaling of the standard metabolism p, and the coefficient of standard metabolism ks, among others. $Stat_{screening}$ exceeded 0.05 for the fisheries yield, LFI, mean weight, and community slope model outputs (Figure 3.5) as the width of the confidence intervals of at least five parameters that had a non-negligible impact on the model outputs, including p, q, λ, σ, and the fishing effort F, exceeded 0.05 (not shown).

When using the normalised derivative-based sensitivity indices Υ_i^*, $Stat_{screening}$ was below the threshold of 0.05 for all seven model outputs excluding the LFI (Figure 3.5). $Stat_{screening}$ exceeded 0.05 for the LFI as the widths of the confidence intervals of p, the volumetric search rate γ, the carrying capacity of the background resource κ, and the width of the prey size preference σ were all greater than 0.07 (not shown). For all other model outputs, the screening process identified two parameters that were consistently associated with higher sensitivities: λ and q. n, p, and F were also associated with higher sensitivities for some of the model outputs, particularly fisheries yield.

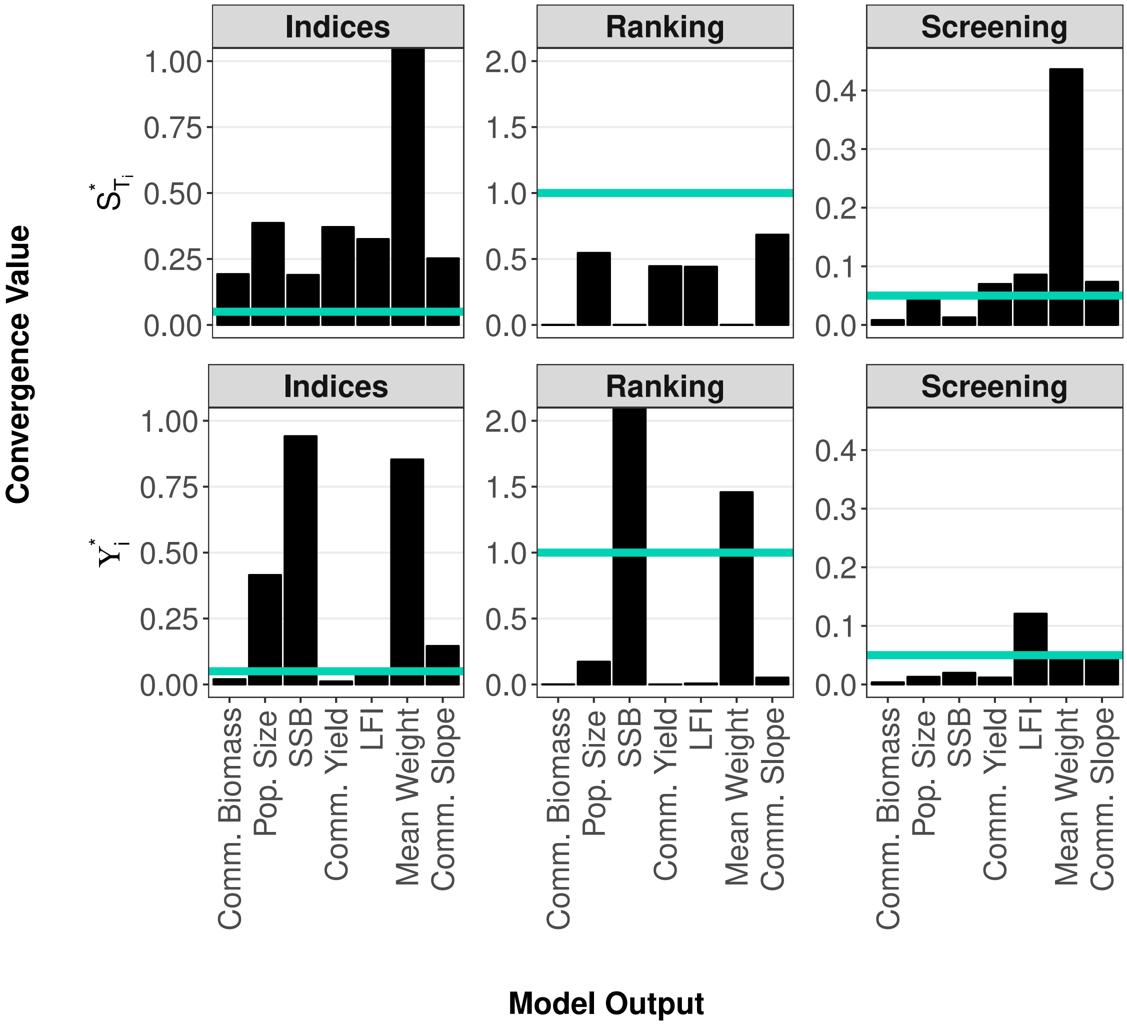


Figure 3.5: The convergence of the variance- (top) and derivative-based (bottom) sensitivity indices and the subsequent ranking and screening of the parameters for the seven outputs of the trait-based *mizer* model. The convergence value refers to the value of $Stat_{indices}$ (left), $\rho_{s,j,k}$ (middle), and $Stat_{screening}$ (right) respectively. The teal line represents the value below which the results are assumed to have reached convergence. Please note that the variance-based $Stat_{indices}$ and derivative-based $\rho_{s,j,k}$ for mean weight (equal to 2497.46 and 8.46 respectively) are not shown in full for plotting purposes.

When using S_{T_i}, $\rho_{s,j,k}$ was below the threshold of one for all of the model outputs (Figure 3.5).

A value of less than one indicates that the average rank of the parameters associated with the greatest sensitivities changed by fewer than one position across all bootstrap resamples and therefore we can be confident in the ranking of parameters for these model outputs. When using Υ_i^*, $\rho_{s,j,k}$ was below the threshold of one for all of the model outputs excluding SSB and mean weight (Figure 3.5). $\rho_{s,j,k}$ was equal to 1.46 for mean weight and 8.46 for SSB (Figure 3.5). The $\rho_{s,j,k}$ of the mean weight indicator exceeded one purely due to a change in the ranking of q from 1st to 3rd position and vice versa in two of the 1000 bootstrap resamples (not shown). On the other hand, the $\rho_{s,j,k}$ associated with SSB exceeded one due to relatively large changes in the rankings of λ, n, p, and F within the top 15 positions (not shown).

When using S_{T_i}, $Stat_{indices}$ was above the threshold of 0.05 for all of the model outputs (Figure 3.5). When using Υ_i, $Stat_{indices}$ was below the threshold of 0.05 for the community biomass, fisheries yield, and LFI, indicating the sensitivity indices reached convergence using the chosen threshold (Figure 3.5).

Overall ranking of the model parameters

When ranking the parameters from high to low sensitivity using the variance-based sensitivity indices S_{T_i}, the exponent of the background resource λ, the scaling of standard metabolism p, the search volume exponent q, and the scaling of the food intake n each appeared in the top five for six of the seven model outputs (Table 3.2). λ was ranked in first position for all model outputs excluding the mean weight and community slope indicators (Table 3.2). Instead, n and p were ranked in first position for the mean weight and community slope respectively (Table 3.2). λ did not appear in the top five parameters for mean weight but was ranked in second position for the community slope (Table 3.2). The assimilation efficiency α also appeared in the top five for population size, the LFI, mean weight, and the community slope, whilst n appeared in the top five for three of the model outputs (Table 3.2).

Table 3.2: The rankings of the five parameters with the greatest derivative- Υ_i and variance-based S_{T_i} sensitivity indices for each of the outputs of the trait-based *mizer* model. ** indicates the rankings did not reach convergence for a given model output.

	Comm. Biomass		**Pop. Size**		**SSB**		**Comm. Yield**	
Rank	Υ_i	S_{T_i}	Υ_i	S_{T_i}	**Υ_i	S_{T_i}	Υ_i	S_{T_i}
1	λ	λ	λ	λ	p	λ	λ	λ
2	q	q	α	q	q	q	q	F
3	n	n	σ	p	λ	n	p	q
4	p	α	κ	ks	F	α	α	p
5	α	p	γ	n	σ	p	ks	n

	LFI		**Mean Weight**		**Comm. Slope**	
Rank	Υ_i	S_{T_i}	**Υ_i	S_{T_i}	Υ_i	S_{T_i}
1	q	λ	λ	n	λ	p
2	λ	p	q	ks	F	λ
3	p	q	n	$w_{pp_{cut}}$	q	q
4	ks	ks	p	w_{min}	n	n
5	α	σ	α	κ	p	ks

Overall, the rankings of the parameters using the normalised derivative-based sensitivity indices Υ_i^* were similar to those described above. For example, λ was the only parameter to appear in the top five for all seven model outputs, whilst p and q appeared in the top five for all of the model outputs excluding population size (Table 3.2). Again, λ was ranked in first position for all of the model outputs excluding SSB and the LFI (Table 3.2). Instead, p and q were ranked in first position for SSB and the LFI respectively, whilst λ was ranked in third and second for these model outputs respectively (Table 3.2). α also appeared in the top five parameters for all model outputs except SSB and community slope (Table 3.2).

A total of 11 different parameters appeared in the top five across all model outputs when using S_{T_i}, whilst only ten appeared in the top five when using Υ_i^* (Table 3.2). The volumetric search rate γ was the only parameter to appear in the top five when using Υ_i^* but not when using S_{T_i} (Table 3.2). Conversely, both the minimum size of the community spectrum w_{min} and the maximum size of the background resource $w_{pp_{cut}}$ appeared in the top five when using S_{T_i} but not when using Υ_i^* (Table 3.2).

3.4.2 Derivative-based sensitivity analysis of the multispecies *mizer* model

In this section, we present the results of the derivative-based sensitivity analysis of the North Sea multispecies *mizer* model. We first describe the sensitivity of each model output to the species-specific parameters, species-independent parameters, and interaction matrix θ, before describing the convergence of the normalised derivative-based sensitivity indices Υ_i^* and the final rankings of the parameters based on Υ_i^*. It is important to note that the derivative-based sensitivity indices Υ_i are used to describe the results of the sensitivity analysis to maintain the link with the variance-based sensitivity indices S_{T_i}, whilst the normalised sensitivity indices Υ_i^* are used to assess the convergence of the sensitivity indices. Furthermore, Υ_i exceeded one in some cases and may therefore not be useful as an upper bound of the variance-based indices S_{T_i}. However, Υ_i may still be used to better understand the influence of each parameter on the model outputs.

Sensitivity to the species-specific parameters

The community biomass, fisheries yield, and SSB were largely insensitive to the species-specific parameters in the model, with the exception of the fishing effort F associated with Atlantic herring and sandeel (Figure 3.6). The population size was sensitive to all of the parameters associated with European plaice and saithe (Υ_i > 0.1), excluding the size class of their recruits W_{min} (Υ_i < 0.01) (Figure 3.6). The population size was also sensitive to a number of the parameters associated with Norway pout, particularly the assimilation efficiency α (Υ_i = 3596) (Figure 3.6).

The LFI, mean weight, and community slope displayed similar patterns of sensitivity across all of the species-specific parameters, with α, F, the coefficient of standard metabolism ks, the volumetric search rate γ, and the width of the prey size preference σ resulting in many of the greatest sensitivities (Figure 3.6). However, the sensitivity indices associated with these parameters varied greatly across different species, with many of the highest sensitivities being associated with the larger fish species, such as European plaice, saithe, and Atlantic cod (Figure 3.6). The community slope indicator was particularly sensitive to many of the parameters associated with saithe, including the reproductive efficiency $eRepro$, the length at which 25% of the stock is selected by the fishing gear L_{25}, and the length-weight converters a and b, all of which resulted in Υ_i exceeding 1500 (Figure 3.6)

Aside from the notable exceptions described above, the model outputs were largely insensitive (Υ_i < 0.01) to the length-weight converters a, predator-prey mass ratios β, reproductive efficiencies $eRepro$, selectivity lengths L_{25} and L_{50}, initial population sizes N_0, maximum

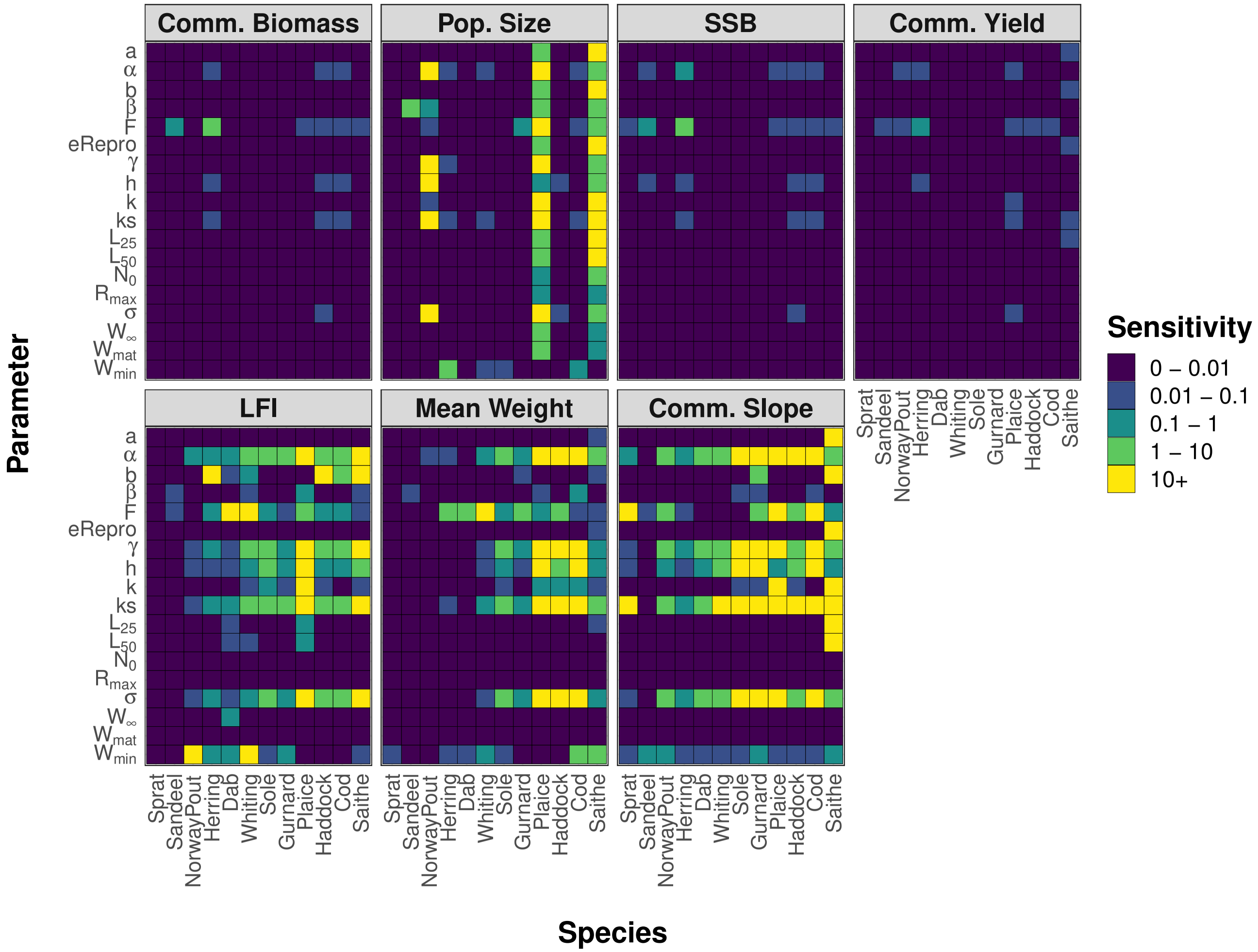


Figure 3.6: The derivative-based sensitivity Υ_i of the model outputs to the species-specific parameters in the North Sea multispecies *mizer* model. Green and yellow indicate that Υ_i exceeds one and is therefore not useful as an upper bound of S_{T_i}, whilst purple indicates that Υ_i is less than 0.01 and the parameter is thus deemed to have a negligible impact on the model output.

recruitment R_{max}, asymptotic weights W_{∞}, and maturation weights W_{mat} (Figure 3.6).

Sensitivity to the species-independent parameters

In terms of the species-independent parameters in the model, all seven model outputs were most sensitive to the exponent of the background resource λ, the scaling of food intake n, and the scaling of standard metabolism p, with the sensitivity indices associated with these parameters ranging from 1.1 to 571373.4 (Figure 3.7). The population size, LFI, mean weight, and community slope were also sensitive to the carrying capacity of the community spectrum κ, the search volume exponent q, and the maximum size of the community spectrum w_{max}, all of which resulted in Υ_i exceeding 23.4 (Figure 3.7). All of the model outputs were insensitive (Υ_i < 0.01) to the average feeding level of individuals feeding mainly on the background resource f_0, the maximum size of the background resource $w_{pp_{cut}}$, and the exponent of the background mortality of the community spectrum $z_{0_{exp}}$ (Figure 3.7).

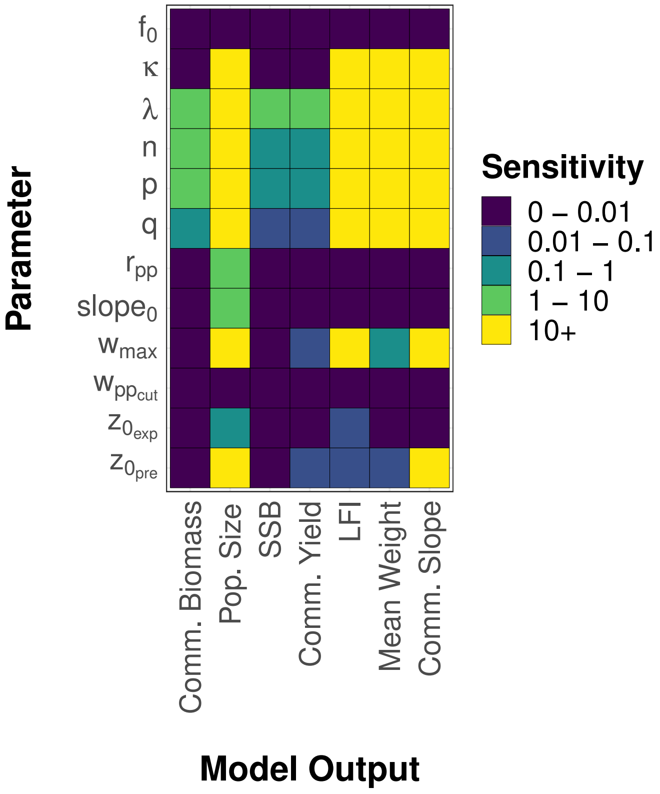


Figure 3.7: The derivative-based sensitivity Υ_i of the model outputs to the species-independent parameters in the North Sea multispecies *mizer* model. Green and yellow indicates that Υ_i exceeds one and is therefore not useful as an upper bound of S_{T_i}, whilst purple indicates that Υ_i is less than 0.01 and the parameter is thus deemed to have a negligible impact on the model output.

Sensitivity to the interaction matrix

All of the model outputs were largely insensitive to the species interaction matrix θ (Figure 3.8). However, the community slope was highly sensitive to the interaction between European plaice and itself (Υ_i = 1593.8) (Figure 3.8). The population size was also highly sensitive to the interactions of European plaice and saithe with themselves and each other (Υ_i > 2.2) (Figure 3.8). These results echo those described previously in which the population size was sensitive to almost all of the species-specific parameters associated with European plaice and saithe.

Convergence of the sensitivity indices

As previously stated, the sensitivity indices, parameter rankings, and parameter screenings are assumed to have reached convergence when $Stat_{indices}$ < 0.05 (Equation 3.14), $\rho_{s,j,k}$ < 1 (Equation 3.15), and $Stat_{screening}$ < 0.05 (Equation 3.16). Based on the normalised derivative-based sensitivity indices Υ_i^*, $Stat_{screening}$ was below the threshold of 0.05 for all seven model outputs except SSB (Figure 3.9). Again, these results indicate that we are successfully able to differentiate between those parameters with a negligible and non-negligible impact on all of the model outputs excluding SSB. The parameter screening did not reach convergence for SSB purely as a result of the width of the confidence interval for the initial population size N_0 of sandeel being greater than 0.05 (not shown). A total of four parameters were consistently associated with higher sensitivities (Υ_i^* > 0.05) across all model outputs: the exponent of the background resource λ, the scaling of standard metabolism p, the search volume exponent q, and the fishing effort F associated with Atlantic herring.

$\rho_{s,j,k}$ was below the threshold of one for SSB, fisheries yield, and the community slope indicator (Figure 3.9). A value of less than one indicates that the average rank of the parameters associated with the greatest sensitivities changed by fewer than one position across all bootstrap resamples and therefore we can be confident in the rankings of the parameters for these model outputs. Conversely, $\rho_{s,j,k}$ was between two and four for the community biomass, population size, LFI, and the mean weight indicator (Figure 3.9). However, the parameter rankings did not reach convergence for the community biomass and mean weight indicator due to relatively minor changes in the rankings of some of the parameters in a small number of the bootstrap resamples. For example, the community biomass had a $\rho_{s,j,k}$ exceeding one purely due to switches in the rankings of λ and the F associated with Atlantic herring between 1st and 4th position in 54 of the 1000 bootstrap resamples (not shown). The mean weight indicator also had a $\rho_{s,j,k}$ exceeding one due to changes in the ranking of the scaling of standard

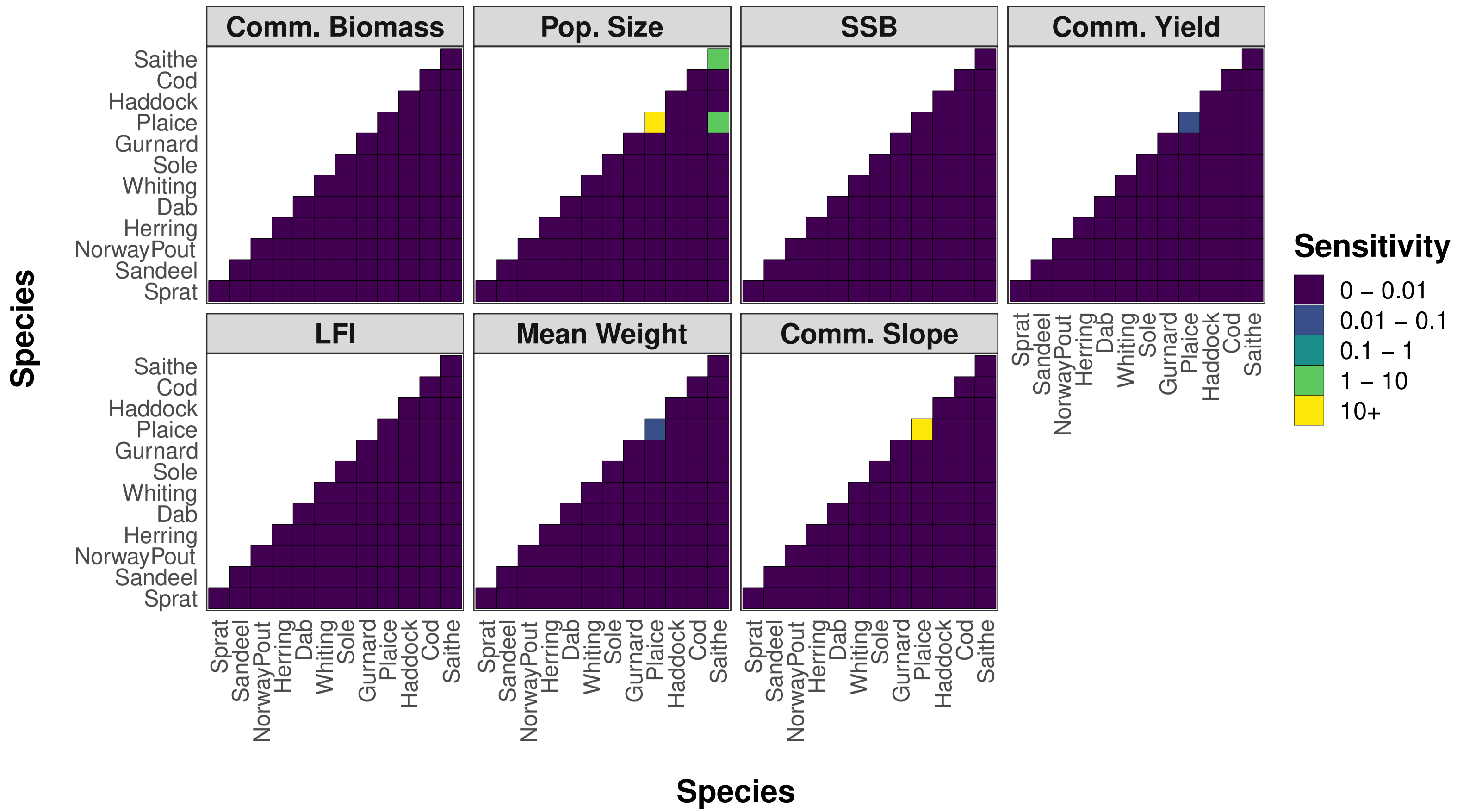


Figure 3.8: The derivative-based sensitivity Υ_i of the model outputs to the interaction matrix in the North Sea multispecies *mizer* model. Green and yellow indicates that Υ_i exceeds one and is therefore not useful as an upper bound of S_{T_i}, whilst purple indicates that Υ_i is less than 0.01 and the parameter is thus deemed to have a negligible impact on the model output.

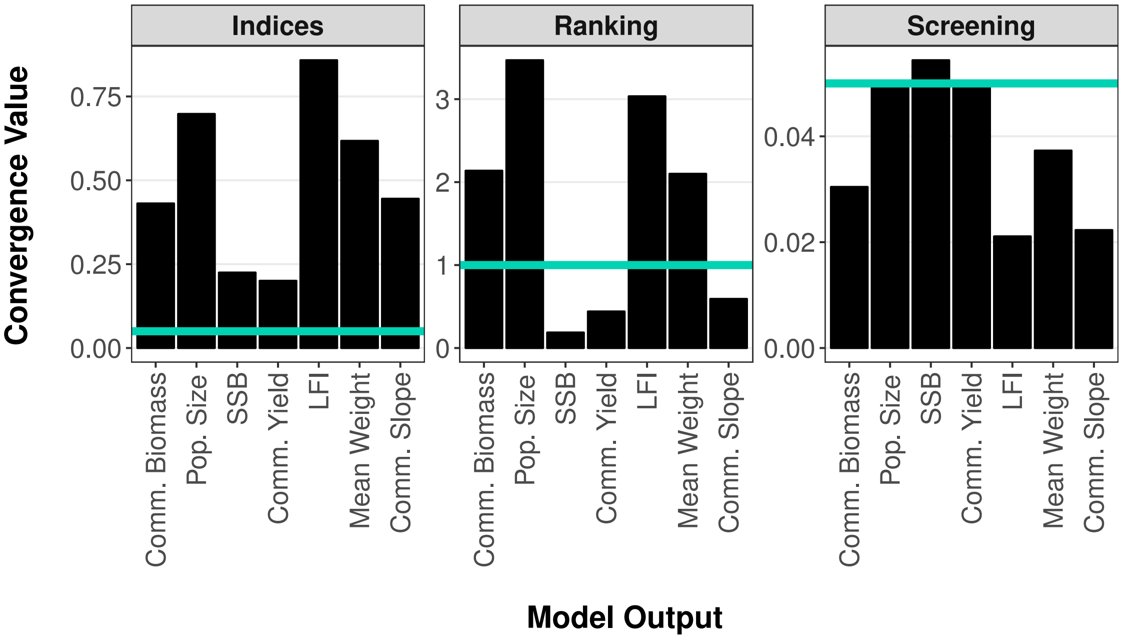


Figure 3.9: The convergence of the derivative-based sensitivity indices Υ_i and the subsequent ranking and screening of the parameters for all seven outputs of the North Sea multispecies *mizer* model. The convergence value refers to the value of $Stat_{indices}$ (left), $\rho_{s,j,k}$ (middle), and $Stat_{screening}$ (right) respectively. The teal line represents the value below which the results are assumed to have reached convergence.

metabolism p from 1st to 5th position and vice versa in 28 of the bootstrap resamples (not shown).

Conversely, the parameter rankings did not reach convergence for population size or the LFI due to much larger changes in the rankings of some of the parameters in a number of bootstrap resamples. For example, changes in the rankings of the assimilation efficiency α of Norway pout and q from 2nd or 3rd position to 19th position or lower in 157 of the bootstrap resamples resulted in the $\rho_{s,j,k}$ of the population size exceeding one (not shown). A change in the ranking of the size class of Norway pout recruits W_{min} from 111th and 163rd to 1st position in two of the bootstrap resamples, as well as a change in the ranking of the maximum size of the community spectrum w_{max} from 53rd and 32nd to 2nd position in two of the bootstrap resamples, was largely to blame for the LFI $\rho_{s,j,k}$ exceeding one (not shown). However, small changes in the rankings of λ, p, and q within the top 13 positions also prevented $\rho_{s,j,k}$ from dropping below one (not shown).

Finally, $Stat_{indices}$ exceeded 0.05 for all seven model outputs (Figure 3.9), suggesting the sensitivity indices did not reach convergence when using the chosen threshold.

Overall ranking of the model parameters

When ranking the parameters from high to low sensitivity using the normalised derivative-based sensitivity indices Υ_i^*, the exponent of the background resource λ was ranked in first position for all model outputs excluding the community biomass and mean weight indicator (Table 3.4). Instead, the fishing effort F associated with Atlantic herring and the scaling of standard metabolism p were ranked in first position for the community biomass and mean weight indicator respectively, whilst λ was ranked in second for both of these model outputs (Table 3.4). λ and p were the only two parameters to appear in the top five positions across all seven model outputs (Table 3.4). The search volume exponent q also appeared in the top five positions for all model outputs except SSB and fisheries yield, whilst the scaling of food intake n appeared in the top five for all model outputs except the LFI and community slope indicator (Table 3.4). The F associated with various different species, including Atlantic herring, sandeel, and European plaice, appeared in the top five for the community biomass, SSB, and fisheries yield model outputs (Table 3.4). The only other species-specific parameters to appear in the top five for any of the model outputs included the assimilation efficiency α of Norway pout, the size class of Norway pout recruits W_{min}, the coefficient of standard metabolism ks for Atlantic cod, the width of the prey size preference σ for European plaice, and the reproductive efficiency $eRepro$ of saithe (Table 3.4).

Table 3.4: The rankings of the five parameters with the greatest derivative-based sensitivity indices for each of the outputs of the North Sea multispecies *mizer* model. ** indicates the rankings did not reach convergence for a given model output.

Rank	****Comm. Biomass**	****Pop. Size**	**SSB**	**Comm. Yield**
1	Herring F	λ	λ	λ
2	λ	q	Herring F	Herring F
3	p	Norway Pout α	Sandeel F	p
4	n	p	p	n
5	Sandeel F	n	n	Plaice F

Rank	****LFI**	****Mean Weight**	**Comm. Slope**
1	λ	p	λ
2	p	λ	p
3	q	q	q
4	Norway Pout W_{min}	Cod ks	Plaice σ
5	w_{max}	n	Saithe $eRepro$

3.5 Discussion

Both variance- and derivative-based methods of sensitivity analysis were used to better understand the sensitivity of the trait-based and multispecies versions of the North Sea mizer model. A total of seven model outputs were considered across both versions of the model, including the community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, and the community slope. Of all of these model outputs, it was the LFI, mean weight, and community slope that most often displayed high sensitivity to the parameters of the trait-based and multispecies *mizer* models. Size-based metrics such as these have been or are currently used as indicators of ecosystem health in various management contexts. For example, the LFI is listed as a food web indicator under the EU Marine Strategy Framework Directive (European Commission, 2008b) and has been used by OSPAR to define an Ecological Quality Objective (EcoQO) for fish communities (Heslenfeld and Enserink, 2008; Greenstreet et al., 2010). Both the LFI and community slope have also been used to assess various management strategies in the North Sea (Nicholson and Jennings, 2004; Blanchard et al., 2014; Thorpe et al., 2015; Marshall et al., 2016). The purpose of using indicators such as these is to enable us to track how changes in fishing mortality affect the health of an ecosystem (Shin and Shannon, 2009; Shin et al., 2018). For an indicator to be successfully able to do this, it must be sensitive to changes in fishing pressure, but insensitive to any other environmental or anthropogenic-induced drivers (ICES, 2004; Greenstreet et al., 2010; Shin et al., 2018). Assuming *mizer* is considered to be an accurate representation of the North Sea, the high sensitivity of the LFI, mean weight, and community slope to changes in fishing effort would support the use of these metrics as indicators of ecosystem health in the North Sea. However, the LFI, mean weight, and community slope were also highly sensitive to parameters that may be heavily impacted by environmental conditions, such as the size of the background resource (e.g. λ) and standard metabolic rates (e.g. p and ks). The results of the sensitivity analyses therefore suggest that we must be cautious when using these metrics as indicators of fishing mortality in the North Sea. This conclusion is supported by Greenstreet et al. (2010), who found that the LFI and mean weight of the North Sea fish community were affected both by changes in fishing pressure and environmentally-driven recruitment variation. Similarly, Blanchard et al. (2005) and Guiet et al. (2016) found that the community slope was affected by changes in natural stressors, despite being previously believed to be a consistent indicator of fishing pressure (Bianchi et al., 2000; Fulton et al., 2004; Shin et al., 2005).

In contrast to the LFI, mean weight, and community slope model outputs, the community biomass, population size, SSB, and fisheries yield were largely insensitive to changes in fishing effort in both the trait-based and multispecies versions of the North Sea *mizer* model, suggesting that these metrics may also not be useful as indicators of the effects of changes in fishing pressure on the ecosystem. Again, this result is supported by Greenstreet et al. (2010), who found that the community-averaged SSB of the North Sea remained almost unchanged between 2001 and 2008 despite the fishing mortality of the seven main demersal fish stocks dropping below precautionary reference points. Although community biomass, population size, SSB, and fisheries yield may not be useful as indicators of fishing pressure, the results of the sensitivity analyses suggest that they may be useful as indicators of environmental drivers. For example, the community biomass, population size, SSB, and fisheries yield of the North Sea multispecies *mizer* model were largely insensitive to all of the parameters in the model excluding λ, n, p, and q; these parameters are used to determine the size of the background resource (e.g. phyto- and zooplankton), as well as search, food intake, and metabolic rates, all of which may be heavily impacted by changes in environmental conditions (Roessig et al., 2004). However, we must again be cautious when analysing the relative sensitivities of the outputs of the *mizer* model as the sensitivity indices did not reach convergence and the analysis was based on potentially unrealistic parameter distributions (i.e. the 'true' values of the parameters may lie outside $\pm$ 10% of the nominal parameter values (see Borrett et al. (2016); Hines et al. (2018); Bentley et al. (2019a,b) for example). Further research is therefore required to identify how the sensitivities of the model outputs change when more realistic parameter distributions are used and when the base sample size of the sensitivity analysis is increased. By increasing the base sample size, we would also be able to more extensively sample the parameter space, thus making the conclusions drawn from this research more robust.

Overall, the outputs of both versions of the *mizer* model were most sensitive to the parameters relating to resource availability and feeding, such as the exponent and carrying capacity of the background resource (λ and κ respectively), the scaling of food intake n, the search volume exponent q, and the assimilation efficiency α. A high sensitivity to such parameters is to be expected given that one of the central assumptions of size-based models such as *mizer* is that the structure of the community is largely determined by trophic interactions (Andersen et al., 2015). However, it is also assumed that these trophic interactions are primarily driven by predator-prey mass ratios (Andersen et al., 2015). It is therefore perhaps surprising that all of the outputs of the trait-based and multispecies versions of the *mizer* model were generally insensitive to the preferred predator-prey mass ratio β. The multispecies model was also relatively insensitive to the species interaction matrix θ, which represents the spatial overlap between each pair of species and is used to determine the extent to which the predator-prey interactions are determined by prey size preferences (Scott et al., 2018). Instead, almost all of the model outputs were sensitive to the width of the prey size preference σ, suggesting that it is changes to the range of preferred prey sizes that drives the sensitivity of the model to changes in the trophic interactions between species, not the mean preferred prey size or the spatial overlap between species.

Generally speaking, both the trait-based and multispecies models tended to exhibit greater sensitivity to the species-independent parameters associated with resource availability and feeding (λ, κ, n, and q) than the species-specific parameters (β, σ, θ, the maximum food intake rate h, and the volumetric search rate γ). Again, this is perhaps unsurprising given that the species-independent parameters have a direct impact on the interactions between every species in the model. Although changes to the species-specific parameters may also affect the interactions between every species in the model via knock-on effects, it would seem that such changes often do not result in the same level of community restructuring that occurs as a result of changes to the species-independent parameters. However, it is possible that the effects of changes in the species-specific parameters may be masked by estimating the sensitivity of the *mizer* model using community-level rather than species-specific model outputs. For example, the community biomass may remain stable despite massive fluctuations in the biomass of individual species. Further research is therefore required to determine the sensitivity of the species-specific outputs of the multispecies model to both the species-independent and species-specific parameters associated with feeding and resource availability.

In addition to the parameters relating to resource availability and feeding, the model outputs were also highly sensitive to the two parameters associated with standard metabolism: p and ks. The high sensitivity of the model outputs to the parameters relating to metabolic rates, resource availability, and feeding are supported by an analysis of the process and observation errors associated with the Haizhou Bay version of the *mizer* model (Zhang et al., 2015). In this example, the 'metabolic scale' parameters, which included ks, q, and n, were found to dominate the uncertainties in the model outputs (Zhang et al., 2015). Additionally, a sensitivity analysis of the Andersen-Ursin multispecies Beverton-Holt model, the model on which *mizer* is conceptually based, showed the outputs were sensitive to the fraction of consumed food that is assimilated (equivalent to α) and the prey size preferences (equivalent to σ) (Livingston, 2013). Both Zhang et al. (2015) and Livingston (2013) therefore emphasised the importance of future research into the parameters that describe diet and food intake, a conclusion that is further supported by our research.

Some effort has already been made to reduce the uncertainties of the parameters relating to food availability and encounter rates. For example, work is already underway to improve estimates of q through a thorough investigation of the scaling relationship between movement and body mass using in situ observations (Griffiths, 2019). However, some of the parameters in the model are not directly measurable in the environment. For instance, we cannot directly measure λ in the real world, but we may be able to reduce the uncertainties associated with the size of the background resource, thereby helping to constrain possible values for both λ, κ, and the growth rate of the background resource r_{pp}. Such goals may be achieved through further monitoring of the North Sea plankton community via large-scale surveys such as the Continuous Plankton Recorder (`sahfos.ac.uk`).

A number of the outputs of the multispecies *mizer* model were also highly sensitive to the fishing effort associated with various different species, including Atlantic herring, sandeel, Norway pout, and European plaice. Perhaps unsurprisingly, fishing rates were also associated with high sensitivity for pelagic and demersal fish populations in a Sobol' variance-based sensitivity analysis of the StrathE2E North Sea marine ecosystem model (Morris et al., 2014). Although there is arguably much less uncertainty associated with fishing activity than the size of the background resource, we may still be able to reduce these uncertainties through improved reporting of fishing activity in the North Sea. Some improvements have been made since the EU Regulation aimed at eliminating illegal, unreported, and unregulated fishing and the reformed Common Fisheries Policy, although there is still more work to be done (European Commission, 2008a, 2013). For example, the European Court of Auditors recently identified a need for increased reporting of catch for vessels below 12m in length (ECA, 2017).

Overall, the high sensitivity of the model outputs to both the size of the background resource and fishing effort highlights the importance of bottom-up (resource-driven) and top-down (consumer-driven) controls on fish populations within the model. Changes in the balance between bottom-up and top-down processes have been shown to result in unstable model dynamics and species extinctions in multiple applications of the Ecopath with Ecosim (EwE) marine ecosystem model (Shannon et al., 2000; Araújo et al., 2006), an occurrence that was also observed in the outputs of *mizer* under some parameter combinations. For example, the high sensitivity of the population size to the parameters associated with European plaice and saithe was driven by large increases in the biomass of these two species, alongside concurrent declines in the biomass of all other species to the point of extinction, in just one of the 1000 base parameter combinations. These results highlight the ability of some parameter combinations to cause unexpected model behaviour. Further research is required to identify the parameter combinations that result in such behaviours to understand why these occur and to help constrain the model in the future.

Such extreme model behaviour caused by a small number of parameter combinations may also be the reason for a number of the sensitivity indices exceeding one. However, it is also possible that the sensitivity indices exceeded one purely due to the presence of non-linearities between the model parameters and the outputs (Lamboni et al., 2013). Although these results indicate that some of the derivative-based indices Υ_i cannot be used as an upper bound for S_{T_i}, we have shown that there is a strong positive relationship between the two different indices, with high values of Υ_i almost always being associated with high values of S_{T_i}. Increasing the number of model evaluations included in both methods of sensitivity analysis may help to reduce the number of indices exceeding one and improve the relationship between Υ_i and S_{T_i}, but doing so would require much larger computational resources than were available for this research. Furthermore, it is unlikely that increasing the number of model evaluations would dramatically change the overall conclusions of this research as the parameters that were consistently associated with the greatest sensitivities would likely remain the same, despite the exact value of the indices changing. This view is further supported by the fact that the parameter rankings reached convergence for many of the model outputs, particularly those associated with the trait-based model (see Section 3.5.1 below).

Applying both methods of sensitivity analysis to the trait-based *mizer* model also highlighted the occasional underestimation of the sensitivity of the model outputs to the parameters with low to mid-levels of influence when using the derivative-based method. Although the largest underestimations occurred as a result of the variance-based sensitivity indices exceeding the threshold of one, underestimations may also have been caused by the choice of 0.0001 as the increment used in the derivative-based sensitivity analysis (see Section 3.3.2). For example, the mean weight of the trait-based *mizer* model was highly sensitive to the maximum size of the background spectrum $w_{pp_{cut}}$ when using the variance-based sensitivity indices, but not when using the derivative-based indices. Because $w_{pp_{cut}}$ is used solely to determine which size bins are part of the background resource and which are part of the community spectra, the derivative-based method would only identify the mean weight as being sensitive to this parameter when the increment of 0.0001 moved $w_{pp_{cut}}$ from one size bin to another. There were no parameter combinations in which this occurred in our sample for the derivative-based sensitivity analysis, but it is possible that a larger increment would have enabled us to better detect the influence of $w_{pp_{cut}}$ on the model outputs, as well as some of the other parameters that were associated with lower sensitivity indices. It is important to note that this issue is therefore an artefact and is not mathematically or biologically relevant.

It is also important to note that we treated all of the parameters as independent and therefore ignored the relationships between them. Because of this, some of the parameters that were deemed to have little impact on the model outputs, such as the species-specific length-weight converters (particularly a), weight at maturity W_{mat}, and asymptotic weights W_{∞}, may indirectly affect the model outputs substantially as they are used to calculate other parameters in the model that have a much larger influence on the model outputs. For example, W_∞ is used to calculate the species-specific maximum food intake rates h and volumetric search rates γ if not specified. Future research is therefore required to determine the impacts of such relationships between parameters on the sensitivities of the model outputs.

Other parameters with low sensitivities, such as the initial population sizes N_0, maximum recruitment R_{max}, reproductive efficiencies $eRepro$, and the average feeding level of individuals feeding mainly on the background resource f_0, may be fixed at their nominal values as changes in the value of these parameters will have a negligible impact on the model outputs. However, this conclusion is upheld only if the true value of the parameter is assumed to lie within $\pm$ 10% of the nominal value, as it is possible that the model outputs may be highly sensitive to changes in the value of these parameters outside of this range.

3.5.1 Convergence of the sensitivity indices

We assessed the convergence of the sensitivity indices and the subsequent parameter rankings and screenings via bootstrapping. The parameter screenings reached convergence for the community biomass, SSB, and population size when using the variance-based sensitivity indices of the trait-based *mizer* model. The parameter screenings also reached convergence for all model outputs excluding the LFI and SSB when using the derivative-based sensitivity indices of the trait-based and North Sea multispecies models respectively. Although the parameter screenings did not reach convergence for all of the model outputs, either the parameter screening or the parameter rankings did reach convergence for all of the model outputs. We are therefore successfully able to accurately identify the parameters with the greatest sensitivities across all of the model outputs, either through screening or ranking.

The parameter rankings reached convergence for all of the trait-based model outputs when using the variance-based method. When using the derivative-based method, the parameter rankings reached convergence for all of the trait-based model outputs excluding SSB and mean weight. For the North Sea multispecies model, the parameter rankings reached convergence for the SSB, fisheries yield, and community slope model outputs. The parameter rankings did not reach convergence for the trait-based mean weight or the multispecies community biomass and mean weight due to relatively small changes in the rankings of the parameters within the top five positions. These small changes in rankings caused $\rho_{s,j,k}$ to exceed the threshold of 0.05 as Equation 3.15 is weighted to reduce the impact of changes in the rankings of the parameters with very low sensitivities; this means that the impact of changes in the rankings of the parameters with the greatest sensitivities is increased. However, such small changes in ranking are relatively unimportant given that the main aim of this research is to identify a small subset of parameters that we may realistically focus on in terms of future research to reduce the uncertainty in the model outputs.

Conversely, the parameter rankings for SSB in the trait-based model and for the population size and LFI in the multispecies model did not reach convergence due to larger changes in the rankings of some of the parameters. These larger changes in rankings were likely caused by the aforementioned parameter combinations that triggered a dramatic change in the abundance and/or biomass of one or more of the species within the model. Furthermore, Sobol' and Kucherenko (2009) proved that the rankings of the parameters associated with the greatest sensitivities in a highly non-linear function may be misleading when based on the derivative-based sensitivity indices. However, the overall conclusions of the present research are likely to be largely unaffected by these issues, as it is clear that the parameters associated with the greatest sensitivities were generally consistent across the two methods of sensitivity analysis and across both the trait-based and multispecies *mizer* models. As previously stated, the exact ordering of the parameters associated with the greatest sensitivities is also not important given the main aim of this research.

The derivative-based sensitivity indices reached convergence for the community biomass, fisheries yield, and LFI of the trait-based model. However, all other sensitivity indices did not reach convergence. In many cases, this lack of convergence was likely caused by variability in the indices of the parameters with very low sensitivity, as the parameter rankings did reach convergence for many of the model outputs. Overall, the lack of convergence of the sensitivity indices is not unsurprising given the large number of uncertain parameters included in the sensitivity analyses. It is particularly unsurprising that the sensitivity indices of the multispecies model did not converge as Morris et al. (2014) were required to run 540,000 model evaluations for the sensitivity indices associated with the StrathE2E model to converge, despite including fewer parameters compared to the present study. It is difficult to know how many model evaluations would be required to reach convergence, but it may become computationally infeasible as some of the model evaluations in this study took hours to reach equilibrium. Furthermore, the convergence of the sensitivity indices is relatively unimportant compared with the convergence of the parameter rankings and screenings, as we do not need to know the exact value of the indices to be able to identify the parameters causing high sensitivity.

3.6 Implications and conclusions

To the best of our knowledge, this study is the first to apply both variance- and derivative-based methods of global sensitivity analysis to a complex marine ecosystem model. We have shown that there is a strong relationship between the variance- and derivative-based sensitivity indices of the trait-based *mizer* model using a base sample size of 1000. These results highlight the ability of the derivative-based method to accurately estimate the variance-based sensitivity indices using a relatively small number of model evaluations. Nevertheless, further research is required to determine the number of model evaluations required for the sensitivity indices, the parameter rankings, and the parameter screenings to converge under both the derivative- and variance-based methods of sensitivity analysis. Although different models will likely require a different number of model evaluations to reach convergence, such research may help to guide those who wish to apply the derivative- or variance-based methods to models of similar complexity. If the difference in the number of model evaluations required for convergence under the two methods of sensitivity analysis is small, we would recommend using the variance-based method to estimate the sensitivity indices. This is because the relatively small increase in the efficiency of the derivative-based method when applied to a model with a comparatively low number of parameters comes at the cost of a reduction in the accuracy of the rankings of the parameters from high to low sensitivity (Sobol' and Kucherenko, 2009). However, the difference in the number of model evaluations required for the derivative- and variance-based sensitivity indices to reach convergence is likely to grow as model complexity increases (De Lozzo and Marrel, 2016). The benefits of using the derivative-based method may therefore only become apparent when applying different methods of sensitivity analysis to highly complex models that include a large number of parameters.

As previous applications of the derivative-based method tend to focus on simple models with a relatively small number of parameters (e.g. Kucherenko and Iooss (2017), Sobol' and Kucherenko (2009), Iooss et al. (2012) and Sudret and Mai (2015)), little is known about the performance of the derivative-based method when applied to more complex models. By applying the derivative-based method to the multispecies *mizer* model, we have shown that this method of sensitivity analysis may be used to successfully estimate the sensitivity indices of a model with many hundreds of parameters. The potential applicability of the derivative-based method may therefore be wide-ranging, particularly in research areas in which complex models are routinely used (as is often the case in environmental modelling) (Cartwright et al., 2016). However, if the derivative-based method is to be applied to a model of similar complexity to the multispecies *mizer* model, we would recommend using a larger base sample size and sampling with Sobol' sequences to allow for the base sample size to be increased without having to run the entire sensitivity analysis again if the indices do not reach convergence (Becker et al., 2018). If the number of model evaluations associated with conducting a derivative-based sensitivity analysis is deemed to be infeasible for a particular model, a similar method to Morris et al. (2014) could be employed; a local method of sensitivity analysis could be used to first screen the parameters and then the derivative-based method could be used to estimate the variance-based sensitivity indices of the parameters that have a non-negligible impact on the model outputs. The derivative-based method could also be used in conjunction with an emulator (or meta-model) to reduce the computational expense of conducting a sensitivity analysis of a complex model (Ratto et al., 2012).

Generally speaking, both the trait-based and multispecies North Sea *mizer* models were most sensitive to the parameters associated with resource availability, feeding, standard metabolic rates, and/or fishing effort. These results are particularly important given the likely impacts of climate change on many of these parameters. For example, climate change is expected to exacerbate the decline in the abundance of phyto- and zooplankton that has occurred in the North Sea over the past 25 years (Capuzzo et al., 2018). Warmer water temperatures may also cause the metabolic rate of each species to increase as a result of elevated biochemical reaction rates (Roessig et al., 2004), although it is possible that species may either move poleward or into deeper waters to offset this effect (Simpson et al., 2013). Furthermore, the oxygen concentration of the North Sea may decline as water temperatures increase (Simpson et al., 2013). Such declines in oxygen concentrations may result in stunted fish growth, thus causing the average and maximum body size of each species to decrease (Roessig et al., 2004; Cheung et al., 2013; Simpson et al., 2013). As body size is strongly correlated with vital rates such as search, intake, and metabolic rates (Andersen et al., 2015), reduced oxygen concentrations may therefore also lead to dramatic changes in the values of the parameters associated with feeding and metabolic rates (Neubauer and Andersen, 2018). Assuming *mizer* is deemed to be an accurate representation of the North Sea, the high sensitivity of the model outputs to the parameters associated with feeding and metabolic rates therefore suggest that climate change is likely to have wide-ranging impacts on the community biomass, population size, SSB, fisheries yield, LFI, mean weight, and community slope of the North Sea fish community in the future. However, the sensitivity indices described here were estimated based on the assumption that each parameter follows a uniform distribution of $\pm$ 10% of their nominal value. More informative distributions must be assigned to each of the parameters if we are to advance our understanding of the sensitivity of each of the model outputs. Although it may be possible to constrain some of the parameters using previously published data and through improved reporting of fishing activity, further research is also required to improve estimates of the size of the background resource and to develop our understanding of the acquisition and assimilation of food by fish in the North Sea. Together, this research will help to ensure multispecies models such as *mizer* are well placed to support ecosystem-based fisheries management in the future.

Chapter 4

Using machine learning to predict the behaviour of the *mizer* marine ecosystem model

4.1 Abstract

Environmental models may include many hundreds or thousands of parameters and take days or even weeks to run a single model evaluation. Predicting how changes in the parameters might affect the model outputs or identifying parameter combinations that do not result in 'extreme' model behaviour, such as species extinctions, may therefore be extremely difficult. The aim of this research is to demonstrate the potential for recently developed methods of machine learning to accurately predict the behaviour of a complex marine ecosystem model. More specifically, we assess the ability of the random forest machine learning algorithm to predict a wide range of outputs from the North Sea multispecies *mizer* size spectrum model, including community coexistence and species biomass, using information on different subsets of the parameters. The results are used to: (1) identify the parameters that are required to maximise the accuracy of the algorithm; (2) highlight interactions between species in the model; (3) identify areas of the parameter space in which community coexistence occurs; and (4) draw comparisons with the global sensitivity analyses described in Chapter 3. Overall, we hope that the results of this research can be used to help better understand and subsequently improve the behaviour of the *mizer* model. This research is vital to increasing the confidence that decision makers have in marine ecosystem models such as *mizer* and ensuring these models continue to develop into tools that can be used to support fisheries management in the future.

4.2 Introduction

Environmental models, such as climate models and marine ecosystem models, have become increasingly complex in recent years due to advances in our scientific understanding of the processes represented by the models, as well as vast improvements in computing power (Plaganyi, 2007; Orth et al., 2015). Some environmental models include many hundreds or thousands of parameters and may take hours or weeks to run a single model evaluation (Plaganyi, 2007; Kaplan and Marshall, 2016). Even models considered to be of low to intermediate complexity may include hundreds of parameters. For example, the North Sea multispecies *mizer* model (described in detail in Chapter 2), a marine ecosystem model of intermediate complexity that is used to simulate the size dynamics of the North Sea fish community (Spence et al., 2016), includes over 300 parameters that describe a community of 12 common and commercially important fish species. A change to any one of these parameters may have unexpected consequences for some or all of the species in the model through the knock-on effects of processes such as predation and competition. Such knock-on effects make it difficult to predict how changes in the parameters may affect the model outputs using traditional methods of data analysis (Curry, 2017; Sedkaoui, 2018). Identifying plausible parameter combinations that do not result in 'extreme' model behaviour, such as species extinctions, may be even more difficult. To overcome these issues, a model may be run using different parameter combinations to identify areas of the parameter space that result in the behaviour of interest. However, the outputs from such an analysis may be large and noisy when using complex models, making it difficult to attribute the causes of different behaviours to specific parameters. Fortunately, recently developed machine learning algorithms may be used to achieve this goal with increased efficiency and reliability compared with traditional data analysis techniques (Sedkaoui, 2018).

Machine Learning (ML) is a branch of artificial intelligence that is used to construct algorithms that learn from and detect patterns in 'big data' (Alpaydin, 2014). The ability of ML methods to learn and adapt makes them applicable to a wide range of tasks, including facial and speech recognition, medical diagnosis, the development of self-driving cars, image compression, bioinformatics, and playing chess (Alpaydin, 2014). ML methods may be grouped into two main categories: supervised and unsupervised learning techniques. Supervised ML techniques, which include decision trees, random forests, support vector machines, and neural networks (Tan and Gilbert, 2003; Mohri et al., 2012), may be particularly useful when attempting to predict the behaviour of an environmental model as they learn classification rules from pre-labelled training data to make predictions about unlabelled testing data (Maglogiannis et al., 2007). A model may therefore be run under multiple parameter combinations and the outputs can be used to train the algorithm to predict the behaviour of the model under a new set of parameter combinations. As supervised methods of ML are often capable of performing classification and regression tasks, they may be used to predict both continuous and discrete model outputs, such as species biomass and species extinctions respectively (Tan and Gilbert, 2003; Mohri et al., 2012).

Tree-based methods are some of the most transparent and easy to use forms of supervised ML techniques (Westreich et al., 2010). The simplest tree-based method, known as a decision tree, works by splitting the training data into increasingly small subsets based on the values of the predictor variables (Tan et al., 2006). The splitting process continues until each subset can be attributed to a particular class label (Tan et al., 2006). The decision tree can then be used to make predictions about the class of unlabelled data points (Tan et al., 2006). Although a single decision tree may suffer from overfitting, ensemble tree-based methods that fit many decision trees, such as random forests, may be used to overcome this issue (Ali et al., 2015). A further benefit of using the random forest algorithm is that various measures of 'importance' can be used to highlight the predictor variables that drive changes in the response variable of interest (Louppe et al., 2013), thus enabling us to identify areas in which to focus future research efforts to reduce the uncertainties in the model outputs. This may be particularly important in marine ecosystem modelling as it tends to be more difficult and expensive to collect data in the marine environment than in terrestrial ecosystems (Murray et al., 2018).

To the best of our knowledge, there are no examples in the literature of random forests (or any other ML algorithm) being used to predict the behaviour of a complex marine ecosystem model. We therefore aim to fill this research gap by exploring the ability of the random forest algorithm to accurately predict the outputs of the North Sea multispecies *mizer* model. Being able to accurately predict the outputs of the *mizer* model using the random forest algorithm would reduce the need to run the full model and enable us to explore the parameter space more efficiently. This in turn would allow us to more easily identify areas of the parameter space that result in specific model behaviours, such as species extinctions, and to determine the parameters that drive such behaviours. To achieve this goal, we train the random forest algorithm to predict the outputs of the *mizer* model using a set of 3000 model evaluations. We use the trained random forests to determine the 'importance' score of each parameter and then compare the performance of the algorithm when trained using different subsets of the most 'important' parameters, thus enabling us to identify which parameters are required to maximise the accuracy of the algorithm. The performance of the algorithm is measured by applying the trained random forests to a test dataset made up of a further 2000 model evaluations, none of which appear in the training dataset. Overall, the results of the analysis will allow us to better understand and subsequently improve the behaviour of the North Sea multispecies *mizer* model. This research is vital to increasing the confidence that decision makers have in marine ecosystem models such as *mizer* and ensuring these models continue to develop into tools that can be used for strategic fisheries management advice in the future.

4.3 Methods

We applied the random forest Machine Learning (ML) algorithm to the North Sea multispecies *mizer* model to assess the ability of the algorithm to predict the behaviour of the model. The *mizer* model is described in detail in Chapter 2, whilst the formulation of the training and testing datasets required as inputs to the random forest algorithm is described in Section 4.3.1 below. Finally, the application of the random forest algorithm to the outputs of the *mizer* model is described in Section 4.3.2.

4.3.1 Training and testing data

The random forest algorithm is a supervised ML technique that requires separate training and testing datasets. To formulate these datasets, we first ran the *mizer* model with 5000 different sets of parameter combinations, which were selected through stratified sampling of the parameter space.

Sampling the parameter space

In order to sample the parameter space of the *mizer* model, we assigned a uniform distribution to each parameter with an upper and lower limit of ±10% of their nominal value (see Chapter 2, Tables 2.3 to 2.5 for a list of the parameters included in the analysis and their nominal values). If the nominal value of a parameter was set to one and it could not be increased further, the upper and lower limits were set to 1 and 0.9 respectively. Conversely, if the nominal value of a parameter was zero and it could not take a negative value, the upper and lower limits were set to 0.1 and 0 respectively. Although the fishing mortality parameter F can be varied at each time step within the model, we chose to maintain effort at a constant level for the duration of each model evaluation. The upper and lower limits of F were set to 1.5 and 0 respectively. A maximum fishing effort of 1.5 was chosen to reflect the mean maximum catch of the 12 modelled species (see Chapter 2, Section 2.4.3) in the North Sea between 1957 and 2011 (`ices.dk`; Blanchard et al. 2014). A stratified sampling technique, known as Latin Hypercube Sampling (LHS), was used to generate 5000 parameter sets using the aforementioned parameter distributions (see Chapter 3, Section 3.3.2 for further details of this method). It is important to note that although some of the parameters are typically calculated using other parameters in the *mizer* model, we treated each parameter as independent and therefore any relationships between the parameters were ignored during sampling (as in Borrett et al. (2016) for example). We chose to ignore these relationships so that we could assess the predictive ability of the random forest using information on individual parameters, rather than information on the combination of multiple parameters.

Model equilibrium

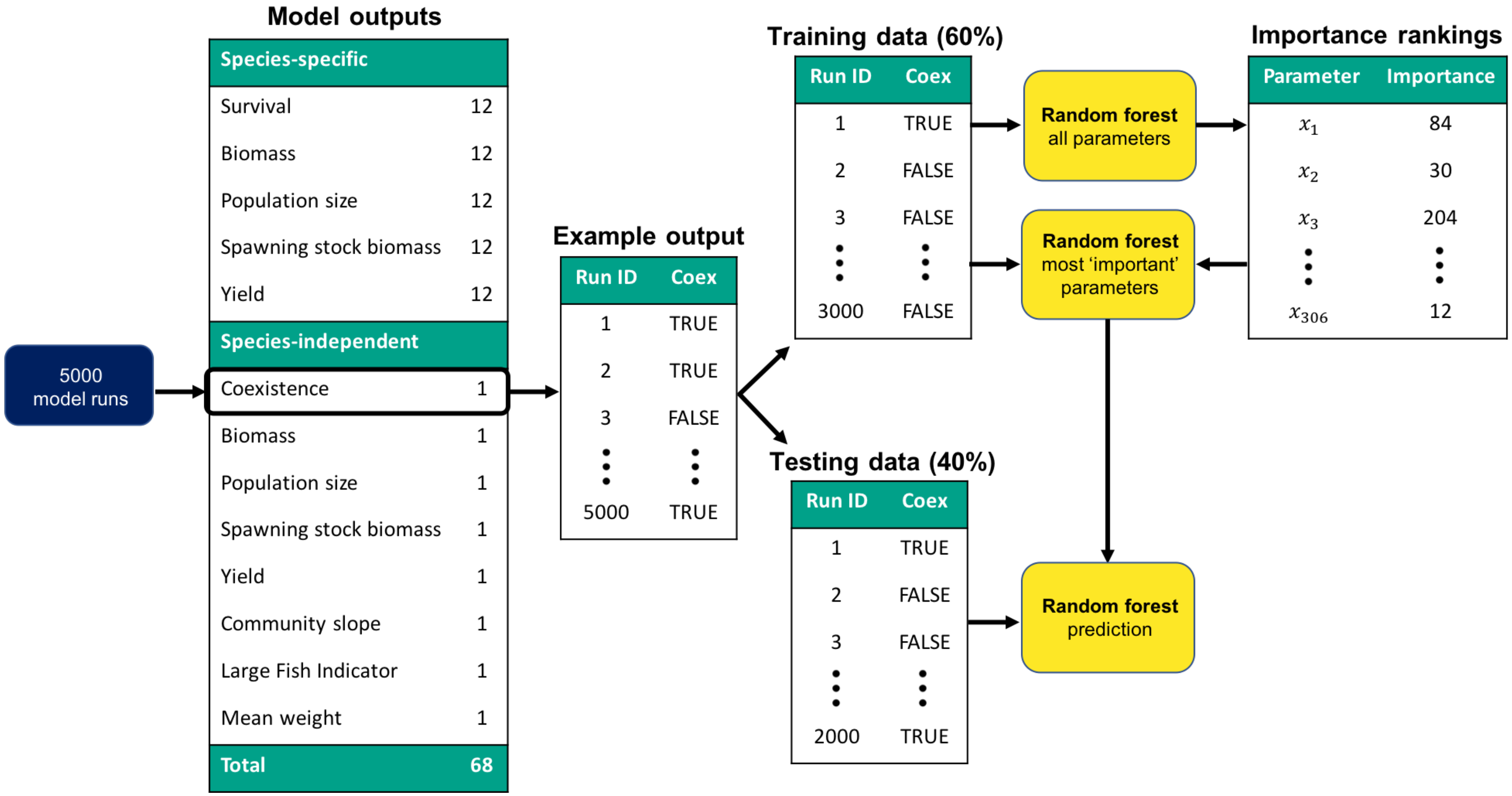
The 5000 parameter sets were evaluated in the *mizer* model until the biomass of each species reached equilibrium. In many cases, the chosen parameter sets resulted in a slow decline in the biomass of a given species to infinitesimal values. We therefore assumed a species had reached equilibrium if the total biomass of the species dropped below the weight of an egg (w_0; 0.001g). The biomass of a species was also deemed to have reached equilibrium if it remained within $\pm 10^{-6}$g of the mean throughout the final 400 time steps (100 model years). However, a large number of model evaluations displayed both regular and irregular periodicity (often referred to as quasiperiodicity; Huggett (2003)) in the biomass of a given species. In cases where there was regular periodicity, we used the `ADCF()` function in the `dCovTS` R package to identify the time lag L (or period length) (Zhou, 2012). The time series was then assumed to have reached equilibrium if the ratio between the final L time steps and the penultimate L time steps was within 1 ± 10^{-6}. Where there was irregular periodicity, we assumed equilibrium had been reached when a linear regression of the biomass of a given species indicated there was no significant change (p > 0.05) in biomass over time. The linear regression was applied over the second half of the time series to avoid the initial 'spin-up' period of the model and was used only when all other checks for equilibrium had failed.

Sampling from the model outputs

The following model outputs were extracted from each of the 5000 model evaluations: species-specific biomass, population size, Spawning Stock Biomass (SSB), and fisheries yields, as well as three community-level indicators: the Large Fish Indicator (LFI), mean weight, and community slope (Figure 4.1). The species-specific model outputs were also aggregated to form four community-level descriptors of the ecosystem (Figure 4.1). These model outputs were selected as they have been widely used in the literature as indicators of ecosystem health in the North Sea (see Nicholson and Jennings (2004), Blanchard et al. (2014), Thorpe et al. (2015), and Marshall et al. (2016) for example). All model outputs were summarised using the mean in the final 400 time steps (100 model years). Species biomass was additionally used to determine a further 12 species-specific model outputs ('species survival') that took the value of zero if the species became extinct in a given model evaluation and one if the species survived (Figure 4.1). The community-level version of this model output ('community coexistence') took the value of zero if one or more species became extinct in a given model evaluation or one if all 12 species survived (Figure 4.1). A species was deemed to have reached extinction when the biomass of the species dropped below the weight of an egg (w_0; 0.001g). All model outputs were divided into training and testing datasets by sampling from the 5000 model evaluations 100 times with replacement (Figure 4.1). By sampling from the model evaluations multiple times, we hoped to prevent the undue influence of parameter combinations that result in 'extreme' model behaviour, such as widespread species extinctions and the subsequent dominance of one or two fish species (see Chapter 3). The training dataset consisted of 60% of the sampled data, whilst the testing dataset consisted of 40% of the sampled data (Figure 4.1).

4.3.2 Application of the random forest algorithm

Random forests work by fitting an ensemble of decision trees to a training dataset (Kocev et al., 2007). Each decision tree is made up of a set of nodes and branches (Figure 4.2). At each node, a random subset of predictor variables is selected and each predictor variable is used to split the training data into two or more groups (Jiang et al., 2007; Kocev et al., 2007). Each split therefore creates two or more branches that lead to separate decision nodes, with each decision node containing a subset of the observations based on the split (Figure 4.2) (Ye, 2013). The predictor variable resulting in the lowest node impurity is chosen as the 'best split' for a given node (Shaikhina et al., 2017). A node is deemed to be pure if the split results in all observations being in one decision node or the other, whilst a node is deemed to be impure if the observations are divided between the decision nodes (Shaikhina et al., 2017). The splitting process is repeated at each decision node until one or more of the following conditions are met: (1) all terminal nodes (also referred to as leaf nodes) are deemed to be pure; (2) the purity of the terminal nodes cannot be increased by a pre-specified minimum amount; or (3) the terminal nodes include a pre-specified minimum number of observations (Zhang, 2016). The final random forest is formed of many of these individual decision trees, each of which is grown using a different subset of predictor variables. When using the random forest algorithm for classification purposes, the majority vote across all trees is used to give an overall prediction for each observation (Liaw and Wiener, 2002). When using the random forest for regression purposes, the mean prediction of each tree is used to give the final prediction for each observation (Liaw and Wiener, 2002).



Model outputs

Species-specific	
Survival	12
Biomass	12
Population size	12
Spawning stock biomass	12
Yield	12
Species-independent	
Coexistence	1
Biomass	1
Population size	1
Spawning stock biomass	1
Yield	1
Community slope	1
Large Fish Indicator	1
Mean weight	1
Total	**68**

Example output

Run ID	Coex
1	TRUE
2	TRUE
3	FALSE
⋮	⋮
5000	TRUE

Training data (60%)

Run ID	Coex
1	TRUE
2	FALSE
3	FALSE
⋮	⋮
3000	FALSE

Testing data (40%)

Run ID	Coex
1	TRUE
2	FALSE
3	FALSE
⋮	⋮
2000	TRUE

Importance rankings

Parameter	Importance
x_1	84
x_2	30
x_3	204
⋮	⋮
x_{306}	12

Figure 4.1: A schematic of the methods used in this research. The *mizer* model was run 5000 times with different combinations of parameter values. A total of 68 model outputs were extracted from each of the model evaluations, which were then split into training and testing datasets with a ratio of 60:40 respectively. This split was conducted 100 times to account for random sampling effects. The random forest algorithm was applied to each training dataset using all 306 parameters in the *mizer* model as predictor variables. The 'importance' of each parameter was extracted from the random forest and the algorithm was then re-applied to each training dataset using different subsets of the most important parameters as predictor variables. Finally, the trained random forests were applied to the testing datasets to assess the predictive ability of the algorithm using different subsets of the most important predictor variables.

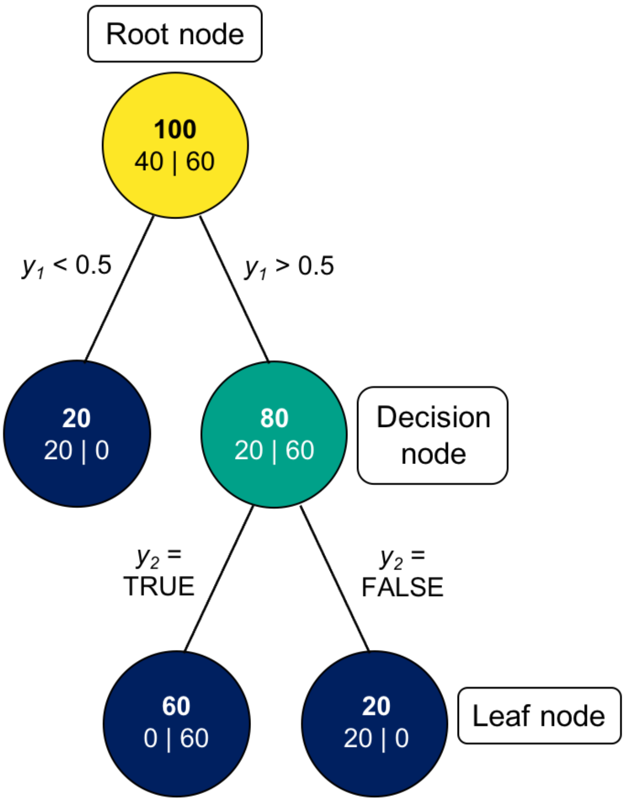


Figure 4.2: A simplified example of one of the many decision trees that forms a random forest. The yellow, teal, and dark blue circles represent root nodes, decision nodes, and leaf nodes respectively, whilst the black lines represent the branches of the decision tree. The number in bold at the top of each circle represents the number of observations at each node. The two numbers at the bottom of each circle represent the number of observations that are classified in each group (e.g. species survival versus species extinction). In this example, the first split divides the training data into two groups based on the value of the predictor variable y_1 and the second split divides the training data based on the value of y_2.

In this research, the random forest algorithm was applied to the outputs of the *mizer* model using the `randomForest` R package (Liaw and Wiener, 2002) with 1501 trees. A value of 1501 was chosen as preliminary results indicated that this number of trees was sufficient to stabilise the error rate of the Out-Of-Bag (OOB) predictions (Liaw and Wiener, 2002). Furthermore, an odd number of trees was selected to ensure ties would not be broken at random (Blouin et al., 2016). The minimum number of observations in the terminal node of a fully-grown tree was set to one for classification tasks and five for regression tasks.

The random forest algorithm was first applied to the training data to identify the 'importance' of each predictor variable (Figure 4.1). In this example, the predictor variables refer to the 306 parameters of the *mizer* model that were included in the analysis. In cases where the random forest algorithm was applied to a classification task, such as predicting species survival, the importance of each parameter was determined using the mean decrease in node impurity as measured by the Gini index (Gini, 1912; Liaw and Wiener, 2002). In cases where the random forest algorithm was applied to a regression task, such as predicting biomass, population size, or fisheries yields, variable importance was determined using the Residual Sum of Squares (RSS). The random forest algorithm was then re-applied to the training datasets using different subsets of the most important parameters (Figure 4.1). The subsets ranged in size from including only one parameter (the one with the greatest importance score) to including up to 300 of the most 'important' parameters. Finally, the trained random forests were applied to the testing datasets to assess the predictive ability of the algorithm (Figure 4.1). These methods were repeated for each of the 100 testing and training datasets.

The accuracy with which the trained random forests were able to predict community coexistence and species survival was measured using Cohen's kappa statistic (referred to simply as the kappa statistic from here on), which takes into account chance agreement between the observations and the predictions (Klimenko, 2017). A kappa statistic of zero indicates that the extent of agreement between the observations and the predictions is no greater than chance; greater values of the kappa statistic indicate increased agreement between the observations and the predictions, with a kappa statistic of one indicating perfect agreement (Cohen, 1960). Landis and Koch (1977) developed a nomenclature to further help to describe the extent of agreement associated with different values of the kappa statistic. Based on this nomenclature, a kappa statistic of between 0.61 and 0.8 indicates substantial agreement between a set of observations and predictions and a value exceeding 0.81 indicates almost perfect agreement (Landis and Koch, 1977).

The accuracy with which the trained random forests were able to predict the continuous outputs of the *mizer* model, such as biomass, population size, and fisheries yields, was measured using the Root Mean Square Error (RMSE). It is important to note that RMSE is an absolute measure of accuracy, whilst the kappa statistic is a relative measure of accuracy. Because of this, it is more difficult to assess the performance of the random forest algorithm using RMSE than it is with the kappa statistic. Nevertheless, the RMSE is measured in the same units as the model outputs and can therefore be interpreted as the standard deviation of the unexplained variance in the observations (Salkind, 2010; Oppenlander and Schaffer, 2017). A RMSE of zero would thus indicate that the random forest algorithm was able to predict a given model output perfectly, whilst a RMSE that is similar to or greater than the standard deviation of the model output would indicate that the algorithm was not able to accurately predict the model output (Salkind, 2010; Oppenlander and Schaffer, 2017). However, a drawback of using RMSE as a measure of accuracy is that comparisons cannot be drawn across the different model outputs due to differences in scale (Hyndman and Koehler, 2006). Despite this, comparisons can still be made across different species within a single model output.

To better understand the subsets of parameters that were required to predict each of the model outputs with the greatest accuracy, we compared the number of parameters that were included in the 'best performing' random forests (i.e. those with the greatest kappa statistic or the lowest Root Mean Square Error (RMSE) for classification and regression tasks respectively) across all 100 testing datasets. We also determined the frequency with which each parameter appeared in the 'best performing' random forests.

4.4 Results

The random forest algorithm was applied to the North Sea multispecies *mizer* model to explore the ability of a supervised Machine Learning (ML) technique to predict the behaviour of the model using different subsets of the most 'important' predictor variables (or model parameters). A total of five species-specific and eight community-level model outputs, as well as 306 parameters, were considered in the analysis. The accuracy with which the random forest algorithm was able to predict each of the model outputs is described in Section 4.4.1 and the frequency with which each parameter appeared in the best performing random forests is described in Section 4.4.2.

4.4.1 Accuracy

In this section, we report the accuracy with which the random forest algorithm was able to predict the outputs of the *mizer* model using different subsets of the most important parameters. We first describe the ability of the random forest algorithm to predict species survival and community coexistence, before describing the ability of the random forest algorithm to predict the continuous community-level and species-specific model outputs, such as biomass, population size, or mean weight.

Classification

The community represented by the North Sea *mizer* model reached coexistence in 1312 (or 26.2%) of the 5000 model evaluations. Norway pout, whiting, and dab were the only species to survive in over 75% of the model evaluations, whilst European plaice was the only species to survive in fewer than 50% of the model evaluations (Figure 4.3). The accuracy with which the random forest algorithm was able to predict species survival and community coexistence was measured using the kappa statistic.

Community coexistence

For community coexistence, the kappa statistic displayed consistent patterns across all 100

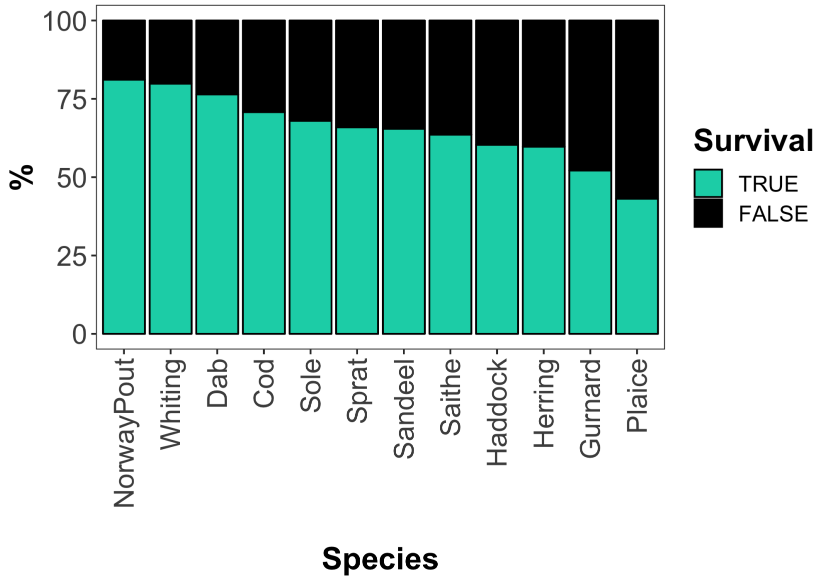


Figure 4.3: The percentage of model evaluations (n = 5000) in which a given species survived or became extinct in the North Sea multispecies *mizer* model. Teal indicates the number of times the species survived, whilst black indicates the number of times the species became extinct.

testing datasets (Figure 4.4). The kappa statistic was lowest when applying the random forest algorithm using only the most important parameter, with a mean (± Standard Deviation (SD)) of 0.34 ± 0.04 (Figure 4.4). However, the kappa statistic increased rapidly with increasing numbers of the most important parameters, reaching peaks of between 0.63 to 0.73 (Figure 4.4). Such high values indicate substantial agreement between the observations and the predictions made by the random forest algorithm (Landis and Koch, 1977). After peaking, the kappa statistic slowly declined with increasing numbers of parameters (Figure 4.4), reaching minima of between 0.03 and 0.19 when including the maximum of 300 parameters in the random forest (not shown). This drop-off in performance is to be expected given that the algorithm selects a random subset of predictor variables with which to split the training data at each node; increasing the number of parameters in the random forest therefore increases the likelihood that the predictor variables that are not useful in predicting community coexistence are included in these random subsets. The median (5th and 95th quantile) number of parameters required to maximise the kappa statistic across all 100 testing datasets was 11 (7, 20) (Figure 4.4).

Species survival

The kappa statistics of the random forests that were used to predict species survival displayed similar patterns to those described above (Figure 4.5). For all species, the kappa statistic was relatively low when very few parameters were included in the random forest (Figure 4.5). The kappa statistic then increased dramatically with increasing numbers of parameters, before slowly declining again (Figure 4.5). The kappa statistic was consistently greatest for sprat, sandeel, and Atlantic herring, with peaks of between 0.90 and 0.96, and consistently lowest

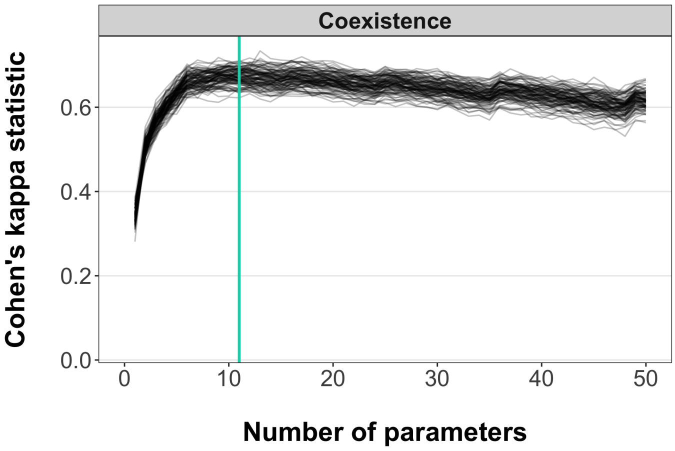


Figure 4.4: The kappa statistics of the random forests that were used to predict community coexistence in the North Sea multispecies *mizer* model with varying numbers of the most important parameters (black lines). The teal line represents the median number of parameters included in the random forests that had the lowest classification error across all 100 testing datasets. Please note the x-axis is limited to 50 parameters for plotting purposes.

for Norway pout, with peaks of between 0.66 and 0.75 (Figure 4.5). Again, such high values indicate substantial to near perfect agreement between the observations and the predictions of the algorithm for all 12 fish species (Landis and Koch, 1977). The median (5th and 95th quantile) number of parameters required to maximise the kappa statistic was between 4 (3, 10) and 10 (8, 14) for all species excluding European plaice and Atlantic herring; for these two species, the median number of parameters required to maximise the kappa statistic was much larger and more variable than all of the other species, with medians (5th and 95th quantiles) of 17.5 (11, 44) and 19 (7, 200) parameters respectively (Figure 4.6). Nevertheless, both European plaice and Atlantic herring were associated with some of the most consistent kappa statistics (Figure 4.5), suggesting the high variability in the number of parameters required to maximise the kappa statistic occurred as a result of very small improvements in the kappa statistic when including much larger subsets of the most important parameters in the random forests. It is therefore likely that far fewer numbers of parameters could be used to predict the survival of Atlantic herring and European plaice with very little loss in agreement between the predictions and the observations.

To test this theory, we compared the kappa statistics of the random forests that included the ’optimum’ number of parameters (i.e. the number of parameters required to maximise the kappa statistic) with the random forests that included just six parameters. For almost all of the species in the *mizer* model, the median kappa statistic of the random forests that included six parameters remained within 0.03 of the median kappa statistic of the random forests that included the optimum number of parameters (Figure 4.7). European plaice was the only species to display a more noticeable decline in the median kappa statistic of the random forests that included the optimum number of parameters versus the random forests that included six pa-

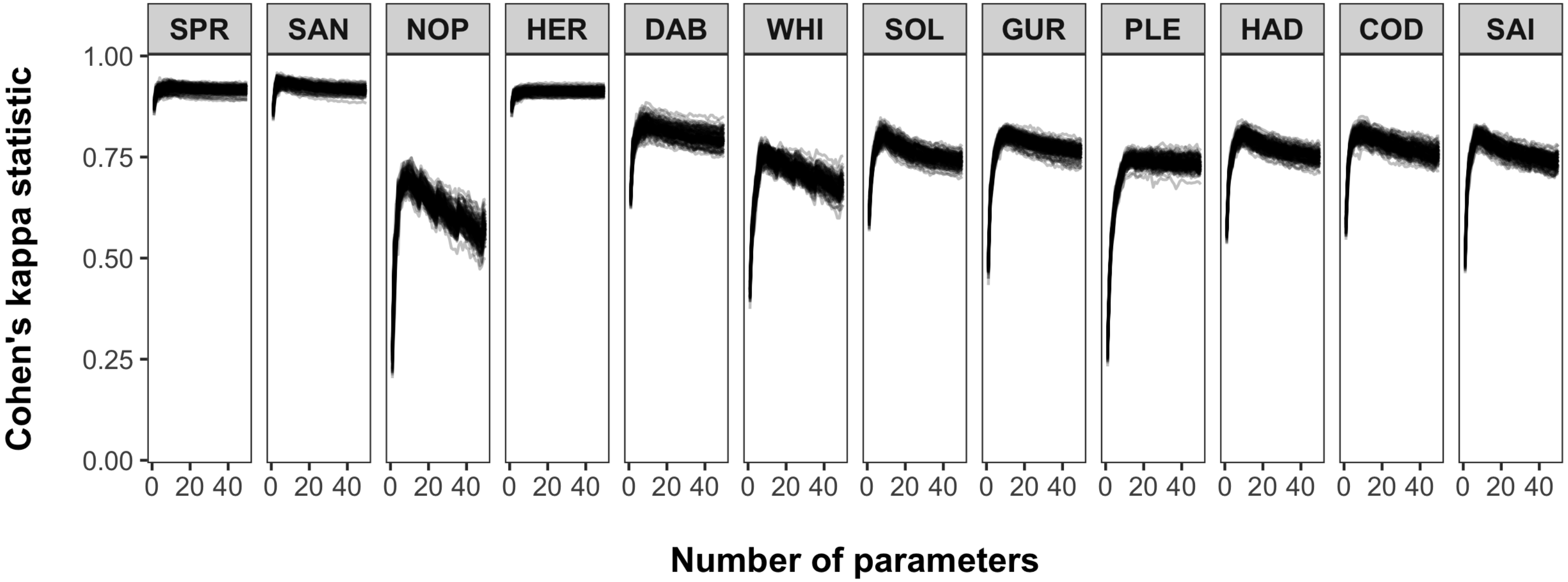


Figure 4.5: The kappa statistics of the random forests that were used to predict species survival in the North Sea multispecies *mizer* model with varying numbers of the most important parameters (black lines). The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe. Please note the x-axis is limited to 50 parameters for plotting purposes.

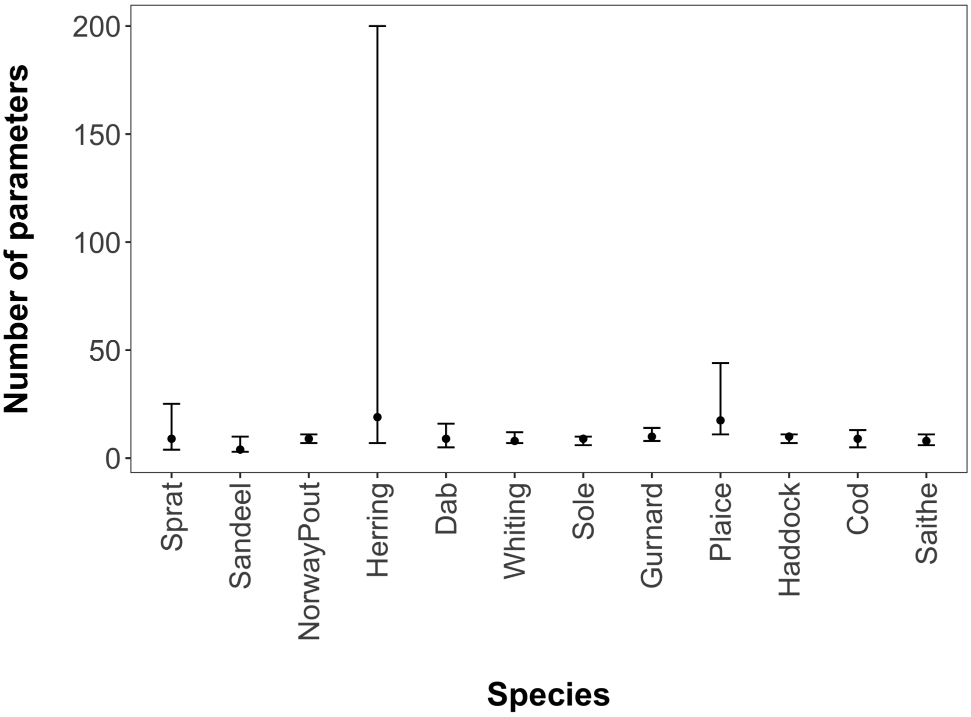


Figure 4.6: The median (5th and 95th quantile) number of parameters required to maximise the kappa statistic of the random forests that were used to predict species survival in the North Sea multispecies *mizer* model.

rameters, with medians (25th and 75th quantile) of 0.75 (0.75, 0.76) and 0.66 (0.65, 0.67) respectively (Figure 4.7). However, the median (25th and 75th quantile) kappa statistic associated with the survival of European plaice increased to 0.73 (0.73, 0.75) when the random forests included ten parameters instead of six (not shown). Overall, these results indicate that the random forest algorithm was able to predict the survival of each species with relatively high accuracy using information on between six and ten of the most important parameters in the model.

Regression

The ability of the random forest algorithm to accurately predict the continuous outputs of the *mizer* model, such as biomass, population size, or mean weight, was measured using Root Mean Square Error (RMSE).

Community-level model outputs

For all of the continuous community-level model outputs, the RMSEs displayed consistent patterns across all 100 testing datasets (Figure 4.8). The RMSEs were greatest when applying the random forest algorithm to each of the community-level model outputs using only the single most important parameter (Figure 4.8). The RMSEs then declined rapidly with increasing numbers of parameters, before beginning to stabilise (Figure 4.8). For all of the

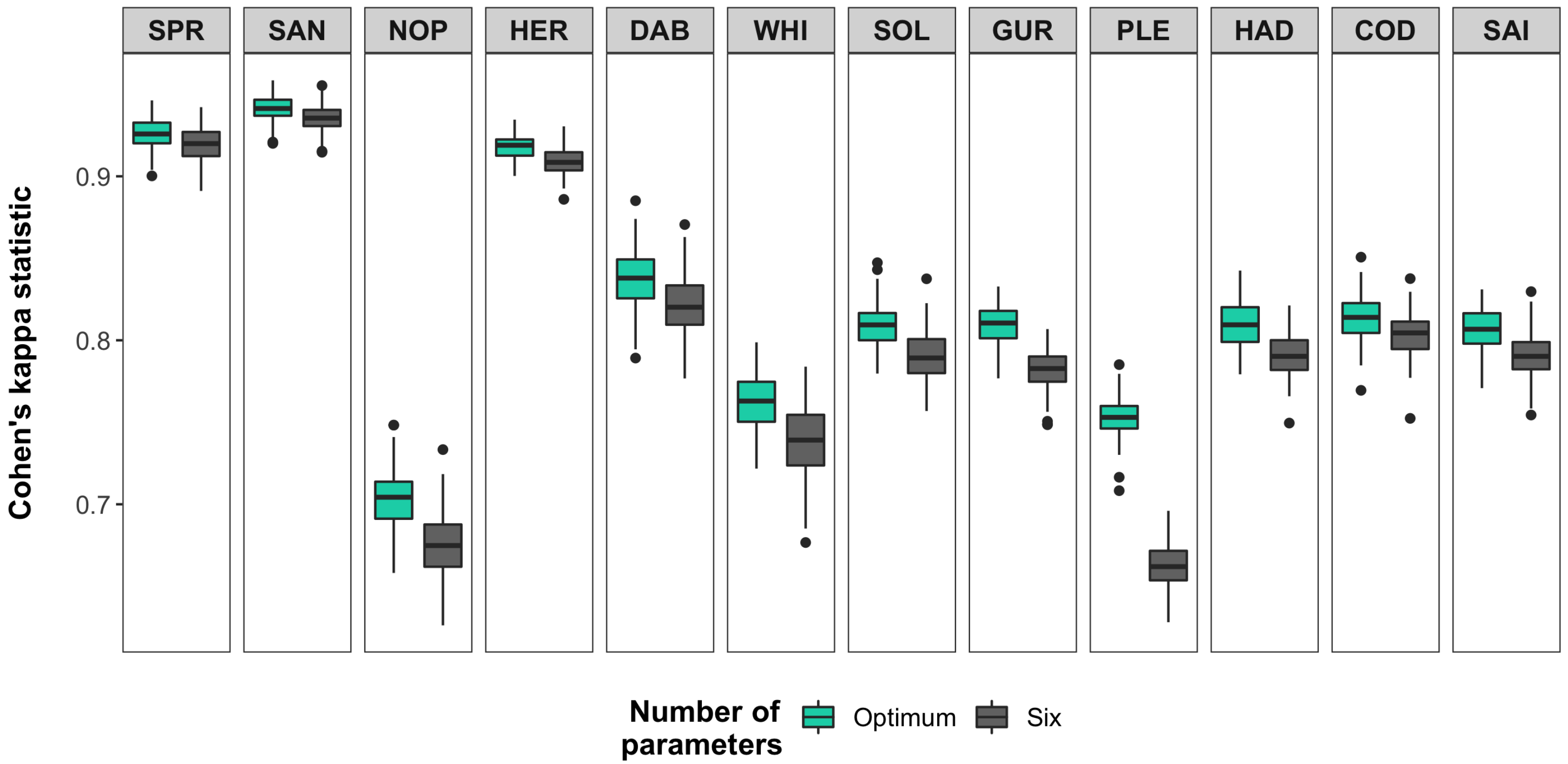


Figure 4.7: The kappa statistics of the random forests that were used to predict species survival in the North Sea multispecies *mizer* model with the 'optimum' number of parameters (i.e. the number of parameters required to maximise the kappa statistic of the random forests; teal) compared with the random forests that included just six parameters (black). The boxes represent the median, 25th, and 75th quantiles. The lower whisker represents the smallest observation that is greater than or equal to the 25th quantile $-$ 1.5 $*$ the InterQuartile Range (IQR). The upper whisker represents the largest observation that is less than or equal to the 75th quantile $+$ 1.5 $*$ the IQR. The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe.

continuous community-level model outputs excluding population size, the mean RMSE of the best performing random forests was smaller than the standard deviation of the model outputs, suggesting the random forest algorithm was able to predict these model outputs with some degree of success (Table 4.1). For example, the standard deviation of the community biomass was 5.27×10^{12}, whilst the mean RMSE (± SD) of the best performing random forests was 2.00×10^{12} (± 2.25×10^{11}) (Table 4.1). Although an RMSE of 2.00×10^{12} may seem high, it is less than 0.5 × the standard deviation of the community biomass. Similar levels of performance were also apparent when the random forest algorithm was used to predict SSB, fisheries yield, and the community slope (Table 4.1). For the LFI and mean weight, the mean RMSE of the best performing random forests was ~0.8 × the standard deviation of the model outputs (Table 4.1). Conversely, the mean RMSE (± SD) of the best performing random forests for population size was larger than the standard deviation, with values of 6.93×10^{15} (± 1.19×10^{15}) and 6.90×10^{15} respectively, thus suggesting the random forest algorithm was not able to accurately predict community population size (Table 4.1).

Table 4.1: The Standard Deviation (SD) of the seven continuous community-level outputs of the *mizer* model compared with the mean RMSE (± SD) of the best performing random forests for each model output. The model outputs include the community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, Large Fish Indicator (LFI), mean weight, and community slope.

Model output	**SD**	**Mean RMSE (± SD)**
Biomass	5.27×10^{12}	2.00×10^{12} (± 2.25×10^{11})
Population Size	6.90×10^{15}	6.93×10^{15} (± 1.19×10^{15})
SSB	5.78×10^{12}	2.65×10^{12} (± 2.94×10^{11})
Yield	1.63×10^{12}	5.33×10^{11} (± 4.99×10^{11})
LFI	0.22	0.18 (± 0.01)
Mean Weight	514.75	414.00 (± 35.3)
Slope	34.40	20.20 (± 2.20)

As previously mentioned, although it is not possible to directly compare the RMSEs associated with each of the model outputs due to differences in scale, comparisons can still be made regarding the number of parameters required to maximise the performance of the algorithm across all of the model outputs. For example, the community biomass, SSB, and community slope model outputs required the fewest parameters to minimise the RMSEs, with a median of six parameters (Figure 4.8). The community slope was the most consistent of these three model outputs, with six parameters required across all 100 testing datasets (Figure 4.8). All other continuous community-level model outputs, excluding population size, required a median (5th and 95th quantile) of between 7 (6, 12) and 12 (9, 12) parameters to minimise the RMSEs

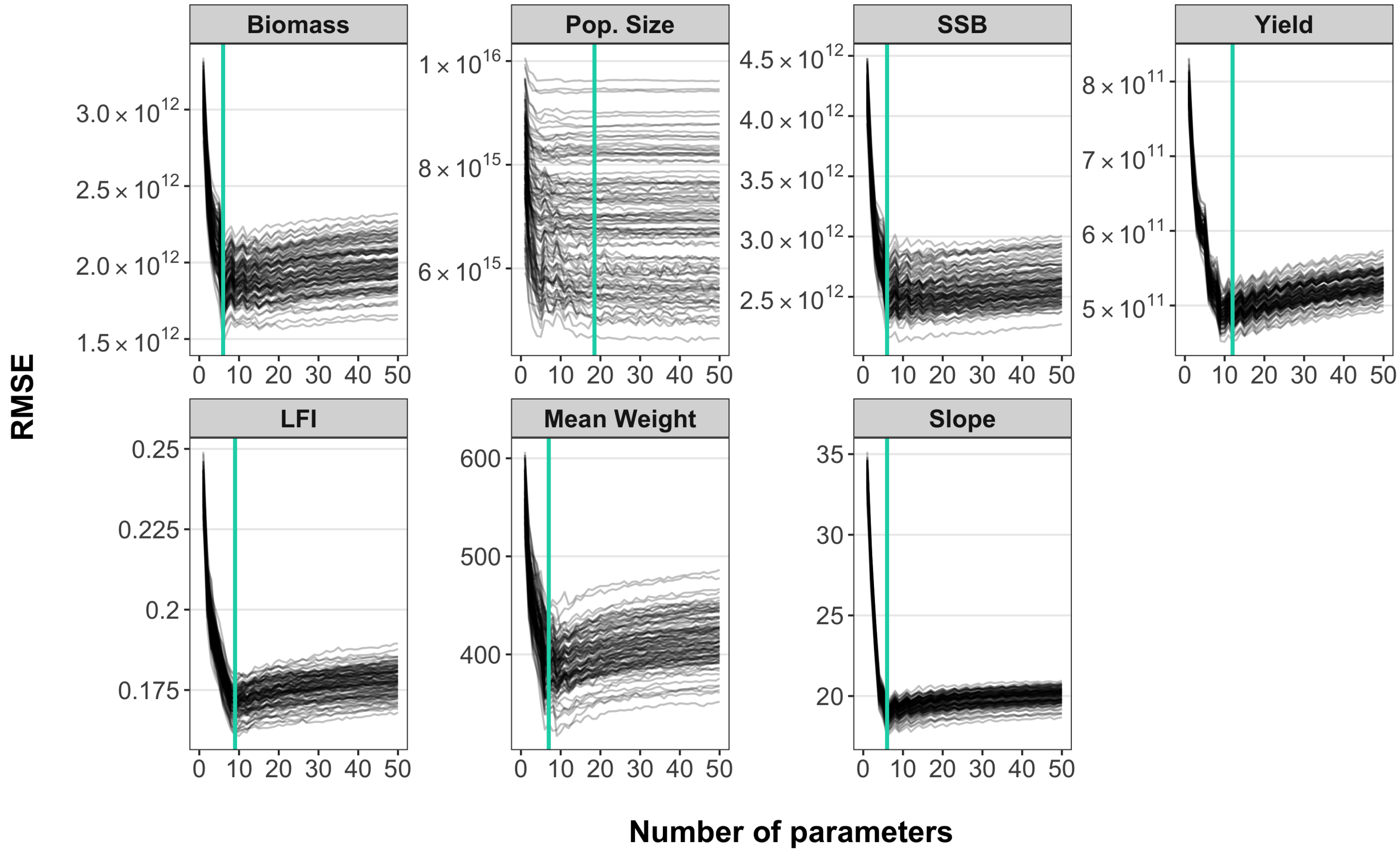


Figure 4.8: The Root Mean Square Error (RMSE) of the random forests that were used to predict the community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, Large Fish Indicator (LFI), mean weight, and community slope of the North Sea multispecies *mizer* model with varying numbers of the most important parameters. The teal line represents the median number of parameters included in the random forests with the lowest error across all 100 testing datasets. Please note the x-axis is limited to 50 parameters for plotting purposes.

of the random forests (Figure 4.8). The community population size required the largest number of parameters to minimise the RMSEs of the random forests, with a median (5th and 95th quantile) of 18.5 (5, 70.5) (Figure 4.8). However, far fewer numbers of parameters could be used to predict community population size with very little loss in agreement between the predictions and the observations. This is true since the median (25th and 75th quantiles) RMSE of the random forests that included the optimum number of parameters was 6.84×10^{15} (5.72×10^{15}, 7.59×10^{15}), but this increased only slightly to 7.02×10^{15} (6.10×10^{15}, 7.84×10^{15}) when the random forests included only the six most important parameters (Figure 4.9). The random forest algorithm could therefore be used to predict community population size using information on just six parameters, although the overall accuracy of the predictions would still be low.

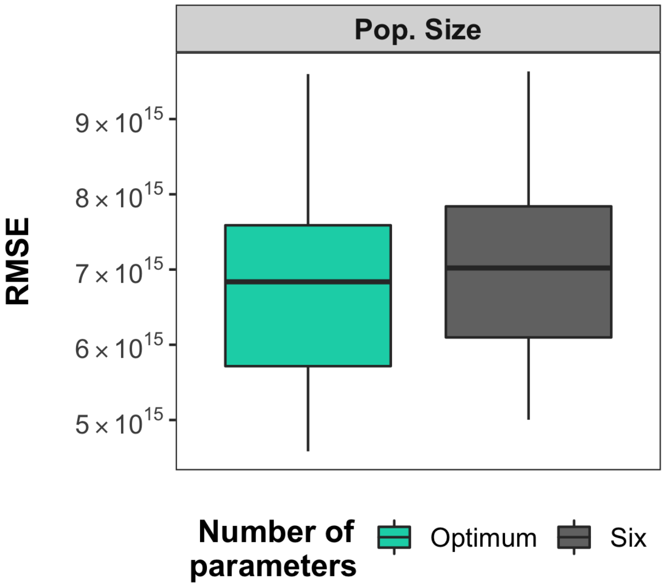


Figure 4.9: The Root Mean Square Error (RMSE) of the random forests that were used to predict community population size in the North Sea multispecies *mizer* model with the 'optimum' number of parameters (i.e. the number of parameters required to minimise the RMSEs of the random forests; teal) compared with the random forests that included just six parameters (black). The boxes represent the median, 25th, and 75th quantiles. The lower whisker represents the smallest observation that is greater than or equal to the 25th quantile $-$ 1.5 $*$ the InterQuartile Range (IQR). The upper whisker represents the largest observation that is less than or equal to the 75th quantile $+$ 1.5 $*$ the IQR.

Species-specific model outputs

The RMSEs of the random forests that were used to predict the continuous species-specific outputs of the *mizer* model also displayed consistent patterns across all 100 testing datasets, with a characteristic decline in RMSE with increasing numbers of parameters, followed by a stabilisation (Figure 4.10). The ranking of each species based on the RMSEs of the random forests was broadly consistent across three of the four model outputs, with dab, grey gurnard, and common sole being associated with the lowest RMSEs and Atlantic herring being associated with the greatest RMSEs for biomass, Spawning Stock Biomass (SSB), and fisheries yields (Figure 4.10). Conversely, the RMSEs of the random forests that were used to predict the population size of each species displayed very different patterns, with Atlantic cod being associated with the lowest RMSEs and Norway pout and European plaice being associated with the greatest RMSEs (Figure 4.10).

For all of the continuous species-specific model outputs, the mean RMSE of the best performing random forests was lower than the standard deviation of the model outputs. For example, the standard deviation of the biomass of dab (the species with the lowest RMSEs) was 4.02×10^{10}, whilst the mean RMSE (± SD) of the best performing random forests was 2.75×10^{10} (± 2.73×10^{9}) (Table 4.2). The mean RMSE of the best performing random forests was therefore ~0.68 × the standard deviation. Similar levels of performance were apparent across all species and model outputs excluding population size. For population size, the mean RMSE of the best performing random forests was almost identical to the standard deviation (Table 4.2). Although these results represent an improvement in accuracy compared with community population size, it is clear that the random forest algorithm was not able to accurately predict species-specific population sizes.

For biomass, SSB, and fisheries yields, there was relatively little variation in the number of parameters required to minimise the RMSEs of the random forests. For all of these model outputs, the median (5th and 95th quantile) number of parameters required to minimise the RMSEs ranged from 6 (6, 12) when predicting the SSB of sprat and sandeel to 15.5 (10, 24) when predicting the biomass of grey gurnard (Figure 4.11). Generally speaking, the random forests that were used to predict species-specific population sizes required information on a much larger number of parameters to minimise the RMSEs, with medians (5th and 95th quantile) ranging from 5 (3, 41.3) for whiting to 47.5 (5, 252) parameters for Norway pout (Figure 4.11). However, far fewer numbers of parameters could again be used to predict species-specific population sizes with very little loss in agreement between the predictions and the observations. Using the population size of Norway pout as an example, the median (25th and 75th quantiles) RMSE of the random forests was 6.39×10^{15} (5.12×10^{15}, 7.06×10^{15}) when including the optimum number of parameters, but this increased by a relatively small amount to 6.58×10^{15} (5.42×10^{15}, 7.16×10^{15}) when including only the six most important parameters (Figure 4.12).

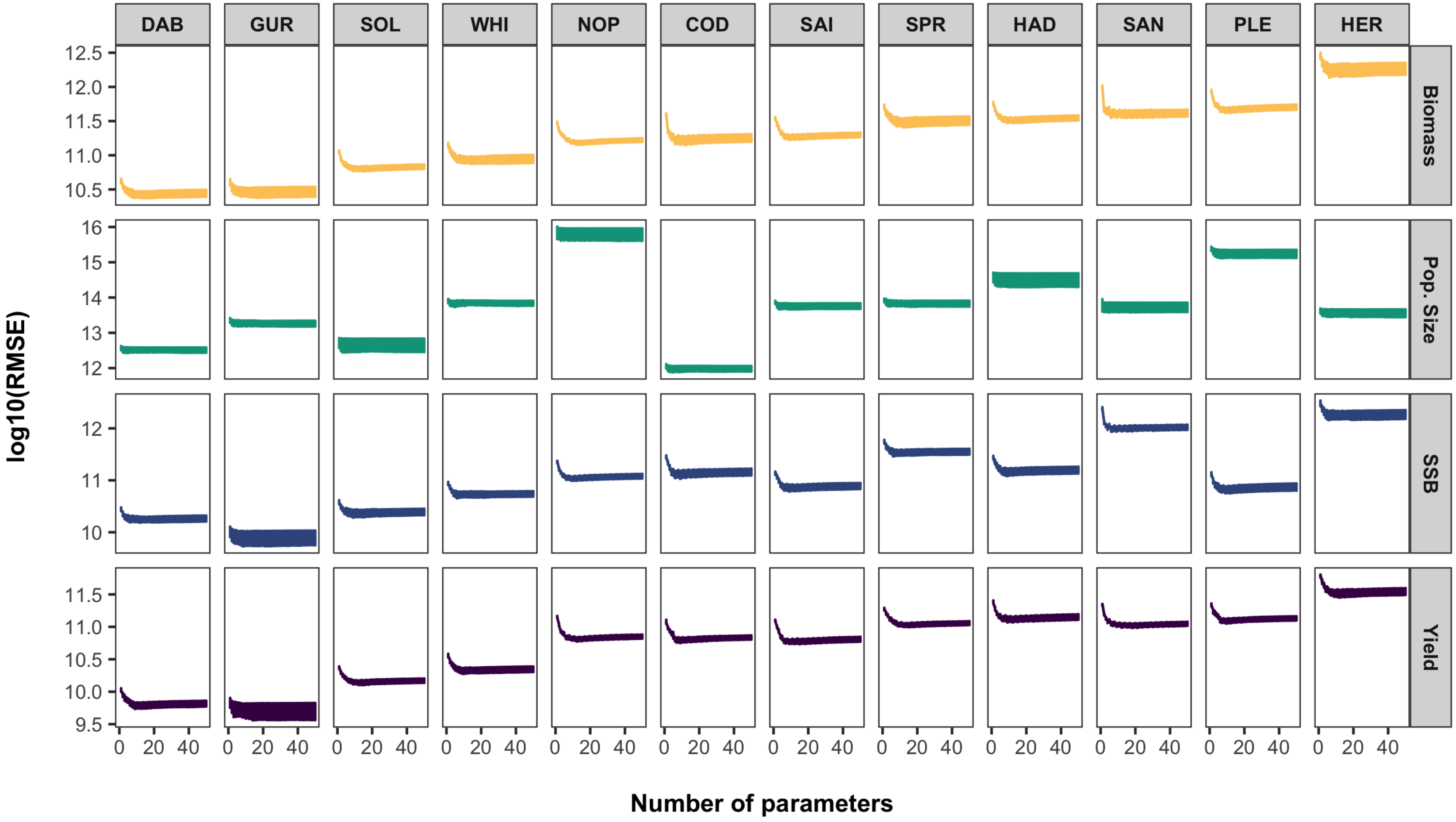


Figure 4.10: The log-transformed Root Mean Square Error (RMSE) of the random forests that were used to predict the species-specific biomass, population size, Spawning Stock Biomass (SSB), and fisheries yield outputs of the North Sea multispecies *mizer* model with varying numbers of the most important parameters. The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe. Please note the x-axis is limited to 50 parameters for plotting purposes.

Table 4.2: The Standard Deviation (SD) of the continuous species-specific outputs of the *mizer* model compared with the mean Root Mean Square Error (RMSE) (± SD) of the best performing random forests for each model output. The model outputs include species-specific biomass, population size, Spawning Stock Biomass (SSB) and fisheries yield. For biomass, SSB, and fisheries yield, the results are given for dab and Atlantic herring as these species were associated with the lowest and greatest RMSEs respectively. For population size, the results are given for Atlantic cod and Norway pout instead.

Model output	**Species with low RMSE**		**Species with high RMSE**	
	SD	Mean RMSE (± SD)	SD	Mean RMSE (± SD)
Biomass	4.02×10^{10}	2.75×10^{10} ($\pm$ 2.73×10^{9})	2.96×10^{12}	1.86×10^{12} ($\pm$ 2.17×10^{11})
Population Size	1.01×10^{12}	9.91×10^{11} ($\pm$ 1.06×10^{11})	6.42×10^{15}	6.36×10^{15} ($\pm$ 1.26×10^{15})
SSB	2.71×10^{10}	1.86×10^{10} ($\pm$ 1.83×10^{9})	3.25×10^{12}	1.93×10^{12} ($\pm$ 2.37×10^{11})
Yield	1.06×10^{10}	6.76×10^{9} ($\pm$ 7.33×10^{8})	7.21×10^{11}	3.61×10^{11} ($\pm$ 4.45×10^{10})

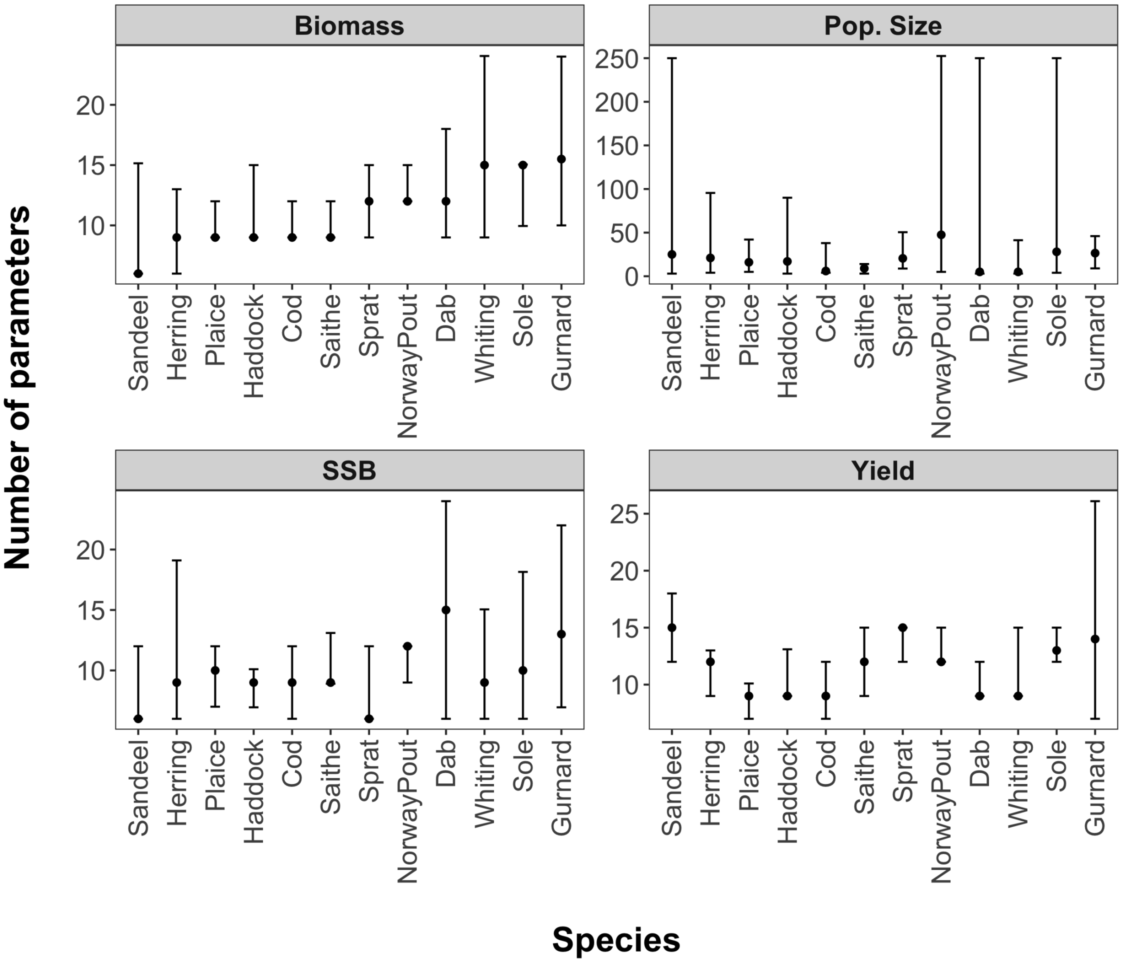


Figure 4.11: The median (5th and 95th quantile) number of parameters required to minimise the Root Mean Square Error (RMSE) of the random forests that were used to predict species-specific biomass, population size, Spawning Stock Biomass (SSB), and fisheries yields in the North Sea multispecies *mizer* model. Please note the different y-axis limits for each model output.

4.4.2 Frequency

In this section, we report the frequency with which each of the parameters appeared in the 100 'best performing' random forests for each model output. We first describe the frequency with which each parameter was present in the best performing random forests for species survival and community coexistence, before describing the frequency with which each parameter was present in the best performing random forests for the continuous community-level and species-specific model outputs, such as biomass, population size, or mean weight.

Classification

Community coexistence

For community coexistence, the species-independent volumetric search rate λ, search volume exponent q, scaling of the food intake n, and the scaling of standard metabolism p (Figure 4.14), as well as the fishing mortality F and assimilation efficiency α of European plaice

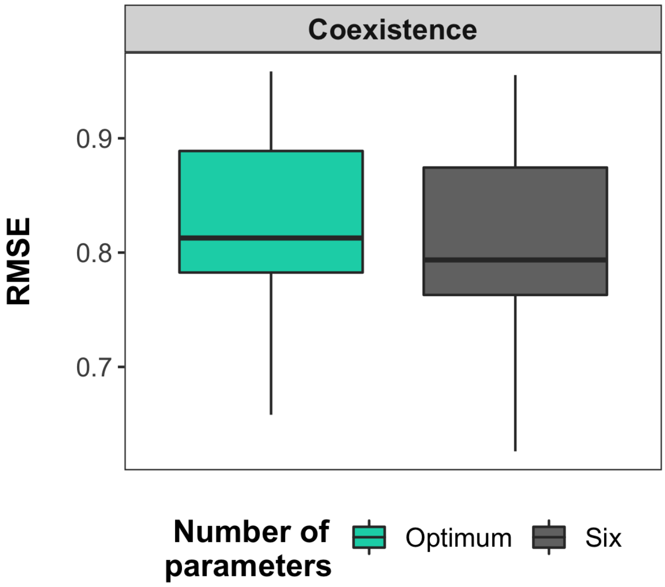


Figure 4.12: The Root Mean Square Error (RMSE) of the random forests that were used to predict the population size of Norway pout (NOP) in the North Sea multispecies *mizer* model with the 'optimum' number of parameters (i.e. the number of parameters required to minimise the RMSEs of the random forests) (teal) compared with the random forests that included just six parameters (black). The boxes represent the median, 25th, and 75th quantiles. The lower whisker represents the smallest observation that is greater than or equal to the 25th quantile $-$ 1.5 $*$ the InterQuartile Range (IQR). The upper whisker represents the largest observation that is less than or equal to the 75th quantile $+$ 1.5 $*$ the IQR.

(Figure 4.13), were present in all 100 of the best performing random forests. The α of grey gurnard and the width of the prey size preference σ for European plaice were also present in over 90 of the 100 best performing random forests (Figure 4.13). The average feeding level of individuals feeding mainly on the background resource f_0, the starting slope of the community spectrum $slope_0$, the maximum size of the community spectrum w_{max}, the maximum size of the background spectrum $w_{pp_{cut}}$, and the exponent of the background mortality of the community spectrum $z_{0_{exp}}$ were not present in any of the 100 best performing random forests (Figure 4.14). None of the parameters in the species interaction matrix θ were present in more than 18 of the best performing random forests (Figure 4.15).

Of the six parameters that were present in all 100 of the best performing random forests, λ displayed the strongest relationship with community coexistence (Figures 4.16 and 4.17). Community coexistence occurred only when the value of λ exceeded 2.06 (or 0.33 when rescaled between 0 and 1; Figures 4.16 and 4.17). A Wilcoxon rank sum test (W) (Wilcoxon, 1945), which is a non-parametric statistical test that is used to determine whether two independent samples are drawn from populations with the same distribution (Triola, 2006), indicated that the mean value of λ was significantly greater for model evaluations in which community coexistence occurred (μ = 2.25) compared with model evaluations in which at least one species became extinct (μ = 2.06; W = 669820, p < 0.001). Despite seemingly showing a weaker

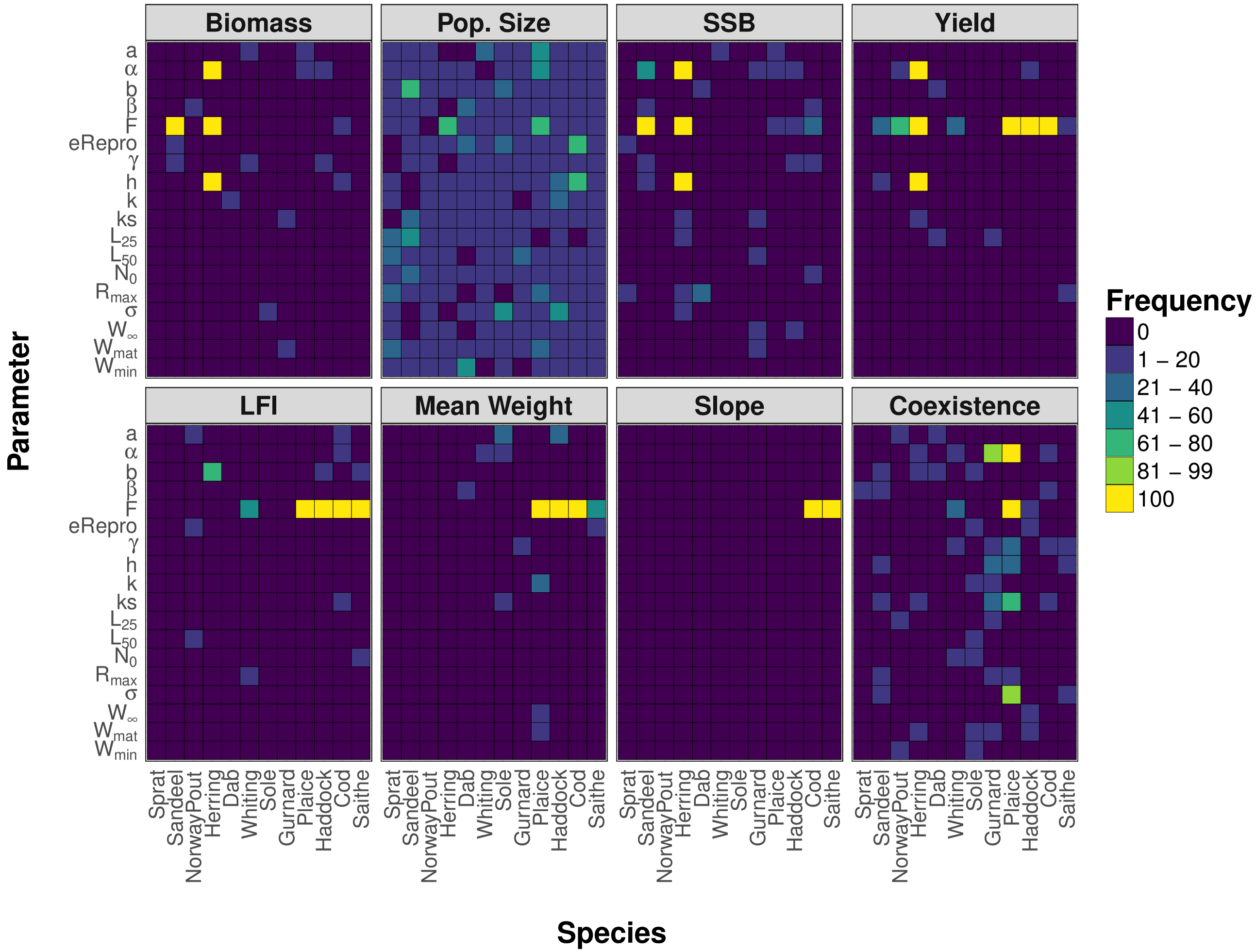


Figure 4.13: The frequency with which each of the species-specific parameters in the North Sea multispecies *mizer* model appeared in the random forests with the lowest error when predicting community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, community slope, and community coexistence. A maximum frequency of 100 was possible as the random forest algorithm was applied to 100 different testing datasets for each model output.

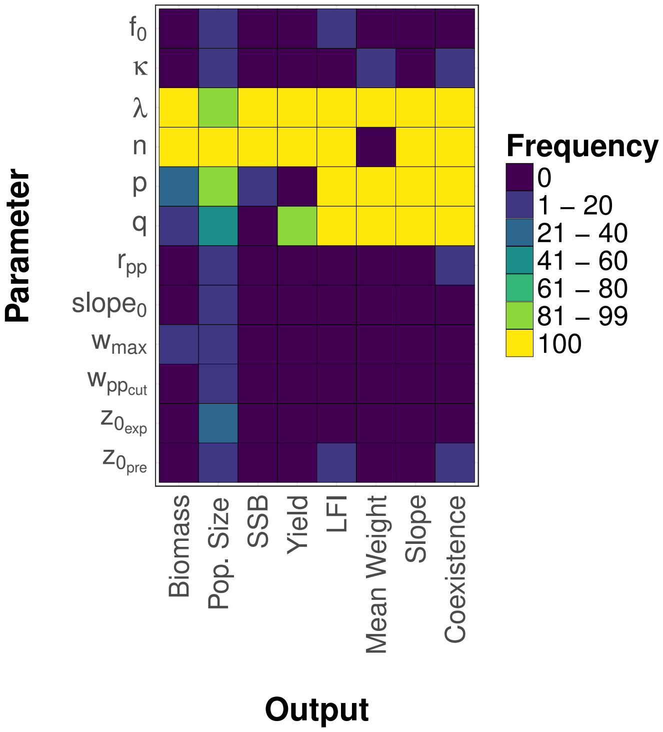


Figure 4.14: The frequency with which each of the species-independent parameters in the North Sea multispecies *mizer* model appeared in the random forests with the lowest error when predicting community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, community slope, and community coexistence. A maximum frequency of 100 was possible as the random forest algorithm was applied to 100 different testing datasets for each model output.

relationship with community coexistence than λ, the value of p, q, n, and the F and α of European plaice also differed significantly (p < 0.001) between model evaluations in which community coexistence either did or did not occur. Community coexistence was additionally affected by interactions between different parameters. For example, community coexistence frequently occurred given any value of n (Figure 4.16) but did not occur when low values of n were combined with high values of p (Figure 4.17). Similar interactions were less evident when low values of n were combined with high or low values of q or the F and α associated with European plaice (Figure 4.17).

Species survival

When predicting species survival, the parameters were divided into three groups: (1) those associated with the species being predicted by the random forest; (2) those associated with a different species than the one being predicted by the random forest; and (3) the species-independent parameters. Overall, the frequency with which each parameter appeared in the best performing random forests was relatively similar across all 12 fish species included in the model (Figure 4.18). Six species-specific (α, F, the maximum food intake rate h, the coeffi-

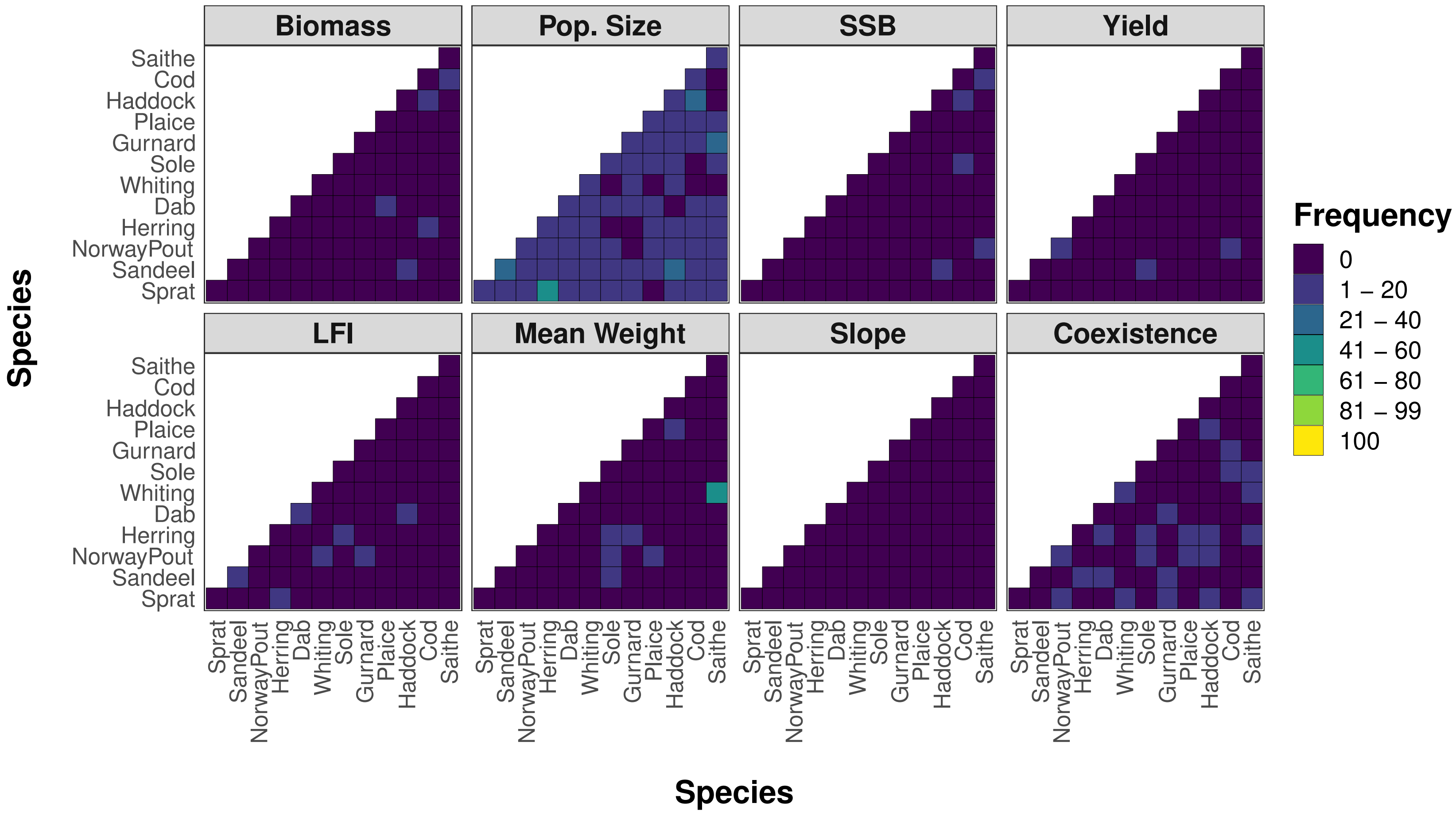


Figure 4.15: The frequency with which each of the parameters in the interaction matrix θ of the North Sea multispecies *mizer* model appeared in the random forests with the lowest error when predicting community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, community slope, and community coexistence. A maximum frequency of 100 was possible as the random forest algorithm was applied to 100 testing datasets for each model output.

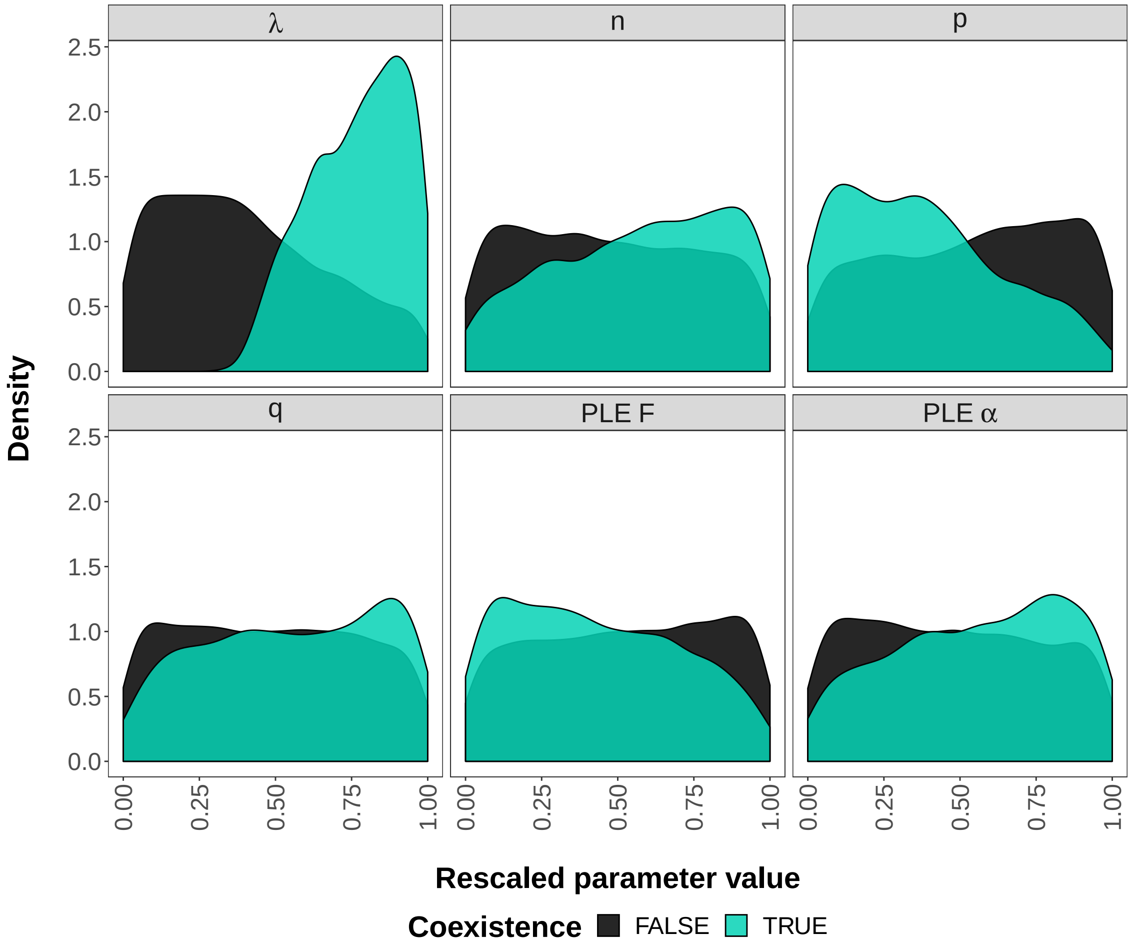


Figure 4.16: The distribution of the six most 'important' parameters for model evaluations in which community coexistence did (teal) or did not (black) occur. The six parameters shown here were the only parameters to appear in all 100 of the best performing random forests as measured by Cohen's kappa statistic. The parameters include the exponent of the background resource λ, the scaling of food intake n, the scaling of standard metabolism p, the search volume exponent q, and the fishing effort F and assimilation efficiency α of European plaice. Please note that the parameters were rescaled between zero and one for plotting purposes.

cient of standard metabolism ks, the width of the prey size preference σ, and the volumetric search rate γ) and four species-independent (λ, n, p, and q) parameters appeared in all 100 of the best performing random forests for at least one species (Figure 4.18). All of the species-specific parameters that consistently appeared in the 100 best performing random forests for species survival were associated with the same species that the random forest was used to predict (Figure 4.18). The species-specific parameters that were associated with a different species to the one the random forest was used to predict were always present in fewer than 71 of the best performing random forests (Figure 4.18). The species-independent parameters λ, n, p, and q were generally present in high frequencies when predicting the survival of the larger species in the model, such as European plaice, saithe, common sole, and haddock, and in the lowest frequencies when predicting the survival of the smaller species in the model, particularly sprat and Atlantic herring (Figure 4.18).

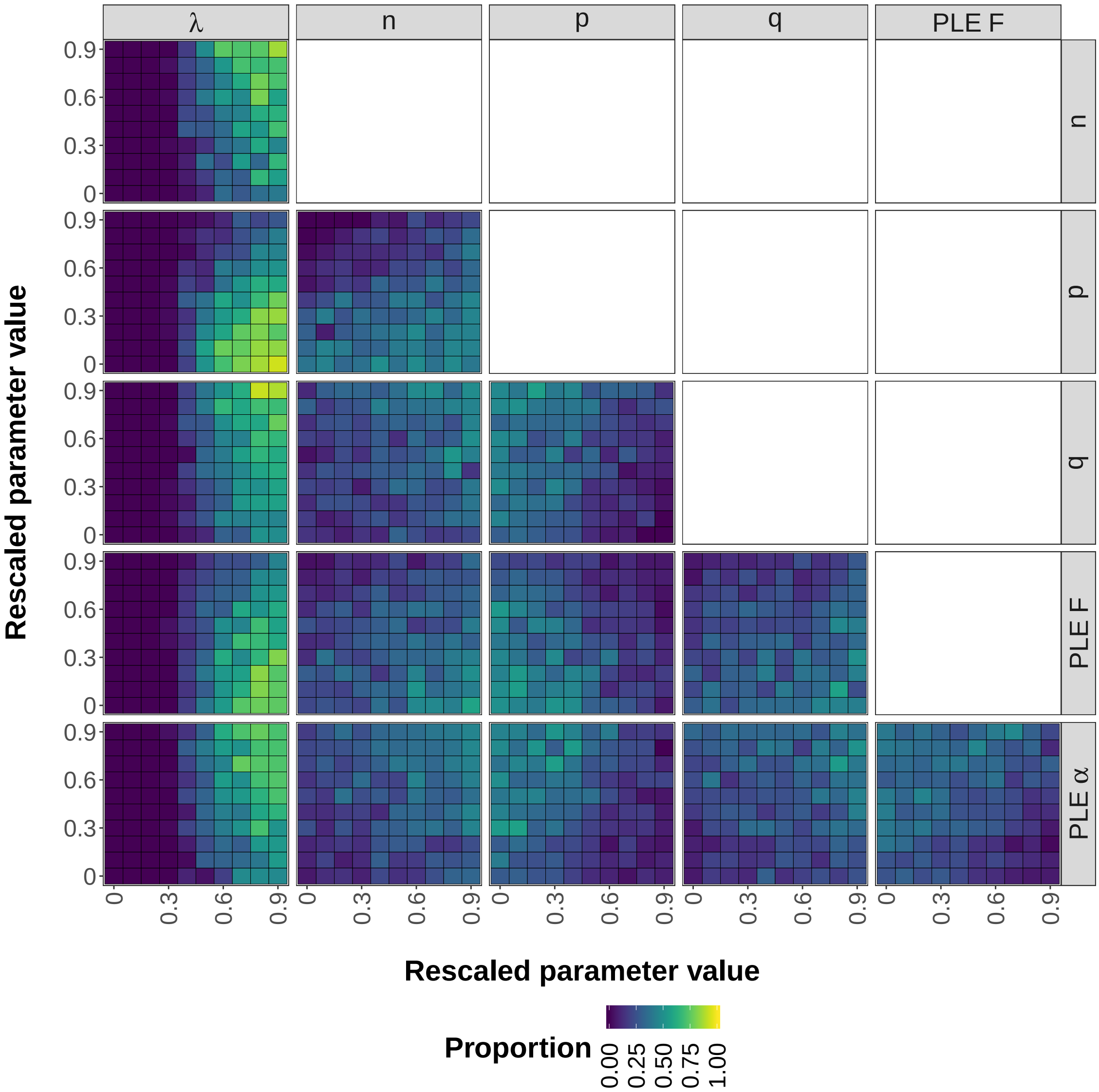


Figure 4.17: A matrix of plots depicting areas of the parameter space in which community coexistence occurred in the North Sea multispecies *mizer* model. Each plot represents the parameter space of one pair of the six most 'important' parameters, which included the exponent of the background resource λ, the scaling of food intake n, the scaling of standard metabolism p, the search volume exponent q, and the fishing effort F and assimilation efficiency α of European plaice. These parameters were the only parameters to appear in all 100 of the best performing random forests as measured by Cohen's kappa statistic. The parameter space is divided up into 100 bins of equal size and colour is used to represent the proportion of parameter combinations that resulted in community coexistence in each bin. Purple represents areas of the parameter space in which none (or very few) of the parameter combinations resulted in community coexistence and yellow represents areas of the parameter space in which all (or most) of the parameter combinations resulted in community coexistence. Please note that the parameters were rescaled between zero and one for plotting purposes.

Regression

Community-level model outputs

Fewer than seven species-specific parameters were present with frequencies of more than 20 when predicting the continuous community-level outputs of the *mizer* model, such as community biomass, fisheries yield, and mean weight (Figure 4.13). However, the F of a number of different species was consistently found in the best performing random forests (Figure 4.13).

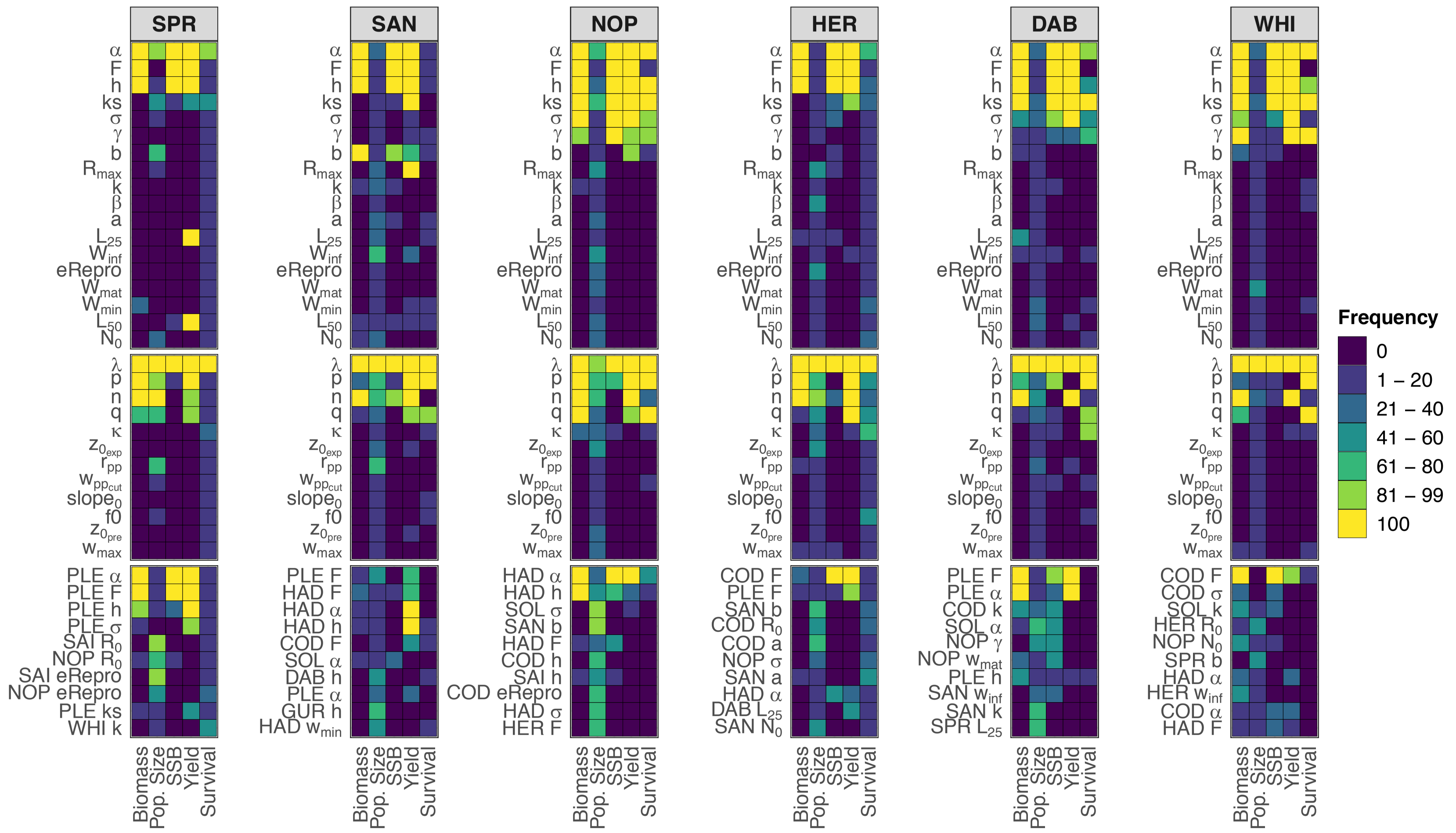


Figure continued overleaf.

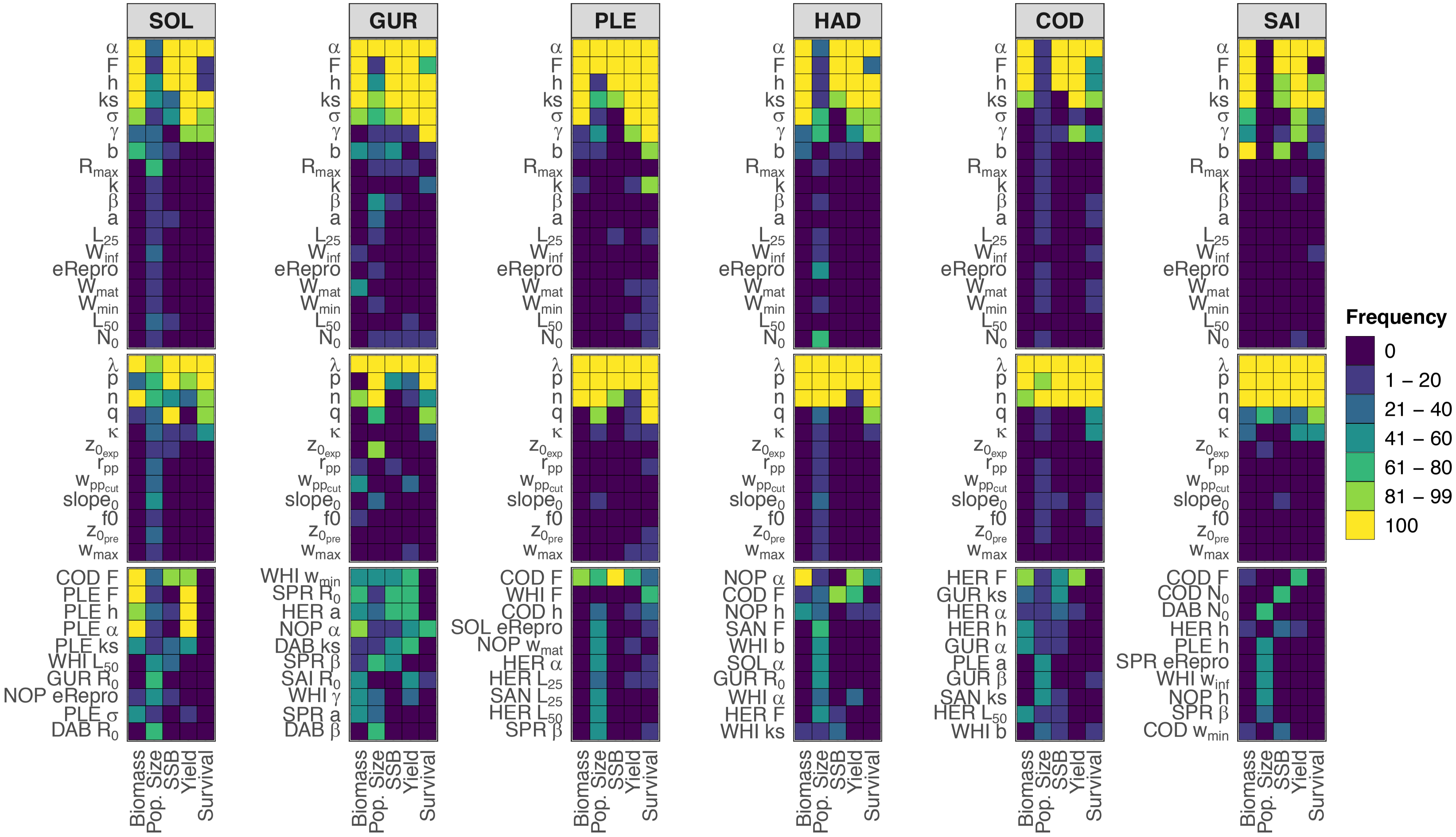


Figure 4.18: The frequency with which the parameters of the North Sea multispecies *mizer* model appeared in the best performing random forests when predicting species biomass, population size, Spawning Stock Biomass (SSB), fisheries yields, and species survival. The parameters in the top panel represent those associated with the same species that the random forest was being used to predict, whilst the parameters in the middle panel represent the species-independent parameters in the model. The parameters in the bottom panel represent the ten parameters that were associated with a different species to the one the random forest was being used to predict that had the greatest frequency across all five model outputs in question. The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe. A maximum frequency of 100 was possible as the random forest algorithm was applied to 100 different testing datasets for each model output.

For example, when predicting fisheries yield, the Large Fish Indicator (LFI), mean weight, and community slope, the F of the larger species in the model, such as European plaice, haddock, Atlantic cod, and saithe, were present in at least 81 of the 100 best performing random forests (Figure 4.13). The F associated with Atlantic herring was also present with high frequency (>65 out of 100) when using the random forest algorithm to predict community biomass, population size, SSB, and fisheries yield (Figure 4.13). Furthermore, both the α and h of Atlantic herring were present in all 100 of the best performing random forests when predicting community biomass, SSB, and fisheries yield (Figure 4.13). None of the species-specific parameters were present in more than 75 of the best performing random forests when predicting the community population size (Figure 4.13). Instead, almost all of the species-specific parameters appeared in the best performing random forests for community population size with very low frequencies (Figure 4.13), despite often not being present in any of the best performing random forests for any other continuous community-level model output.

In terms of the species-independent parameters in the *mizer* model, λ was present in all 100 of the best performing random forests for every continuous community-level model output excluding population size (Figure 4.14). For community population size, λ was present in 95 of the 100 best performing random forests (Figure 4.14). n was also present in all 100 of the best performing random forests for every continuous community-level model output excluding mean weight, whilst p and q were present in all 100 of the best performing random forests when predicting the LFI, mean weight, and community slope (Figure 4.14). All other species-independent parameters were present in fewer than 39 of the 100 best performing random forests (Figure 4.14). None of the parameters in the species interaction matrix θ were present in more than 48 of the best performing random forests for any of the continuous community-level model outputs (Figure 4.15).

Species-specific model outputs

When predicting the continuous species-specific outputs of the *mizer* model, the parameters were again divided into three groups: (1) those associated with the species being predicted by the random forest; (2) those associated with a different species to the one being predicted by the random forest; and (3) the species-independent parameters. The parameters in the first group that were most often present in the best performing random forests were relatively consistent across the different species and model outputs (Figure 4.18). For example, α, F, and h were present in at least 98 of the 100 best performing random forests when predicting the biomass, SSB, and fisheries yields of every species in the model (Figure 4.18). The coefficient of standard metabolism ks and the width of the prey size preference σ were also present with high frequency for many of the species in the model, with the exception of sprat, sandeel, and Atlantic cod (Figure 4.18). When predicting the population size of each species, the best performing random forests tended to include a wider range of parameters with lower frequencies than the other model outputs (Figure 4.18). This pattern was evident across both the species-specific and species-independent parameters for all species excluding sprat, European plaice and saithe (Figure 4.18). The species-specific parameters that were present with the lowest frequencies when predicting population sizes included the activity coefficient k, the length at which 25% of the stock is selected by the fishing gear L_{25}, and the size class of recruits W_{min} (Figure 4.18). For all other continuous species-specific model outputs, the species-specific parameters that were present in the best performing random forests with the lowest frequencies included the length-weight converter a, initial population size N_0, and the reproductive efficiency $eRepro$ of each species (Figure 4.18).

Few of the species-specific parameters that were associated with a different species to the one being predicted were present in the best performing random forests with high frequency (Figure 4.18). However, the F of Atlantic cod and/or European plaice was often present in the best performing random forests when predicting the biomass, SSB, or fisheries yields of almost all of the species in the *mizer* model (Figure 4.18). Other parameters associated with European plaice, such as α and h, were also often present in the best performing random forests when predicting the biomass, SSB, and fisheries yield of sprat, as well as the biomass and fisheries yields of dab and common sole (Figure 4.18). Furthermore, the α and h of haddock were frequently present in the best performing random forests for all of the model outputs associated with Norway pout (Figure 4.18). Similarly, the α of Norway pout was always present in the best performing random forests when predicting the biomass of haddock (Figure 4.18).

Again, λ, n, p, and q were the species-independent parameters that were most often present in the best performing random forests when predicting the continuous species-specific outputs of the *mizer* model (Figure 4.18). More specifically, λ was present in all 100 of the best performing random forests for all species and model outputs excluding the population size of Norway pout and common sole (Figure 4.18). n and p were present in the best performing random forests with particularly high frequency for all of the model outputs that were associated with the larger species in the model, such as saithe, Atlantic cod, European plaice, and haddock (Figure 4.18). The average feeding level of individuals feeding mainly on the background resource f_0, the coefficient of the background mortality of the community spectrum $z_{0_{pre}}$, the exponent of the background mortality of the community spectrum $z_{0_{exp}}$, and the maximum size of the community spectrum w_{max} were all present in fewer than eight of the 100 best performing random forests across all species and model outputs excluding population size

(Figure 4.18).

Summaries of the parameters that appeared in all 100 of the best performing random forests when predicting the community and species-specific model outputs are given in Tables 4.3 and 4.4 respectively.

Table 4.3: The species-independent and species-specific parameters of the North Sea multispecies *mizer* model that appeared in all 100 of the best performing random forests when predicting community biomass, population size, Spawning Stock Biomass (SSB), fisheries yield, the Large Fish Indicator (LFI), mean weight, community slope, and community coexistence. The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe.

Model output	**Species-independent**	**Species-specific**
Biomass	λ, n	HER α, HER F, HER h SAN F
Pop. Size	n	
SSB	λ, n	HER α, HER F, HER h SAN F
Yield	λ, n	COD F, HAD F, HER α HER F, HER h, PLE F
LFI	λ, n, p, q	COD F, HAD F, PLE F SAI F
Mean Weight	λ, p, q	COD F, HAD F, PLE F
Slope	λ, n, p, q	COD F, SAI F
Coexistence	λ, n, p, q	PLE α, PLE F

Table 4.4: The species-independent and species-specific parameters of the North Sea multispecies *mizer* model that appeared in all 100 of the best performing random forests when predicting species-specific biomass, population size, Spawning Stock Biomass (SSB), fisheries yields, and species survival. The species-specific parameters are divided into those associated with the species the random forest was being used to predict and those associated with a different species to the one the random forest was being used to predict. The following abbreviations are used to represent each species: SPR - sprat, SAN - sandeel, NOP - Norway pout, HER - Atlantic herring, DAB - dab, WHI - whiting, SOL - common sole, GUR - grey gurnard, PLE - European plaice, HAD - haddock, COD - Atlantic cod, SAI - saithe (cont. overleaf).

Species	Type of parameter	Biomass	Pop. Size	SSB	Yield	Survival
Sprat	Species-independent	λ, n, p	λ, n	λ	λ, p	λ
	Same species	α		α, F, h	α	
	Different species	PLE α, PLE F		PLE α, PLE F	PLE α, PLE F, PLE h	
Sandeel	Species-independent	λ, n	λ	λ	λ, n, p	$\lambda\ p$
	Same species	α, b, F, h		α, F, h	α	
	Different species				HAD α, HAD h	
Norway pout	Species-independent					λ, p, q
	Same species	α, F, h, ks		α, F, γ, h, ks	α, F, h, ks, σ	α, h, ks
	Different species	HAD α, HAD h		HAD α	HAD α	
Herring	Species-independent	λ, n, p	λ	λ	λ, n	λ
	Same species	α, F, h		α, F, h	α, F, h	
	Different species			COD F	COD F	
Dab	Species-independent	λ, n	λ	λ	λ	λ, p
	Same species	α, F, h, ks		α, F, h, ks	α, F, h, ks, σ	ks
	Different species					
Whiting	Species-independent	λ, n	λ	λ	λ, n	λ, p, q
	Same species	α, F, γ		α, F, h, ks	α, F, γ, h	α, γ, ks
	Different species	COD F		COD F		

Sole	Species-independent Same species Different species	λ, n α COD F, PLE α, PLE F		λ, p, q α, F, h	λ α, F PLE α, PLE F, PLE h	λ, p α, ks
Gurnard	Species-independent Same species Different species	λ α, F, h, ks	λ, n, p α	λ α, F, h, ks	λ α, F, h, ks	λ $\alpha, \gamma, h, ks, \sigma$
Plaice	Species-independent Same species Different species	λ, n, p α, F, h	λ, n, p α, F	λ, p α, F, h COD F	λ, p α, F, h, ks	λ, n, p, q α, F
Haddock	Species-independent Same species Different species	 α, F, h, ks, σ NOP α	λ, n, p	λ, n, p α, F, h	λ, p α, F, h, ks	λ, n, p α, h, ks
Cod	Species-independent Same species Different species	λ, p α, F, h	λ, n	λ, n, p α, F, h	λ, n α, F, h, ks	λ, n, p α
Saithe	Species-independent Same species Different species	λ, n, p α, b, F	λ, n, p	λ, n, p α, F	λ, n, p α, F, h	λ, n, p α, ks

4.5 Discussion

We applied the random forest algorithm to the North Sea multispecies *mizer* model to assess the ability of the algorithm to predict the behaviour of the model using different subsets of the 'most important' predictor variables (or model parameters). The random forest algorithm was particularly successful at predicting both community coexistence and species survival. For species survival, the random forest algorithm was most accurate when used to predict the survival of sprat, sandeel, and Atlantic herring, and least accurate when used to predict the survival of Norway pout. Although we might expect the species with the greatest survival rates to be associated with increased levels of accuracy due to the larger amount of information available to the algorithm regarding the parameter combinations that resulted in the survival of the species, Norway pout survived in over 80% of the model evaluations, whilst sprat, sandeel, and Atlantic herring survived in 60-66% of the model evaluations. It would therefore seem that the algorithm performed best when provided with a more balanced set of model evaluations. This theory is further supported by the fact that European plaice survived in the fewest number of model evaluations (43%) and was also associated with comparatively low accuracy, which suggests that the algorithm was provided with too few examples of parameter combinations that resulted in the survival of the species to make highly accurate predictions.

The random forest algorithm also performed relatively well when used to predict community and species-specific biomass, SSB, and fisheries yield, as well as the LFI, mean weight, and community slope, but it was not able to accurately predict community and species-specific population sizes, most likely due to the much greater variability associated with these model outputs. For biomass, SSB, and fisheries yield, there was a consistent pattern in the ranking of each species based on the RMSEs of the random forest algorithm. For example, species of intermediate size, such as dab, grey gurnard, and common sole, were consistently associated with the lowest RMSEs (or greatest accuracy), whilst the smaller species in the model, such as sandeel and Atlantic herring, were consistently associated with the greatest RMSEs (or lowest accuracy). Similarities in the predictive ability of the random forest algorithm across these model outputs is to be expected given the interdependencies of the biomass, SSB, and fisheries yield model outputs, all of which are weight-based. The reduced accuracy with which the random forest algorithm was able to predict the biomass, SSB, and fisheries yields of the smaller fish species in the model may be caused by the increased impact of predation pressure on these species compared with the larger species in the model. Increased predation pressure may make the smaller fish species in the model more difficult to predict because of the greater complexity in their species interaction networks; such complexity occurs as a result of the interrelationships between predation, competition, and fishing pressure (Wong and Candolin, 2015).

The RMSEs of the random forests that were used to predict the population size of each species displayed a remarkably different pattern to biomass, SSB, and fisheries yields, with Atlantic cod being associated with the lowest RMSEs and Norway pout being associated with the greatest RMSEs. This difference is to be expected given that the highly interdependent weight-based outputs of the *mizer* model are likely to respond to changes in the parameters in similar ways, whilst the population size, which is measured in terms of numbers of individuals, may respond very differently. To illustrate this point, the biomass (and hence the SSB and fisheries yield) of a species may fluctuate greatly without there being any change in the population size (and vice versa). The random forest algorithm may therefore be expected to show similar patterns in accuracy across all of the weight-based model outputs, but a different pattern when used to predict the population size. Furthermore, the RMSEs of the random forests that were used to predict population size may be lowest for Atlantic cod and greatest for Norway pout due to differences in the strength of association between the population size of each species and the parameters of the *mizer* model, a factor that is discussed further below.

Overall, we found that the algorithm was successfully able to predict the majority of the model outputs with relatively high accuracy using information on fewer than ten of the 306 parameters. These results highlight the driving nature of a small subset of model parameters; a fact that is perhaps best exemplified by the community slope, which required just six parameters to minimise the RMSEs of the random forests for all 100 testing datasets. This level of consistency was unparalleled across all other model outputs, indicating a weaker association between these outputs and any given set of parameters in the *mizer* model. Both community and species-specific population sizes required the largest number of parameters to maximise the accuracy of the predictions. However, we have shown that far fewer parameters could be used to predict population sizes with very little loss in agreement between the observations and the predictions made by the random forest algorithm; the same is also true for all other model outputs that required increased numbers of parameters to maximise accuracy. If the subsets of parameters included in the best performing random forests are consistent across all of the model outputs, as might be expected given the interrelationships between them, it may therefore be possible to predict all of the outputs of the *mizer* model with relatively high accuracy using information on fewer than 5% of the parameters. Such consistencies in the parameters that appeared in the best performing random forests with the greatest frequencies across all of the model outputs are explored further in Section 4.5.1 below.

4.5.1 Frequency

The six parameters that were required to predict the community slope with the greatest accuracy across all 100 testing datasets included the exponent of the background resource spectrum λ, the scaling of food intake n, the scaling of standard metabolism p, the search volume exponent q, and the fishing mortality F associated with Atlantic cod and saithe (Table 4.3). The parameters that were present in the best performing random forests with the greatest frequencies across all other model outputs were similar to those listed above but with a few additions. For example, the assimilation efficiency α, maximum food intake rate h, volumetric search rate γ, width of the prey size preference σ, and the coefficient of standard metabolism ks were also present in the best performing random forests with high frequency for one or more of the model outputs (Tables 4.3 and 4.4). It is clear from these results that there is a strong association between the outputs of the *mizer* model and the parameters associated with the acquisition and assimilation of food and the standard metabolism of each species. The importance of these parameters is unsurprising given that any changes to feeding and metabolic rates would impact the ability of all 12 fish species to grow and reproduce (Steele et al., 2001; Lall and Tibbetts, 2009). A change in any one of these parameters could therefore have huge knock-on effects across the entire community through changes in predator-prey interactions and competition for resources (commonly referred to as a 'trophic cascade') (Steele et al., 2001; Baum and Worm, 2009). Understanding these knock-on effects is of great importance given that climate change is expected to affect feeding and metabolic rates in the future (Roessig et al., 2004). However, single-species models are typically used to help develop management strategies for the North Sea marine ecosystem, despite being unable to simulate a trophic cascade in full. Increasing the contribution of ecosystem models, such as *mizer*, to the decision-making process is therefore vital to ensuring that these knock-on effects can be assessed and accounted for when developing management strategies for the future (Hyder et al., 2015).

In addition to the parameters associated with the acquisition and assimilation of food and the standard metabolism, the fishing mortality F of a number of different species, particularly the larger fish species, were present in the best performing random forests with high frequency for most of the model outputs. The importance of F is to be expected given that the removal of the largest fish species would not only affect the population size of the fished species through declines in the number of reproductively-active individuals, but would also impact the productivity of the smaller fish species through a reduction in predation pressure (Steele et al., 2001; Baum and Worm, 2009). Assuming the *mizer* model can be deemed an accurate representation of the real world, the importance of the parameters related to both food intake and fishing mortality thus highlights the importance of bottom-up (resource-driven) and top-down (consumer-driven) processes on fish populations in the North Sea (Cury et al., 2008; Lynam et al., 2017). Again, using models such as *mizer* to discern the relative influence of bottom-up and top-down processes on different species, as well as to understand the community-wide impacts of changes in either one of these processes, is thus necessary to ensure that management efforts are not misplaced.

The importance of the parameters associated with the larger fish species in the model was perhaps most apparent for the species-specific outputs. For these outputs, the parameters were divided into three groups: (1) those associated with the species being predicted by the random forest; (2) those associated with a different species to the one being predicted by the random forest; and (3) the species-independent parameters. Unsurprisingly, the parameters in the first group were generally present in the best performing random forests with much greater frequency than the parameters in the second group due to their direct control on the outputs being predicted by the random forest. However, a number of parameters associated with European plaice and/or Atlantic cod, including F, α, h, and σ, appeared in the best performing random forests when predicting the biomass, SSB, or fisheries yields of almost all of the species in the model. European plaice and Atlantic cod may be particularly important due to the aforementioned knock-on effects of changes in the feeding rates of large predators on lower trophic levels (Steele et al., 2001), either by direct changes to the parameters related to feeding or through the removal of these predators via fishing. Parameters associated with haddock, such as α and h, also appeared in the best performing random forests when predicting the biomass, SSB, or fisheries yield of Norway pout and vice versa. It is possible that this relationship is driven largely by the strength of the interaction θ between these species in the *mizer* model, which is rivalled only by the interaction between dab and European plaice (Blanchard et al., 2014).

For both community and species-specific population size, the species-specific parameters related to feeding, metabolism, and fishing mortality were generally present in the best performing random forests with much lower frequencies when compared with the random forests that were used to predict the other model outputs. Instead, almost all of the species-specific parameters appeared in the best performing random forests with low frequency, despite most of these parameters not being present in any of the best performing random forests for any other model output. Community coexistence, as well as the survival of sprat and Atlantic herring, also displayed similar patterns in parameter frequency to population size but to a lesser extent. These results suggest that there is a much looser association between these model outputs and any given set of parameters, which may help to explain the increased number of parameters required to predict the survival of Atlantic herring, as well as community and species-specific population size. A looser association between the model parameters and population size may be expected given that the *mizer* model is a size-based model and thus all of the parameters in the model are related to size (or weight) rather than numbers of individuals (Scott et al., 2014). The model is also fitted to fisheries landings data (Spence et al., 2016), which is measured in terms of biomass, thus strengthening the association between the model parameters and the weight-based model outputs. Although community coexistence is related to weight, it may have a weaker association with the model parameters as it is not a direct model output - it is calculated based on the value of another model output. Furthermore, successfully predicting coexistence requires the random forest algorithm to be capable of accurately predicting extremely low species biomass, which may not be the case (see Appendix A).

Although community coexistence may have a looser association with the parameters of the *mizer* model when compared with some of the other model outputs, λ, n, p, q, and the α and F associated with European plaice were present in all 100 of the best performing random forests for community coexistence (Table 4.3). Of these six parameters, it was λ that displayed the strongest relationship with this model output, with coexistence occurring only when the value of λ was above 2.06. A brief exploration of the baseline North Sea *mizer* model indicates that a change in λ from the nominal value of 2.33 to 2.06 causes a dramatic decline (of up to approximately 3.4 $\times$ 10^{30} at the start of the model evaluation) in plankton abundance in the smallest size classes and a comparably small increase (of up to approximately 1.3 $\times$ 10^{9} at the start of the model evaluation) in plankton abundance in the largest size classes. It is therefore likely that when λ is equal to 2.06, plankton numbers drop below some critical threshold beneath which it is not possible to sustain all 12 fish populations in the model. Not only does the knowledge of this threshold help us to better understand the inner workings of the model, it may also be useful for decision makers if the model is used to support policy in the future. Knowledge of this threshold may be particularly important to decision makers given that the abundance of plankton has declined significantly in the North Sea over the past 25 years (Capuzzo et al., 2018). If the declining trend in plankton abundance continues into the future, as may be expected under climate change (Capuzzo et al., 2018), the relevance of this threshold will only increase.

Such a defined threshold was not apparent for any other parameter, although community coexistence did not occur when low values of n were combined with high values of p. These results indicate that there were parameter combinations in which food intake rates could not sustain standard metabolic rates, resulting in the extinction of one or more species in the model. However, further research is required to better understand the parameter combinations that result in the extinction of one or more species using more realistic parameter distributions than were used in this research (see Section 4.5.3 for further details).

The parameters that were present in the best performing random forests with the lowest frequencies were again similar across all of the community-level and species-specific model outputs and were largely associated with the species interaction matrix θ, the initial population size of each species N_0, the starting slope of the community spectrum $slope_0$, the maximum size of the background and community spectra ($w_{pp_{cut}}$ and w_{max} respectively), and the natural and predation mortality rates of the background and community spectra (f_0, $z_{0_{pre}}$, and $z_{0_{exp}}$). It is perhaps unsurprising that parameters such as N_0 and $slope_0$ were less important in helping to successfully predict the outputs of *mizer* as they only play a role at the very start of each model evaluation. Furthermore, although the size of the background resource was found to be extremely important in ensuring the random forest algorithm was able to accurately predict all of the model outputs, it is clear that this is driven almost entirely by λ rather than the parameters associated with the maximum size $w_{pp_{cut}}$ or mortality rates f_0 of this resource. Similarly, parameters relating to the acquisition and assimilation of food outside of the background resource were also shown to be important in improving the accuracy of the random forest algorithm across all model outputs. As the interaction matrix is used to help quantify food encounter rates in the model (Andersen et al., 2015), we might expect θ to be present in a much larger number of the best performing random forests. However, it is clear that the importance of the parameters related to food intake are largely driven by n, q, and γ instead of the interaction matrix.

4.5.2 Comparison with global sensitivity analysis

The parameters that most often appeared in the best-performing random forests (namely λ, n, p, q, F, and α) were very similar to those associated with the greatest sensitivity indices in a derivative-based sensitivity analysis of the North Sea multispecies *mizer* model (see Chapter 3). Such similarities highlight the potential ability of the random forest algorithm to identify the parameters that drive changes in different model outputs with far fewer model evaluations than a global sensitivity analysis. The methods described in this research may therefore be particularly beneficial for models that have previously been deemed to be too computationally expensive to conduct a sensitivity analysis, either due to the number of parameters included in the model or as a result of long model run times.

In order to confirm this theory, further comparisons must be made between the two methods when applied to different models. Such comparisons may be relatively inexpensive to complete for marine ecosystem models that have already been run under many different parameter combinations, either for a sensitivity analysis (e.g. StrathE2E; Morris et al. (2014)) or to determine parameter combinations that result in historically plausible model outputs (e.g. LeMANS; Thorpe et al. (2015, 2016, 2017)). Not only would this research help us to better understand the potential benefits of using machine learning in marine ecosystem modelling, it would also enable us to assess the generality of our conclusions, particularly in terms of the areas in which to focus future research efforts to reduce the uncertainties in the model outputs.

However, it is important to note that although the methods described in this research may require far fewer model evaluations to complete than a global sensitivity analysis, it may still not be computationally feasible to apply these methods to some of the most complex marine ecosystem models. For example, models such as ERSEM (`https://www.pml.ac.uk/Modelling_at_PML/Models/ERSEM`) and Atlantis (`https://research.csiro.au/atlantis/`) may take much longer than *mizer* to complete a single model evaluation. It is therefore unlikely that a large number of model evaluations could be run in a reasonable amount of time to explore the parameter space in enough detail to successfully train the random forest algorithm. Nevertheless, it may be possible to use the random forest algorithm to explore the behaviour of these models in small subsections of the parameter space or in specific components of the models.

4.5.3 Limitations

The main limiting factor during this research was the high computational costs associated with running both the *mizer* model and the random forest algorithm. The large number of parameters included in the *mizer* model, as well as the often long model run times, prevented us from being able to explore the entire parameter space of the model. In an attempt to overcome this issue, we used Latin Hypercube Sampling (LHS) to generate a stratified sample of the parameter space. LHS generally provides a better representation of the parameter space than random sampling by preventing clustering and helping to ensure the edges of the parameter space are included in the sample (Agarwal et al., 2012; O'Sullivan and Perry, 2013). However, some regions of the parameter space may still be undersampled when using LHS (Agarwal et al., 2012; O'Sullivan and Perry, 2013). Despite this, we believe the 5000 parameter combinations used in this research offer a good starting point with which to explore the ability of the random forest algorithm to predict the outputs of the *mizer* model. Further research is required to determine whether the conclusions of this research remain the same when a larger set of parameter combinations is used to train the random forest algorithm.

Using realistic parameter distributions (instead of uniform distributions with upper and lower limits of ±10% of the nominal parameter values) would additionally enable us to better understand the ability of the random forest algorithm to accurately predict the outputs that are of most interest to scientists, decision makers, and the general public. Using realistic parameter distributions may be especially important to decision makers as the algorithm cannot be used to explore the consequences of parameter values that exceed the range of values given in the training data (Müller et al., 2016). Therefore, if the realistic distribution of a parameter exceeds ±10% it would not be possible to use the trained random forests to predict the outputs of the model and to explore the impacts of different management strategies on the North Sea fish community.

As was previously noted in a global sensitivity analysis of the *mizer* model (see Chapter 3), a small number of the parameter combinations also result in 'extreme' model behaviour, such as widespread species extinctions and the subsequent dominance of one or two fish species. This extreme model behaviour has the potential to inflate the number of parameters required to maximise the accuracy of the random forest algorithm and to increase the frequencies with which each parameter is present in the best performing random forests. In an attempt to prevent extreme model evaluations from having an undue influence on the results of the analysis, we applied the random forest algorithm to 100 different testing datasets, each of which consisted of a random sample of 40% of the model evaluations. In doing so, we reduced the likelihood of an extreme model evaluation appearing in each of the testing datasets, thereby helping to ensure the results were less skewed by extreme model behaviour. However, this issue may be prevented entirely by identifying a set of parameter combinations that produce historically plausible model outputs, as has already been achieved for the North Sea LeMANS marine ecosystem model (see Thorpe et al. (2015, 2016, 2017) for example). The random forests described in this research may help to achieve this goal by screening large numbers of parameter combinations to give an initial indication as to whether the combinations may result in plausible model behaviour.

Finally, we chose to use random forests for their relative simplicity and ease of application. However, it is possible that other forms of ML may be capable of predicting the outputs of the *mizer* model with greater accuracy than the random forest algorithm. A preliminary exploration of the ability of Support Vector Machines (SVMs) to successfully predict the outputs of the *mizer* model indicated that SVMs were not able to outperform the random forest algorithm using the default settings (not shown). Although several parameters may be tuned to optimise the performance of the SVM algorithm, we chose not to explore this avenue of research as it would have required much greater computational resources to conduct the analysis than the random forest algorithm (Li and Kong, 2014). Further research is therefore required to determine if tuned SVMs, or any other supervised ML algorithms, are successfully able to predict the outputs of the *mizer* model with greater accuracy than the random forest algorithm. The accuracy and efficiency of traditional methods of data analysis, such as logistic regression, should also be explored to ensure the most appropriate method is used to predict the outputs of the *mizer* model in the future.

4.6 Implications and conclusions

To the best of our knowledge, this study is the first to apply the random forest algorithm to predict the behaviour of a complex marine ecosystem model with more than 300 parameters. The algorithm was successfully able to predict most of the outputs of the *mizer* model using information on fewer than ten parameters. Nevertheless, improvements in accuracy would need to be made if the trained random forests were to be used for predictive purposes in the future. Such improvements could be made by training the random forest algorithm with a larger set of model evaluations or by fine tuning the parameters of the algorithm. Further research is therefore required to explore the potential benefits of using increased computational resources to improve the accuracy of the random forest algorithm, as well as to identify whether any other method of ML may be better able to predict some or all of the outputs of the *mizer* model.

The parameters that were most often found in the best performing random forests were largely consistent across all of the model outputs. The parameters with the greatest frequencies tended to include six of the species-specific (F, α, h, γ, σ, and ks) and four of the species-independent (λ, n, p, and q) parameters, all of which were related to feeding, metabolic, and fishing mortality rates. The importance of these parameters is supported by a global sensitivity analysis of the *mizer* model (see Chapter 3), which identified the need to focus future research efforts on reducing the uncertainties associated with the size of the background resource (i.e. the planktonic community), the acquisition and assimilation of food, and fishing mortality rates. Large-scale monitoring of the planktonic community through surveys such as the Continuous Plankton Recorder (`sahfos.co.uk`), increased availability of data related to food intake, such as the collated stomach content analysis data that was recently made available by the Centre for Environment, Fisheries and Aquaculture Science (`cefas.co.uk`; Pinnegar et al. (2015)), and improved reporting of fishing activity would help to achieve this goal.

Overall, the importance of exploring the ability of the random forest algorithm to predict the outputs of the *mizer* model cannot be overstated. It can take days or even weeks for some of the model evaluations to reach equilibrium. Being able to accurately predict the outputs of the model using information on a small number of the parameters may reduce the need to run the full model and help to screen potential parameter combinations for historically plausible model outputs, thereby helping to lessen the costs associated with marine ecosystem modelling both in terms of human and computational resources; this may be especially important for scientists and decision makers that have limited funding and/or short deadlines. The methods described here may also be used to highlight important interactions between species in the model. For example, European plaice and Atlantic cod were identified as influential drivers of the biomass, SSB, and/or fisheries yields of almost all of the lower trophic level species in the *mizer* model. If deemed to be biologically important, these interactions must be accounted for when making management decisions. To do this, we need to continue to move from single-species to ecosystem-based management of the North Sea. Doing so will help to prevent unintended consequences of management decisions on non-target species and thus help to ensure that fishing activity remains sustainable (Uusitalo et al., 2015). However, further research is required to place more informative distributions on the most important parameters, such as those associated with feeding, metabolic, and fishing mortality rates. Not only would this research help us to further understand and improve the behaviour of the *mizer* model, it may also help to increase the confidence that decision makers have in marine ecosystem models, both of which are vital to ensuring the *mizer* model can be used to support fisheries management in the future.

Chapter 5

Uncertainty in projections of global and regional sea surface temperature and salinity

5.1 Abstract

Environmental models are increasingly being used to identify possible changes in environmental conditions in response to climate change over the next century. However, these models often suffer from large uncertainties; understanding the implications of these uncertainties is vital to ensuring the successful development of robust management solutions for the future, but disentangling the effects of multiple sources of uncertainty can be extremely difficult. Fortunately, multi-model ensembles, which combine the outputs of multiple structurally different models run under a common set of scenarios, may be used to disentangle the effects of three types of uncertainty: internal variability, model, and scenario uncertainty. In this research, we use the methods of Hawkins and Sutton (2009) to quantify the relative contributions of internal variability, model, and scenario uncertainty to the total variance of the projections of global and regional Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) from over 10 of the latest state-of-the-art climate models developed during phase five of the Coupled Model Intercomparison Project (`cmip.llnl.gov/cmip5/`). The results of this research are used to: (1) discuss spatio-temporal changes in the dominance of each type of uncertainty; (2) discuss the potential to reduce the uncertainty in the model projections; and (3) identify regions and time periods in which the projections are most certain and are thus most useful to decision makers in terms of adaptation planning. Overall, we hope that this research can be used to improve the representation of SST and SSS in climate models in the future, thus helping to ensure ocean heat and CO_2 uptake, sea level rise, and coupled ocean-atmosphere phenomena such as ENSO are correctly specified in the models.

5.2 Introduction

Identifying possible changes in environmental conditions in response to climate change over the next century is of great societal and economic importance due to the potentially wide-ranging impacts on resources, particularly in terms of food and water security (Villarini and Vecchi, 2012). To identify these possible changes, we must make use of projections from environmental models. These models often suffer from large uncertainties, yet management decisions must still be made in spite of this. Recognising and understanding the implications of these uncertainties is therefore vital to ensuring the successful development of robust management solutions for the future (Walker et al., 2003). A thorough evaluation of uncertainties may also enable the prioritisation of model components in which uncertainties may be reduced via further research, as well as help to identify irreducible uncertainties (Walker et al., 2003; Villarini and Vecchi, 2012).

The uncertainties in projections of future environmental conditions can usefully be divided into three broad categories: internal variability, model, and scenario uncertainty. In climate models, internal variability represents the inherent variability in the system that is not caused by external anthropogenic forcing (Hawkins and Sutton, 2009). For example, internal variability may arise through the chaotic nature of natural phenomena such as the El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), and the Pacific Decadal Oscillation (PDO), all of which can have a dramatic impact on atmospheric and oceanic conditions in their respective regions (Cheung et al., 2016). Internal variability may be of particular interest to decision makers as is it can mask the signal of anthropogenic-induced change over relatively short time frames (Hawkins and Sutton, 2009, 2011). Model uncertainty is described by Beven (1993, 1996, 2001, 2006) as an issue of 'equifinality', whereby a number of equally plausible model structures may generate very different predictions for the future. This type of uncertainty arises through the choice of variables or processes deemed necessary to include or exclude in a model, how these components are represented mathematically, and the relationships between these variables and the model inputs and outputs (Walker et al., 2003; Ascough II et al., 2008). Scenario uncertainty is most often associated with the unknown nature of future conditions, including the degree of socio-economic development, technological advances, and the impacts of - and societal responses to - climate change. For example, uncertainties regarding future greenhouse gas emissions will impact climate model projections via uncertainties in radiative forcing (Hawkins and Sutton, 2009).

Internal variability, model, and scenario uncertainties may be disentangled from one another by combining the outputs of multiple structurally different models, run under a common set of future scenarios, into a multi-model ensemble (Gårdmark et al., 2013). This disentangling is feasible since any variations in the outputs of different models within a single scenario will be caused solely by differences in the structure of the models (Wang et al., 2011; Gårdmark et al., 2013). Conversely, variations in the output of a single model under multiple scenarios will be caused solely by the propagation of uncertainties regarding possible changes in external forcing in the future (Knutti and Sedláček, 2013). Internal variability may then be estimated by running each model in the ensemble using multiple sets of initial conditions (see Deser et al. (2014) and Cheung et al. (2016) for example), although other options are available if an initial condition ensemble such as this is not available (see Hawkins and Sutton (2009) for example). However, few research areas have well-developed ensembles with which to explore these uncertainties. Fortunately, such ensembles are becoming increasingly popular in climate science as a result of accruing evidence suggesting seasonal climate predictions from multi-model ensembles are almost always more accurate than predictions from a single model (Schmittner et al., 2005). In particular, the Coupled Model Intercomparison Project (CMIP) of the World Climate Research Programme (WCRP) has been working for over two decades to develop a climate ensemble that currently includes over 30 different models, all of which have been run under a set of common scenarios for the future. The CMIP ensemble thus enables an exploration of the uncertainties associated with a wide range of environmental variables.

Hawkins and Sutton (2009) developed a novel approach to both quantifying and visualising the relative contributions of internal variability, model, and scenario uncertainties to the total variance of the multi-model ensemble developed during phase three of CMIP (CMIP3; `wcrp-climate.org/wgcm-cmip/wgcm-cmip3`). This method involves smoothing the model projections and estimating the contribution of each of the three types of uncertainty using: (1) the multi-model mean of the variance of the model residuals from the model fits, independent of lead time; (2) the multi-scenario mean of the variance of the model fits; and (3) the variance of the multi-model mean of the smooth fits (Hawkins and Sutton, 2009). Although simple, this method has proven to be effective in communicating the uncertainties associated with a wide range of different variables from global climate models, including surface air temperature (Hawkins and Sutton, 2009), precipitation (Hawkins and Sutton, 2011), sea surface temperature (Villarini and Vecchi, 2012; Cheung et al., 2016), sea level (Little et al., 2015), tropical storm frequency (Villarini and Vecchi, 2012), and the strength of the Atlantic Meridional Overturning Circulation (Reintges et al., 2017).

The aim of this research is to expand upon this growing body of literature by quantifying the contribution of internal variability, model, and scenario uncertainty to the total variance of projections of Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) using the latest state-of-the-art global climate models developed during phase five of CMIP (CMIP5; `cmip.`

llnl.gov/cmip5/). SST and SSS are of particular interest as the accurate representation of these variables is vital to ensuring ocean heat and CO_2 uptake, sea level rise, and coupled ocean-atmosphere phenomena such as ENSO are correctly specified in the models (IPCC, 2014a). Although Villarini and Vecchi (2012) and Cheung et al. (2016) have previously applied similar methods to those described in Hawkins and Sutton (2009, 2011) to projections of SST, the former focused solely on the tropics, while the latter focused on global means alongside two basin-scale examples from the Northeast Atlantic and Northeast Pacific. Here, we focus instead on both global and regional projections of SST and SSS. Including regional projections from around the world will enable us to better understand the spatio-temporal changes in the contribution of each type of uncertainty to the total variance of the projections. Quantifying the uncertainties at a regional level may also be of far greater relevance to decision makers (Hawkins and Sutton, 2009), as they tend to act at a regional rather than global scale. We further add to the work of Villarini and Vecchi (2012) and Cheung et al. (2016) by quantifying the signal-to-noise ratio of the projections in different regions, allowing us to identify regions and time periods in which the projections are most certain and are thus most useful to decision makers in terms of adaptation planning (Hawkins and Sutton, 2009).

5.3 Methods

To evaluate the spatio-temporal changes in the contribution of internal variability, model, and scenario uncertainty to the total variance of the projections of absolute Sea Surface Temperature (SST) and Sea Surface Salinity (SSS), we used the multi-model ensemble produced in phase five of the Coupled Model Intercomparison Project (CMIP5). This ensemble also forms part of the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2014a). Each model in the ensemble was used to simulate historical climatological and oceanographic conditions between 1850 and 2005 and then run under three scenarios for the future, known as Representative Concentration Pathways (RCPs), between 2006 and 2100. The three RCPs, referred to as RCP 2.6, RCP 4.5, and RCP 8.5, span the range of currently available estimates for the predicted level of radiative forcing that is expected to occur by the end of the century (van Vuuren et al., 2011). RCP 2.6 represents a scenario in which radiative forcing increases from <2W.m^2 at the start of the century to 3W.m^2 (~490ppm CO_2 eq) in the coming decades, followed by a decline to 2.6W.m^2 by the end of the century (van Vuuren et al., 2011). RCP 4.5 represents an intermediate pathway, with an increase in radiative forcing to 4.5W.m^2 (~650 ppm CO_2 eq) by 2100 (van Vuuren et al., 2011). Finally, RCP 8.5 represents a scenario in which radiative forcing substantially increases to 8.5W.m^2 (~1370 ppm CO_2 eq) by 2100 (van Vuuren et al., 2011).

For the purposes of this research, we selected models from the CMIP5 multi-model ensemble based on the availability of the appropriate data with consistent global coverage spanning the years between 1950 and 2099. This meant that we were able to use the projections from 14 different models for SST and 11 different models for SSS (see Appendix B for further details of the selected models). Please note that the following analysis was also applied using only the nine models that were available for both SST and SSS (see Appendix B), but the results were very similar to those presented in Section 5.4 and are therefore not discussed further.

Following the methods of Hawkins and Sutton (2009), we extracted the global annual mean SST and SSS from the historical simulations and from each of the three RCP model runs using the IPCC AR5 online database (`dkrz.de`). It is important to note that some of the models were run under each scenario multiple times, with each run producing slightly different model outputs. However, to ensure that all models were treated equally we chose to use only one set of outputs per model, selecting the first set of outputs in which all of the required scenarios were available (Hawkins and Sutton, 2009, 2011).

To determine the contribution of internal variability, model, and scenario uncertainty to the total variance of the global projections of SST and SSS, we fit a fourth-order polynomial to the output of each model to give a smooth fit (Hawkins and Sutton, 2009). A fourth-order polynomial was selected to ensure the non-linear response of SST and SSS to changes in radiative forcing could be captured whilst also smoothing the data to enable internal variability to be quantified (Reintges et al., 2017). The raw outputs X of a given model m under scenario s in the year t may be expressed as:

$$X_{m,s,t} = x_{m,s,t} + i_{m,s} + \varepsilon_{m,s,t} \tag{5.1}$$

where x represents the smooth fit of the fourth-order polynomial, i represents the mean SST or SSS in the reference period, and ε represents the residuals of the projections from the smooth fit. For the purposes of this research, i was defined as the mean of x between 1971 and 2000 in accordance with Hawkins and Sutton (2009, 2011). The models were not weighted by their ability to simulate historical trends in SST or SSS and instead we chose to treat all models as equally plausible representations of reality (Hawkins and Sutton, 2009). All models were also assumed to be independent as in Hawkins and Sutton (2011).

The uncertainty associated with internal variability IV was assumed to be equal to the multi-model mean of the variance of the residuals ε:

$$IV = \sum_{m} \text{var}_{s,t}(\varepsilon_{m,s,t}) \tag{5.2}$$

where $\varepsilon_{m,s,t}$ is smoothed before the variance is calculated and IV is constant in time. Although studies such as Boer (2009) have highlighted the potential for internal variability to increase slightly over decadal timescales, the effect is often negligible and we therefore chose to keep IV constant for simplicity (Hawkins and Sutton, 2011).

The model uncertainty M was assumed to be equal to the variance of the smooth fits, averaged across all three RCP scenarios:

$$M(t) = \frac{1}{N_s} \sum_s \text{var}_m(x_{m,s,t}) \tag{5.3}$$

where N_s represents the number of scenarios included in the analysis.

The scenario uncertainty S was assumed to be equal to the variance of the multi-model mean of the smooth fits:

$$S(t) = \text{var}_s \left(\frac{1}{N_m} \sum_m x_{m,s,t} \right) \tag{5.4}$$

As each type of uncertainty was assumed to be independent, the total variance T was equal to:

$$T(t) = IV + M(t) + S(t) \tag{5.5}$$

The mean change of all of the model outputs from the reference period, denoted as G, was equal to:

$$G(t) = \frac{1}{N_s} \sum_{m,s} x_{m,s,t} \tag{5.6}$$

The fractional uncertainty F (90% confidence level) was thus:

$$F(t) = \frac{1.65\sqrt{T(t)}}{G(t)} \tag{5.7}$$

and the Signal-to-Noise Ratio (SNR), which represents the robustness of the projections, was defined as the reciprocal of the fractional uncertainty.

To better understand the spatial changes in the contribution of each type of uncertainty through time, we applied the same method as described above to different regions across the globe. To do this, we extracted the annual mean SST or SSS from the multi-model ensemble for 54 distinct regions, which were selected based on Longhurst's widely accepted partitioning of the global ocean into biogeographical provinces (Longhurst (2007); see Appendix C for further details). Each province represents an area with different environmental and biological conditions, thus making this classification system particularly appropriate for use in this research.

5.4 Results and discussion

Projections of global Sea Surface Temperature (SST) indicate that warming is expected to occur in surface waters across the globe, with increases in absolute temperature ranging from 0.14% under RCP 2.6 to 1.33% under RCP 8.5 compared with the mean temperature in the reference period of between 1971 and 2000 (Figure 5.1). The opposite trend is apparent in the projections of global Sea Surface Salinity (SSS), in which declines of between 0.16% and 1.39% psu are expected to occur by 2099 (Figure 5.1). The projections for both SST and SSS remain similar under all three scenarios in the first half of the 21st century, before starting to diverge due to the delayed effect of emissions on the global climate (Figure 5.1; Hawkins and Sutton (2009)). The divergence between scenarios is most apparent in the projections of SST, where there is little overlap between the three scenarios by the end of the century (Figure 5.1). Conversely, there is much greater overlap between the projections of SSS under each of the three different scenarios, largely due to the increased spread between models (Figure 5.1). For example, the projected change in SSS under RCP 8.5 ranges from -0.35% to -1.39% by 2099, whilst the projected change in SST under the same scenario ranges from 0.63% to 1.33% by 2099 (Figure 5.1). By quantifying the spread between the different models and scenarios, along with the internal variability of the models, we can identify how the contribution of each type of uncertainty changes with lead time for both SST and SSS.

5.4.1 Uncertainty at a global scale

For projections of global SST, internal variability contributes a maximum of 17% to the total variance of the projections at short lead times, but this drops off to less than 1% by the end of the century (Figure 5.2c). These results differ from Cheung et al. (2016), who found the contribution of internal variability to be approximately 50% of the total variance at the start of the century. This difference is caused by contrasting methodologies, with Cheung et al. (2016) estimating the contribution of internal variability using the standard deviation of the outputs of a single CMIP5 model that was run under multiple sets of initial conditions. Although the methods used by Cheung et al. (2016) may provide a more accurate representation of the internal variability of a single model than the methods used here, they may not accurately reflect the internal variability of all of the models in the ensemble. However, our methods may

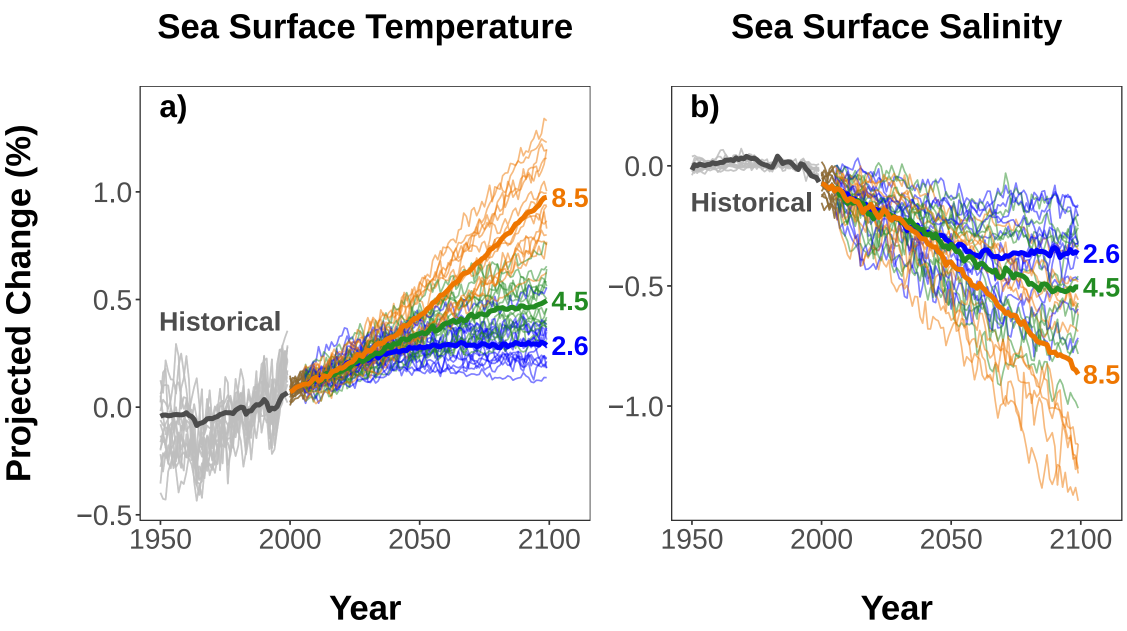


Figure 5.1: Projections of historical and future global annual Sea Surface Temperature (SST; a) and Sea Surface Salinity (SSS; b) from the selected CMIP5 climate models between 1950 and 2099. The projections are shown as the percentage change in SST or SSS from the reference period (1971-2000) to allow for direct comparisons to be made between the two variables. Historical projections (grey) cover the time period between 1950 and 2006, whilst the future projections cover the time period between 2006 and 2099. All models were forced with three Representative Concentration Pathways (RCPs), known as RCP 2.6 (blue), RCP 4.5 (green), and RCP 8.5 (orange). RCP 2.6 represents a low greenhouse gas emissions scenario, whilst RCP 4.5 and RCP 8.5 represent intermediate and high emissions scenarios respectively. The projections from individual models are shown as thin lines, whilst the multi-model mean is shown as a bold line.

also underestimate the contribution of internal variability, an issue that is discussed further in Section 5.4.3. Excluding the difference in contribution from internal variability at the start of the century, the findings of Cheung et al. (2016) are largely similar to those presented in Figure 5.2c, with the contribution of internal variability dropping to less than 5% for projections of SST by the 2090s.

Internal variability plays a larger role for projections of global SSS than SST at short lead times, contributing 28% to the total variance at the start of the century before again dropping to less than 1% by the 2090s (Figure 5.2d). This decrease in the percentage contribution of internal variability to the total variance is to be expected given that our methods assume internal variability does not change over time, whilst both model and scenario uncertainties increase with time (Figure 5.2b). It is also perhaps unsurprising that the contribution of internal variability to the total variance of the projections is greater for SSS than for SST, as it has previously been shown that internal variability plays a more important role in global and regional projections of precipitation than it does for surface air temperature (Räisänen, 2001; Murphy et al., 2004; Hawkins and Sutton, 2009, 2011). As precipitation is an important driver of SSS (along with evaporation, river run-off, and ice melt) and surface air temperatures are highly interrelated with SST, such similarities are to be expected (IPCC, 2007).

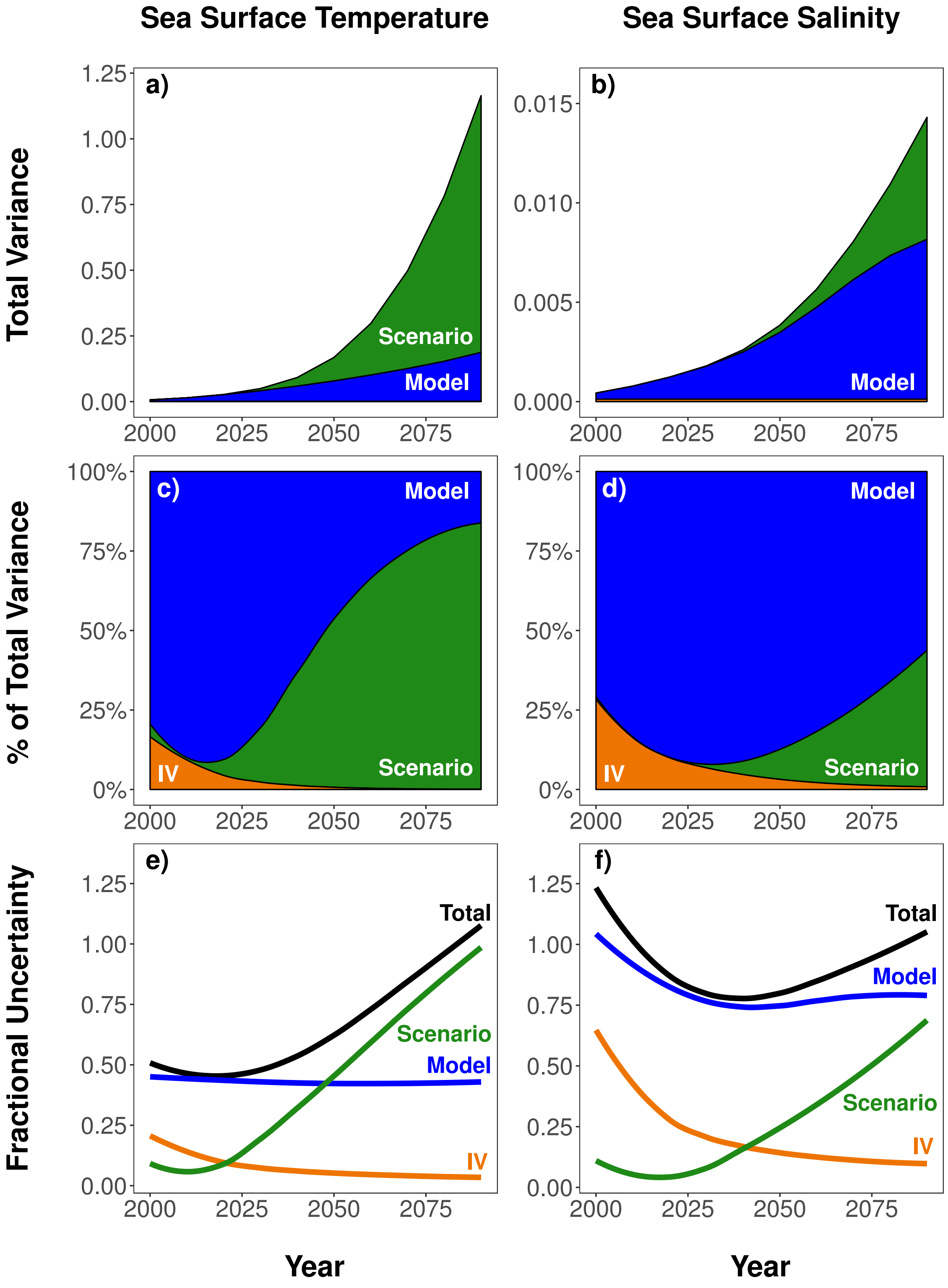


Figure 5.2: The contribution of internal variability (IV; orange), model (blue), and scenario (green) uncertainty to the total variance of the projections of future global, decadal sea surface temperature (left) and sea surface salinity (right) from the selected CMIP5 climate models run under three scenarios of future greenhouse gas emissions between 2006 and 2099. The contribution of each type of uncertainty to the total variance is depicted in a and b) in absolute terms (Kelvin (K)^2 and psu^2 for sea surface temperature and sea surface salinity respectively); in c and d) as a percentage; and in e and f) as the fractional uncertainty, defined as the 90% confidence level of the contribution of each type of uncertainty to the total variance of the projections divided by the mean change of the projections from the reference period of 1971 to 2000.

For both SST and SSS, the total variance of the projections increases exponentially towards the end of the century as model and scenario uncertainties grow (Figure 5.2a and b). Model uncertainty dominates the total variance of the projections at short lead times, with contributions of up to 92% prior to the 2040s (Figure 5.2c and d). For SST, the dominant contributor to the fractional uncertainty of the projections switches from model to scenario uncertainty in the 2050s as the spread between the three scenarios becomes larger than the spread between models (Figure 5.2e). By the end of the century, scenario uncertainty dominates the total variance of the projections of SST, with a contribution of over 80% (Figure 5.2c). For SSS, the contribution of scenario uncertainty to the total variance of the projections also increases with lead time, with contributions of less than 1% in the 2010s compared with 43% in the 2090s (Figure 5.2f). In contrast to SST, the uncertainty associated with projections of SSS continues to be dominated by model uncertainty at the end of the century, with a contribution of 56%, indicating the spread between models remains higher than the spread between scenarios (Figure 5.2f). The total fractional uncertainty of the projections is minimised in the 2020s for SST and in the 2040s for SSS, occurring immediately prior to the large increases in scenario uncertainty (Figure 5.2e and f). Although perhaps less important at a global scale than at a regional scale, the identification of this minimum may be particularly useful for planning purposes as it highlights the point in time in which the projections are most robust (Hawkins and Sutton, 2009).

Again, the contributions of both model and scenario uncertainty to the total variance of the projections of SST and SSS are similar to that of surface air temperature and precipitation (as described in Hawkins and Sutton (2009, 2011)) respectively. For example, the switch in dominance between model and scenario uncertainty occurs at the same lead time for both surface air temperature and SST. However, the variance attributed to scenario uncertainty is much larger for projections of SST than surface air temperature by the end of the century, with a value of more than 0.6K^2 for SST (Figure 5.2a) and between 0.3 and 0.4 Kelvin (K)^2 for surface air temperature (Hawkins and Sutton, 2009). Although such a direct comparison between the variance attributed to scenario uncertainty for SSS and precipitation is not possible given the information provided in Hawkins and Sutton (2011), the fractional uncertainty associated with the different scenarios is also much greater for SSS than for precipitation by the end of the century, with values of 0.69 (Figure 5.2b) and between 0.3 and 0.4 respectively (Hawkins and Sutton, 2011).

The increased scenario uncertainty in projections of SST and SSS compared with surface air temperature and precipitation is most likely caused by a difference in the scenarios used in the present research compared with those used in Hawkins and Sutton (2009, 2011). In this research, we used an ensemble of models that had been run under three Representative Concentration Pathways (RCPs), which represent a wider range of scenarios than those discussed in Hawkins and Sutton (2009, 2011). Hawkins and Sutton (2009, 2011) used an older version of the multi-model ensemble, which was produced in phase three of the Coupled Model Intercomparison Project (CMIP3) and made use of the scenarios discussed in the Special Report on Emissions Scenarios (SRES) (IPCC, 2000). The SRES scenarios B1 and A2 are comparable to RCP 4.5 and RCP 8.5 respectively, whilst the SRES scenario A1B represents a scenario somewhere between SRES B1 and SRES A2 (van Vuuren et al., 2011). This means there are no SRES scenarios that are comparable to RCP 2.6, which was recently introduced to represent a scenario in which emissions are reduced by the end of the century as a result of climate policy (van Vuuren et al., 2011; Knutti and Sedláček, 2013). Although the switch from the SRES scenarios to the RCP scenarios might increase the uncertainty associated with the outputs of the climate models, it does not mean that we have become more uncertain about how the climate will change in the future (Knutti and Sedláček, 2013). Instead, the increased scenario uncertainty reflects a choice between different economic scenarios (Knutti and Sedláček, 2013).

Because of the increased scenario uncertainty in the projections of SST and SSS compared with surface air temperature and precipitation, we might expect the percentage contribution of scenario uncertainty to the total variance of the projections to also be greater for SST and SSS by the end of the century. Although this seems to be the case for projections of SSS, it does not apply to SST as the variance attributed to model uncertainty is also larger for SST than surface air temperature, with values of 0.19K^2 (Figure 5.2a) and 0.1K^2 (Hawkins and Sutton, 2009) in the 2090s respectively. Potential reasons for the larger model uncertainty in projections of SST compared with surface air temperatures may include: 1) large regional biases in the projections of SST in some of the models, which result in a large spread between models at the global level (Wang et al., 2014; Cheung et al., 2016); 2) the inclusion of models with increased spread in the present research compared with the models used in Hawkins and Sutton (2009); and 3) model development between CMIP3 and CMIP5 that has resulted in an increase in the spread between models shared by this research and the research of Hawkins and Sutton (2009) (Yan et al., 2013). Large regional biases in the projections of SSS may also explain the increased contribution of model uncertainty to the total variance of the projections of SSS compared with SST, as discussed in Section 5.4.2 below.

5.4.2 Uncertainty at a regional scale

By dividing the projections into 54 biogeographical provinces based on Longhurst's scheme (Longhurst, 2007), we can identify regions that are particularly affected by each type of uncertainty at different lead times. It is this regional information that may be of the greatest interest to decision makers as they rarely act at a global level. For both SST and SSS, internal variability and model uncertainties dominate in all regions in the early part of the century, whilst model and/or scenario uncertainties dominate in the mid- to late part of the century (Figure 5.3 and 5.4). The contribution of internal variability to the total variance of the projections is much greater at the start of the century at a regional scale compared with at a global scale. For example, internal variability contributes up to 62% of the total variance in regional projections of SST (Figure 5.3), but just 8% at a global level (Figure 5.2c). Similarly, internal variability contributes up to 79% of the total variance in regional projections of SSS (Figure 5.4), but 16% at a global level (Figure 5.2d). This result is to be expected given the increased variability of environmental conditions at smaller spatial scales.

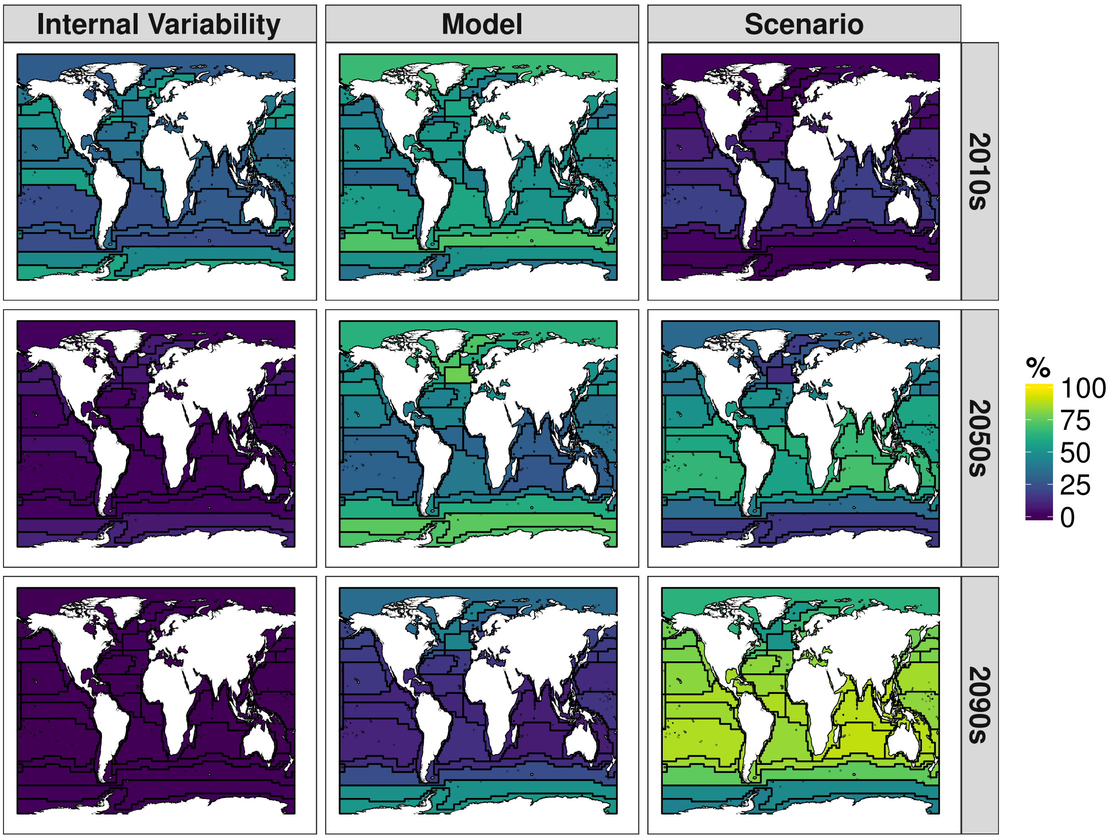


Figure 5.3: The percentage contribution of internal variability (left), model (middle), and scenario (right) uncertainty to the total variance of the projections of future regional, decadal sea surface temperature from 14 of the CMIP5 climate models under three scenarios of future greenhouse gas emissions. The 2010s (top), 2050s (middle), and 2090s (bottom) are plotted to allow for comparisons to be made between the start, middle, and end of the century. The black lines delineate the 54 biogeographic provinces described by Longhurst (2007)

For projections of SST, a strong latitudinal gradient becomes apparent in the 2050s, with model uncertainty dominating in the polar regions and scenario uncertainty dominating in tropical and temperate regions (Figure 5.3). This pattern continues to the end of the century, although the importance of model uncertainty declines in all regions despite remaining an

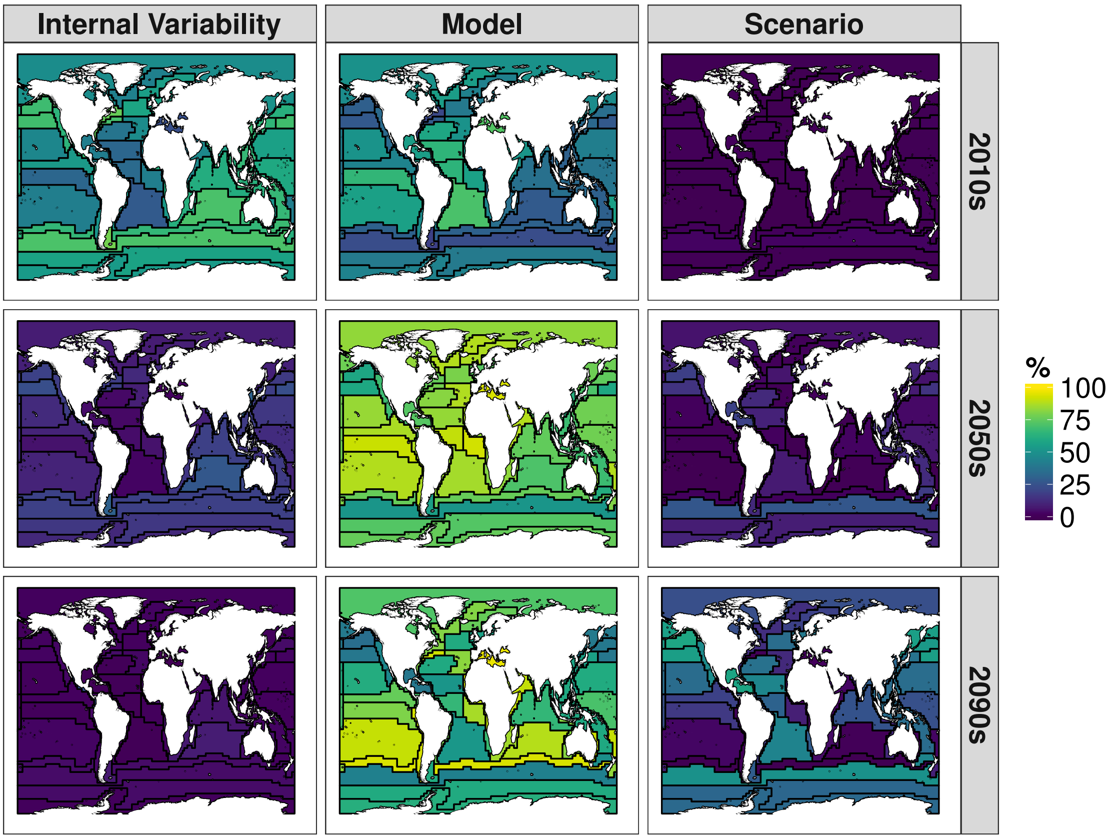


Figure 5.4: The percentage contribution of internal variability (left), model (middle), and scenario (right) uncertainty to the total variance of the projections of future regional, decadal sea surface salinity from 11 of the CMIP5 climate models under three scenarios of future greenhouse gas emissions. The 2010s (top), 2050s (middle), and 2090s (bottom) are plotted to allow for comparisons to be made between the start, middle, and end of the century. The black lines delineate the 54 biogeographic provinces described by Longhurst (2007)

important contributor to the total variance of the projections in the polar regions (Figure 5.3). Outside the polar regions, model uncertainty is greatest in the North Atlantic Drift province, with a contribution of 45% in the 2090s compared with a maximum of 27% in the surrounding temperate regions (Figure 5.3).

A latitudinal gradient is also evident in the Signal-to-Noise Ratio (SNR) of the projections of SST. The SNR represents the robustness of the projections and can be used to highlight regions in which the climate projections provide the most ’added value’ (Hawkins and Sutton, 2009). A SNR significantly exceeding one indicates a region in which the projections may be particularly useful for planning purposes, as the signal can be easily detected above the noise (Hawkins and Sutton, 2009). In the 2010s, all regions excluding the Antarctic, Austral Polar, Atlantic Arctic, and North Atlantic Drift provinces have a SNR exceeding one (Figure 5.5). The regions with the greatest SNRs for projections of SST are predominately found in the tropics, particularly in the Indian Ocean (Figure 5.5). In the 2050s, a total of 49 provinces continue to have a SNR exceeding one (Figure 5.5). For all regions, the SNR is maximised (and hence the fractional uncertainty is minimised) between the 2020s and 2050s (see Appendix

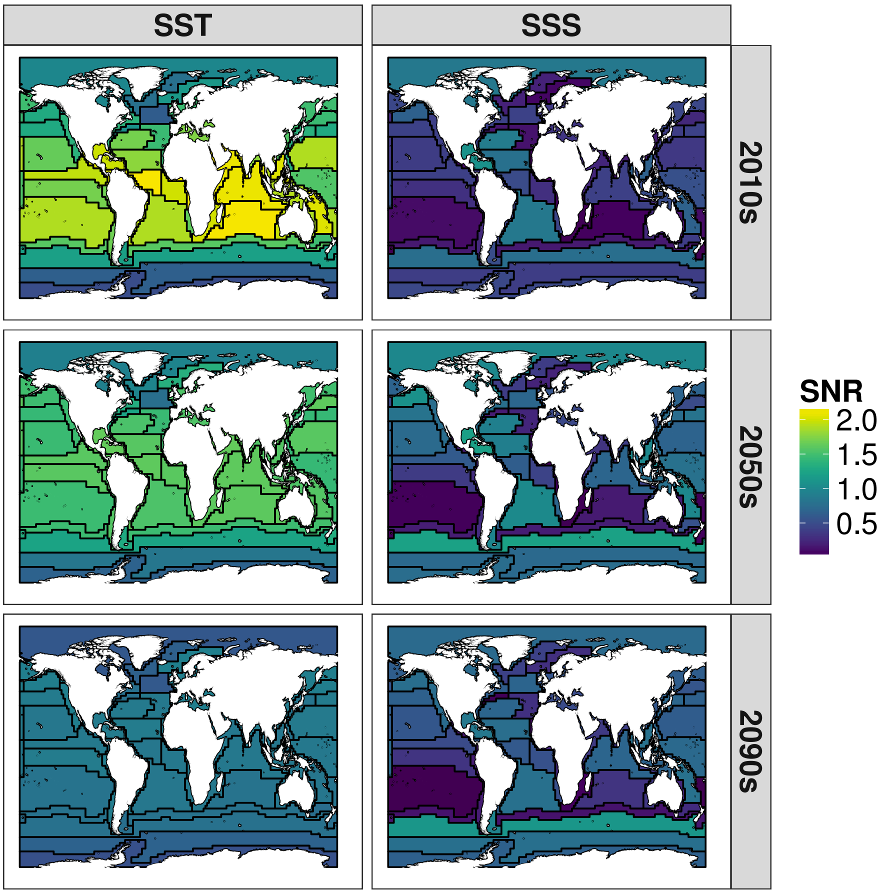


Figure 5.5: The absolute Signal-to-Noise Ratio (SNR) for projections of future regional, decadal Sea Surface Temperature (SST; left) and Sea Surface Salinity (SSS; right) (90% confidence levels). The 2010s (top), 2050s (middle), and 2090s (bottom) are plotted to allow for comparisons to be made between the start, middle, and end of the century. The black lines delineate the 54 biogeographic provinces described by Longhurst (2007). A high SNR indicates high confidence in the projections for a given region, whilst a low SNR indicates low confidence in the projections for a given region.

D), indicating a high level of confidence in the projections of SST in the first half of the century. By the 2090s, none of the 54 Longhurst regions have a SNR exceeding one (Figure 5.5). Such a decline in SNR in the later part of the century is to be expected given the exponential increase in scenario uncertainty (and the resulting increase in the fractional uncertainty as depicted in Figure 5.2e) in the projections of SST over time.

Based on the findings of Hawkins and Sutton (2009), the latitudinal gradient in the uncertainty of the projections of SST is also shared by surface air temperature, with model uncertainty dominating in the polar regions and in the North Atlantic and scenario uncertainty dominating in tropical and temperate regions by the end of the century. The SNRs associated with surface air temperatures are also lowest in the polar regions and in the North Atlantic, becoming increasingly greater towards the tropics (Hawkins and Sutton, 2009). Hawkins and Sutton (2009) suggested the polar regions may have increased model uncertainties and a lower SNR due to the large uncertainties associated with climate feedbacks at high latitudes. The authors also suggested that the SNR of the North Atlantic may be particularly low compared with other temperate regions as a result of the large uncertainties associated with the potential impacts of climate change on water circulation in the Atlantic Ocean, namely the Atlantic Meridional Overturning Circulation (AMOC) (see Cheng et al. (2013), Wang et al. (2014) and Reintges et al. (2017) for example; Hawkins and Sutton (2009)). Such uncertainties may be exacerbated by the unknown impacts of climate change on the North Atlantic Oscillation (NAO), which can have a large impact on both surface and deep water ocean conditions and is known to be a large contributor to internal variability in the CMIP climate models (Sarafanov, 2009; Deser et al., 2017).

For projections of SSS, there is no apparent latitudinal gradient in the contribution of each type of uncertainty to the total variance (Figure 5.4). Scenario uncertainty contributes very little to the total variance in both the 2010s and 2050s, especially when compared with projections of SST (Figure 5.3 and 5.4). For example, scenario uncertainty contributes a maximum of 30% to the total variance of the projections of SSS in the 2050s (Figure 5.4), but contributes up to 70% of the total variance of the projections of SST (Figure 5.3). Instead, model uncertainty dominates in all regions in the 2050s, with contributions of at least 48.5% (Figure 5.4). Although becoming more important over time, scenario uncertainty dominates in just eight of the 54 Longhurst provinces by the end of the century, whilst model uncertainty dominates in all other regions (Figure 5.4).

Model uncertainty may be particularly important in projections of SSS due to large regional biases in some of the models. For example, an analysis of the CMIP3 models highlighted regional biases of up to ±2.5 psu (Terray et al., 2012; IPCC, 2014a). Such regional biases may in part be caused by the relative lack of observations of SSS compared with SST, as well as the loose association between variations in SSS and the driving forces of precipitation, evaporation, sea ice, and river run-off (IPCC, 2014a). This loose association makes it more difficult to accurately represent the relationships between these variables in the climate models (IPCC, 2014a). Excessive simulated precipitation in some regions, particularly in the Southern Hemisphere tropics (Hwang and Frierson, 2013) and over southern Africa and the Indian Ocean (Lazenby et al., 2016), may explain the high model uncertainties in these regions in the 2050s and 2090s respectively (Figure 5.4). The eight regions in which scenario uncertainty dominates at the end of the century are largely found in the North Pacific (namely the Gulf of Alaska and Bering Sea), the Gulf of Mexico and Caribbean Sea, and the Subantarctic province (Figure 5.4). All of these regions are associated with the surface currents of the thermohaline circulation, which is partially driven by salinity. These results may therefore highlight the large uncertainties regarding how climate change may impact global water circulation in the future (Schmittner et al., 2005).

In general, the SNRs of the projections of SSS are lower and are maximised at a later time point than the SNRs of the projections of SST (Figure 5.5), indicating the climate change signal in the model is much weaker for SSS than it is for SST. For most regions, the SNRs of the projections of SSS are maximised in the 2060s and 2070s (see Appendix D and in nine regions the SNR is still increasing at the end of the century (Figure 5.5). Again, these results are similar to those of Hawkins and Sutton (2011), who found that the SNRs of the projections of future changes in precipitation were lower and were maximised at a later time point than those of surface air temperature. For projections of SSS, just seven of the 54 Longhurst provinces have a SNR exceeding one in any of the three decades shown in Figure 5.5, most of which occur in the polar and subpolar regions. The Caribbean Sea and Gulf of Mexico, as well as the South Atlantic Gyral province and Archipelagic Deep Basins province, also exhibit high SNRs (Figure 5.5). The New Zealand coastal province, Guinea coastal current province, and the South Pacific subtropical gyre province are associated with particularly low SNRs (Figure 5.5). However, the coastal provinces may have a low SNR due to the increased contribution of internal variability to the total variance of the projections at smaller spatial scales (Hawkins and Sutton, 2009; Cheung et al., 2016), whilst the low SNR of the subtropical gyres may be driven by uncertainties associated with the extent to which these gyres are likely to intensify and/or shift in the future (Pontes et al., 2016).

5.4.3 Limitations

The methods used in this research assume internal variability does not change over time and can be calculated from the residuals of the projections from a smooth fit (Hawkins and Sutton, 2009, 2011). As previously mentioned, although there is some evidence to suggest that internal variability may increase over time (see Boer (2009) for example), it is unlikely to affect the overall conclusions of this research as the effect of this increase is often negligible (Hawkins and Sutton, 2011). However, by smoothing the data with a fourth-order polynomial, we effectively remove any internal fluctuations in climate that act over longer time periods (i.e. longer than 15 to 30 years) (Deser et al., 2014). As such, we may be underestimating the contribution of internal variability to the total variance of the projections. We also assume that the models selected for inclusion in the analysis are independent and represent the full spread of all possible models (Hawkins and Sutton, 2011). Independence is unlikely given that some or all of the models will have been parameterised with the same datasets, potentially resulting in common biases (Cheung et al., 2016). The models also represent an 'ensemble of opportunity' rather than an ensemble that has been strategically selected to explore the full range of possible model structures (IPCC, 2007; Cheung et al., 2016). Our estimate of the contribution of model uncertainty may therefore also be interpreted as a lower bound of the true value

(Hawkins and Sutton, 2011). Finally, we included only three possible scenarios for future changes in radiative forcing. Although the scenarios span the range of available estimates given in recent scientific literature (van Vuuren et al., 2011), the upper and lower bounds of these estimates will change in the future in response to technological and/or political developments that impact our progress towards emissions targets. If such changes are reflected in the scenarios used in upcoming model intercomparisons, the relative contribution of scenario uncertainty to the total variance of the projections may either increase or decrease. Despite these limitations, the methods used in this research are expected to give a qualitatively robust approximation of the uncertainties in the available projections, particularly over the next few decades (Hawkins and Sutton, 2009, 2011).

5.4.4 Reducing uncertainty

Reducing the uncertainty in climate projections is of great importance to decision makers, particularly at short lead times of less than a decade or so (Hawkins and Sutton, 2009). Although very little can be done to reduce scenario uncertainties, model uncertainties may be reduced via investments in observational data (e.g. through the Global Ocean Observing System Roemmich et al. (2009)), as well as through model development (Hawkins and Sutton, 2009). In particular, improvements in the representation of global water circulation, as well as reductions in cloud and thermocline depth errors, should help to reduce the model uncertainties associated with projections of SST and SSS (Schmittner et al., 2005; Villarini and Vecchi, 2012; Hwang and Frierson, 2013; IPCC, 2014a). Growing interest in regionally tuning the climate models may also lead to reductions in model biases in the projections of SST and SSS in some regions (Mulholland et al., 2017). Such reductions in uncertainty could have a large impact on the total variance of short-term projections of both SST and SSS, but it is SSS that would likely benefit most from a reduction in model uncertainty across all lead times. The regions with the greatest potential for improvement in model uncertainty include the polar regions for projections of SST and the Southern Hemisphere tropics for projections of SSS. However, it is important to note that changes to the structure of the models in response to new information may result in an increase in the uncertainty of the projections (Knutti and Sedláček, 2013); this should not be regarded as a step backwards and should instead be considered as an increase in confidence in the models due to improved realism (Knutti and Sedláček, 2013).

Investments in observational data may also help to reduce the internal variability uncertainty in short-term projections through initialisation of the climate models (Smith et al., 2007; Hawkins and Sutton, 2009). Again, projections of SSS would likely benefit most from a reduction in internal variability uncertainty compared with SST. Both the internal variability and model uncertainties associated with projections of SSS may therefore be reduced by increasing the availability of long-term global salinity data, such as those from the Array for Real-time Geostrophic Oceanography (ARGO) network (IPCC, 2014a). For projections of both SST and SSS, the Southern Ocean may benefit greatly from a reduction in internal variability uncertainty. For projections of SST, the North Pacific and Tasman Sea would also likely benefit most from reductions in internal variability, whilst the Indian Ocean would benefit from a reduction in internal variability in the projections of SSS. However, a large part of the uncertainty associated with internal variability will likely be irreducible due to the chaotic nature of natural phenomena (Hawkins and Sutton, 2009; Knutti and Sedláček, 2013; Villarini and Vecchi, 2012).

5.5 Summary and conclusions

The aim of this research was to quantify the contribution of internal variability, model, and scenario uncertainty to the total variance of the projections of SST and SSS from over 10 different global climate models. The results showed that for both SST and SSS, internal variability and model uncertainties dominate in the early part of the century, with scenario uncertainties becoming increasingly more important in the mid- to late part of the century. Internal variability uncertainty is of particular importance in projections of SSS and at smaller spatial scales. Uncertainties in the projections of SST exhibit a strong latitudinal gradient in the mid- to late part of the century, with scenario uncertainty dominating in tropical and temperate regions and model uncertainty dominating in the polar regions. No such latitudinal gradient exists for projections of SSS, with model uncertainty dominating in almost all regions in the mid- to late part of the century. As indicated by the SNR, projections of SST are most robust in the early- to mid-part of the century, particularly in the tropics. Projections of SSS are far less robust, with the lowest SNRs found in the New Zealand and Guinea coastal provinces and the South Pacific subtropical gyre.

Importantly, uncertainties in projections of SST and SSS over the next few decades are perhaps of greatest relevance to decision makers. During this time period, it is the potentially reducible internal variability and model uncertainties that limit our ability to project changes in SST and SSS with a high degree confidence. Investments in observational data and model development could help to greatly reduce these uncertainties and subsequently increase confidence in the projections, thus ensuring the models are well-placed to support management in the future (Hawkins and Sutton, 2009; Cheung et al., 2016). Such investments would arguably be most beneficial to projections of SSS, which suffer most from internal variability and model uncertainties. Improving the availability of long-term global salinity data would not only help to reduce internal variability uncertainty, but would also help to reduce model uncertainty by allowing more detailed comparisons to be made between the model outputs and the observations; in doing so, model biases or errors may be more easily identified and dealt with (Hawkins and Sutton, 2009). However, the costs and benefits of investing in improving the climate models that are involved in CMIP needs to be carefully weighed against the costs of adaptation in the face of large uncertainties (Hawkins and Sutton, 2009). Considering the potentially wide-ranging impacts of changes in SST and SSS over the next century, as well as the high costs that will undoubtedly be associated with adapting to these changes, it is likely that the benefits of investing in reducing the uncertainty in the projections of SST and SSS will far outweigh the costs. Conducting a cost-benefit analysis should therefore be of high priority for future research to ensure investments in model improvements can be targeted where the potential gains are greatest (Hawkins and Sutton, 2009).

Chapter 6

Visualising uncertainty in multi-model ensembles

6.1 Abstract

Multi-Model Ensembles (MMEs) are increasingly being used to better understand how our environment is likely to change in the future. However, predicting the future is complicated and various types of models can make very different predictions. In the past, a lack of effective communication of such variable model outputs, both to decision makers and the general public, has been blamed for ineffective management decisions. To help combat this issue, we conducted an in-depth online survey aimed at identifying the most effective methods for visually communicating the outputs of 15 state-of-the-art climate models developed during phase five of the Coupled Intercomparison Model Project. We measure the accuracy, confidence, and ease with which the survey participants were able to interpret 10 visualisations, all of which depict the same data in slightly different ways, as well as their subjective preferences for each visualisation. We use the results of the survey to: (1) rank each of the visualisations based on their performance; (2) discuss possible reasons for the poor performance of certain visualisation types; and (3) discuss the effects of the demographics of the participants on the performance of each visualisation. Overall, we hope that the results can be used to help generate guidelines for visually communicating the outputs of MMEs across a wide range of research areas. These guidelines can then be used to target visualisations at specific audiences to maximise their impact, whilst also minimising the potential for misinterpretations. This in turn will help to increase the societal impact of the models and ensure they are well-placed to support management in the future.

6.2 Introduction

Understanding the likelihood of alternative future states is a key challenge for managing natural systems in the face of global change. A wide range of environmental models have been used to help develop management solutions, but the differing structures and assumptions of individual models can lead to wildly different predictions (Hansen and Hoffman, 2011; Huang et al., 2018). An increasingly popular way to deal with the shortcomings of individual models is to combine the outputs of multiple structurally different models that have been run under a common set of scenarios for the future into a Multi-Model Ensemble (MME). MMEs have been successfully used to increase the skill and reliability of model predictions in a wide range of research areas (Tebaldi and Knutti, 2007), including climate science (see Giorgi and Mearns (2003) and Palmer et al. (2005) for example) and ecosystem modelling (see Dormann et al. (2008) and Spence et al. (2018) for example). However, these increases in skill and reliability may come at the cost of greater uncertainties in the outputs as, for example, different models within the ensemble may give contrasting predictions for the future (Hansen and Hoffman, 2011). In the past, a lack of effective communication of such variable model outputs, both to decision makers and the general public, has been blamed for ineffective management decisions (Janssen et al., 2005). This in turn has contributed to public distrust of scientific evidence, particularly in regards to climate science (Frewer, 2004). Improving the communication of uncertainties to non-specialist audiences is therefore vital to ensuring environmental models continue to make a significant contribution to the decision-making process.

Previous research into the successful communication of the uncertainties associated with environmental models has largely been focused on written and verbal forms of communication (see Patt and Schrag (2003), Patt and Dessai (2005), Morgan (2009) and Mastrandrea et al. (2010) for example), whilst comparatively little research has been conducted to identify the most effective methods for visually communicating these uncertainties (Spiegelhalter and Riesch, 2011). As a result, many of the techniques used for visualising the outputs of environmental models ignore the presence of uncertainties or are used to depict only one source of uncertainty at a time (MacEachren et al., 2005; Brodlie et al., 2012). This is particularly problematic when attempting to communicate the outputs of MMEs, which typically require a visual representation of changes in both model and scenario uncertainties over time. Whilst animated and interactive visualisations could be used to communicate multiple uncertainties at the same time, these methods are not appropriate for media requiring static images, and interactive visualisations may require a greater level of skill or expertise to use than static or animated visualisations (Spiegelhalter and Riesch, 2011). By focusing on how best to communicate the outputs of MMEs using static visualisations, we may be able to improve engagement and trust across a broader cross-section of society than would be possible using animated or interactive visualisations.

Examples of static visualisations that are often used to communicate uncertainty in environmental modelling include line, dot, and box plots. These visualisation methods typically depict a summary of the data, such as an average, and an estimate of the uncertainty through the use of uncertainty bands (or envelopes) or error bars. Although dot and box plots have previously been shown to be effective at communicating uncertain snowfall forecasts (Ibrekk and Morgan, 1987), relatively little is known about the ability of the general public to interpret these visualisations. It is also possible that more modern visualisation methods, such as infographics and cascade plots (see Wilby and Dessai (2010) and Hawkins (2014) for example), may outperform dot and box plots when used to communicate the outputs of MMEs to decision makers and/or the general public. Traditional methods of visualisation that are less frequently used in environmental modelling, such as radar (or spider) and heat plots, may also perform well when adapted to depict the outputs of MMEs. However, there is a lack of empirical research comparing the performance of these visualisation methods when used to communicate the outputs of MMEs to different groups of people, making it difficult for researchers to maximise the impact of these model ensembles.

Various methods may be used to assess the performance of different methods of visualising uncertainty (see Kinkeldey et al. (2014) for a review). Typically, the effectiveness of a particular visualisation is determined by measuring the accuracy and/or self-assessed confidence with which a set of individuals are able to interpret the visualisation (Kinkeldey et al., 2014). User preferences and subjective measures of ease of use are also often used to compare the performance of different visualisation methods (Kinkeldey et al., 2014). However, we are not aware of any research that has combined all of these measures of performance to determine the most effective methods for visually communicating the outputs of MMEs. In this study, we aim to fill this research gap by conducting an in-depth online survey that measures the accuracy, confidence, and ease with which the participants are able to interpret 10 different visualisations (see Figure 6.1), all of which depict the same set of data from a state-of-the-art climate MME (see `cmip.llnl.gov/cmip5/`), as well as their subjective preferences for each of the visualisations. As the effectiveness of each visualisation method may depend on factors such as the numeracy and scientific literacy of the audience (Spiegelhalter and Riesch, 2011), we also take into account the education level, background, and expertise of the participants when determining the performance of each visualisation. Overall, we hope that the results of this research can be used to help generate guidelines for visually communicating the outputs of MMEs across a wide range of different research areas. These guidelines can then be used to target visualisations at specific audiences to maximise their impact, whilst also minimising the potential for misinterpretations. This in turn will help to increase the societal impact of the models and ensure they are well-placed to support management in the future.

6.3 Methods

To better understand the effectiveness of different methods of visualising Multi-Model Ensembles (MMEs), we developed a survey using the Qualtrics online survey software (qualtrics.com).

6.3.1 The data

The survey was focused on projections of surface air temperature from the MME produced in phase five of the Coupled Model Intercomparison Project (CMIP5) (cmip.llnl.gov/cmip5/). This model ensemble was also used by the Intergovernmental Panel on Climate Change (IPCC) for the Fifth Assessment Report (AR5) (IPCC, 2014a). Each model in the ensemble was used to simulate historical surface air temperatures between 1850 and 2005 and then run under three different greenhouse gas emissions scenarios for the future, known as Representative Concentration Pathways (RCPs), between 2006 and 2100. The three RCPs, referred to as RCP 2.6, RCP 4.5, and RCP 8.5, span the range of currently available estimates for the predicted level of radiative forcing that is expected to occur by the end of the century (van Vuuren et al., 2011). RCP 2.6 represents a scenario in which radiative forcing increases from <2W.m^2 at the start of the century to ~3W.m^2 (~490ppm CO_2 eq) in the coming decades, followed by a decline to 2.6W.m^2 by the end of the century (van Vuuren et al., 2011). RCP 4.5 represents an intermediate pathway, with an increase in radiative forcing to 4.5W.m^2 (~650 ppm CO_2 eq) by 2100 (van Vuuren et al., 2011). Finally, RCP 8.5 represents a scenario in which radiative forcing substantially increases to 8.5W.m^2 (~1370 ppm CO_2 eq) by 2100 (van Vuuren et al., 2011).

For the purposes of this research, we selected models from the CMIP5 MME based on the availability of the appropriate data with consistent global coverage spanning the years between 1850 and 2099. This meant that we were able to use the projections from 15 different models (see Appendix E for further details of the selected models). We extracted the annual global mean surface air temperature projections (~1.25 - 2m above ground) between 2000 and 2099 from the IPCC AR5 online database (dkrz.de) for each model and RCP scenario. It is important to note that some of the models were run under each scenario multiple times, with each run producing slightly different model outputs. However, we chose to use only one set of outputs per model to ensure all models were treated equally, selecting the first set of outputs in which all of the required scenarios were available (Hawkins and Sutton, 2009, 2011). We also extracted the annual global mean surface air temperature projections for each model between 1850 and 1900 and took the mean as a pre-industrialisation reference temperature. The projected change in global mean surface air temperature (referred to simply as 'temperature' from here on) expected to occur in each year was then quantified by comparing the 2000 to 2099 model outputs with the pre-industrialisation reference temperature.

6.3.2 The visualisations

The projected temperature change data was used to create ten visualisations that depict the same data in different ways. The visualisations included two different versions of a line plot (line1 and line2), two different versions of a dot plot (dot1 and dot2), two different versions of a box plot (box1 and box2), a radar plot, a cascade plot, a heat plot, and an infographic (Figure 6.1; see Appendix F for larger versions of each visualisation and their accompanying legends). These visualisation methods were chosen to represent plots that are frequently used in the scientific literature and in the media, as well as some that are more unusual and may be less familiar to a wider audience. Please note that some of the visualisations were based on the work of Prof. Ed Hawkins and Dr. Rowan Sutton from the National Centre for Atmospheric Sciences (NCAS) (see `climate-lab-book.ac.uk` for more information).

Some of the selected visualisation methods allowed for the depiction of projected temperature changes in every decade between 2000 and 2099, whilst other methods were limited to depicting the data in a smaller number of decades. To guarantee that comparisons could be made across all visualisation types, we ensured that each visualisation depicted the data in at least three decades: the 2010s, 2050s, and 2090s. The culturally-ingrained traffic light colour system (i.e. red, yellow, and green) was used in all ten visualisations to maintain consistency. However, the selected colour scheme may make it more difficult for those who experience deuteranopia or protanopia (red-green colour blindness) to distinguish between the different colours. Because of this, we added the option for the participants to request visualisations with a more suitable colour palette if required, although none of the participants selected this option.

6.3.3 The survey

In the first section of the survey, the participants were asked to provide some basic information about themselves, including their age, gender, location, level of education, and expertise in working with environmental models and/or their outputs. The participants were also asked

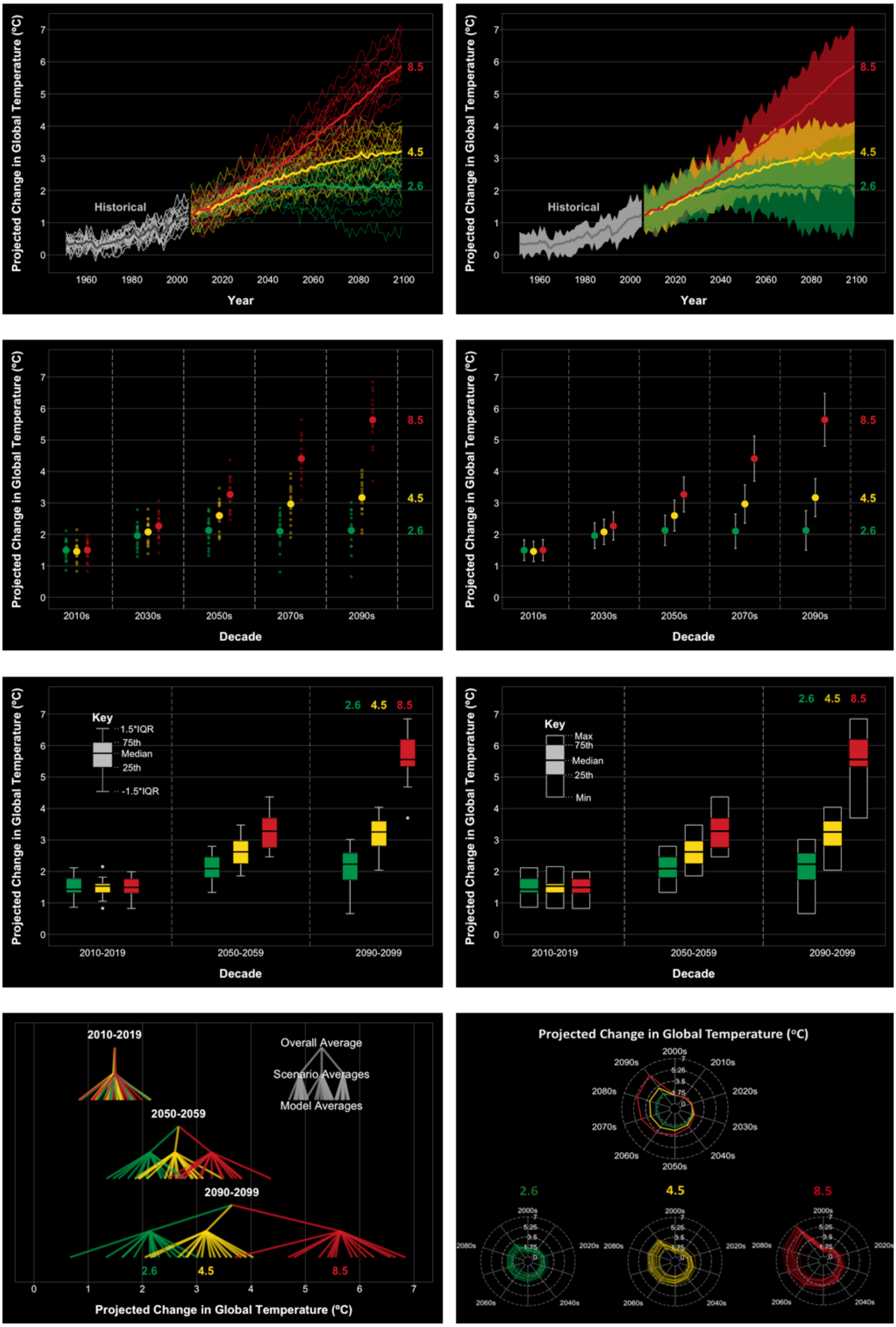


Figure continued overleaf.

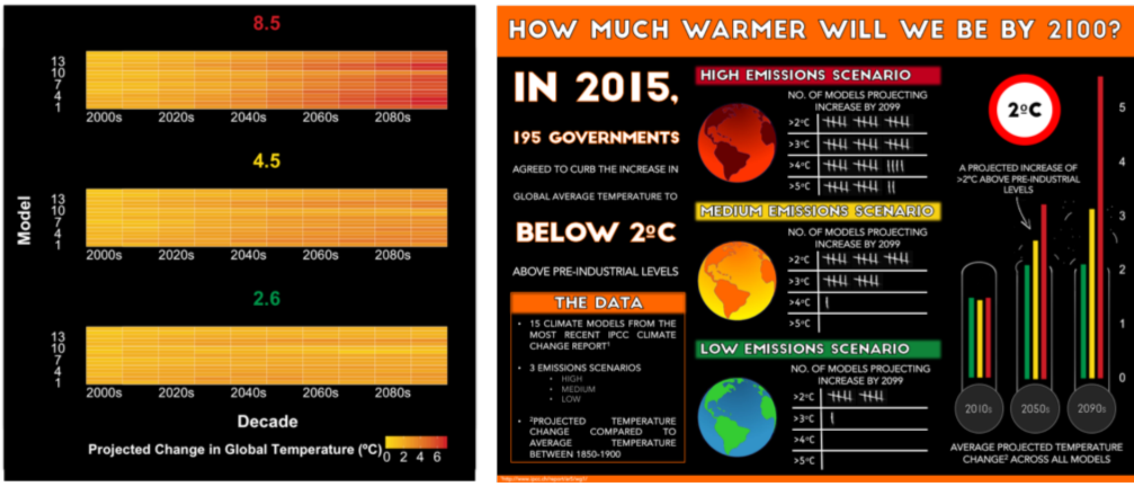


Figure 6.1: The ten visualisations that were included in the survey. From left to right and top to bottom: the line1, line2, dot1, dot2, box1, box2, cascade, radar, heat, and infographic plots. See Appendix F for larger versions of each visualisation and their accompanying legends.

about their background in terms of whether they considered themselves to be a member of the general public, a scientist, or a decision maker and/or environmental manager.

In the second section of the survey, the participants were shown a randomly selected visualisation (referred to as visualisation A) and asked if they had encountered a similar visualisation prior to completing the survey. The participants were then asked to estimate the average global temperature change projected to occur by the end of a randomly selected decade (2010s, 2050s, or 2090s) under a randomly selected scenario (RCP 2.6, RCP 4.5, or RCP 8.5). We used the term ’average’ instead of ’mean’ to ensure the survey remained accessible to a wider audience, although we accept that some of the participants may have provided the median when shown visualisations such as the box plots. However, this issue is relatively unimportant given that the largest difference between the mean and median in every combination of decade and scenario was 0.1°C. Each participant was then required to comment on their confidence in the answer they provided, as well as the ease with which they were able to identify the answer, on a Likert scale (Likert, 1932) (disagree, somewhat disagree, neutral, somewhat agree, agree). The participants were also asked to give a qualitative description of the level of uncertainty in a second randomly selected decade and scenario on a Likert scale (very low, low, moderate, high, very high), before being asked to estimate the minimum and maximum global temperature change projected to occur by the end of the same decade and scenario (if possible using the visualisation provided). In the last part of section two, the survey participants were again asked to comment on their confidence in the answer they provided and the ease with which they were able to identify the answer.

In the third section of the survey, the participants were shown a second randomly selected visualisation (referred to as visualisation B) alongside visualisation A and asked to choose which of the two visualisations they preferred across five different categories: the ability to view changes in temperature over time, the ability to view changes in uncertainty over time, the ability to retrieve specific values (such as the mean, minimum, and maximum), visual appeal, and overall ease of understanding. The participants were given the option of preferring visualisation A, preferring visualisation B, or having no preference for either A or B. In cases where the participant showed no preference for visualisation A or B, they were given the option of selecting ’both the same’ or ’neither’.

Together, the second and third sections of the survey formed a single ’block’. The survey was designed so that the participants could decide how many of these blocks they wished to complete. Each individual was given the option to exit the survey at the end of a block or to continue the survey by starting a new block containing a different set of randomly selected visualisations. The participants could complete a maximum of five blocks and the visualisations, scenarios, and decades were randomised for each participant using JavaScript random number generation.

The survey received full ethical approval from the University of Sheffield’s Department of Animal and Plant Sciences, in accordance with the University of Sheffield’s Research Ethics Approval Procedure. The survey was distributed internally at the University of Sheffield and publicly through numerous channels including Twitter, various mailing lists, and personal contacts. A total of 380 individuals participated in the survey.

6.3.4 Demographics

The majority of the participants were aged between 18 and 34 (n = 258), with relatively few over the age of 55 (n = 26) (Figure 6.2). There were slightly more female participants than males (n = 202 and n = 174 respectively) (Figure 6.2) and the majority of the participants were based in the United Kingdom (n = 306), although individuals from a total of 21 countries participated in the survey (not shown). 318 participants held a university-level qualification, whilst 60 participants held GCSE, A-Level, or vocational qualifications (Figure 6.2). 18 participants considered themselves to be a decision maker or environmental manager and 190 considered themselves to be a scientist (Figure 6.2). 131 participants had previously worked with environmental models and/or their outputs and 49 of these participants had five or more years of experience (Figure 6.2). 285 participants considered themselves to have little to no expertise in working with environmental models and/or their outputs, whilst 44 participants considered themselves to be an expert (Figure 6.2).

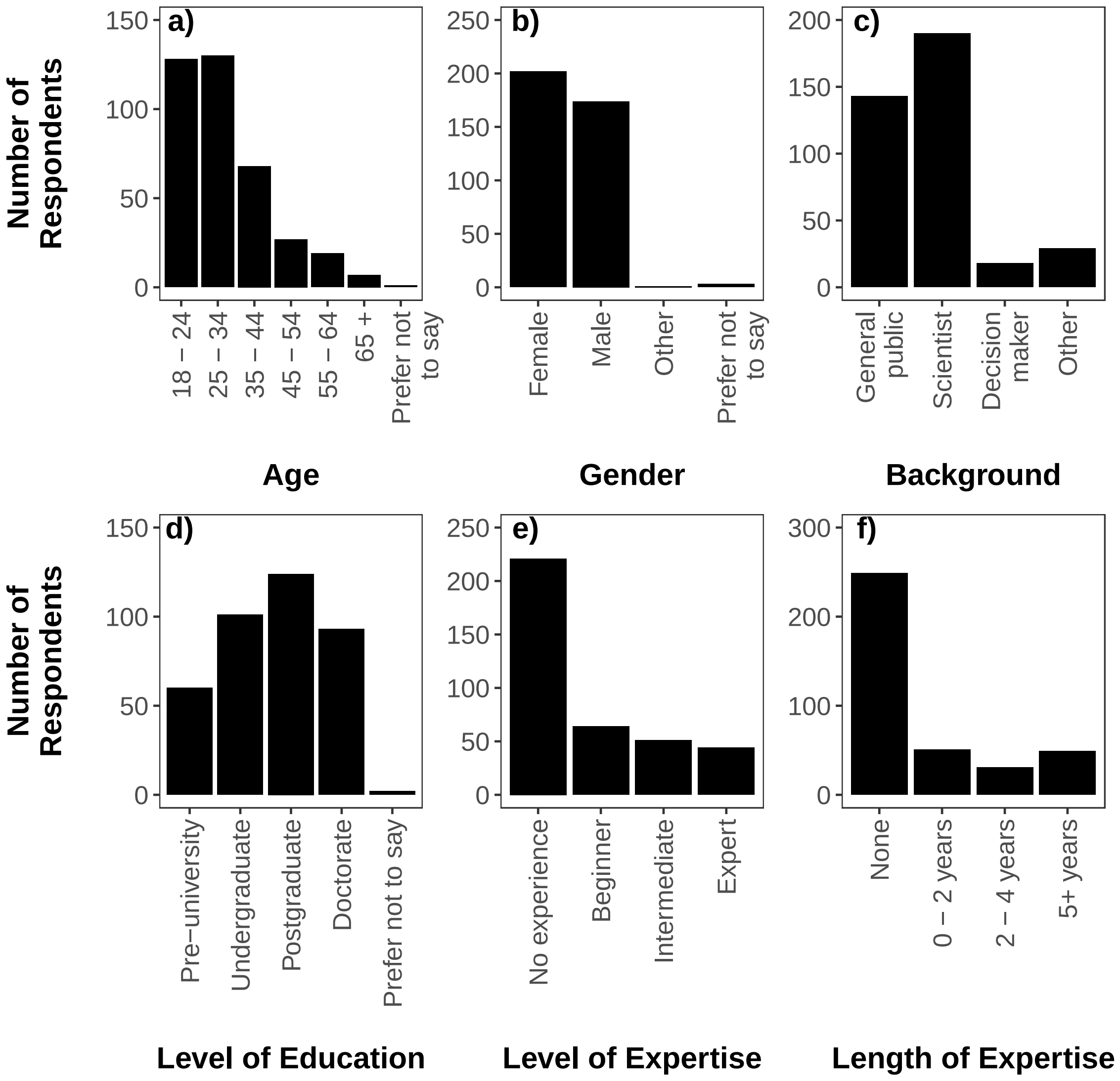


Figure 6.2: The demographics of the survey participants (n = 380), including a) age, b) gender, c) background, d) level of education, e) level of expertise in working with environmental models and/or their outputs, and f) length of expertise in working with environmental models and/or their outputs.

6.3.5 Statistical analysis

Various methods of statistical analysis were used to better understand the accuracy, confidence, and ease with which the survey participants were able to interpret the visualisations, as well as their preferences for each visualisation across a number of different categories. The participants' background, level of education, and expertise in working with environmental models and/or their outputs were included in the statistical analyses to determine whether specific groups of people were better able to interpret the visualisations or whether they preferred different types of visualisations.

To perform the statistical analyses, we removed participants from the data that selected 'prefer not to say' for level of education (n = 2) due to small sample sizes and because these participants could not be used to better understand the preferences of different groups of people.

We also removed the participants that selected 'other' for background (n = 29) as the majority of those who selected this option could have been placed either in the general public, scientist, or decision maker and/or environmental manager groups based on the corresponding text part of their answer. However, we chose not to assign a background to these individuals to prevent introducing personal subjectivity into the analyses. Furthermore, although the participants were given the option of selecting more than one category for background, we extracted only the 'highest' category for each participant assuming an ordering of general public < scientist < decision maker and/or environmental manager. This ordering was selected as decision makers and environmental managers can also be both scientists and members of the general public at the same time, whilst scientists that do not consider themselves to be a decision maker or environmental manager can also be considered as a member of the general public. Finally, we combined secondary (n = 2), post-secondary (n = 51), and vocational (n = 2) levels of education into a 'pre-university' group due to small sample sizes. All statistical analyses were performed using the R statistical computing software (R Core Team, 2018).

Accuracy

The accuracy with which the participants were able to determine the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade was analysed using Generalised Linear Mixed Models (GLMMs) (Breslow and Clayton, 1993). GLMMs can be used to better understand the relationship between a response variable (e.g. accuracy) and a set of predictor variables (e.g. level of education, visualisation type etc.) when the error structure is not gaussian (Bolker et al., 2009). GLMMs also allow both fixed and random effects to be included in the model, thus enabling the quantification of the variation in the accuracy with which individual participants were able to identify the mean, minimum, and maximum projected temperature change across multiple blocks of questions (Bolker et al., 2009). Using the notation of IDRE (2016), the general form of a GLMM can be written in matrix notation as:

$$\log(\mathbb{E}(\mathbf{y})) = X\beta + Zu \tag{6.1}$$

where $\mathbf{y}$ represents the response variable, X represents the model matrix (including p predictor variables), β represents the regression coefficients of the fixed effects, Z represents the design matrix for q random effects, and u represents the regression coefficients associated with the random effects (IDRE, 2016). In our case, the response variables were the absolute difference (x10) between the participants' estimates of the mean, minimum, and maximum projected temperature change and the 'true' values given by the climate models. The fixed effects included visualisation type, decade, scenario, background, level of education, and length of expertise in working with environmental models and/or their outputs. Participant ID was included as a random effect to take into account the fact that the participants were able to answer the same set of questions in up to five blocks.

To fit GLMMs to the survey data, we first checked whether the participants' estimates of the maximum temperature change projected to occur in a given scenario and decade were greater than their estimates of the minimum temperature change projected to occur in the same scenario and decade. Where this was not the case, we swapped the estimates of the minimum and maximum projected temperature change. We then quantified the absolute difference between the participants' estimates of the mean, minimum, and maximum projected temperature change with the true values given by the climate models. As previously mentioned, we asked the participants to estimate the 'average' projected temperature change in a given decade and scenario to ensure the survey was accessible to a wide audience, but it should be noted that the estimates provided by the participants were compared with the mean of the MME (not the median or mode). The absolute difference was then multiplied by 10 to convert the data into positive integers, thus allowing us to fit quasipoisson GLMMs to account for the overdispersion in the data (Ver Hoef and Boveng, 2007). The GLMMs were fit with the `glmmPQL()` function in the `MASS` R package (Venables and Ripley, 2002). Please note that as the participants were given the option of selecting 'not applicable' when they were unable to estimate the minimum and maximum projected temperature change using the visualisation provided, there were slightly fewer estimates with which to analyse the accuracy of the participants when asked to estimate the minimum and maximum (343 unique participants with a total of 965 individual responses) compared with the mean (345 unique participants with a total of 1036 individual responses) (see Appendix G).

One of the main disadvantages of using the `glmmPQL()` function is that this method computes penalised quasi-likelihoods instead of true likelihoods (Bolker et al., 2009). Because of this, likelihood ratio tests could not be used to test the significance of including participant ID as a random effect. Furthermore, Akaike's Information Criterion (AIC) (Akaike, 1973), which is widely used as a measure of model fit, could not be used for model selection purposes. Quasi-AIC (QAIC) can be used as measure of model fit instead, but this is often frowned upon by statisticians who believe that quasi-methods should not report likelihoods at all (Bolker et al., 2009). However, a brief comparison of the QAICs of the full models (including all possible predictor variables) and the best-fitting models without the random effect indicated that there was little difference in the goodness-of-fit. For example, when analysing the accuracy with which the participants were able to estimate the mean projected temperature change in a given scenario and decade, the QAIC of the best-fitting model was 1484.8, whilst the QAIC of the full model was 1488.4. We also tested models that included the interactions between all predictor variables and visualisation type, but the complexity of the model was too great and there was little evidence to suggest the model fit improved. Because of this, we decided to continue with the full models (minus the interactions) without relying on QAIC to search for the best-fitting models. The final model fits were checked by plotting the fitted values against the standardised residuals.

To aid the interpretation of the outputs of the GLMMs, the reference levels used for each predictor variable were chosen to represent a ’typical’ (or average) participant. In this case, the typical participant was a postgraduate scientist with no experience in working with environmental models and/or their outputs. The typical participant was asked to estimate the mean, minimum, and maximum temperature change projected to occur in the 2050s in scenario 4.5 using the box1 plot. As there were no typical visualisations in the survey, the box1 plot was selected as it tended to fall in the middle of the visualisations when ranked based on the mean absolute difference (x10) between the participants’ estimates of the mean, minimum, and maximum projected temperature change and the true value given by the climate models. It is important to note that because the results of the analysis are presented relative to the ’typical’ response, they apply only when all other predictor variables are fixed at the reference levels.

The predictions of the GLMMs are presented as ratios between the inaccuracies associated with the typical response and the inaccuracies associated with all other levels of the predictor variables (referred to as Absolute Difference (AD) ratios from here on), thereby allowing an assessment of the variability in the differences in accuracy across all levels of the predictor variables, although the raw model predictions are also given in Appendix H. An AD ratio of greater than one suggests the participants in the group in question tended to be more accurate than those in the reference group, whilst an AD ratio of less than one suggests the participants in the group in question tended to be less accurate than those in the reference group. As there are currently no widely accepted methods for incorporating the uncertainty in the random effects at present (Bates et al., 2014), the standard errors (and hence 95% confidence intervals) of the predictions of the GLMMs should be treated as lower bounds of the uncertainty. Fortunately, this issue is relatively unimportant in this research given that we are largely interested in the mean accuracy of different groups of participants, rather than the variability in the accuracy of individual participants.

Confidence and ease

Ordinal Logistic Regression (OLR) was used to better understand the confidence with which the participants were able to identify the mean or the minimum and maximum temperature change projected to occur in a given decade and scenario, as well as the ease with which they were able to identify the answers. To do this, we applied Mixed Proportional Odds Models (MPOM; also known as ordered logit models) to the Likert scale data using the `clmm2()` function in the `ordinal` R package (Christensen, 2018). MPOMs are specifically designed to handle ordinal response variables (Fullerton and Xu, 2012) and allow both fixed and random effects to be included in the model. MPOMs typically estimate the cumulative probability of being in level j of the Likert scale or less (Schmidt, 2012). Using the notation of Schmidt (2012), the general form of a MPOM can be written as:

$$\text{logit}[P(Y_i \leq j)] = \gamma_j - (Z_{t[i]}u_t + X_i\beta) \tag{6.2}$$

for $j = 1, \ldots, J-1$, where γ_j represents the threshold for level j in the Likert scale, J represents the total number of levels in the Likert scale, u_t represents the regression coefficients associated with the random effects (which are assumed to be normally distributed and centred on zero), $Z_{t[i]}$ represents the design matrix of the random effects for the observations i nested in participant t, X represents the model matrix (including p predictor variables and the intercept), and β represents the regression coefficients of the fixed effects (Hedeker and Gibbons, 2006; Schmidt, 2012). In our case, the response variables were the confidence and ease with which the participants were able to identify the mean or the minimum and maximum temperature change projected to occur in a given decade and scenario. The fixed and random effects were the same as those given in the previous section with the addition of previous encounters as a fixed effect.

Proportional odds models assume that the relationships between all pairs of levels in the Likert scale are the same, i.e. the coefficients that describe the relationship between 'disagree' and all higher levels in the Likert scale are the same as those that describe the relationship between 'somewhat disagree' and all higher levels in the Likert scale (Aiello and McFarland, 2014; Momeni et al., 2018). If the assumption of proportional odds is met, only one set of coefficients must be estimated for each of the predictor variables (Liu, 2015). However, if the assumption of proportional odds is not met then multiple sets of coefficients must be estimated to describe the relationship between each pair of levels (Liu, 2015). As the assumption of proportional odds is rarely met using real-world data (Aiello and McFarland, 2014), we used the `nominal_test()` function in the `ordinal` R package (Christensen, 2018) to perform a Likelihood Ratio Test (LRT) of the proportional odds assumption for both the confidence and ease Likert scale data individually. As the `nominal_test()` function can only be used on a model that does not include random effects, we fit (non-mixed) proportional odds models to the confidence and ease data using the `clm()` function in the `ordinal` R package (Christensen, 2018) before applying the `nominal_test()` function.

The LRTs indicated that background, level of education, and/or time of expertise did not meet the assumption of proportional odds (p < 0.05) for at least one of the measures of confidence and ease (Table 6.1). We therefore fit Mixed Partial Proportional Odds Models (MPPOMs) to the confidence and ease Likert scale data, treating the predictor variables that did not meet the assumption of proportional odds as nominal effects. The general form of a MPPOM may be written as:

$$\text{logit}[P(Y_i \leq j)] = \gamma_j - (Z_{t[i]}u_t + X_i\beta + v_i\alpha_j) \tag{6.3}$$

where v_i represents the observations i nested in participant t for the h predictor variables that do not meet the assumption of proportional odds, and α_j represents the regression coefficients of the h predictor variables that do not meet the assumption of proportional odds (Hedeker and Gibbons, 2006).

As previously mentioned, the participants were given the option of selecting 'not applicable' when they were unable to estimate the minimum and maximum projected temperature change using the visualisation provided (see Appendix G). When this option was selected, confidence and ease were also set to 'not applicable' and therefore there were fewer observations of confidence and ease when the participants were asked to estimate the minimum and maximum projected temperature change compared with the mean.

As there are no automated methods of model selection available for MPPOMs, we fit the full models (minus the interactions) to both the confidence and ease data (with the corresponding nominal effects) instead of searching for the best-fitting models. The condition number of Hessian can be used to identify whether the model is ill defined (Christensen, 2018), with values of over 10^4 or 10^6 indicating potential problems with optimisation, unidentifiable parameters, and a need to simplify the models (Christensen, 2015). The condition number of Hessian was below 7602 for all of the MPPOMs used in this research, suggesting the models were appropriately defined.

Again, the reference levels used for each of the predictor variables were chosen to represent a 'typical' (or average) participant as in the previous section. The regression coefficients are given for the ordinal and nominal effects separately as the ordinal effects are presented as regression coefficients (β), whilst the nominal predictor variables are presented as threshold coefficients (γ_j) (Christensen, 2015). Regression coefficients above one indicate that the participants in the group in question were more likely to select one of the higher categories in the Likert scale (e.g. 'somewhat agree' or 'agree') than the reference group, whilst a regression coefficient of less than one indicates the participants in the group in question were less likely to select one of the higher Likert scale categories than the reference group (Christensen,

Table 6.1: The Likelihood Ratio Test (LRT) statistics that were used to determine which of the predictor variables met the assumption of proportional odds when they were used to analyse the confidence and ease with which the participants were able to estimate the mean or the minimum and maximum temperature change projected to occur in a given scenario and decade. A p-value of less than 0.05 indicates the assumption was not met and the predictor variable should be included in the mixed partial proportional odds models as a nominal effect. *p < 0.05, **p < 0.01, ***p < 0.001.

Predictor variable	**Confidence (mean)**		**Confidence (min and max)**		**Ease (mean)**		**Ease (min and max)**	
	LRT	**p-value**	**LRT**	**p-value**	**LRT**	**p-value**	**LRT**	**p-value**
Visualisation	25.62	0.54	26.85	0.47	15.05	0.97	34.03	0.17
Decade	5.84	0.44	1.72	0.94	3.28	0.77	2.73	0.84
Scenario	7.01	0.32	3.74	0.71	7.43	0.28	7.40	0.29
Previous encounter	4.50	0.61	6.48	0.37	8.21	0.22	0.71	0.99
Background	12.84	0.05*	6.57	0.36	3.31	0.77	5.56	0.47
Education	17.25	0.04*	21.07	0.01*	9.52	0.39	17.53	0.04*
Expertise (time)	19.89	0.02*	17.28	0.04*	20.52	0.01*	18.49	0.03*

2015). The 95% confidence intervals of the regression coefficients should again be treated as underestimations as the random effects cannot incorporated in this measure of uncertainty at present. Furthermore, it was not possible for us to estimate the standard errors (and hence 95% confidence intervals) of the predictions of the MPPOMs and therefore the predictions are given without a measure of uncertainty.

Preference

Bradley-Terry models (Bradley and Terry, 1952) were used to rank the visualisations based on the pairwise preference comparisons in the survey. Bradley-Terry models are probability models that assume that the odds that visualisation i is preferred over visualisation j $(i, j \in \{1, \dots, K\})$ is α_i/α_j, where α_i and α_j represent the score (or 'ability') of the visualisations (Turner and Firth, 2012). Bradley-Terry models may also be written in the form:

$$\text{logit}[P(i \text{ beats } j)] = \lambda_i - \lambda_j \qquad (6.4)$$

where $\lambda_i = \log \alpha_i$ for all i (Turner and Firth, 2012). Assuming all pairwise comparisons are independent, the parameters $\{\lambda_i\}$ may be estimated using maximum likelihood (Turner and Firth, 2012).

We applied Bradley-Terry models to the survey preference data using the `BTm()` function in the `BradleyTerry2` R package (Turner and Firth, 2012). A single Bradley-Terry model was fit to each of the preference categories (i.e. ability to view changes in temperature and uncertainty over time, visual appeal etc.) using bias-reduced maximum likelihood (Turner and Firth, 2012). Importantly, we treated cases where participants showed no preference for either of the two visualisations given to them (i.e. the participant selected 'both the same' or 'neither') as a half win and half loss for both of the visualisations as described in Turner and Firth (2012). Again, visualisation type, decade, scenario, background, level of education, and length of expertise were included in the BT models as predictor variables. However, the participants that considered themselves to be decision makers and/or environmental managers were removed from the main part of the analysis as the small sample size of this group resulted in issues with the model fit. Despite this, we were successfully able to fit individual BT models to the survey data provided by members of the general public, scientists, and decision makers and/or environmental managers separately in order to better understand the visualisation preferences of those with different backgrounds (see Appendix I).

The best-fitting model for each preference category was determined by comparing the AIC of the null model with models that included each one of the predictor variables separately. If the AIC of any one of the models that included one of the predictor variables was lower than the AIC of the null model, the model with the lowest AIC was re-fit to the survey data with a second predictor variable to identify whether the AIC could be further improved. This process continued until the AIC did not improve when increased numbers of predictor variables were included in the model. In the main part of the analysis, background or level of education were the only predictor variables that were included in the models with the lowest AIC scores for all of the preference categories. When the survey data was divided into three different datasets based on the background of the participants, the best-fitting models were the null models for all preference categories (see Appendix I). We chose to use the best-fitting models for each preference category instead of the full models as the AICs of the best-fitting models were much lower than the full models. For example, the AIC of the best-fitting model for visual appeal was 1248.8, whilst the AIC of the full model was 1296.3. A comparison of the best-fitting and full models for each preference category also indicated that the parameter estimates were broadly similar but that the standard errors of the best-fitting models were much smaller than the full models. Again, the final model fit was checked by plotting the fitted values against the residuals.

The predicted 'ability' of each visualisation (referred to as 'preference' from this point forth) was extracted from the best-fitting models using the `BTabilities()` function in the `BradleyTerry2` R package (Turner and Firth, 2012). A greater preference score indicates the visualisation was preferred more often than a visualisation with a lower preference score. The `qvcalc()` function from the `qvcalc` R package (Firth, 2017) was used to estimate the 'quasi standard errors' of the predicted preference scores for each visualisation. The quasi standard errors were then used to determine 95% 'comparison' intervals, which can be interpreted as if the estimates of visualisation preference were independent, thus allowing comparisons to be made across all visualisations rather than comparisons with only the reference (Turner and Firth, 2012). However, the `qvcalc()` function cannot be used to estimate the quasi standard errors of the preference scores of the visualisations at different levels of the predictor variables (e.g. between scientists and the general public) and therefore 95% confidence intervals, which are based on (non-quasi) standard errors, were used instead. Fortunately, the distinction between 95% comparison intervals and 95% confidence intervals is relatively unimportant given that the main aim of this research is to compare the general trends in visualisation preferences between scientists and the general public and between those with different levels of education, rather than to quantify the exact value of the comparison intervals. The predictions of the BT models are presented relative to the box1 plot, which was selected as the reference level in accordance with the previous sections. By convention, the reference level is given a preference score of zero in the BT models, but not in the predictions of the model. Nevertheless, it is the difference in the preference scores between visualisations that is important here, not the exact value of the preference score of each visualisation.

6.3.6 Rankings

The parameter estimates of the above statistical models were used to rank each of the visualisations based on the accuracy, confidence/ease, and preferences displayed by the survey participants. For accuracy, we produced three sets of rankings (i.e. one for the mean, minimum, and maximum projected temperature change) and combined them into a 'final' ranking for accuracy using the `RankAggreg()` function in the `RankAggreg` R package (Pihur et al., 2018). The Cross-Entropy Monte Carlo algorithm (Rubinstein, 1999) was used to aggregate the three rankings by searching for the final ranking that minimised the 'distance' between itself and the three original rankings. We used Kendall's tau (Kendall, 1938) as a measure of distance, where distance represents the extent of disagreement between rankings (see Pihur et al. (2009) for further details). To check that the rankings were robust to the choice of algorithm and distance measure, we compared the rankings of the visualisations using the Genetic algorithm (Goldberg, 1989) to aggregate the rankings and Spearman's footrule (Spearman, 1904) as a measure of distance. The final rankings were identical when using the two different algorithms and were almost identical when using the two measures of distance; the only differences between Kendall's tau and Spearman's footrule occurred between two visualisations in the middle of the rankings and therefore the choice of method would not affect the visualisations deemed to be the best or worst performers.

The above methodology was repeated to determine the final rankings of the visualisations for confidence/ease and for preferences. To do this, we produced four sets of rankings for confidence/ease (i.e. one for the confidence and ease associated with the mean and the minimum and maximum) and five sets of rankings for preference (i.e. one for each preference category) using the parameter estimates of the MPPOMs and BT models described in Section 6.4.3 and 6.4.4). The 'overall' ranking of the visualisations was determined by combining the final rankings of the visualisations for accuracy, confidence/ease, and preferences. We aggregated the final rankings of the visualisations instead of the 12 individual rankings to ensure that accuracy, confidence/ease, and preferences were treated equally, rather than the overall ranking being weighted towards preferences and confidence/ease.

6.4 Results

We used an online survey to better understand the effectiveness of different methods of visualising the outputs of Multi-Model Ensembles (MMEs). The number of times each visualisation had previously been encountered by the survey participants is described in Section 6.4.1. The accuracy, confidence, and ease with which different groups of participants were able to interpret each visualisation is discussed in Sections 6.4.2 and 6.4.3 and the participants' preferences for each visualisation across a number of different categories is discussed in Section 6.4.4. The final rankings of the visualisations based on the accuracy, confidence, ease, and preferences displayed by the participants are described in Section 6.4.5 .

6.4.1 Previous encounters

Of all of the visualisations in the survey, the participants were most familiar with the box1 plot, with over 70% of the participants having previously encountered a similar visualisation (Figure 6.3). Over 50% of the survey participants had also encountered visualisations similar to the dot2, line1, and line2 plots prior to completing the survey (Figure 6.3). The cascade plot was by far the least familiar of all of the visualisations in the survey, with over 90% of the participants having not seen a similar visualisation in the past (Figure 6.3). Over 50% of the survey participants had also not previously encountered visualisations that were similar to the heat plot or the infographic (Figure 6.3).

6.4.2 Accuracy

Visualisation type

The fitted values of the Generalised Linear Mixed Models (GLMMs), which are based on changing a single covariate from its 'typical' value, are shown in Figure 6.4. Based on the coefficients of the GLMMs, visualisation type had a significant effect (p < 0.05) on all three measures of accuracy (i.e. the absolute difference (x10) between the participants' estimates of the mean, minimum, and maximum projected temperature change and the true values given by the climate models) (Figure 6.4). More specifically, the participants were significantly more accurate (p < 0.05) when they were asked to estimate the mean, minimum, and maximum projected temperature change using the dot1 plot compared with the reference box1 plot (Figure 6.4). Conversely, the participants were significantly less accurate (p < 0.001) when they were asked to estimate the mean and/or the minimum and maximum projected temperature change using the heat plot and the infographic compared with the reference box1 plot (Figure

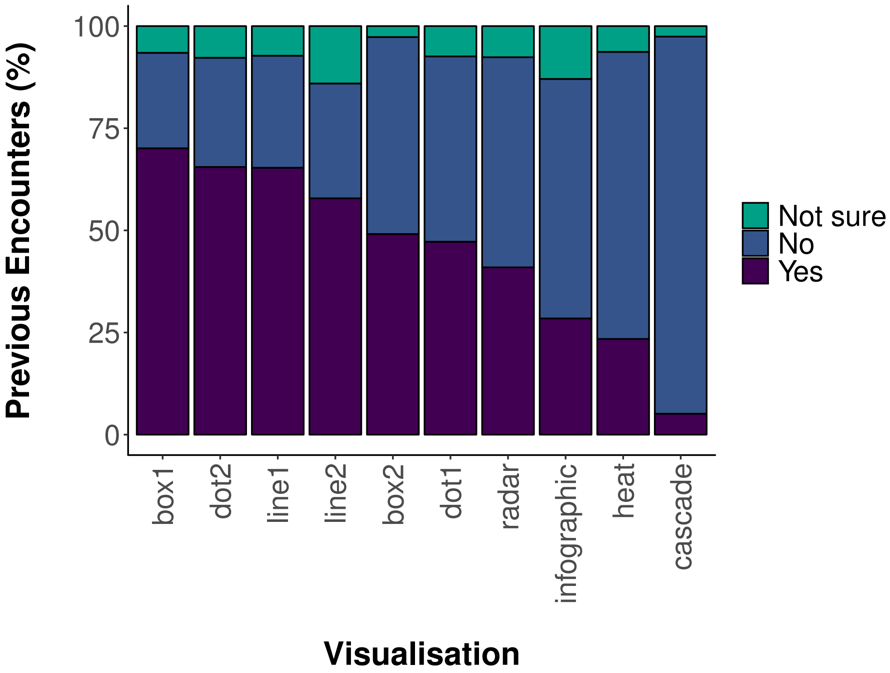


Figure 6.3: The proportion of participants (%) that had previously encountered each visualisation type prior to completing the survey. Teal indicates the proportion of participants that were not sure whether they had previously encountered the visualisation, whilst purple and blue indicate the proportion of participants that had or had not previously encountered the visualisation respectively.

6.4).

The predictions of the GLMM that was used to analyse the absolute difference between the participants' estimates of the mean projected temperature change and the true value given by the climate models (represented as ratios relative to the 'typical' response) suggest that the heat plot was outperformed by all other visualisation types, as the Absolute Difference (AD) ratio (95% Confidence Interval (CI)) of the heat plot was 2.27 (1.73, 2.99), whilst the AD ratios for all other visualisation types fell between 0.63 (0.44, 0.88) and 1.25 (0.93, 1.69) (Figure 6.5). The dot1 plot was associated with the greatest accuracy, with an AD ratio of 0.63 (0.44, 0.89) (Figure 6.5). In addition to outperforming the heat plot, the dot1 plot also outperformed the box2, line2, and radar plots, all of which had AD ratios (95% CI) ranging from 1.23 (0.91, 1.67) to 1.25 (0.93, 1.69) (Figure 6.5). There were no discernible differences in the AD ratios of the box plots, line plots, the dot2 plot, the cascade plot, or the infographic (Figure 6.5).

The predictions of the GLMMs that were used to analyse the accuracy with which the participants were able to estimate the minimum and maximum temperature change projected to occur in a given scenario and decade were very much alike and were broadly similar to those associated with the mean projected temperature change, although there were some notable differences (see below). For example, the dot1 plot was again associated with the lowest inaccuracies for both the minimum and maximum projected temperature change, with AD ratios (95% CI) of 0.75 (0.58, 0.98) and 0.68 (0.50, 0.91) respectively (Figure 6.5).

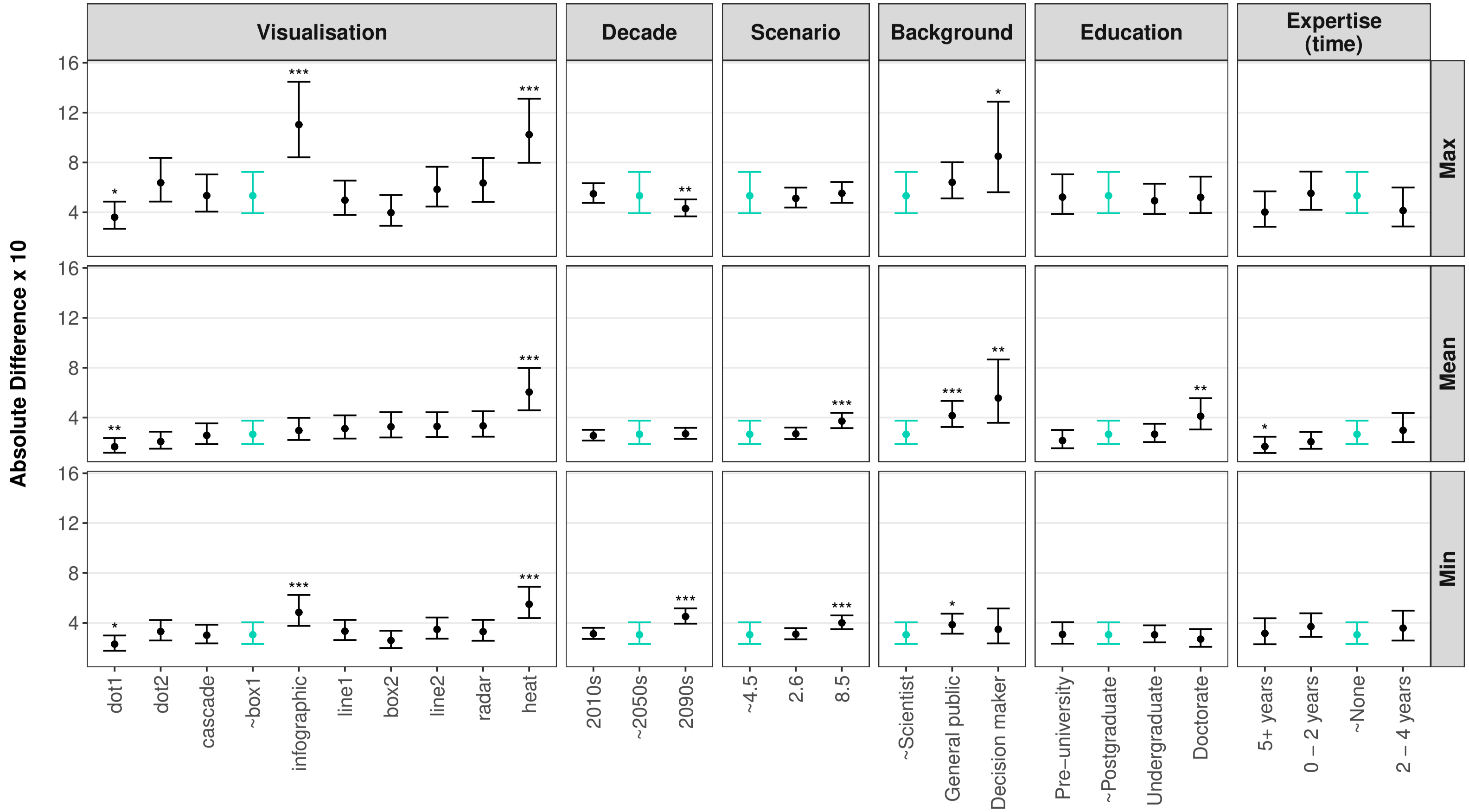


Figure 6.4: The fitted values (95% confidence interval) of the Generalised Linear Mixed Models that were used to analyse the absolute difference (x10) between the participants' estimates of the minimum (bottom), mean (middle), and maximum (top) temperature change projected to occur in a given scenario and decade and the true values given by the climate models. The coefficients are shown for visualisation type, decade, scenario, background, level of education, and expertise in working with environmental models and/or their outputs. One level of each predictor variable is highlighted in teal and marked with a tilde to indicate the reference levels used to represent the 'typical' response (see Section 6.3.5 for further details). *p < 0.05, **p < 0.01, ***p < 0.001.

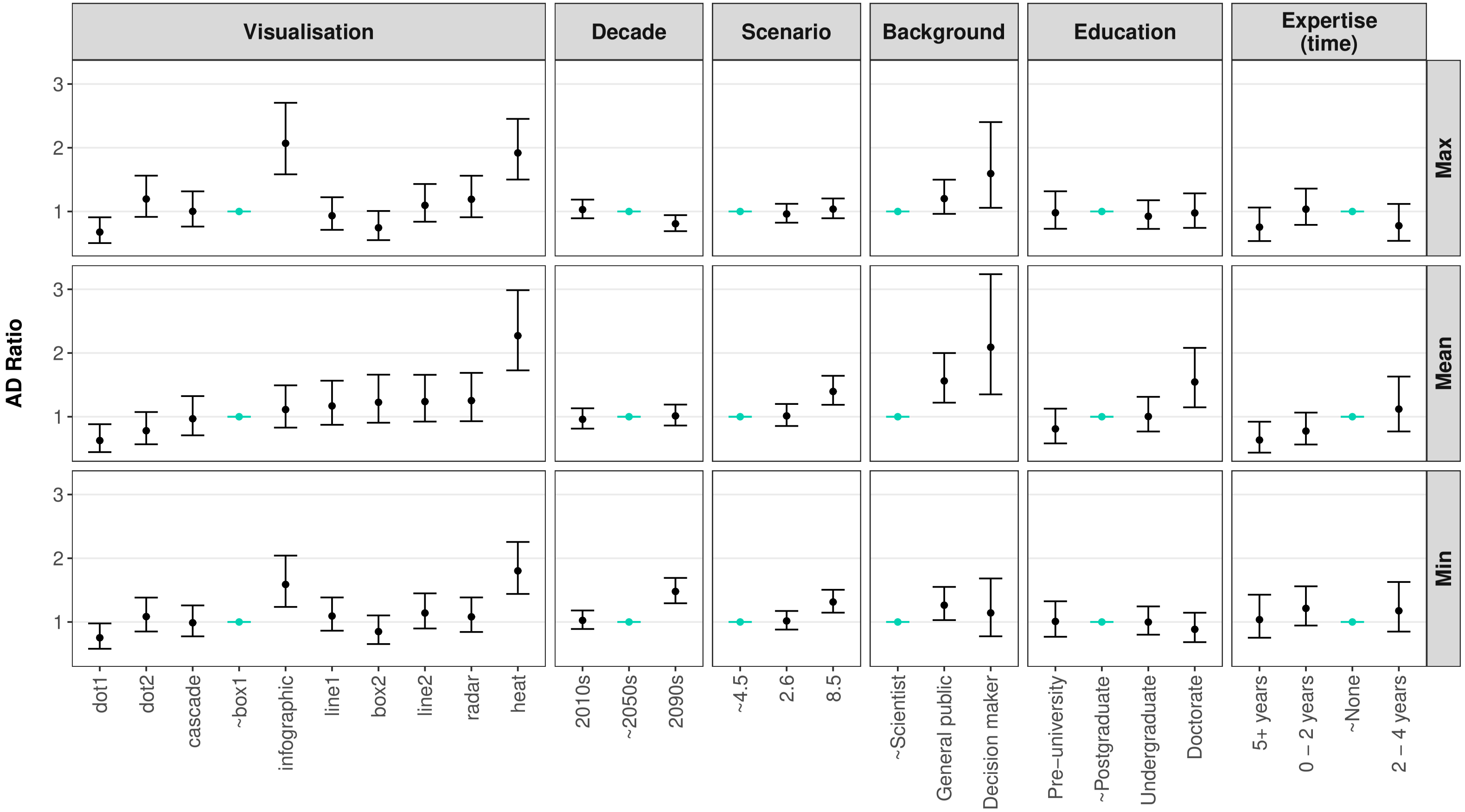


Figure 6.5: The predictions (95% confidence interval) of the Generalised Linear Mixed Models that were used to analyse the absolute difference (x10) between the participants' estimates of the minimum (bottom), mean (middle), and maximum (top) temperature change projected to occur in a given scenario and decade and the true values given by the climate models. The predictions are presented as ratios between the inaccuracies associated with the 'typical' response (as described in Section 6.3.5) compared with the inaccuracies associated with all other levels of the predictor variables (referred to as Absolute Difference (AD) ratios), although the raw predictions are also given in Appendix H. The reference levels associated with the 'typical' response are highlighted in teal and marked with a tilde. Predictions are given for visualisation type, decade, scenario, background, level of education, and expertise in working with environmental models and/or their outputs.

When the participants were asked to estimate the minimum projected temperature change, the dot1 plot outperformed both the infographic and heat plots, which had AD ratios (95% CI) of 1.59 (1.24, 2.04) and the 1.80 (1.44, 2.26) respectively (Figure 6.5). The box, cascade, radar, dot2, and line1 plots also outperformed the infographic and/or heat plot, with AD ratios (95% CI) ranging from 0.85 (0.65, 1.10) to 1.09 (0.86, 1.38) (Figure 6.5). When the survey participants were asked to estimate the maximum projected temperature change, the dot1 plot outperformed the radar, dot2, heat, and infographic plots, all of which had AD ratios (95% CI) ranging from 1.19 (0.91, 1.56) to 2.07 (1.58, 2.71) (Figure 6.5). Similar to the above, the box, line, cascade, radar, and dot2 plots also outperformed the heat plot and/or the infographic, with AD ratios (95% CI) ranging from 0.74 (0.55, 1.01) to 1.20 (0.92, 1.56) (Figure 6.5). There were no discernible differences between the AD ratios of the box plots, line plots, dot1 plot, and the cascade plot when the participants were asked to estimate either the minimum or maximum projected temperature change (Figure 6.5).

Decade and scenario

The decade that was given to each participant had no effect (p > 0.05) on the accuracy with which they were able to estimate the mean projected temperature change, but it had a significant effect (p < 0.01) on the accuracy with which they were able to estimate the minimum and maximum projected temperature change (Figure 6.4). For example, the participants were less accurate when they were asked to estimate the minimum projected temperature change in the 2090s compared with the reference 2050s, as the AD ratio (95% CI) associated with the 2090s was 1.48 (1.29, 1.69) (Figure 6.5). However, the opposite trend was apparent when the participants were asked to estimate the maximum projected temperature change; the AD ratio (95% CI) associated with the 2090s was 0.81 (0.69, 0.94), suggesting that the participants tended to be more accurate when they were asked to estimate the maximum projected temperature change in the 2090s compared with the 2050s (Figure 6.5). There were no apparent differences in the accuracy with which the participants were able to estimate the mean or the maximum temperature change projected to occur in the 2010s compared with the 2050s or the 2090s (Figure 6.5). Conversely, the participants tended to be more accurate when asked to estimate the minimum projected temperature change in the 2010s compared with the 2090s, as the AD ratios (95% CI) associated with the 2010s and 2090s were 1.03 (0.89, 1.18) and 1.48 (1.29, 1.69) respectively (Figure 6.5).

The scenario that was given to each participant had no effect (p > 0.05) on the accuracy with which the participants were able to estimate the maximum projected temperature change, but it had a significant effect (p < 0.001) on the accuracy with which they were able to estimate the mean and the minimum projected temperature change (Figure 6.4). The participants tended to be less accurate when they were asked to estimate the mean or the minimum projected temperature change in scenario 8.5 compared with scenario 4.5, as the AD ratios (95% CI) associated with scenario 8.5 were 1.40 (1.19, 1.64) for the mean and 1.31 (1.15, 1.51) for the minimum projected temperature change (Figure 6.5). There were no notable differences between the accuracy with which the participants were able to estimate the mean, minimum, and maximum temperature change projected to occur in scenario 2.6 compared with scenario 8.5.

Background

The background of the participant (i.e. general public, scientist, or decision maker/environmental manager) had a significant effect (p < 0.05) on the accuracy with which they were able to estimate the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade (Figure 6.4). When asked to estimate the mean projected temperature change, both the decision makers/environmental managers and the general public were associated with greater inaccuracies than the reference scientist group, with AD ratios (95% CI) of 2.09 (1.35, 3.24) and 1.56 (1.22, 2.00) respectively (Figure 6.5). The general public also tended to be less accurate than the reference scientist group when they were asked to estimate the minimum projected temperature change, whilst the decision makers/environmental managers were less accurate than the reference scientist group when they were asked to estimate the maximum projected temperature change, with AD ratios (95% CI) of 1.26 (1.03, 1.55) and 1.59 (1.06, 2.40) respectively (Figure 6.5). There were no apparent differences between the accuracy of the decision makers/environmental managers and the general public when they were asked to estimate the mean, minimum, or maximum temperature change projected to occur in a given scenario or decade.

Education

The education level of the participants had no effect (p > 0.05) on the accuracy with which they were able to estimate the minimum and maximum temperature change projected to occur in a given scenario and decade, although it did have a significant effect (p < 0.01) on their ability to estimate the mean projected temperature change (Figure 6.4). For example, participants with a doctorate degree had an AD ratio (95% CI) of 1.55 (1.15, 2.08), suggesting they were less accurate than the reference postgraduate group (Figure 6.5). The participants that had not attended university (i.e. those with GCSE, A-Level, or vocational training) had an AD ratio of (95% CI) of 0.81 (0.58, 1.13), suggesting that this group of individuals was also more accurate than those with a doctorate degree (Figure 6.5). There were no notable differences in the accuracy with which the participants with pre-university, undergraduate, or postgraduate levels of education were able to estimate the mean projected temperature change.

Expertise

The participants' expertise in working with environmental models and/or their outputs had no effect (p > 0.05) on the accuracy with which they were able to estimate the minimum or maximum temperature change projected to occur in a given scenario and decade (Figure 6.4). However, when the participants were asked to estimate the mean projected temperature change, those with more than five years of experience were significantly more (p < 0.05) accurate than those with no experience, as they had an AD ratio (95% CI) of 0.63 (0.44, 0.92) (Figure 6.4 and 6.5).

Random effects

It was not possible to test the significance of including participant ID as a random effect in the GLMMs using a Likelihood Ratio Test as the method used here computes quasi-likelihoods instead of true likelihoods (Bolker et al., 2009). Despite this, it is clear that some of the participants were inherently more accurate than others when estimating the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade. For example, the standard deviation of the random intercepts (95% CI), which represents the amount of within-treatment variability that is explained by participant ID, was between 0.56 (0.50, 0.64) and 0.67 (0.59, 0.76) for all three GLMMs, whilst the residual within-treatment standard deviation, which represents the amount of within-treatment variability that is not explained by participant ID, was between 1.78 (1.69, 1.88) and 2.10 (2.00, 2.21) (Table 6.2).

Table 6.2: The Standard Deviation (SD) of the random intercepts and the residual within-treatment SD (95% Confidence Interval (CI)) of the three Generalised Linear Mixed Models that were used to analyse the absolute difference (x10) between the participants' estimates of the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade and the true value given by the climate models. The SD of the random intercepts represents the amount of within-treatment variability that is explained by participant ID, whilst the residual SD represents the unexplained within-treatment variability. A SD of zero for the random intercepts would indicate that the participants did not vary consistently across treatments.

Response variable	Intercept SD (95% CI)	Residual SD (95% CI)
Mean	0.67 (0.59, 0.77)	2.10 (2.00, 2.21)
Minimum	0.56 (0.50, 0.64)	1.78 (1.69, 1.88)
Maximum	0.61 (0.54, 0.70)	2.09 (1.98, 2.20)

6.4.3 Confidence and ease

Visualisation type

The type of visualisation that was shown to the participants had a significant effect (p < 0.05) on the confidence and ease with which they were able to estimate the mean or the minimum and maximum temperature change projected to occur in a given scenario and decade (Figure 6.6). For example, the participants were between 2.30 (95% CI: 1.27, 4.15) and 2.82 (95% CI: 1.54, 5.16) times more likely (p < 0.01) to select a higher rating for confidence and ease when they were asked to estimate the mean projected temperature change using the dot1 or dot2 plots compared with the reference box1 plot (Figure 6.6). The participants were also 2.11 (95% CI: 1.19, 3.74) times more likely to give a higher rating for confidence when they were asked to estimate the mean temperature change using the line1 plot compared with the box1 plot (Figure 6.6).

On the other hand, the survey participants were significantly less likely (p < 0.01) to select a higher rating for confidence and ease when they were asked to estimate either the mean or the minimum and maximum projected temperature change using the radar and heat plots, both of which had odds ratios (95% CI) of between 0.05 (0.03, 0.10) and 0.39 (0.21, 0.70) (Figure 6.6). The participants were also significantly less likely (p < 0.001) to select a higher rating for confidence and ease when they were asked to estimate the minimum and maximum projected temperature change using the line2 and infographic plots; the same was also true when the participants were asked to comment on the ease with which they were able to estimate the minimum and maximum using the line1 plot (Figure 6.6). In all of these cases, the odds ratios (95% CI) of the line and infographic plots remained within 0.14 (0.07, 0.28) and 0.30 (0.17, 0.55) (Figure 6.6).

The predictions of the MPPOMs further support the conclusions described above. For example, the survey participants were most likely to give a positive response (i.e. somewhat agree or agree) when they were asked to estimate the mean projected temperature change in a given scenario and decade using the dot and line plots. To illustrate this point, the predicted probability of a positive confidence rating was between 0.85 and 0.92 when the participants were asked to estimate the mean projected temperature change using the dot and line plots (Figure 6.7). However, the probability of a positive response remained above 0.79 for all visualisation types excluding the radar and heat plots, which were associated with probabilities of 0.61 and 0.23 respectively (Figure 6.7). The heat plot was the only visualisation type to be associated with a greater probability of a negative rating than a positive rating for confidence (Figure 6.7). Similar results were also apparent when the participants were asked to comment

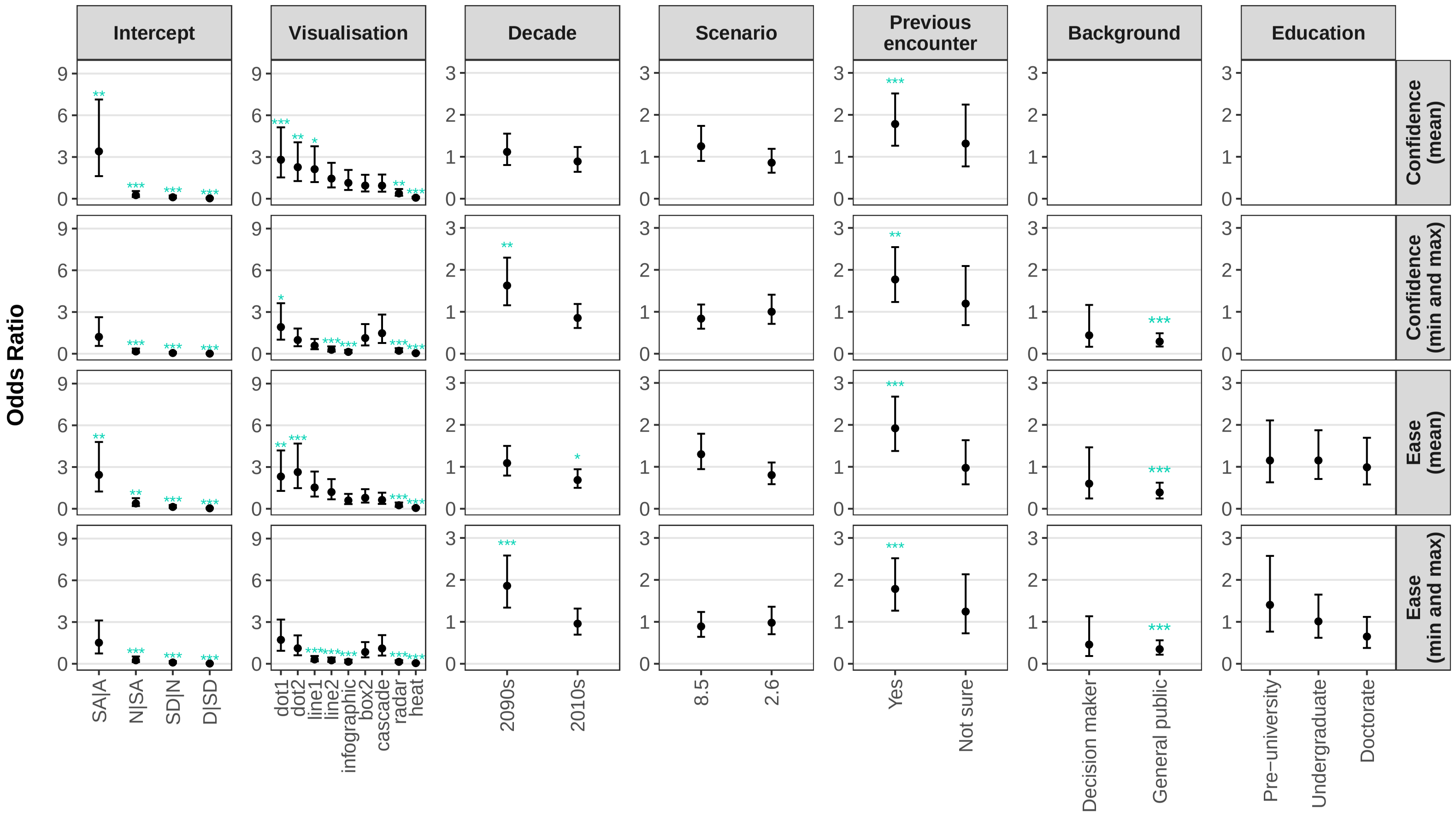


Figure 6.6: The regression coefficients (or 'odds ratios') (95% confidence interval) of the Mixed Partial Proportional Odds Models (MPPOMs) that were used to analyse the confidence and ease with which the participants were able to estimate the mean and minimum/maximum temperature change projected to occur in a given scenario and decade. The levels of the predictor variables that were included in the 'typical' response (as described in Section 6.3.5) are included in the coefficients associated with the intercept (left), whilst the levels of the predictor variables that were not included in the 'typical' response are given separately. SA|A refers to the threshold between somewhat agree and agree, N|SA refers to the threshold between neutral and somewhat agree, SD|N refers to the threshold between somewhat disagree and neutral, and D|SD refers to the threshold between disagree and somewhat disagree. The regression coefficients are given for visualisation type, decade, scenario, previous encounters, background, and level of education. Please note the differences in the y-axis limits between each of the predictor variables. Blank spaces indicate that the predictor variable did not meet the assumption of proportional odds and was included in the MPPOM as a nominal effect instead (see Figure 6.9 instead). *p < 0.05, **p < 0.01, ***p < 0.001.

on the ease with which they were able to estimate the mean projected temperature change (see Figure 6.8).

Conversely, the participants were most likely to give a positive response for confidence and ease when they were asked to estimate the minimum and maximum projected temperature change using the dot1, cascade, box2, dot2 and box1 plots. Using confidence as an example, the predicted probability of a positive rating was between 0.86 and 0.92 for all of these visualisation types (Figure 6.7). The participants were more likely to give a neutral or negative rating than a positive rating for confidence when they were asked to estimate the minimum and maximum using the infographic and heat plots, with probabilities of 0.56 and 0.80 respectively (Figure 6.7). The radar and line2 plots were also associated with a relatively high probability of a neutral or negative confidence rating, with probabilities of 0.45 and 0.37 respectively (Figure 6.7). Again, similar results were apparent when the participants were asked to comment on the ease with which they were able to estimate the minimum and maximum projected temperature change (see Figure 6.8).

Decade and scenario

The scenario that was given to each of the participants had no effect (p > 0.05) on the confidence or ease with which they were able to estimate the mean or the minimum and maximum projected temperature change (Figure 6.6). The decade that was given to each participant also had no effect (p > 0.05) on the confidence with which the participants were able to estimate the mean projected temperature change, but it had a significant effect (p < 0.05) on all other measures of confidence and ease (Figure 6.6). For example, the participants were 1.66 (95% CI: 1.18, 2.34) times more likely to select one of the higher Likert scale categories for confidence and 1.92 (95% CI: 1.39, 2.67) times more likely to do the same for ease when they were asked to estimate the minimum and maximum temperature change projected to occur in the 2090s compared with the reference level of the 2050s (Figure 6.6). Conversely, the participants were less likely to select a higher rating for ease when they were asked to estimate the mean in the 2010s compared with the 2050s, with an odds ratio (95% CI) of 0.70 (0.51, 0.96) (Figure 6.6). When the participants were asked to estimate the minimum and maximum projected temperature change in the 2090s, the predicted probability of a positive rating for confidence was 0.91, but this dropped to 0.85 and 0.84 when they were asked to estimate the same value in the 2050s and 2010s respectively (Figure 6.7). Similar patterns were also apparent when the participants were asked to comment on the ease with which they were able to estimate the mean or the minimum and maximum projected temperature change (Figure 6.7).

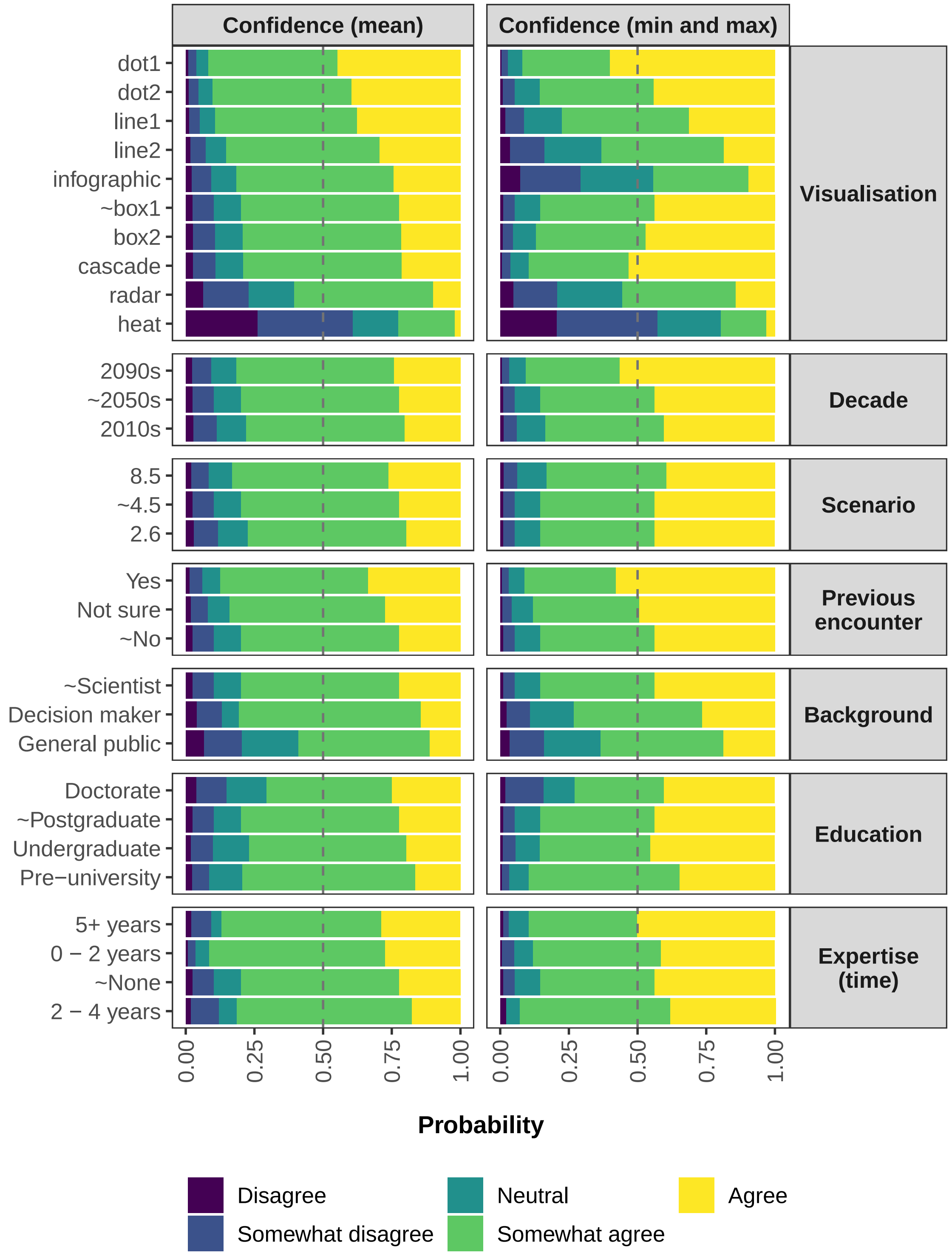


Figure 6.7: The predictions of the Mixed Partial Proportional Odds Models that were used to analyse the confidence with which the participants were able to estimate the mean (left) or minimum/maximum (right) temperature change projected to occur in a given scenario and decade. The predictions are given as probabilities for each level of the Likert scale, ranging from 'disagree' in purple to 'agree' in yellow. The predictions are made relative to the 'typical' response (as described in Section 6.3.5) and the reference levels associated with the 'typical' response are marked with a tilde. Predictions are given for visualisation type, decade, scenario, previous encounters, background, level of education, and expertise in working with environmental models and/or their outputs.

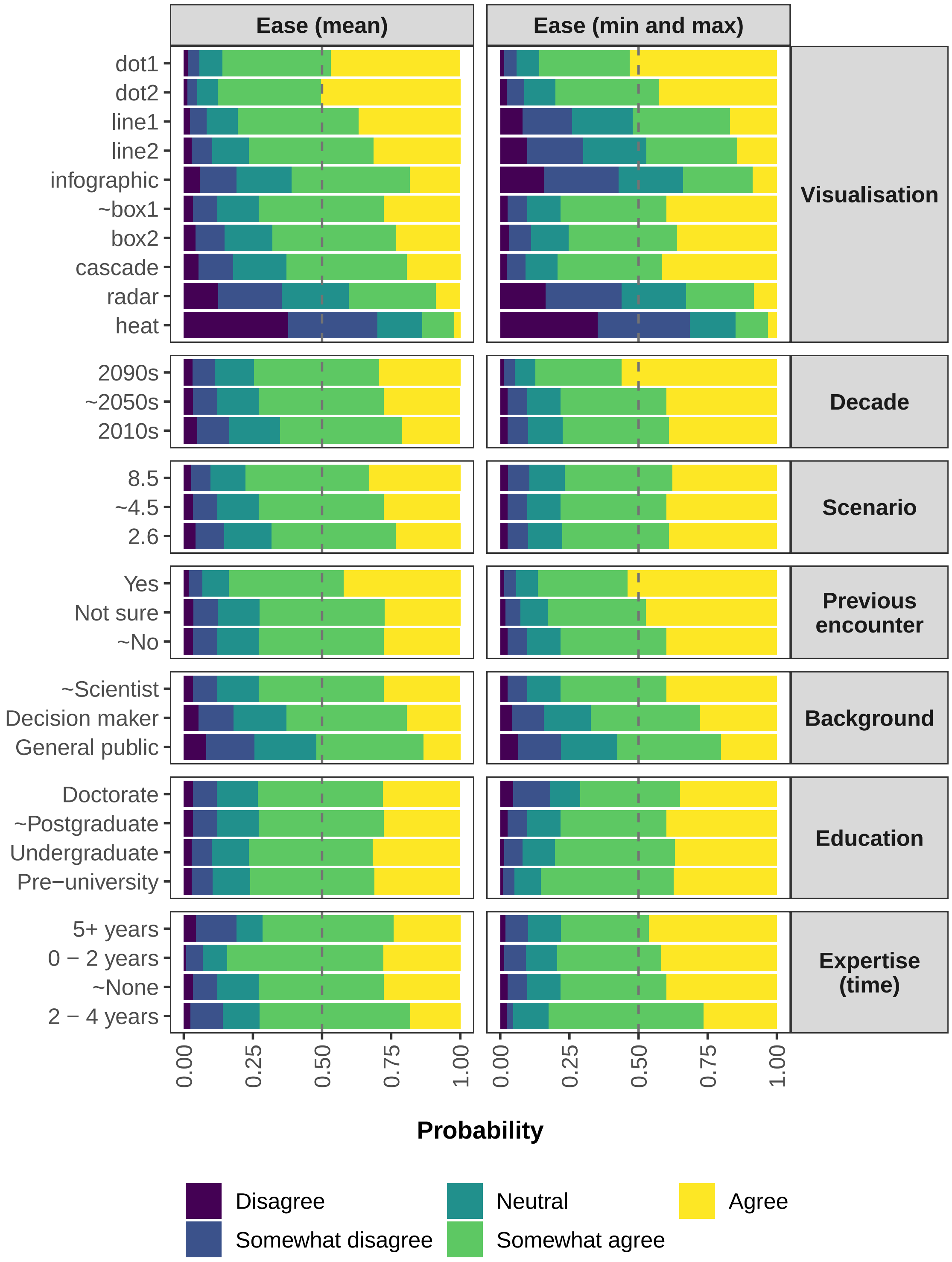


Figure 6.8: The predictions of the Mixed Partial Proportional Odds Models that were used to analyse the ease with which the participants were able to estimate the mean (left) or minimum/maximum (right) temperature change projected to occur in a given scenario and decade. The predictions are given as probabilities for each level of the Likert scale, ranging from 'disagree' in purple to 'agree' in yellow. The predictions are made relative to the 'typical' response (as described in Section 6.3.5) and the reference levels associated with the 'typical' response are marked with a tilde. Predictions are given for visualisation type, decade, scenario, previous encounters, background, level of education, and expertise in working with environmental models and/or their outputs.

Previous encounters

The familiarity of the visualisation that was given to each participant had a significant effect (p < 0.01) on the confidence and ease with which they were able to estimate the mean and the minimum and maximum temperature change projected to occur in a given scenario and decade (Figure 6.6). More specifically, the participants were between 1.76 (95% CI: 1.25, 2.48) and 1.92 (95% CI: 1.38, 2.67) times more likely to select one of the higher ratings for confidence and ease when they were asked to interpret a visualisation type that they were already familiar with compared with a visualisation type they had never previously encountered prior to completing the survey (Figure 6.6). There were no apparent differences (p > 0.05) in the confidence and ease with which the participants were able to estimate the mean or the minimum and maximum projected temperature change between those who were not sure if they had previously encountered a similar visualisation prior to completing the survey and those who either had or had not previously encountered a similar visualisation (Figure 6.6).

The most dramatic difference between the predicted probability of a participant giving a positive rating when asked to interpret a familiar visualisation compared with an unfamiliar visualisation occurred when the participants were asked to comment on the ease with which they were able to estimate the mean projected temperature change, with respective probabilities of 0.84 and 0.73 (Figure 6.8). However, the differences in the probability of a positive rating between familiar and unfamiliar visualisations were relatively similar across all measures of confidence and ease (Figures 6.7 and 6.8).

Background

Background (i.e. general public, scientist, or decision maker/environmental manager) met the assumption of proportional odds for all measures of confidence and ease excluding the confidence with which the participants were able to estimate the mean temperature change projected to occur in a given scenario and decade. In all cases where background met the assumption of proportional odds, members of the general public were significantly less likely (p < 0.001) to select one of the higher Likert scale categories for confidence and ease than the reference scientist group (Figure 6.6). In all of these examples, the odds ratio (95% CI) of a member of the general public selecting one of the higher Likert scale categories was between 0.30 (0.18, 0.49) and 0.40 (0.26, 0.63) (Figure 6.6). The most dramatic difference between the reference scientist group and the general public occurred when the participants were asked to comment on the confidence with which they were able to estimate the minimum and maximum projected temperature change; the predicted probability of the reference scientists giving a

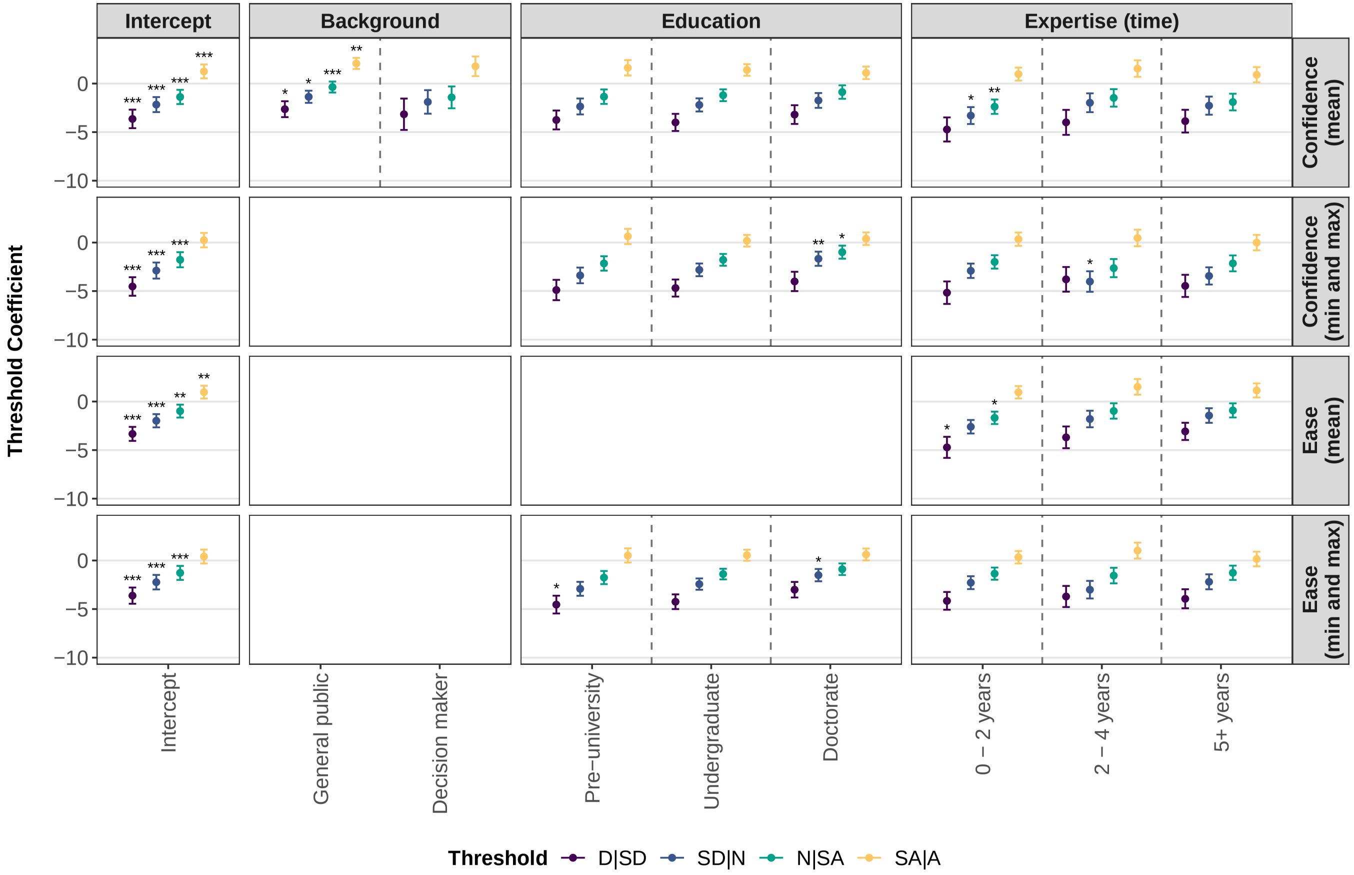


Figure 6.9: The threshold coefficients (95% confidence interval) of the Mixed Partial Proportional Odds Models (MPPOMs) that were used to analyse the confidence and ease with which the participants were able to estimate the mean and minimum/maximum temperature change projected to occur in a given scenario and decade. The thresholds between each level of the Likert scale are presented in purple, blue, teal, and yellow, with purple representing the threshold between Disagree (D) and Somewhat Disagree (SD) and yellow representing the threshold between Somewhat Agree (SA) and Agree (A). The levels of the predictor variables that were included in the 'typical' response (as described in Section 6.3.5) are included in the coefficients associated with the intercepts (left), whilst the levels of the predictor variables that were not included in the 'typical' response are given separately. The threshold coefficients are given for background, level of education, and expertise in working with environmental models and/or their outputs as these predictor variables were the only ones to break the assumption of proportional odds. Blank spaces indicate the predictor variable met the assumption of proportional odds and was included in the MPPOM as an ordinal effect instead (see Figure 6.9). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

positive rating was 0.86, whilst the probability of a member of the general public giving a positive rating was just 0.64 (Figure 6.7). In this example, decision makers and environmental managers were associated with a probability of 0.73 (Figure 6.7). Similar patterns were also present across both measures of ease.

When background was incorporated in the MPPOM that was used to analyse the confidence with which the participants were able to estimate the mean projected temperature change as a nominal effect, the threshold coefficients between all levels of the Likert scale were significantly greater (p < 0.05) for members of the general public than the reference scientist group (Figure 6.9). The threshold between neutral and somewhat agree displayed the greatest difference, occurring at -0.36 (95% CI: -0.92, 0.20) for the general public and at -1.38 (95% CI: -2.11, -0.65) for the reference scientist group (Figure 6.9). The threshold between somewhat disagree and neutral displayed the smallest difference, occurring at -1.36 (95% CI: -1.99, 0.74) for the general public and at -2.17 (95% CI: -2.94, -1.39) for scientists (Figure 6.9). Overall, the predicted probability of a positive rating was 0.80 for scientists, 0.59 for members of the general public, and 0.81 for decision makers and environmental managers (Figure 6.7). Despite the relatively large difference between the probability of a positive response from members of the general public and decision makers/environmental managers, there were no significant differences in the odds ratios or threshold coefficients of these two groups across all measures of confidence and ease (Figure 6.6). However, the confidence intervals associated with the odds ratios of the decision makers/environmental managers were relatively large, likely as a result of the small number of individuals in this group (n = 18).

Level of education

Education only met the assumption of proportional odds when the participants were asked to comment on the ease with which they were able to estimate the mean projected temperature change. In this particular case, there were no significant differences (p > 0.05) between the Likert scale categories of those with pre-university, undergraduate, or doctoral training when compared with the reference postgraduate group (Figure 6.6). When education was included in the MPPOM that was used to analyse the confidence with which the participants were able to estimate the minimum and maximum projected temperature change as a nominal effect, there were significant differences (p < 0.05) in the thresholds between somewhat disagree and neutral and between neutral and somewhat agree for those with doctoral training compared with the postgraduate reference group (Figure 6.9). For example, the threshold (95% CI) between somewhat disagree and neutral occurred at -2.17 (2.94, -1.39) for the postgraduate group, but at -1.67 (-2.40, -0.94) for those with doctoral training (Figure 6.9). A similar pattern was present when the participants were asked to comment on the ease with which they were able to estimate the minimum and maximum projected temperature change, although to a lesser extent. For this particular measure of ease, the threshold between disagree and somewhat disagree was also significantly lower (p < 0.05) for those with pre-university education compared with the reference postgraduate group, with the threshold (95% CI) occurring at -4.54 (-5.46, -3.62) for the former and at -3.62 (-4.45, -2.78) for the latter (Figure 6.9). There were no significant differences (p > 0.05) between the threshold coefficients associated with each level of the Likert scale when the participants were asked to comment on the confidence with which they were able to estimate the mean projected temperature change (Figure 6.9).

Focusing on the predictions of the MPPOMs, the probability of a positive rating for confidence and ease was relatively consistent across all levels of education. For example, when the participants were asked to comment on the ease with which they were able to estimate the mean projected temperature change, the predicted probability of a positive response was between 0.73 and 0.77 across all levels of education (Figure 6.7). The probability of a positive response was slightly more variable when the participants were asked to comment on the confidence and ease with which they were able to estimate the minimum and maximum projected temperature change, although it remained within 0.71 and 0.90 across all levels of education (Figures 6.7 and 6.8). In the majority of cases, those without a university education were most likely to give a positive response, whilst those with doctoral training were least likely to give a positive response.

Expertise

Expertise did not meet the assumption of proportional odds for any of the measures of confidence and ease and therefore it was included in all of the MPPOMs as a nominal effect. When the participants were asked to comment on the confidence with which they were able to estimate the minimum and maximum, the threshold coefficient between somewhat disagree and neutral was significantly lower (p < 0.05) for those with two to four years of experience compared with the reference group that had no previous experience, with the threshold (95% CI) occurring at -4.02 (-5.07, -2.96) for the former and at -2.88 (-3.71, -2.06) for the latter (Figure 6.9). When the participants were asked to comment on the confidence and ease with which they were able to estimate the mean projected temperature change, the threshold coefficients between disagree and somewhat disagree, somewhat disagree and neutral, and/or between neutral and somewhat agree occurred between 0.69 and 1.38 units lower (p < 0.05) for those with up to two years of experience compared with the reference group that had no previous experience (Figure 6.9). When the participants were asked to comment on the ease with which they were able to estimate the minimum and maximum projected temperature change, there were no notable differences in the thresholds between each of the Likert scale categories based on level of expertise.

Overall, the probability of a positive rating for confidence and ease was relatively consistent across all levels of expertise. For example, when the participants were asked to comment on the ease with which they were able to estimate the minimum and maximum projected temperature change, the predicted probability of a positive rating was between 0.78 and 0.83 across all levels of expertise (Figures 6.7 and 6.8). However, the participants with more than zero but less than two years of experience tended to be more likely to give a positive rating for confidence and ease when they were asked to estimate the mean compared with the participants that had more or less experience. Using ease as an example, the predicted probability of a participant with more than zero but less than two years of experience giving a positive rating was 0.84, whilst the probability of a positive rating was between 0.71 and 0.73 for all other levels of expertise (Figure 6.8).

The mean versus the minimum and maximum

Interestingly, the probability of a participant selecting 'agree' was often greater when they were asked to comment on the confidence and ease with which they were able to estimate the minimum and maximum temperature change projected to occur in a given scenario and decade compared with the mean. For example, the predicted probability of a participant selecting 'agree' when asked to comment on the confidence with which they were able to estimate the minimum and maximum projected temperature change using a visualisation that was familiar to them was 0.58, but this dropped to 0.34 when they were asked to estimate the mean (Figure 6.7). Similarly, the predicted probability of a participant selecting 'agree' to describe the ease with which they were able to estimate the minimum and maximum projected temperature change in the 2090s was 0.56, but this fell to 0.29 when they were asked to estimate the mean (Figure 6.8). These are just a couple of examples of cases in which the probability of selecting 'agree' was greater when the participants were asked to comment on the confidence and ease with which they were able to estimate the minimum and maximum projected temperature change compared with the mean, but similar examples can also be found across almost all of the predictor variables.

Random effects

We used Likelihood Ratio Tests (LRTs) and IntraClass Correlation (ICC) scores to determine the importance of including participant ID as a random effect in the MPPOMs that were used to analyse both the confidence and ease with which the participants were able to estimate the mean or the minimum and maximum temperature change projected to occur in a given scenario and decade. A LRT p-value of less than 0.05 indicates the fit of the MPPOM significantly improved when participant ID was included as a random effect. An intraclass correlation score close to one indicates high within-cluster similarity, whilst a value close to zero indicates low within-cluster similarity (Schmidt, 2012).

The LRTs indicated that the goodness-of-fit of the MPPOMs significantly improved ($p < 0.001$) when participant ID was included as a random effect (Table 6.3). The ICC scores (95% CI) of the MPPOMs were all between 0.55 and 0.62 (Table 6.3), suggesting that the participants tended to give relatively similar answers across all blocks of questions.

Table 6.3: The Likelihood Ratio Test (LRT) statistics and the IntraClass Correlation (ICC) scores associated with including participant ID as a random effect in the mixed partial proportional odds models that were used to analyse the confidence and ease with which the survey participants were able to estimate the mean or the minimum and maximum temperature change projected to occur in a given scenario and decade. A LRT p-value of less than 0.05 indicates that the fit of the model significantly improved when participant ID was included in the models as a random effect. An ICC score of close to one indicates a high within-cluster similarity, whilst a score close to zero indicates a low within-cluster similarity. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Response variable	LRT statistic	p-value	ICC score
Confidence (mean)	92.75	<0.001***	0.62
Confidence (min and max)	81.74	<0.001***	0.62
Ease (mean)	73.53	<0.001***	0.57
Ease (min and max)	59.45	<0.001***	0.55

6.4.4 Preference

Visualisation type

When using the visualisations to view changes in mean temperature over time, the survey participants displayed the greatest preference for the two line plots and the dot2 plot, all of which had preference scores that were significantly greater ($p < 0.01$) than the reference box1 plot (and hence the radar, infographic, heat, and cascade plots) (Figure 6.10). More specifically, the preference scores of the line1, line2, and dot2 plots (95% comparison interval) were between 0.72 (0.32, 1.11) and 1.24 (0.82, 1.66), whilst the preference score of the box1 plot was 0.00 (-0.41, 0.41) (Figure 6.10). The radar, heat, infographic, and cascade plots were the least preferred visualisation types, with preference scores that were significantly lower (p < 0.01) than the box1 plot (and hence all other visualisation types) (Figure 6.10). For example, the preference scores (95% comparison interval) of the radar and cascade plots were -0.60 (-1.01, -0.20) and -0.84 (-1.26, -0.43) respectively (Figure 6.10). There were no significant differences (p > 0.05) in the preference scores of the dot and box plots (Figure 6.10).

The rankings of each visualisation based on the ability to view changes in uncertainty over time were slightly different to those described above for temperature. For example, the dot and box plots had the greatest preference scores (95% comparison interval), with values ranging from -0.02 (-0.39, 0.35) to 0.10 (-0.29, 0.48) (Figure 6.10). However, there were no significant differences (p > 0.05) between the preference scores of the line plots, dot plots, box2 plot, or the cascade plot when compared with the reference box1 plot (Figure 6.10). The radar, infographic, and heat plots were the least preferred visualisations and had preference scores that were significantly lower (p > 0.001) than the reference box1 plot (and hence the dot1 and box2 plots), with preference scores (95% comparison interval) ranging from -1.09 (-1.48, -0.69) to -1.59 (-2.04, -1.15) (Figure 6.10).

When used to retrieve specific values (such as the mean, minimum, and maximum), the survey participants displayed the greatest preference for the box2 and box1 plots, with scores (95% comparison interval) of 0.38 (-0.04, 0.81) and 0.00 (-0.44, 0.44) respectively (Figure 6.10). The preference scores of the box plots were significantly greater (p < 0.01) than the preference scores of all other visualisation types excluding the two dot plots, which had scores (95% comparison interval) of -0.20 (-0.60, -0.21) and -0.24 (-0.64, 0.15) respectively (Figure 6.10). The cascade, line, and radar plots displayed intermediate but overlapping preference scores, with values (95% comparison interval) of between -0.92 (-1.28, -0.57) and -1.36 (-1.75, -0.97) (Figure 6.10). The heat plot was once again associated with the lowest preference score, with a value (95% comparison interval) of -2.41 (-2.91, -1.90), and was outperformed by all other visualisation types excluding the infographic, which had a preference score of -1.84 (-2.27, -1.41) (Figure 6.10).

For visual appeal, the reference box1 plot had the lowest preference score, with a value (95% comparison interval) of 0.00 (-0.67, 0.67) (Figure 6.10). However, there was a great deal of overlap between the preference scores of all of the visualisation types; the dot1 and dot2 plots were the only visualisations with a significantly greater (p < 0.05) preference score than the reference box1 plot, with scores of 2.02 (1.14, 2.90) and 1.06 (0.27, 1.85) respectively (Figure 6.10). There were no other notable differences in the preference scores of each visualisation based on visual appeal.

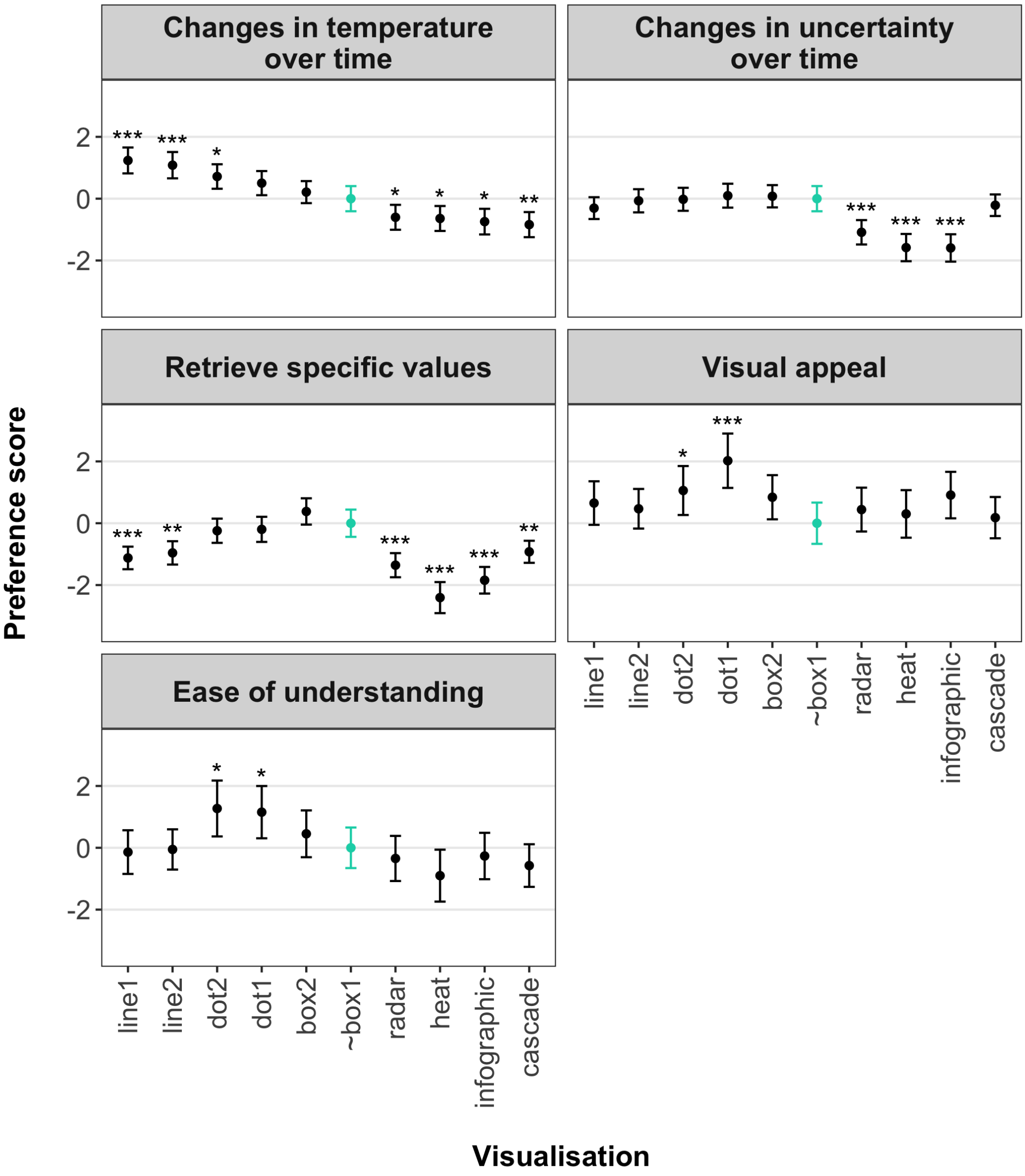


Figure 6.10: The preference scores (95% 'comparison' intervals) of each visualisation across five different preference categories: the ability to view changes in temperature over time, the ability to view changes in uncertainty over time, the ability to retrieve specific values (such as the mean, minimum, or maximum), visual appeal, and overall ease of understanding. The 95% comparison intervals are estimated using quasi standard errors to allow for comparisons to be made across all of the visualisations. The preference scores of each visualisation are presented relative to the box1 plot (highlighted in teal and marked with a tilde), which was defined as the reference visualisation in the BT models and is therefore given a preference score of zero. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

The visualisations with the greatest preference scores for overall ease of understanding were the dot2 and dot1 plots, with scores of 1.27 (0.37, 2.17) and 1.15 (0.31, 2.00) respectively (Figure 6.10). The dot plots were the only visualisation types with significantly greater ($p <$ 0.05) preference scores than the reference box1 plot (and hence the line, infographic, radar, cascade, and heat plots), although there was some overlap in the preference scores of all of the visualisation types excluding the cascade and heat plots (Figure 6.10). The cascade and heat plots had the lowest preference scores, with values of -0.57 (1.26, 0.11) and -0.90 (-1.74, 0.06) respectively (Figure 6.10).

Background and education

Background or education were included in the best-fitting BT models across all five preference categories. However, there were no notable differences in the preference scores of each visualisation based on background (see Appendix I) and there were only two notable differences between the preference scores of each visualisation based on the level of education of the participants: (1) for overall ease of understanding, those with a doctorate degree tended to prefer the line1 plot more than those with an undergraduate degree, with preference scores of 0.79 (0.65, 0.94) and 0.38 (0.20, 0.56) respectively (Figure 6.11); and (2) for visual appeal, those without a university education tended to prefer the dot1 plot more than those with an undergraduate degree, with preference scores of 0.88 (0.77, 1.00) and 0.55 (0.35, 0.75) respectively (Figure 6.11).

6.4.5 Rankings

The ten visualisations included in the survey were ranked based on participant accuracy, confidence/ease, and preferences. The dot1 and heat plots were consistently ranked in first and last position across all categories respectively (Figure 6.12). The dot2 plot performed relatively poorly for accuracy, but was second best for confidence/ease and preferences, thus resulting in an overall ranking of second place (Figure 6.12). Both of the box plots were in the top four for accuracy and preferences, but dropped down to 6th and 7th position for confidence/ease (Figure 6.12). However, this drop in performance had little effect on the overall rankings of the box plots, which placed in 3rd and 4th (Figure 6.12). The cascade plot was ranked in 3rd and 4th position for accuracy and confidence/ease respectively, but in 9th for preferences, resulting in an overall ranking of 6th (Figure 6.12). The line plots displayed an intermediate level of performance across all categories, remaining between 3rd and 7th throughout (Figure 6.12). Finally, the radar plot and infographic were ranked in the bottom four across all categories (Figure 6.12).

6.5 Discussion

An online survey was conducted to identify the most effective methods for visually communicating the outputs of Multi-Model Ensembles (MMEs). We measured the accuracy, confidence, and ease with which the participants were able to interpret different types of visualisation, as well as their preferences for each visualisation across a number of categories. The performance of each visualisation is discussed in Section 6.5.1, whilst the effects of decade, scenario, previous encounters, and the demographics of the participants are discussed in

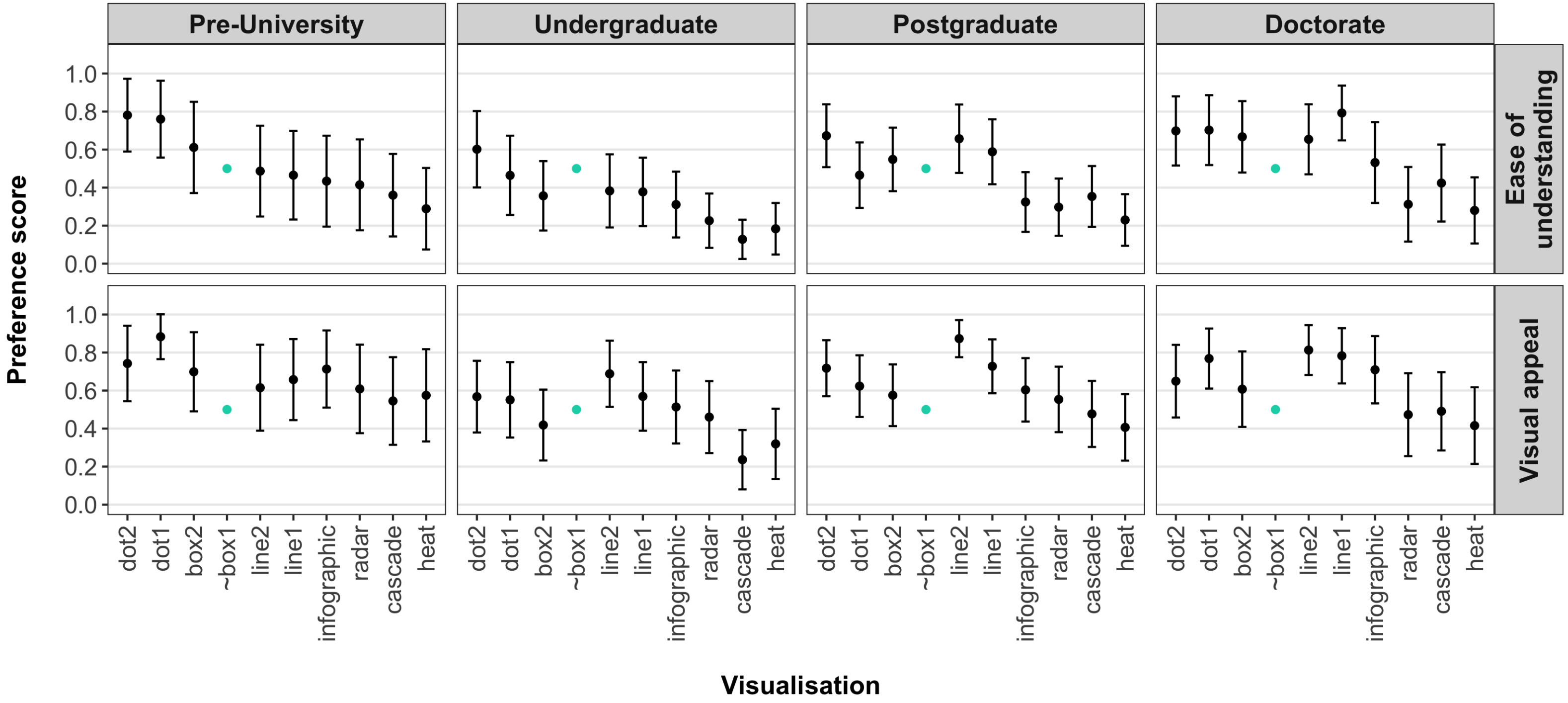


Figure 6.11: A comparison of the predicted preference scores (95% confidence intervals) of each visualisation as determined by those with pre-university, undergraduate, postgraduate, and doctoral training. The preference scores are given for the following two categories: overall ease of understanding and visual appeal. Comparison intervals are not used in this plot as we were unable to estimate the quasi standard errors of the visualisations across different levels of the predictor variables. The preference scores of each visualisation are presented relative to the box1 plot (highlighted in teal and marked with a tilde), which was defined as the reference visualisation in the BT models.

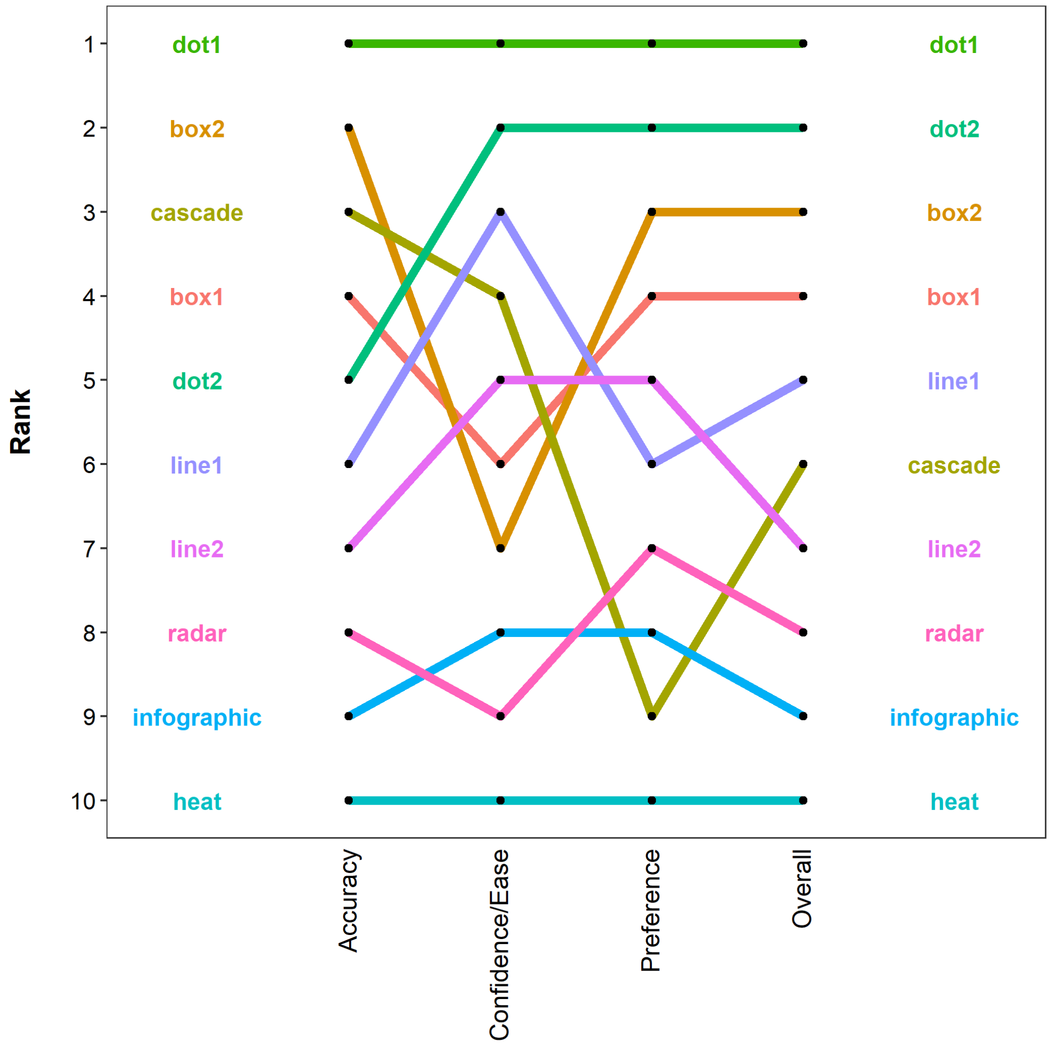


Figure 6.12: The rankings of the visualisations based on participant accuracy, confidence/ease, and preferences. An ’overall’ ranking is given based on the aggregation of the rankings for each of these categories (see Section 6.3.6 for further details).

Section 6.5.2.

6.5.1 Visualisation type

The type of visualisation that was shown to the participant had a significant effect on all measures of accuracy, confidence, ease, and preferences. Of all of the visualisations that were included in the survey, it was the dot plots that tended to perform the best, ranking highly across almost all measures of performance. The only category in which the dot plots were not in 1st and 2nd position was accuracy, as the dot2 plot dropped down to 5th. However, this drop in performance was driven by the low accuracy with which the participants were able to estimate the minimum and maximum projected temperature change, not the mean. It is likely that this drop in accuracy occurred as a result of the uncertainty in the model outputs being represented using error bars (Figure 6.1 and F.4). As the error bars were determined using the standard deviation of the projections, it would not have been possible for the participants to correctly identify the minimum and maximum, although they may have been able to provide a rough approximation of these values. Nevertheless, the dot2 plot ranked highly when used to view changes in uncertainty over time and to retrieve specific values, whilst fewer than 10% of the participants that were shown the dot2 plot indicated that they were unable to estimate the minimum and maximum. Those participants that did try to give an estimate may have misinterpreted what the error bars represented. For example, it is possible that some of the participants assumed that the caps of the error bars depicted the minimum and maximum projected temperature change, which would have resulted in an overestimation of the minimum and an underestimation of the maximum. Such misinterpretations have been widely recognised in the scientific literature (see Belia et al. (2005); Cumming et al. (2007); Hullman et al. (2015) for example), with some researchers suggesting that error bars may in fact be harmful (Correll and Gleicher, 2014). This sentiment is supported by the results of this research, which showed that the participants' ratings of confidence and ease remained relatively high despite the aforementioned drop in accuracy. The dot2 plot may therefore be unintentionally misleading, a potentially dangerous trait given that it is likely to result in an underestimation of the uncertainty in the model outputs.

A similar issue might also be expected to affect the accuracy with which the participants were able to interpret the box1 plot, which included error bars that represented the minimum and maximum projected temperature change that was within 1.5 × the inter-quartile range (Figure 6.1 and F.5). However, this was not the case, perhaps due to the relatively small difference between the upper and lower limits of the error bars and the 'true' minimum and maximum given by the climate models. Had the difference been larger, we may still be less concerned about the box1 plot than the dot2 plot as it was ranked lower for confidence and ease than for accuracy, but unfortunately this drop in ranking was driven by an increase in the number of negative ratings (i.e. 'disagree' or 'somewhat disagree') when the participants were asked to estimate the mean, not the minimum and maximum. Such low ratings for confidence and ease likely occurred as a result of confusion about whether the question, which included the term 'average', required the mean or the median. Based on the extended feedback given by the participants, it is clear that some tried to estimate the mean using the mid-point of the 25th and 75th percentiles, whilst others gave the median. Again, this issue would have had a relatively minor impact on the overall ranking of the box plots as the difference between the mean and the median of the model projections was small, but may have been more problematic if the difference had been much larger. To avoid this issue in the future, the type of average used in the visualisation should be made as clear as possible. We also recommend providing definitions of the mean, median, and mode to ensure that misinterpretations are avoided.

Overall, it is clear that caution must be exercised when using error bars to communicate uncertainty in the outputs of MMEs. It should not be assumed that the end users will be able to estimate the full extent of the uncertainty in the model outputs unless the minimum and maximum are made explicit. Despite these issues, the dot and box plots were ranked in the top four overall. The success of these visualisations is supported by a survey aimed at identifying the most effective visualisations for communicating uncertain snowfall forecasts (Ibrekk and Morgan, 1987). In this study, Ibrekk and Morgan (1987) found that the participants were most accurately able to estimate the forecaster's single 'best' estimate when using a dot or box plot (Ibrekk and Morgan, 1987). However, the performance of these visualisations was compared with a bar chart, a pie chart, a box plot, a conventional probability density function, a conventional cumulative distribution function, and three less widely used representations of a probability density function (Ibrekk and Morgan, 1987). None of the visualisations depicted more than one type of uncertainty and more modern visualisations methods, such as infographics, were not considered (Ibrekk and Morgan, 1987). The survey participants also consisted of 45 individuals from a single environmental education facility near Pittsburgh and may thus not be considered a representative sample of the population (Ibrekk and Morgan, 1987). Our results therefore further the work of Ibrekk and Morgan (1987) by highlighting the effectiveness of the dot and box plots across a much wider audience, across a greater number of measures of performance, and when communicating more than one type of uncertainty.

The two line plots that were included in the survey consistently displayed an intermediate level of performance, with rankings of between 3rd and 7th for accuracy, confidence, ease, and preferences. It is perhaps surprising that the line plots were not ranked higher given that they are generally believed to be easy to interpret and have previously been proven to be successful at conveying trends in environmental variables over time (Lipkus and Hollands, 1999; Spiegelhalter et al., 2011). However, it would seem that the relatively low rankings of the line plots occurred as a result of the participants struggling to quantify the uncertainty in the model outputs. In particular, the participants tended to find it more difficult and were less confident when they were asked to use the line plots to estimate the minimum and maximum projected temperature change instead of the mean. This finding is especially interesting given that the opposite was true for almost all of the other visualisations. Such a drop in the performance of the line plots when the participants were asked to estimate the minimum and maximum likely occurred as a result of over-plotting, a factor that can make it more difficult or even impossible to identify specific values (Few, 2008). This would also explain why the line plots performed relatively poorly in many of the preference categories, particularly the ability to view changes in uncertainty over time, the ability to retrieve specific values, and overall ease of understanding, but performed well when used simply to view changes in temperature over time. This theory is supported by a similar visualisation survey conducted by Daron et al. (2015), which suggested that over-plotting may have impacted the participants' ability to assess future changes in rainfall. The line plots may therefore perform better in situations where there are fewer model runs to display or when the scenarios diverge more dramatically. A different colour scheme and/or bolder lines may also help to improve the performance of the line plots by making the differences between each of the model runs more obvious.

Despite the participants being highly confident when using the line plots to estimate the mean projected temperature change, the mean absolute difference (x10) between the participants' estimates and the 'true' values given by the climate models was relatively high for these two visualisations. Nevertheless, the 95% confidence intervals of the parameter estimates associated with the line plots overlapped with all of the other visualisation types excluding the heat plot, suggesting that most of the visualisations performed equally well when the participants were asked to estimate the mean. However, we ranked the visualisations based solely on the mean parameter estimates of the Generalised Linear Mixed Models (GLMMs) and did not take into account the error surrounding these estimates. Negligible differences in the performance of each visualisation may thus have a substantial impact on the final rankings. The final rankings should therefore not be considered alone, they must be used in conjunction with the outputs of the underlying statistical models that were used to analyse the survey data.

The cascade plot showed perhaps the greatest variation in rankings across the different measures of performance, placing in 3rd and 4th for accuracy and confidence/ease respectively, but in 9th for preference. Despite ranking highly for confidence and ease it is clear that the participants found it relatively difficult to estimate the mean using this visualisation. This is perhaps unsurprising given that the visualisation was specifically designed to represent changes in uncertainty, rather than to represent changes in the mean (Hawkins, 2014), a fact that is supported by the comparatively good performance of the cascade plot when used to view changes in uncertainty over time. The cascade plot may have performed poorly across many of the other preference categories (particularly the ability to view changes in temperature over time, visual appeal, and overall ease of understanding) as preferences have often been shown to be strongly related to familiarity (see Elting et al. (1999); Lorenz et al. (2015); Quispel et al. (2016) for example). As the cascade plot was the least familiar of all of the visualisations, it is therefore more likely to be one of the least preferred visualisations. Increased exposure to this type of visualisation may therefore help to improve the performance of the cascade plot across many of the different preference categories, particularly overall ease of understanding and visual appeal.

The radar plot performed poorly across all measures of performance, with rankings consistently at or below 7th position. Again, such poor performance may be relatively unexpected given that radar plots are typically used to display multivariate data (see Saary (2008); Vaughan and Gough (2016) for example), rather than to communicate changes in a single environmental variable through time. Similar to the cascade plot, the radar plot was also one of the most unfamiliar visualisations in the survey and would therefore be more likely to be one of the least preferred visualisations. Furthermore, the cognitive load (or mental effort) required to interpret the radar plot may be far greater than for many of the other visualisation types due to the axes pointing in different directions (Peltier, 2013). As increased cognitive loads have previously been shown to have a negative impact on the accuracy with which individuals may interpret a visualisation (Lohse, 1997), this issue may help to explain the negative response of the participants to the radar plot.

The infographic was consistently found in the bottom three when the visualisations were ranked according to their performance in each category. However, it is likely that the infographic performed poorly as we purposefully chose not to provide enough information in this visualisation to estimate the uncertainty in the projections in the 2010s or 2050s. The participants would also not have been able to estimate the minimum and maximum projected temperature change in the 2090s to the required level of accuracy (one decimal place); this is because the uncertainty was presented as a frequency table depicting the number of models projecting increases of more than 2, 3, 4, and 5°C by 2099 (Figure 6.1 and F.10). We did this to test whether the participants were correctly able to identify this fact or whether they attempted to extract the minimum and maximum using the information provided. 40% of the participants that were shown the infographic indicated that they were unable to estimate the minimum and/or maximum (see Appendix G); those that did try to extract specific values often provided integers (i.e. 2, 3, 4, and 5°C) as a 'best guess' approximation of the minimum and/or maximum and set their ratings of confidence and ease to disagree or somewhat disagree. Nevertheless, approximately 50% of those who provided estimates of the minimum and maximum gave non-integer estimates, suggesting that there may have been some ambiguity in the phrasing of the question that resulted in a different interpretation than what was intended. However, studies have shown that seemingly straightforward questions can be interpreted by different respondents in very different ways (see Suessbrick et al. (2000) for example), suggesting that this issue is not unique to this survey. It is also possible that the participants misinterpreted the visualisation itself. For example, some may have used the information given on the mean projected temperature change instead of the minimum and maximum.

Overall, the lack of detailed information regarding the uncertainties in the projections would have negatively impacted the accuracy, confidence, and ease with which the participants were able to interpret the infographic, as well as their preferences for this visualisation. Despite this, the performance of the infographic improved when the participants were asked to estimate the mean projected temperature change. Feedback from the survey participants suggested that the performance of the infographic could have been further improved by providing grid lines that could be used to more accurately determine the mean. The infographic also ranked relatively highly for visual appeal, highlighting the power of this type of visualisation to grab the attention of an audience despite not necessarily communicating the message clearly. A similar issue occurred when the National Hurricane Center (NHC) developed the 'cone of uncertainty' infographic to depict the probable tracks of hurricanes in the North Atlantic in 2004 (Broad et al., 2007). This infographic became widely used in the media in Florida and many people claimed that their decision to evacuate was heavily influenced by this visualisation (Broad et al., 2007). However, a post-hurricane survey by the NHC indicated that many individuals misinterpreted the visualisation and either underestimated or ignored the uncertainty in the infographic (Broad et al., 2007). Such potential for misinterpretations has important implications for the use of infographics in the communication of MMEs in the future; each infographic must be carefully designed to ensure that the visualisation is engaging but that the intended message is delivered with clarity and integrity. Achieving a balance between visual appeal, clarity, and integrity may be best achieved through interdisciplinary collaborations between natural scientists, social scientists, and graphic designers or communication experts (Grainger et al., 2016). The end users should also be included in such collaborations to ensure that the impact of the visualisation can be maximised, whilst misinterpretations can be minimised (Grainger et al., 2016).

Finally, the heat plot was deemed to be the worst visualisation across all measures of accuracy, confidence, ease, and preferences. Such poor performance may be driven by the difficulties associated with extracting specific values from a continuous scale bar (Few, 2017). This issue may also be exacerbated by the use of a sequential colour scale that had relatively little variation in colour between the minimum and maximum values. A diverging colour scheme may have improved the performance of the heat plot by dividing the scale bar into three easily identifiable regions (low, medium, and high), thus providing more visual cues with which to interpret the visualisation (Moreland, 2009).

6.5.2 Demographics, previous encounters, decade, and scenario

The demographics of the participants, including their background, level of education, and level of expertise in working with environmental models and/or their outputs, were taken into account to identify whether different groups of people were better able to interpret certain visualisations or whether their preferences differed. The familiarity of the visualisation, as well as the scenario and decade given to each participant, were also taken into consideration. The decade that was given to each participant often had a significant effect on the accuracy, confidence, and ease with which they were able to estimate the mean and/or the minimum and maximum projected temperature change. In most cases, the participants were more accurate, more confident, and found it easier to interpret the visualisations in the 2090s compared with the 2010s and 2050s. This is perhaps surprising given that the 2090s were often furthest away from the axes labels, potentially making it more difficult to follow the grid lines (where applicable) to the correct value. Nevertheless, it is likely that the increased spread in the model outputs during the 2090s made it easier for the participants to differentiate between the mean, the minimum, and the maximum. This increase in spread would have been particularly useful when the participants were asked to interpret the line and cascade plots, which suffered more from over-plotting in the 2010s and 2050s than in the 2090s. As previously mentioned, over-plotting may make it more difficult to extract specific values from a visualisation and may therefore explain the decrease in the accuracy, confidence, and ease with which the participants were able to interpret the data in the decades where the over-plotting was greatest (Few, 2008).

Although the participants tended to be more confident and find it easier to estimate the mean, minimum, and maximum projected temperature change in the 2090s compared with the 2010s and 2050s, the accuracy with which they were able to estimate the minimum projected temperature change was lowest in the 2090s. It is possible that this drop in accuracy occurred as the minimum projected temperature change in scenario 8.5 was represented by an outlier. This outlier may not have been obvious due to over-plotting (e.g. the line1 plot) or the use of error bars without a representation of the outliers in the data (e.g. the dot2 plot). The presence of an outlier may also help to explain why the participants tended to be less accurate when they were asked to estimate the minimum projected temperature change in scenario 8.5 compared with scenarios 2.6 and 4.5. Interestingly, the participants were also less accurate when they were asked to estimate the mean projected temperature change in scenario 8.5. It is possible that the drop in accuracy associated with scenario 8.5 was additionally caused by: (1) the use of the colour red to represent scenario 8.5, as red can be difficult to distinguish on a dark background (Byrne and Braha, 2012) and; (2) the increased difference between the projected temperature change at the start and end of each decade in scenario 8.5 compared with scenarios 2.6 and 4.5. The latter may be particularly important as the extended feedback given by the participants highlighted that there was some confusion about which part of the decade to extract the values from when using the line plots.

The familiarity of the visualisations had a significant effect on the confidence and ease with which the participants were able to identify the mean, minimum, and maximum projected temperature change. This is unsurprising given that the participants would likely be better equipped to interpret a visualisation that they had previously encountered compared with one they had never seen before. Although not measured during this research, numerous studies have also linked familiarity with visualisation preferences, particularly in terms of visual appeal (see Daron et al. (2015), Alrehiely et al. (2018) and Saket et al. (2018) for example). Despite this, the final rankings of the visualisations did not always follow the ordering of the visualisations that were most familiar to the survey participants. For example, the dot1 plot performed well across all measures of confidence, ease, and preferences, but was less familiar to the participants than the dot2, box, and line plots. However, the similarities between the dot1 and dot2 plots may have made it easier for the participants to interpret the dot1 plot compared with a visualisation that showed no similarities to any of the other more familiar visualisations. The cascade plot was also the least familiar visualisation, but it performed well for accuracy, confidence, and ease; this result is important as it proves that an unfamiliar visualisation can outperform traditional visualisation types if used correctly. Nevertheless, the cascade plot performed comparatively poorly for preferences, particularly for ease of understanding and visual appeal, thus supporting the conclusion that familiarity is at least somewhat related to preferences.

The education level of the participants had relatively little impact on the accuracy, confidence, and ease with which they were able to interpret the visualisations, as well as their preferences for different visualisations across each of the five categories. Interestingly, those with doctoral training tended to be less accurate than those with pre-university, undergraduate, or postgraduate levels of education when they were asked to estimate the mean projected temperature change. These results suggest that education levels are not necessarily linked to graphic (or visual) literacy. However, it is possible that those with doctoral training simply spent less time answering the questions, resulting in rushed responses that were slightly less accurate. Unfortunately, this theory is difficult to prove as the only information we collected regarding the time taken for each participant to complete the survey was relatively unreliable (see Appendix J for further details). It is also possible that those with doctoral training may have completed a PhD in a subject that did not require the use of visualisations similar to those given in the survey. A greater amount of time may therefore have passed since they were required to interpret visualisations in this way, thus making it more difficult to accurately estimate the required values; this may further help to explain why those with doctoral training often found it more difficult to identify the minimum and maximum projected temperature change than those with a different level of education, although this may also be caused by this group of participants being more aware of their inability to correctly identify the minimum and maximum using the dot2, box1, and infographic plots.

Decision makers, environmental managers, and the general public tended to be less accurate, less confident, and find it more difficult to estimate the mean, minimum, and maximum projected temperature change than scientists. This is to be expected given that scientists likely spend far more time creating and interpreting visualisations similar to those used in the survey than the other two groups. However, it is important to note that the uncertainty surrounding the estimates of the decision makers and environmental managers was relatively large due to the small sample size associated with this group. A much larger sample size would be required to make robust conclusions about the accuracy, confidence, and ease with which decision makers and environmental managers are able to interpret these visualisations, as well as their preferences for different visualisations.

The participants' expertise in working with environmental models and/or their outputs had very little effect on the accuracy with which they were able to interpret the mean, minimum, or maximum projected temperature change. However, those with over five years of experience tended to be slightly more accurate when estimating the mean than those with less experience. Again, this is to be expected given that the participants that had more experience would likely have spent a much greater amount of time developing and interpreting visualisations similar to those included in the survey. Nevertheless, the participants that had zero to two years of experience tended to be more confident and find it easier to estimate the mean projected temperature change than those with more experience. It is therefore possible that the participants that had less experience in working with environmental models were overconfident in their ability, or that those with a greater level of experience were warier about attaching high levels of confidence and ease to their answers. These results may therefore be a prime example of the well-established 'overconfidence effect', which states that a person's subjective confidence in the correctness of their response to a given question tends to be greater than an objective measure of their accuracy (Pallier et al., 2002). Interestingly, the overconfidence effect is thought to occur only when the participants are asked to complete a task that is difficult or unfamiliar to them (Larrick et al., 2007), which may explain why those with the least experience were the most confident but not the most accurate. However, the difference in the probability of a positive rating for confidence and ease between participants with differing levels of experience was relatively small, and the same effect was not apparent when the participants were asked to estimate the minimum and maximum projected temperature change. We also did not make a direct comparison between the objective measures of accuracy and the subjective measures of confidence in the survey. Further research is therefore required to better understand the impact of the 'overconfidence effect' on the results of this research.

Generally speaking, the demographics of the participants had relatively little impact on their accuracy, confidence, ease, and preferences, especially when compared with the effect of visualisation type. However, the importance of including participant ID in the statistical models that were used to analyse the results of the survey suggests that some of the participants were inherently more accurate, more confident, or found the questions easier to answer than others. Nevertheless, there was also a substantial amount of within-treatment variation that was not explained by participant ID and was not accounted for by the predictor variables that were included in the statistical analyses. It is therefore likely that there were other underlying factors that affected the accuracy, confidence, and ease with which the participants were able to interpret the visualisations that were not taken into consideration in the survey. Such factors may include the colour perception, numeracy, visual literacy, or cognitive ability of the participants (Friel et al., 2001; Sterba and Bláha, 2015), the time available to the participant to answer the questions (Peebles and Ali, 2015), or the type and size of the screen used to complete the survey. Although background or level of education were included in the best-fitting models for each of the preference categories, it would seem that the familiarity of the visualisation may also have played an important role in determining participant preferences. Furthermore, personal values and cultural differences are likely to affect visualisation preferences, a fact that has previously been noted in other research areas, such as website and product design (Kastanakis and Voyer, 2014; Reinecke and Gajos, 2014), but were again not accounted for in the survey.

6.5.3 Limitations

One of the main limitations of this research is that the results may under-represent certain demographic groups, particularly those over the age of 55, those with GCSE and vocational qualifications, those from outside the UK, and those who consider themselves to be decision makers and/or environmental managers. The results may also only be representative of individuals that are willing to engage in research associated with climate science, as well as those who are willing to complete a relatively lengthy survey. Although a paid research panel could have been used to ensure that a certain number of individuals in each demographic group completed the survey, this would have required far greater resources than were available for this project and would not necessarily have resulted in a truly representative sample. Nevertheless, we believe that the results of this research offer a good starting point to identifying how best to present the outputs of MMEs, particularly given that the number of participants in this research was much larger than in similar visualisation surveys in other research areas (see Ibrekk and Morgan (1987), Aerts et al. (2003), and Lorenz et al. (2015) for example).

The survey presented here may also suffer from a similar issue as the survey conducted by Daron et al. (2015), as all of the visualisations were based on the same set of data. It is therefore possible that the accuracy, confidence, and ease with which the participants were able to interpret each visualisation could have improved over time, particularly if they were asked to interpret multiple visualisations in the same decade(s) and scenario(s). However, ~25% of the participants completed only one block of questions and would therefore not have been affected by this issue. Of the remaining ~75% that did complete more than one block of questions, ~63% were asked to interpret multiple visualisations in the same decade(s) and scenario(s), but few gave the same answer again. Furthermore, none of the participants highlighted the fact that the data was the same in the extended feedback portion of the survey and it is therefore unlikely that this issue would have had a noticeable effect on the results. Nevertheless, we would recommend using multiple different sets of model outputs or simulated data to avoid this issue in the future.

A further limitation of this study relates to the use of the culturally-ingrained traffic light colour scheme in all ten visualisations. As previously mentioned, the use of this colour palette may have made it more difficult for those who experience deuteranopia or protanopia (red-green colour blindness) to distinguish between the different colours and hence to differentiate between the three scenarios. Although this issue may have had a negative effect on the performance of all of the visualisations, it would likely have had the greatest effect on the line1 and cascade plots due to the overlap between the different colours in these visualisations. To overcome this issue, we provided the option for participants to request the visualisations with a more suitable colour palette if required, although none of the participants selected this option. Nevertheless, it was clear from the extended feedback of the participants that this was still an important issue for those who do not experience any form of colour blindness. We would therefore recommend avoiding this colour scheme in the future.

Additionally, we tried to include as many different types of visualisation as possible in the survey, but the list was not exhaustive and we did not consider interactive or animated visualisations. It is therefore possible that other visualisation types may have been more successful than the dot and box plots. There may also be other measures of performance that may be important but that were not considered in this research. For example, we could have measured the accuracy of the participants by asking them to identify the number of models that projected an increase in temperature of a certain amount in a given decade and scenario, or we could have asked the participants to indicate their preference for a particular visualisation based on the amount of time required to interpret the visualisation. Including these measures of performance could have changed the rankings of each visualisation substantially, but we believe the measures we chose to include in the survey were sufficient given the context of this research. Finally, it is clear from the extended feedback given by the participants that some interpreted the survey questions in a different way than what was intended, which may have biased some of the results. However, this issue is difficult to avoid and is not unique to this survey (Suessbrick et al., 2000), but it highlights the importance of thoroughly testing the questions on various different groups of people before releasing the survey more widely.

6.6 Summary and conclusions

Based on the findings of this research, it would seem that the dot and box plots may be the most effective methods for visualising the outputs of an ensemble of models that have been run under multiple different scenarios. Nevertheless, caution should be exercised when using error bars to represent uncertainty; if the end users are expected to be able to extract the minimum and maximum of the projections, then these values must be made explicit and should be clearly labelled (see the box2 plot for example). Care must also be taken when using the term 'average' as it can result in confusion about whether it represents the mean, the median, or the mode. Although line plots are frequently used to represent changes in environmental variables through time, we found that the survey participants struggled to estimate the minimum and maximum of the model outputs using this type of visualisation, but this was likely due to over-plotting. The cascade plot performed poorly for preferences and was the least familiar of all of the visualisations, thus supporting the theory that preferences are at least partially driven by familiarity. However, the cascade plot outperformed many of the more traditional methods of visualisation for accuracy, confidence, and ease, proving that unfamiliar visualisations can be more effective when used correctly. The infographic used in this research performed poorly across all measures of performance, but this was largely due to the purposeful inclusion of a design flaw that prevented the participants from being able to accurately estimate the minimum and maximum of the projections. Despite this, the infographic performed well for visual appeal, highlighting the importance of carefully designing these types of visualisations to ensure the intended message is delivered with clarity and integrity. The heat and radar plots also performed poorly across all measures of performance, suggesting that they may not be effective when depicting the outputs of multiple models run under different scenarios. However, this finding is only applicable in the context of this research and it does not mean that these visualisation methods may not be useful for other purposes in environmental modelling.

Generally speaking, the demographics of the participants had relatively little impact on their accuracy, confidence, ease, or preferences, although a much larger number of participants would be required to make robust conclusions about the differences between demographic groups. Nevertheless, it is clear that the background of the participants was important. In particular, the extended feedback given by some of the participants suggested that none of the visualisations in the survey would be appropriate for communicating with the general public. However, we would like to make it clear that the point of this survey was to determine the effectiveness of a wide range of basic visualisation types; we do not necessarily condone the use of these particular visualisations for any given purpose. Importantly, the results will enable us to develop more suitable visualisations by helping us to identify the attributes that make a visualisation effective. We can then build on these basic visualisation types to produce more suitable methods of communicating to different groups of people.

Overall, we hope to use the results of the survey to generate guidelines that can be used to improve the visualisation of MMEs across multiple areas of research, including both climate modelling and marine ecosystem modelling. Doing so will enable us to target audiences with visualisations that will both capture their interest and prevent misinterpretations of the data, as well as help to increase the societal impact of the models and ensure they are well-placed to support management decisions in the future.

Chapter 7

Discussion

The main aim of this thesis is to help improve our understanding and communication of the uncertainties associated with projections of future conditions in marine ecosystems. In doing so, we may be able to provide decision makers and the general public with a more realistic insight into the potential impacts of natural and anthropogenic change in the marine environment, thus aiding the development of robust management solutions for the future. As the outputs of environmental models tend to suffer from many different types of uncertainty, various techniques must be used to provide a detailed exploration of the impacts of all possible sources of uncertainty on the projections of a given model or set of models. Although it is not feasible to provide a comprehensive exploration of all sources of uncertainty in one thesis (see Section 7.4), I have attempted to quantify or qualitatively describe as many different types of uncertainty as possible given the available resources. To do this, I have employed sensitivity analysis, machine learning, Multi-Model Ensembles (MMEs), and an in-depth online survey to explore parameter, internal variability, model, scenario, and communication uncertainties. In this chapter, I summarise the key findings (see Section 7.1), management implications (see Section 7.2), unique contributions (see Section 7.3), and limitations of this research (see Section 7.4), before highlighting areas of future research that may be required to further advance our understanding of the uncertainties associated with projections of future conditions in marine ecosystems (see Section 7.5).

7.1 Key findings

The key findings of the research described in this thesis are as follows:

I. The outputs of two versions of a widely-used marine ecosystem model, known as the trait-based and multispecies *mizer* models, are most sensitive to the parameters associated with resource availability (e.g. σ, λ, κ, and q), feeding (e.g. α and n), standard metabolic rates (e.g. ks and p), and/or fishing effort (F) (see Chapter 3). The importance of these parameters is further supported by the results of Chapter 4, which found that the best-performing random forests tended to require information on a similar set of parameters to predict the outputs of the multispecies *mizer* model with the greatest accuracy. This finding is important as it highlights areas in which to focus future field or experimental research to reduce the uncertainties associated with these parameters and thus reduce the overall uncertainty of the model outputs. For example, further research into the size of the North Sea plankton community may help to reduce the uncertainties associated with parameters such as λ and κ, as well as the links between plankton and higher trophic level species.

II. The random forest algorithm is able to predict most of the outputs of the multispecies *mizer* model with relatively high accuracy using information on less than ten of the 306 parameters (see Chapter 4). The random forest algorithm is particularly successful at predicting species survival and community coexistence, but less successful when predicting community and species-specific population size. Being able to accurately predict the outputs of the *mizer* model reduces the need to run the full model (which may take hours to reach equilibrium) and enables us to screen potential parameter combinations for historically plausible model outputs more efficiently, thereby helping to lessen the costs associated with marine ecosystem modelling in terms of human and computational resources.

III. Internal variability and model uncertainties dominate the total variance of the projections of Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) from a state-of-the-art climate MME over the next few decades, with scenario uncertainty becoming increasingly more important in the mid- to late part of the century (see Chapter 5). Uncertainties in the projections of SST exhibit a strong latitudinal gradient in the mid- to late part of the century, with scenario uncertainty dominating in tropical and temperate regions and model uncertainty dominating in the polar regions. No such latitudinal gradient exists for projections of SSS, with model uncertainty dominating in almost all regions in the mid- to late part of the century. Projections of SST are most robust in the early- to mid-part of the century, particularly in the tropics. Projections of SSS are far less robust, especially in the New Zealand and Guinea coastal provinces and the South Pacific subtropical gyre. Again, these findings are important as they enable us to identify areas in which to focus future research to reduce the uncertainties in the projections. For example, the results of Chapter 5 highlight the importance of improving the representation of global water circulation in order to help reduce the model uncertainty associated with projections of SST and SSS.

IV. The demographics of an individual has relatively little effect on the accuracy, confidence, and ease with which they are able to interpret different methods of visualising the outputs of MMEs and their uncertainties, as well as their preferences for different visualisation types (see Chapter 6). Dot and box plots may be the most effective methods for visually communicating the outputs of MMEs to a wide audience, but caution should be exercised when using error bars to represent the uncertainty in the model outputs. Line plots may also be effective when overplotting is not an issue. Less familiar visualisations, such as cascade plots, may be effective in some instances, but are likely to be less preferred by the end users than some of the more traditional methods of data visualisation. Conversely, end users may display increased preference for an infographic due to their visual appeal despite not necessarily communicating the intended message clearly. It is therefore extremely important to ensure infographics are carefully designed to maximise the impact of the visualisation, whilst minimising the potential for misinterpretations. Finally, heat and radar plots may not be effective at communicating the outputs of MMEs. Together, these findings are vital to improving the communication of MMEs and their uncertainties to decision makers and the general public in the future.

7.2 Management implications

The research described in this thesis has implications for fisheries management in the future. At present, single-species models are one of the main tools used by ICES to provide short-term tactical fisheries management advice to the EU. However, single-species models usually do not take into account the fundamentally-important dynamics of lower trophic level species (i.e. phytoplankton, zooplankton, and non-predatory fish species) and their interactions with higher trophic level predators (Lynam et al., 2016; Stäbler et al., 2019) as required for ecosystem-based fisheries management. Such interactions typically influence the productivity of an ecosystem over longer timescales than are currently accurately captured by single-species models (Lynam et al., 2016; Stäbler et al., 2019). Ecosystem models that are capable of taking into account these interactions may therefore be more informative than single-species models at longer timescales and may thus be useful for strategic management advice in the future (Lynam et al., 2016). However, the uptake and use of marine ecosystem models and their products has been limited in UK and EU policy development in the past due to a lack of trust and understanding of the behaviour of these models, as well as a lack of visibility, availability, and user-friendliness of the models and their products (Hyder et al., 2015; Lynam et al., 2016).

To improve uptake, we need to increase the reliability, credibility, and visibility of marine ecosystem models both through model development and through improved communication between modellers and decision makers (Hyder et al., 2015). We also need to be transparent about the limitations and uncertainties of the models. Uncertainty is of particular importance given that there are a wide variety of different types of uncertainty in marine ecosystem modelling, many of which are large and can have wide-ranging impacts on the projections. If we do not deal with these uncertainties explicitly, we risk making uninformed management decisions that put the ecosystem at risk of stock collapse. The results of the sensitivity analysis and machine learning conducted in Chapters 3 and 4 thus act as an important starting point to providing more transparent information to scientists, decision makers, and the general public regarding the behaviour of marine ecosystem models such as *mizer* under various parameter combinations, as well as the potential impacts of parameter uncertainty on the projections of the model. Chapters 3 and 4 also highlight important interactions between species, which can be useful information when prioritising components of the ecosystem for management purposes (Bentley et al., 2019b). Although these chapters focus solely on one marine ecosystem model, the methods can be applied to any marine ecosystem model. It is therefore hoped that the research presented in this thesis will encourage those working with other marine ecosystem models to be more transparent about the uncertainty in their model outputs.

Additionally, Chapters 3 and 4 involved an exploration of the impacts of parameter uncertainties on various proposed indicators of ecosystem health, including the Large Fish Indicator (LFI) and Mean Maximum Length (MML). These indicators are extremely important in the decision-making process as they enable us to monitor the environment, detect the impacts of a given policy or management measure on an ecosystem (Reed et al., 2016; Shin et al., 2018), and effectively communicate this information to the general public (Halouani et al., 2019). The use of such indicators is mandated by national and international legislation, many of which call for an ecosystem-based approach to fisheries management, including the EU Common Fisheries Policy Directive (CFP; European Commission (2013)) and the EU Marine Strategy Framework Directive (MSFD; European Commission (2008b)) (Hyder et al., 2015; Meier et al., 2019). However, very little is known about the robustness of different indicators of ecosystem health, particularly in terms of their specificity to various drivers of change (Shin et al., 2018).

For an indicator to be effectively able to detect human-induced changes in the environment, it should be (predictably) responsive to changes in drivers such as fishing but relatively unresponsive to other drivers such as changing environmental conditions (Shin et al., 2018). Although not initially designed to determine the specificity of different indicators of ecosystem health, the sensitivity analysis and machine learning conducted in Chapters 3 and 4 suggest that size-based indicators, such as the LFI, mean weight, and community slope, are sensitive to fishing effort, the size of the background resource (phyto- and zooplankton) and standard metabolic rates in the North Sea. These results are important in a management context as it suggests that some of the indicators that have been selected to detect changes in fishing also respond to parameters that are heavily impacted by changes in environmental conditions. This kind of research can only be conducted using marine ecosystem models such as *mizer* as we are unable to conduct ecosystem-wide experiments to determine the effects of varying fishing and environmental conditions in the field (Hyder et al., 2015; Halouani et al., 2019). The research described in Chapters 3 and 4 thus enables a better understanding of the indicators that may be best suited to supporting fisheries management in the future (Shin et al., 2018). However, further research is required to: (1) determine how long it takes for an indicator to respond different drivers to ensure management decisions can be made on appropriate timescales (Shin et al., 2018); and (2) identify indicator thresholds that can be used to trigger specific management actions (Halouani et al., 2019).

Chapter 5 involved an assessment of the differing contributions of internal variability, model, and scenario uncertainty to the total variance of two key parameters for environmental policy. By partitioning the uncertainty in the model outputs and by quantifying the signal-to-noise ratio of the projections, we can provide managers with a better understanding of the confidence in the projections (Payne et al., 2015; Lynam et al., 2016). This type of analysis, along with the analyses described in Chapters 3 and 4, also supports management practices by highlighting areas in which we can reduce uncertainty through increased data collection and/or model development. For example, increased availability of long-term global salinity may help to reduce both internal variability and model uncertainty in projections of future SST and SSS across the globe. By knowing where to focus future research we may be able to reduce the costs associated with managing the marine environment. This information may be particularly useful at present given that the funding available for observing and managing the marine environment has decreased in relative terms in recent years (Hyder et al., 2015), whilst the costs associated with both collecting data at sea and implementing and enforcing various management measures remains high (Mangin et al., 2018; Murray et al., 2018). By also identifying locations and time periods in which the projections are most certain, the results of Chapter 5 can be used to highlight areas that may be good candidates for immediate adaptation planning (Hawkins and Sutton, 2009). Areas with larger signal-to-noise ratios may require more expensive adaptation plans that include some level of tolerance for more extreme events (Hawkins and Sutton, 2009). However, the analysis described in Chapter 5 can be used to determine how much of the uncertainty in these locations is potentially reducible. This information can then be used by managers to determine whether it may be more cost-effective to invest in data collection and/or model development or to implement more expensive but highly-tolerant adaptation plans.

Finally, Chapter 6 involved an exploration of the effectiveness of various methods of visually communicating uncertainty to different audiences, including decision makers and the general public. Communication remains an important barrier between specialist and non-specialist audiences and it can have a huge impact on the way in which decisions are made (Hyder et al., 2015; Lynam et al., 2016). Marine ecosystem models can be extremely complex, making it difficult to visualise the outputs, particularly when attempting to incorporate various different types of uncertainty in the visualisation. In the past, a lack of effective communication of uncertainty has been blamed for ineffective management decisions (Janssen et al., 2005) and has contributed to public distrust of scientific evidence, particularly in regards to climate science (Frewer, 2004). The results of Chapter 6 are therefore vital to improving the communication of uncertainties to non-specialist audiences and ensuring these models can make a significant contribution to the decision-making process in the future.

Based on the results of the survey in Chapter 6, specific recommendations for visualising the outputs of complex environmental models and their associated uncertainties might include the following:

I. Choose colour schemes wisely. Avoid using colour schemes that might be difficult to interpret for those who experience colour blindness. Avoid using a black background if using other colours that will be difficult to distinguish against a dark background.

II. Consider using both familiar (e.g. dot, line, and box) and unfamiliar (e.g. cascade) visualisation types. Familiar visualisations may maximise uptake, but new methods of visualisation may outperform more familiar techniques in some circumstances. However, new methods of visualisation should be widely tested before implementation.

III. Use dot or box plots when developing visualisations that require the users to extract specific values from the visualisation.

IV. Use infographics to grab the attention of an audience but ensure it is designed in a way that avoids potential misinterpretations.

V. Use line plots only when overplotting is unlikely to be an issue.

VI. Do not use radar plots when attempting to communicate changes in an environmental variable through time.

VII. Do not use heat plots that have a sequential colour scale with relatively little variation in colour between the minimum and maximum values if you want users to be able to extract specific values from the visualisation.

VIII. If using error bars, label them carefully on the visualisation and provide a detailed description of what the error bars represent.

IX. If communicating to non-specialist audiences, the term "average" can be used for accessibility, but this should be followed by a statement that indicates whether this represents the mean or median to avoid confusion or misinterpretations.

Overall, it is hoped that the research described in this thesis can be used to help build confidence and trust in marine ecosystem models such as *mizer* through increased transparency, improved model behaviour (as a result of data collection and/or model development), and better communication between specialist and non-specialist audiences. However, there is still much to be done before marine ecosystem models can be fully incorporated into the management process alongside single-species models. The modelling community must make a concerted effort to quantify and communicate uncertainties, whilst decision makers must clearly communicate their requirements and the role that uncertainties play in policy formulation to ensure both sides of the science-policy interface are working synergistically towards the same goal.

7.3 Advances on the state of the art

The research described in this thesis provides a unique contribution to science in various different ways. For example, as far as we are aware Chapter 3 is the first to apply a derivative-based global sensitivity analysis to a complex marine ecosystem model. It is also the first to provide a direct comparison of the derivative-based and Sobol' variance-based sensitivity indices of an environmental model with more than 20 parameters, thus helping to confirm that the derivative-based method can be successfully used to estimate the upper bounds of the Sobol' variance-based sensitivity indices of a complex model using a relatively small number of model evaluations. Although various other methods of sensitivity analyses have been applied to marine ecosystem models in the past, they either do not apply a global approach (see Niiranen et al. (2012) and Livingston (2013) for example), do not run each model evaluation to equilibrium (see Morris et al. (2014) and Zhang et al. (2015) for example), and/or quantify the sensitivity of the model outputs to groups of parameters (see Zhang et al. (2015) for example), thereby making it difficult to attribute model output sensitivity to specific parameters. By considering the sensitivity of multiple model outputs to individual parameters based on equilibrated model evaluations, Chapter 3 provides one of the most extensive sensitivity analysis of a marine ecosystem model to date. Chapter 3 is also one of very few examples of a sensitivity analysis that includes a detailed discussion of the convergence of the sensitivity indices.

The main inspiration for Chapter 4 came from Lucas et al. (2013), who used a machine learning algorithm known as a Support Vector Machine to successfully quantify and predict the probability of simulation crashes in climate modelling. Although applied in different contexts, the concept is the same in Chapter 4 as it is in Lucas et al. (2013) - we both use information on the parameter values that result in certain model behaviours (i.e. simulation crash versus no simulation crash and coexistence versus extinction for example) to train a machine learning algorithm to predict the behaviour of the model under unseen parameter combinations. Perhaps surprisingly, the work of Lucas et al. (2013) is the only published example that we are aware of that uses machine learning in this way. Other examples of machine learning in marine science focus on analysing images, videos, and acoustic recordings (see Mahmood et al. (2016) and Abadi (2018) for example), determining the ecological status of a given area (see Cordier et al. (2017) for example), spatial planning (including habitat mapping and the analysis of conflicting uses; see Galparsoro et al. (2015) and Coccoli et al. (2018) for example), and filling in gaps in fisheries data (see Fernandes et al. (2015) for example) (ICES, 2018). Chapter 4 may therefore be the first of its kind in marine science, but the methods may be applicable to models from across a wide range of research areas.

Chapter 5 was based on the work of Hawkins and Sutton (2009, 2011), who developed the method used to quantify the relative contributions of internal variability, model, and scenario uncertainties to the total variance of the projections of a MME and applied it to projections of surface air temperature and precipitation. This method has become increasingly popular in recent years due to its relative simplicity and effective visualisation of the results, with recent applications including projections of SST (Villarini and Vecchi, 2012; Cheung et al., 2016), sea level (Little et al., 2015), tropical storm frequency (Villarini and Vecchi, 2012), and the strength of the Atlantic Meridional Overturning Circulation (Reintges et al., 2017). Chapter 5 builds on this growing body of literature by applying the methods of Hawkins and Sutton (2009, 2011) to global and regional projections of SST and SSS. Although Villarini and Vecchi (2012) and Cheung et al. (2016) have previously applied a similar method to projections of SST, the former focused solely on the tropics, while the latter focused on global means alongside two basin-scale examples from the Northeast Atlantic and Northeast Pacific. Chapter 5 thus expands on this work by focusing on projections from across the globe and at a much finer spatial resolution than in past literature. Chapter 5 is also the first example that we are aware of to apply the methods of Hawkins and Sutton (2009, 2011) using Longhurst's widely accepted partitioning of the global ocean into biogeographical provinces (Longhurst, 2007), which is perhaps a more relevant way to delineate the ocean than simply dividing the globe into rectangles of equal size (as was the case in Hawkins and Sutton (2009, 2011) for example). Finally, Chapter 5 further adds to the work of Villarini and Vecchi (2012) and Cheung et al. (2016) by quantifying the signal-to-noise ratio of the projections in different regions, thus allowing us to identify regions and time periods in which the projections are most certain and are thus most useful to decision makers in terms of adaptation planning (Hawkins and Sutton, 2009). Chapter 5 therefore provides a unique perspective on spatio-temporal changes in the contributions of internal variability, model, and scenario uncertainties to the total variance of the projections of two key parameters for marine environmental policy and the results are extremely relevant to both decision makers and the scientific community.

Finally, the visualisation survey in Chapter 6 took inspiration from a number of sources, most notably the research of Ibrekk and Morgan (1987) and Daron et al. (2015), as well as the literature review produced by Kinkeldey et al. (2014). Although it is clear from these sources that somewhat similar visualisation surveys have been conducted before, these surveys typically either focus on the ability of individuals from similar backgrounds (usually highly educated individuals) to interpret visualisations that depict uncertainty (see Ibrekk and Morgan (1987) for example), they do not take into account more than one type of uncertainty (see Ibrekk and Morgan (1987) and Daron et al. (2015) for example), they focus on geospatial information and maps (see Kardos et al. (2007) for example), and/or they take into account only a limited set of performance measures (see Kardos et al. (2007) for example). The visualisation survey in Chapter 6 thus fills an important gap in the literature by measuring the accuracy, confidence, and ease with which participants from different backgrounds are able to interpret 10 different visualisations, all of which depict multiple types of uncertainty in the outputs of MMEs, as well as the participants' preferences for each visualisation across a number of different categories. Chapter 6 therefore likely represents one of the most in-depth analyses of the effectiveness of different methods of visualising the outputs of MMEs that is currently available.

Overall, Chapters 3 to 6 each provide a unique contribution to marine and climate science. With the results being specific either to the *mizer* model or the CMIP5 climate MME, it may seem that each chapter has a relatively narrow scope. However, the implications of the research described in Chapters 3 to 6 may be far more wide-ranging than might first appear, with potential applicability to a huge range of different models. For example, the sensitivity analyses and machine learning in Chapters 3 and 4 prove that computationally-intensive analyses that were once considered to be too time-consuming are now possible with the use of High Performance Computing (HPC) - this conclusion is not only relevant in marine ecosystem modelling, but also in any research area that involves the use of complex models. The methods described in Chapters 5 and 6 may also be applied to MMEs from various fields of research. The results of the survey may have a particularly wide-ranging impact as it is unlikely that the conclusions of this research would change substantially when using a different

MME; the results may thus be used to improve communication between scientists, decision makers, and the general public across a number of different research areas. The scientific contributions of each chapter are therefore not limited solely to their direct application, they have much broader implications for modelling in general.

7.4 Limitations

As the limitations of the research described in each chapter have already been discussed in detail previously, we provide only a brief summary of the most important limitations of each chapter below. An overview of the limitations of the thesis as a whole are then described at the end of this section.

The main limitation of the sensitivity analyses and machine learning described in Chapters 3 and 4 was driven by the high computational cost associated with running the trait-based and/or multispecies *mizer* model(s) under many different parameter combinations. Because of this issue, we were limited in the number of model evaluations that we could realistically complete in the time frame available. Ideally, we would need a much greater number of model evaluations to fully explore the parameter space of the model(s), to ensure the sensitivity indices reached convergence, and to better train the random forest algorithm to predict the outputs of the multispecies *mizer* model. However, the parameter rankings and/or screenings did reach convergence for most of the model outputs considered in Chapter 3 and the random forest algorithm was successfully able to predict many of the outputs considered in Chapter 4, suggesting that the number of model evaluations in each chapter was sufficient to provide at least a good starting point with which to explore the impacts of parameter uncertainties on the model outputs.

In Chapter 5, we likely underestimate the relative contribution of internal variability to the total variance of the projections of SST and SSS. This is because by smoothing the projections with a fourth-order polynomial we effectively remove any internal fluctuations in climate that act over longer time periods (i.e. more than 15 to 30 years) (Deser et al., 2014). We also do not take into account the fact that internal variability may increase over time (Boer, 2009). Additionally, we likely underestimate the contribution of model uncertainty as we assume that each model in the ensemble is independent and the selected models represent the full spread of all possible models (Hawkins and Sutton, 2011). Nevertheless, the methods used are expected to give a qualitatively robust approximation of the uncertainties in the available projections, particularly over the next few decades (Hawkins and Sutton, 2009, 2011).

The main limitation of the survey in Chapter 6 is that the results may under-represent certain demographic groups, particularly those over the age of 55, those with GCSE and vocational training, and those from outside the UK. However, it is difficult to obtain a representative sample in any survey, particularly one that is relatively lengthy. Furthermore, it is possible that there are other types of visualisation that we did not consider in the survey but that might have performed better than the dot and box plots. There may also be additional measures of performance that are important but that we did not take into account in the survey, such as the time taken to interpret certain aspects of the visualisation. Despite this, we believe the results again offer a good starting point with which to identify how best to communicate the outputs of MMEs to decision makers and the general public, particularly given that the visualisations deemed to be the most and least effective methods were consistent across all measures of performance.

As previously suggested, the thesis as a whole is limited by our inability to describe all of the possible sources of uncertainty in marine ecosystem modelling. However, doing so would be an enormous (if not impossible) task for any research group and would require a huge amount of resources, likely on a similar scale to those used by the Intergovernmental Panel on Climate Change (IPCC), which makes use of the expertise of over a thousand contributors (IPCC, 2014b). A full exploration of uncertainty would also not be possible given the presence of 'deep uncertainties', which (by definition) we are ignorant of at present (Spiegelhalter and Riesch, 2011). A further limitation of this research is that we were forced to make use of a climate MME in Chapters 5 and 6 instead of a marine ecosystem MME. The results of the research presented in these chapters may therefore differ when applied to marine ecosystem models. However, we purposefully chose to focus on SST and SSS in Chapter 5 to ensure the results were still directly relevant to marine science. It is also unlikely that the results of the visualisation survey would substantially differ when using a marine ecosystem MME, but further research is required to confirm this theory (see Section 7.5 below).

7.5 Future research

As previously mentioned, one of the main limiting factors of this research was a lack of computing power. Assuming infinite computing power, we could improve the research in Chapter 3 by increasing the number of model evaluations that were used to estimate the sensitivity indices to ensure convergence is reached. Although other methods of sensitivity analysis have been applied to various marine ecosystem models in the past (see Morris et al. (2014) for example), the same method of sensitivity analysis should be applied to as many models as possible (e.g. Ecopath with Ecosim (EwE), StrathE2E, and FishSUMS) to allow for a direct comparison between each model and to ensure that we focus future data collection in areas in which it is possible to make the greatest reductions in uncertainty across all of the available models. Although infinite computing power would reduce the need to use machine learning to predict the outputs of the *mizer* model (or any other marine ecosystem model), similar explorations of the modelled interactions between species, as well as the areas of the parameter space that drive certain model behaviours, would still be required to ensure the models accurately represent the ecosystem in question.

By conducting further field and experimental research to better understand the modelled ecosystem, we may also be able to place more informative distributions on the parameters of the model(s), rather than using a uniform distribution with upper and lower limits of $\pm$ 10% of the nominal value. Assuming the model(s) are considered to be appropriate representations of the real world, we can then begin to quantify the probability of certain events occurring under different scenarios of the future. However, doing so also requires a concerted effort towards developing realistic scenarios that can be run in all of the available models. By running a number of different models under each scenario, we may then be able to apply the methods described in Chapters 5 and 6 to a marine ecosystem MME rather than a climate MME.

In regards to the visualisation survey in Chapter 6, there is a huge potential for further research into the possible benefits of using interactive and animated visualisations to communicate the outputs of MMEs to non-specialist audiences. Such research may include another online survey or one-on-one interviews and group sessions with individuals representing different audiences with diverse backgrounds. Although a number of marine ecosystem models in the UK are not spatially-explicit at present, including *mizer*, there is an increasing interest in using models with a spatial component to inform fisheries management (Grüss et al., 2019). Further effort should therefore also be placed into identifying how best to communicate the outputs of spatially-explicit models and their associated uncertainties using static, interactive, and animated maps.

By incorporating the information gleaned from this thesis and from any future research into the uncertainties associated with marine ecosystem modelling, we may be able to develop a comprehensive uncertainty matrix (see Walker et al. (2003) for example) that states the importance (or level; e.g. high, medium, or low) of each type of uncertainty and the methods that have been used or are required to describe and/or reduce the impacts of these uncertainties on the model outputs (Hamel and Bryant, 2017). The uncertainty matrix could then be used to not only track the progress of efforts to reduce the uncertainties in marine ecosystem modelling, but also to ensure that appropriate and consistent methodologies are applied to different models (Hamel and Bryant, 2017). A simplified version of the uncertainty matrix may also be used to further improve the communication of uncertainties to decision makers and the general public (Hamel and Bryant, 2017), a key factor to increasing the contribution of marine ecosystem models in the management process in the future.

7.6 Concluding remarks

The uncertainties in projections of marine ecosystem models have largely been ignored in the past, likely due to the high computational costs associated with both running the models and conducting the analyses required to quantitatively or qualitatively describe each source of uncertainty. By ignoring these uncertainties, we make it difficult for decision makers and the general public to place any trust in the projections given by these models, creating a barrier to incorporating marine ecosystem models into the management landscape. Such barriers have resulted in managers continuing to use single-species models, which typically do not take into account the complex relationships between different species and processes within an ecosystem (Möllmann et al., 2013). If we continue to manage the marine environment in this way, we risk suffering from severe ecological and economic consequences, including species extinctions and unintended fishery collapse, as a result of misguided management decisions (Uusitalo et al., 2015). It is therefore vital that we begin to make a concerted effort to better understand the uncertainties associated with marine ecosystem models, thus enabling us to produce more accurate projections and to slowly increase trust in these models.

In this thesis, we have made great strides towards better understanding the uncertainties in projections of future conditions in the marine environment, particularly those given by the *mizer* model. Although much of this research is focused on just one of many marine ecosystem models, it proves that computationally-intensive analyses that were once considered to be too time-consuming to apply to complex environmental models are now possible with the use of High Performance Computing (HPC). We have also shown that relatively simple methods that do not require HPC, such as those described in Chapter 5, can be used to gain a huge amount of information regarding spatio-temporal changes in the contributions of different types of uncertainty to the total variance of the projections from multiple different models. Finally, the results of the visualisation survey in Chapter 6 can now be used to improve communication between scientists, decision makers, and the general public. However, there is still much to be done before marine ecosystem models can be fully incorporated into the management process. Importantly, the onus should not be placed solely on the modellers themselves to improve the models; decision makers must also clearly communicate their requirements and preferences to scientists, as well as the role uncertainties play in policy formulation, to ensure both sides of the science-policy interface are working synergistically towards the same goal. We can then begin to use marine ecosystem models alongside single-species models to help ensure that the exploitation of marine resources remains sustainable and that the health of our marine environment is not jeopardised by ineffective management practices.

References

Abadi, S. (2018). Using machine learning in ocean noise analysis during marine seismic reflection surveys. *The Journal of the Acoustical Society of America*, 144(3):1744–1744.

Aerts, J. C., Clarke, K. C., and Keuper, A. D. (2003). Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science*, 30(3):249–261.

Agarwal, S., Abdalla, F. B., Feldman, H. A., Lahav, O., and Thomas, S. A. (2012). PkANN—I. Non-linear matter power spectrum interpolation through artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 424(2):1409–1418.

Aiello, L. M. and McFarland, D. (2014). *Proceedings of the 6th International Conference in Social Informatic. Barcelona, Spain, 11-13 November*. Springer.

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory. Tsahkadsor, Armenia, USSR, 2-8 September*, pages 267–281.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., and Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, 7(12):16398–16421.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT Press.

Alrehiely, M., Eslambolchilar, P., and Borgo, R. (2018). *Evaluating different visualization designs for personal health data*. Proceedings of the 32nd International BCS Human Computer Interaction Conference, Belfast, UK, 4-6 July.

Andersen, K. H. and Beyer, J. E. (2006). Asymptotic size determines species abundance in the marine size spectrum. *The American Naturalist*, 168(1):54–61.

Andersen, K. H., Jacobsen, N. S., and Farnsworth, K. D. (2015). The theoretical foundations for size spectrum models of fish communities. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):575–588.

Andersen, K. H. and Pedersen, M. (2010). Damped trophic cascades driven by fishing in model marine ecosystems. *Proceedings of the Royal Society B: Biological Sciences*, 277(1682):795–802.

Andersen, K. P. and Ursin, E. (1977). A multispecies extension to the Beverton and Holt theory of fishing: with accounts of phosphorus circulation and primary production. *Meddelelser fra Danmarks Fiskeri-og Havundersøgelser*, 7:319 – 435.

Araújo, J. N., Mackinson, S., Stanford, R. J., Sims, D. W., Southward, A. J., Hawkins, S. J., Ellis, J. R., and Hart, P. J. B. (2006). Modelling food web interactions, variation in plankton production, and fisheries in the western English Channel ecosystem. *Marine Ecology Progress Series*, 309:175–187.

Arhonditsis, G. B., Adams-Vanharn, B. A., Nielsen, L., Stow, C. A., and Reckhow, K. H. (2006). Evaluation of the current state of mechanistic aquatic biogeochemical modeling: citation analysis and future perspectives. *Environmental Science & Technology*, 40(21):6547–6554.

Ascough II, J., Maier, H., Ravalico, J., and Strudley, M. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, 219(3-4):383–399.

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., and Brooks, C. (2010). Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta, Georgia, USA, 10-15 April*, pages 2573–2582. ACM.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., and Grothendieck, G. (2014). Package 'lme4'. *R Foundation for Statistical Computing*.

Baum, J. K. and Worm, B. (2009). Cascading top-down effects of changing oceanic predator abundances. *Journal of Animal Ecology*, 78(4):699–714.

Beaumont, N., Austen, M., Atkins, J., Burdon, D., Degraer, S., Dentinho, T., Derous, S., Holm, P., Horton, T., Van Ierland, E., et al. (2007). Identification, definition and quantification of goods and services provided by marine biodiversity: implications for the ecosystem approach. *Marine Pollution Bulletin*, 54(3):253–265.

Becker, W., Tarantola, S., and Deman, G. (2018). Sensitivity analysis approaches to high-dimensional screening problems at low sample size. *Journal of Statistical Computation and Simulation*, 88(11):2089–2110.

Belia, S., Fidler, F., Williams, J., and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389.

Bentley, J. W., Hines, D., Borrett, S., Serpetti, N., Fox, C., Reid, D. G., and Heymans, J. J. (2019a). Diet uncertainty analysis strengthens model-derived indicators of food web structure and function. *Ecological Indicators*, 98:239–250.

Bentley, J. W., Serpetti, N., Fox, C., Heymans, J. J., and Reid, D. G. (2019b). Fishers' knowledge improves the accuracy of food web model predictions. *ICES Journal of Marine Science*.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, 16(1):41–51.

Beven, K. (1996). Equifinality and uncertainty in geomorphological modelling. In *The Scientific Nature of Geomorphology: Proceedings of the 27th Binghamton Symposium in Geomorphology. Binghamton, New York, USA, 27-29 September*.

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences Discussions*, 5(1):1–12.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1-2):18–36.

Beven, K. (2012). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience*, 344(2):77–88.

Bianchi, G., Gislason, H., Graham, K., Hill, L., Jin, X., Koranteng, K., Manickchand-Heileman, S., Paya, I., Sainsbury, K., Sanchez, F., et al. (2000). Impact of fishing on size composition and diversity of demersal fish communities. *ICES Journal of Marine Science*, 57(3):558–571.

Blanchard, J. L., Andersen, K. H., Scott, F., Hintzen, N. T., Piet, G., and Jennings, S. (2014). Evaluating targets and trade-offs among fisheries and conservation objectives using a multispecies size spectrum model. *Journal of Applied Ecology*, 51(3):612–622.

Blanchard, J. L., Dulvy, N. K., Jennings, S., Ellis, J. R., Pinnegar, J. K., Tidd, A., and Kell, L. T. (2005). Do climate and fishing influence size-based indicators of Celtic Sea fish community structure? *ICES Journal of Marine Science*, 62(3):405–411.

Blanchard, J. L., Heneghan, R. F., Everett, J. D., Trebilco, R., and Richardson, A. J. (2017). From bacteria to whales: using functional size spectra to model marine ecosystems. *Trends in Ecology & Evolution*, 32(3):174–186.

Blanchard, J. L., Jennings, S., Law, R., Castle, M. D., McCloghrie, P., Rochet, M.-J., and Benoît, E. (2009). How does abundance scale with body size in coupled size-structured food webs? *Journal of Animal Ecology*, 78(1):270–280.

Blouin, K. D., Flannigan, M. D., Wang, X., and Kochtubajda, B. (2016). Ensemble lightning prediction models for the province of Alberta, Canada. *International Journal of Wildland Fire*, 25(4):421–432.

Boer, G. J. (2009). Changes in interannual variability and decadal potential predictability under global warming. *Journal of Climate*, 22(11):3098–3109.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Borrett, S. R., Carter, M., and Hines, D. E. (2016). Six general ecosystem properties are more intense in biogeochemical cycling networks than food webs. *Journal of Complex Networks*, 4(4):575–603.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brander, K., Neuheimer, A., Andersen, K. H., and Hartvig, M. (2013). Overconfidence in model projections. *ICES Journal of Marine Science*, 70(6):1065–1068.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A., Karnon, J., Sculpher, M. J., and Paltiel, A. D. (2012). Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. *Medical Decision Making*, 32(5):722–732.

Broad, K., Leiserowitz, A., Weinkle, J., and Steketee, M. (2007). Misinterpretations of the “cone of uncertainty” in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, 88(5):651–668.

Brodlie, K., Osorio, R. A., and Lopes, A. (2012). A review of uncertainty in data visualization. In *Expanding the frontiers of visual analytics and visualization*. Springer.

Brugnach, M., Dewulf, A., Pahl-Wostl, C., and Taillieu, T. (2008). Toward a relational concept of uncertainty: about knowing too little, knowing too differently, and accepting not to know. *Ecology and Society*, 13(2).

Burgman, M. (2005). *Risks and decisions for conservation and environmental management*. Cambridge University Press.

Butenschön, M., Clark, J., Aldridge, J. N., Allen, J. I., Artioli, Y., Blackford, J., Bruggeman, J., Cazenave, P., Ciavatta, S., Kay, S., et al. (2016). ERSEM 15.06: A generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development*, 9(4):1293–1339.

Byrne, B. and Braha, Y. (2012). *Creative motion graphic titling: titling with motion graphics for film, video, and the web*. Focal Press.

Canales, T. M., Law, R., and Blanchard, J. L. (2015). Shifts in plankton size spectra modulate growth and coexistence of anchovy and sardine in upwelling systems. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):611–621.

Cantelaube, P. and Terres, J.-M. (2005). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3):476–487.

Capuzzo, E., Lynam, C. P., Barry, J., Stephens, D., Forster, R. M., Greenwood, N., McQuatters-Gollop, A., Silva, T., van Leeuwen, S. M., and Engelhard, G. H. (2018). A decline in primary production in the North Sea over 25 years, associated with reductions in zooplankton abundance and fish stock recruitment. *Global Change Biology*, 24(1):e352–e364.

Cariboni, J., Gatelli, D., Liska, R., and Saltelli, A. (2007). The role of sensitivity analysis in ecological modelling. *Ecological modelling*, 203(1-2):167–182.

Cartwright, S. J., Bowgen, K. M., Collop, C., Hyder, K., Nabe-Nielsen, J., Stafford, R., Stillman, R. A., Thorpe, R. B., and Sibly, R. M. (2016). Communicating complex ecological models to non-scientist end users. *Ecological Modelling*, 338:51–59.

Chen, W., Jin, R., and Sudjianto, A. (2004). Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. *Journal of Mechanical Design*, 127(5):875–886.

Cheng, W., Chiang, J. C., and Zhang, D. (2013). Atlantic meridional overturning circulation (AMOC) in CMIP5 models: RCP and historical simulations. *Journal of Climate*, 26(18):7187–7197.

Cheung, W. W., Sarmiento, J. L., Dunne, J., Frölicher, T. L., Lam, V. W., Palomares, M. D., Watson, R., and Pauly, D. (2013). Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems. *Nature Climate Change*, 3(3):254.

Cheung, W. W. L., Frölicher, T. L., Asch, R. G., Jones, M. C., Pinsky, M. L., Reygondeau, G., Rodgers, K. B., Rykaczewski, R. R., Sarmiento, J. L., Stock, C., and Watson, J. R. (2016). Building confidence in projections of the responses of living marine resources to climate change. *ICES Journal of Marine Science*, 73(5):1283–1296.

Christensen, J. H. and Christensen, O. B. (2007). A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, 81(1):7–30.

Christensen, R. H. B. (2015). A tutorial on fitting Cumulative Link Models with the ordinal package. https://cran.r-project.org/web/packages/ordinal/vignettes/.

Christensen, R. H. B. (2018). ordinal: regression models for ordinal data. R package version 2018.4-19. https://CRAN.R-project.org/package=ordinal/.

Coccoli, C., Galparsoro, I., Murillas, A., Pınarbaşı, K., and Fernandes, J. A. (2018). Conflict analysis and reallocation opportunities in the framework of marine spatial planning: A novel, spatially explicit Bayesian belief network approach for artisanal fishing and aquaculture. *Marine Policy*, 94:119–131.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., and Pawlowski, J. (2017). Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environmental Science & Technology*, 51(16):9118–9126.

Correll, M. and Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151.

Crain, C. M., Kroeker, K., and Halpern, B. S. (2008). Interactive and cumulative effects of multiple human stressors in marine systems. *Ecology Letters*, 11(12):1304–1315.

Cumming, G., Fidler, F., and Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, 177(1):7–11.

Curry, J. (2017). *Climate models for the layman*. The Global Warming Policy Foundation.

Cury, P. M., Shin, Y.-J., Planque, B., Durant, J. M., Fromentin, J.-M., Kramer-Schadt, S., Stenseth, N. C., Travers, M., and Grimm, V. (2008). Ecosystem oceanography for global change in fisheries. *Trends in Ecology & Evolution*, 23(6):338–346.

Daish, A. (2011). Uncertainty in models for decision making in conservation. Master's thesis, Department of Life Sciences, Imperial College London.

Daron, J. D., Lorenz, S., Wolski, P., Blamey, R. C., and Jack, C. (2015). Interpreting climate data visualisations to inform adaptation decisions. *Climate Risk Management*, 10:17–26.

Datta, S. and Blanchard, J. L. (2016). The effects of seasonal processes on size spectrum dynamics. *Canadian Journal of Fisheries and Aquatic Sciences*, 610:598–610.

De Lozzo, M. and Marrel, A. (2016). Estimation of the derivative-based global sensitivity measures using a Gaussian process metamodel. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):708–738.

Deser, C., Hurrell, J. W., and Phillips, A. S. (2017). The role of the North Atlantic Oscillation in European climate projections. *Climate Dynamics*, 49(9):3141–3157.

Deser, C., Phillips, A. S., Alexander, M. A., and Smoliak, B. V. (2014). Projecting North American climate over the next 50 years: uncertainty due to internal variability. *Journal of Climate*, 27(6):2271–2296.

Donders, W. P., Huberts, W., van de Vosse, F. N., and Delhaas, T. (2015). Personalization of models with many model parameters: an efficient sensitivity analysis approach. *International Journal for Numerical Methods in Biomedical Engineering*, 31(10):1–18.

Dormann, C. F., Schweiger, O., Arens, P., Augenstein, I., Aviron, S., Bailey, D., Baudry, J., Billeter, R., Bugter, R., Bukacek, R., et al. (2008). Prediction uncertainty of environmental change effects on temperate European biodiversity. *Ecology Letters*, 11(3):235–244.

ECA (2017). *Special Report No 08/2017: EU fisheries controls: more efforts needed*. European Court of Auditors.

Elting, L. S., Martin, C. G., Cantor, S. B., and Rubenstein, E. B. (1999). Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. *British Medical Journal*, 318(7197):1527–1531.

European Commission (2008a). Council regulation (EC) No 1005/2008 of 29 September 2008 establishing a Community system to prevent, deter and eliminate illegal, unreported and unregulated fishing, amending Regulations (EEC) No 2847/93, (EC) No 1936/2001 and (EC) No 601/2004 and repealing Regulations (EC) No 1093/94 and (EC) No 1447/1999. *Official Journal of the European Union*, 51:1–32.

European Commission (2008b). Directive 2008/56/EC of the European Parliament and of the Council, of 17 June 2008, establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). *Official Journal of the European Union*, 164:19–40.

European Commission (2013). Regulation (EU) No 1380/2013 of the European Parliament and of the Council of 11 December 2013 on the Common Fisheries Policy, amending Council Regulations (EC) No 1954/2003 and (EC) No 1224/2009 and repealing Council Regulations (EC) No 2371/2002 and (EC) No 639/2004 and Council Decision 2004/585/EC. *Official Journal of the European Union*, 56:22–61.

Fan, J., Wu, J., Kong, W., Zhang, Y., Li, M., Zhang, Y., Meng, W., et al. (2017). Predicting bio-indicators of aquatic ecosystems using the support vector machine model in the Taizi River, China. *Sustainability*, 9(6):892.

Fenchel, T. (1974). Intrinsic rate of natural increase: The relationship with body size. *Oecologia*, 14(4):317–326.

Fernandes, J. A., Irigoien, X., Lozano, J. A., Inza, I., Goikoetxea, N., and Pérez, A. (2015). Evaluating machine-learning techniques for recruitment forecasting of seven North East Atlantic fish species. *Ecological Informatics*, 25:35–42.

Few, S. (2008). *Solutions to the problem of over-plotting in graphs*. Visual Business Intelligence Newsletter.

Few, S. (2017). *Heatmaps: to bin or not to bin?* Visual Business Intelligence Newsletter.

Firth, D. (2017). qvcalc: quasi variances for factor effects in statistical models. R package version 0.9-1. https://CRAN.R-project.org/package=qvcalc.

Frewer, L. (2004). The public and effective risk communication. *Toxicology Letters*, 149(1-3):391–397.

Friel, S. N., Curcio, F. R., and Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2):124–158.

Fu, C., Travers-Trolet, M., Velez, L., Grüss, A., Bundy, A., Shannon, L. J., Fulton, E. A., Akoglu, E., Houle, J. E., Coll, M., et al. (2018). Risky business: the combined effects of fishing and changes in primary productivity on fish communities. *Ecological Modelling*, 368:265–276.

Fullerton, A. S. and Xu, J. (2012). The proportional odds with partial proportionality constraints model for ordinal response variables. *Social Science Research*, 41(1):182–198.

Fulton, E. A., Fuller, M., Smith, A., and Punt, A. (2004). *Ecological indicators of the ecosystem effects of fishing*. CSIRO Division of Marine Research, Australian Fisheries Management Authority.

Furrer, R., Sain, S. R., Nychka, D., and Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics*, 14(3):249–266.

Galparsoro, I., Rodríguez, J. G., Menchaca, I., Quincoces, I., Garmendia, J. M., and Borja, Á. (2015). Benthic habitat mapping on the Basque continental shelf (SE Bay of Biscay) and its application to the European Marine Strategy Framework Directive. *Journal of Sea Research*, 100:70–76.

Gårdmark, A., Lindegren, M., Neuenfeldt, S., Blenckner, T., Heikinheimo, O., Müller-Karulis, B., Niiranen, S., Tomczak, M. T., Aro, E., Wikström, A., et al. (2013). Biological ensemble modeling to evaluate potential futures of living marine resources. *Ecological Applications*, 23(4):742–754.

Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15):2627–2636.

Ghahramani, Z. (2004). *Unsupervised learning*. University College London, UK.

Giacomini, H. C., Shuter, B. J., and Baum, J. K. (2016). Size-based approaches to aquatic ecosystems and fisheries science: a symposium in honour of Rob Peters. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):471–476.

Gini, C. (1912). Variability and mutability. *Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*.

Giorgi, F. and Mearns, L. O. (2003). Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophysical Research Letters*, 30(12):1–4.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc.

Grainger, S., Mao, F., and Buytaert, W. (2016). Environmental data visualisation for non-scientific contexts: literature review and design framework. *Environmental Modelling & Software*, 85:299–318.

Greene, A. M., Goddard, L., and Lall, U. (2006). Probabilistic multimodel regional temperature change projections. *Journal of Climate*, 19(17):4326–4343.

Greenstreet, S. P., Rogers, S. I., Rice, J. C., Piet, G. J., Guirey, E. J., Fraser, H. M., and Fryer, R. J. (2010). Development of the EcoQO for the North Sea fish community. *ICES Journal of Marine Science*, 68(1):1–11.

Griffiths, C. (2019). *Using electronic tagging data to investigate the individual-, population- and community-level consequences of movement in free-roaming marine fish*. PhD thesis, University of Sheffield, UK.

Grüss, A., Drexler, M. D., Chancellor, E., Ainsworth, C. H., Gleason, J. S., Tirpak, J. M., Love, M. S., and Babcock, E. A. (2019). Representing species distributions in spatially-explicit ecosystem models from presence-only data. *Fisheries Research*, 210:89–105.

Guiet, J., Poggiale, J.-C., and Maury, O. (2016). Modelling the community size-spectrum: recent developments and new directions. *Ecological Modelling*, 337:4–14.

Hall, S. J., Collie, J. S., Duplisea, D. E., Jennings, S., Bravington, M., and Link, J. (2006). A length-based multispecies model for evaluating community responses to fishing. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(6):1344–1359.

Halouani, G., Le Loc'h, F., Shin, Y.-J., Velez, L., Hattab, T., Romdhane, M. S., and Lasram, F. B. R. (2019). An end-to-end model to evaluate the sensitivity of ecosystem indicators to track fishing impacts. *Ecological Indicators*, 98:121–130.

Hamel, P. and Bryant, B. P. (2017). Uncertainty assessment in ecosystem services analyses: seven challenges and practical responses. *Ecosystem Services*, 24:1–15.

Hansen, L. J. and Hoffman, J. R. (2011). *Climate savvy: adapting conservation and resource management to a changing world*. Island Press.

Hartvig, M. and Andersen, K. H. (2013). Coexistence of structured populations with size-based prey selection. *Theoretical Population Biology*, 89:24 – 33.

Hartvig, M., Andersen, K. H., and Beyer, J. E. (2011). Food web framework for size-structured populations. *Journal of Theoretical Biology*, 272(1):113–122.

Hawkins, E. (2014). The cascade of uncertainty in climate projections. https://www.climate-lab-book.ac.uk/2014/cascade-of-uncertainty/ [Accessed: 23/08/2018].

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8):1095–1108.

Hawkins, E. and Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, 37(1):407–418.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley & Sons.

Heslenfeld, P. and Enserink, E. L. (2008). OSPAR Ecological Quality Objectives: the utility of health indicators for the North Sea. *ICES Journal of Marine Science*, 65(8):1392–1397.

Hill, S. L., Watters, G. M., Punt, A. E., McAllister, M. K., Quéré, C. L., and Turner, J. (2007). Model uncertainty in the ecosystem approach to fisheries. *Fish and Fisheries*, 8(4):315–336.

Hines, D. E., Ray, S., and Borrett, S. R. (2018). Uncertainty analyses for Ecological Network Analysis enable stronger inferences. *Environmental Modelling & Software*, 101:117–127.

Holling, C. S. (1959). The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. *The Canadian Entomologist*, 91(5):293–320.

Huang, Q., Fleming, C., Robb, B., Lothspeich, A., and Songer, M. (2018). How different are species distribution model predictions?—application of a new measure of dissimilarity and level of significance to giant panda ailuropoda melanoleuca. *Ecological Informatics*, 46:114–124.

Huggett, R. (2003). *Environmental change: the evolving ecosphere*. Routledge.

Hullman, J., Resnick, P., and Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS ONE*, 10(11):e0142444.

Hwang, Y.-T. and Frierson, D. M. W. (2013). Link between the double-Intertropical Convergence Zone problem and cloud biases over the Southern Ocean. *Proceedings of the National Academy of Sciences*, 110(13):4935–4940.

Hyder, K., Rossberg, A. G., Allen, J. I., Austen, M. C., Barciela, R. M., Bannister, H. J., Blackwell, P. G., Blanchard, J. L., Burrows, M. T., Defriez, E., et al. (2015). Making modelling count-increasing the contribution of shelf-seas community and ecosystem models to policy development and management. *Marine Policy*, 61:291–302.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Ibrekk, H. and Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, 7(4):519–529.

ICES (2004). *Report of the Working Group on Ecosystem Effects of Fishing Activities (WGECO)*. ICES CM 2018/ACOM:27. Copenhagen, Denmark, 14-21 April.

ICES (2018). *Report of the Workshop on Machine Learning in Marine Science (WKMLEARN)*. ICES CM 2018/EOSG:20. Copenhagen, Denmark, 16-20 April.

IDRE (2016). *Introduction to generalised linear mixed models*. Retrieved from the Institute for Digital Research and Education, `https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/` [Accessed: 25/07/2018].

Iooss, B., Janon, A., Pujol, G., with contributions from Khalid Boumhaout, Veiga, S. D., Delage, T., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, P., Nelson, B. L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T., and Weber, F. (2018). *sensitivity: Global sensitivity analysis of model outputs*. R package version 1.15.2.

Iooss, B. and Lemaître, P. (2014). A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems: algorithms and applications*. Springer.

Iooss, B., Popelin, A.-L., Blatman, G., Ciric, C., Gamboa, F., Lacaze, S., and Lamboni, M.

(2012). Some new insights in derivative-based global sensitivity measures. In *Proceedings of PSAM 11 & ESREL 2012 Conference. Helsinki, Finland, 25-29 June*.

IPCC (2000). *Special Report on Emissions Scenarios (SRES), a special report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

IPCC (2007). *Climate Change 2007: The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

IPCC (2014a). *Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

IPCC (2014b). Concluding instalment of the Fifth Assessment Report: Climate change threatens irreversible and dangerous impacts, but options exist to limit its effects. Available at: https://www.lastampa.it/rw/Pub/Prod/PDF/copenhagen.pdf.

Jacobsen, N. S., Gislason, H., and Andersen, K. H. (2013). The consequences of balanced harvesting of fish communities. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1775).

Jansen, M. J. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, 117(1):35 – 43.

Jansen, M. J. W., Rossing, W. A. H., and Daamen, R. A. (1994). *Monte Carlo estimation of uncertainty contributions from several independent multivariate sources*, pages 334–343. Springer.

Janssen, P. H., Petersen, A. C., van der Sluijs, J. P., Risbey, J. S., and Ravetz, J. R. (2005). A guidance for assessing and communicating uncertainties. *Water Science and Technology*, 52(6):125–131.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35:339–344.

Jones, M. C. and Cheung, W. W. (2014). Multi-model ensemble projections of climate change effects on global marine biodiversity. *ICES Journal of Marine Science*, 72(3):741–752.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., and Arriaga-Weiss, S. (2010). Classification in conservation biology: a comparison of five machine-learning methods. *Ecological Informatics*, 5(6):441–450.

Kaplan, I. C. and Marshall, K. N. (2016). A guinea pig's tale: learning to review end-to-end marine ecosystem models for management applications. *ICES Journal of Marine Science*, 73:1715–1724.

Kardos, J., Benwell, G. L., and Moore, A. B. (2007). Assessing different approaches to visualise spatial and attribute uncertainty in socioeconomic data using the hexagonal or rhombus (HoR) trustree. *Computers, Environment and Urban Systems*, 31(1):91–106.

Kastanakis, M. N. and Voyer, B. G. (2014). The effect of culture on perception and cognition: a conceptual framework. *Journal of Business Research*, 67(4):425–433.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kinkeldey, C., MacEachren, A. M., and Schiewe, J. (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51(4):372–386.

Kjellström, E., Nikulin, G., Hansson, U., Strandberg, G., and Ullerstig, A. (2011). 21st century changes in the European climate: uncertainties derived from an ensemble of regional climate model simulations. *Tellus A: Dynamic Meteorology and Oceanography*, 63(1):24–40.

Klimenko, K. (2017). *Computer-aided drug design of broad-spectrum antiviral compounds*. PhD thesis, Université de Strasbourg, Strasbourg, France.

Kloprogge, P., van der Sluijs, J. P., and Wardekker, J. A. (2007). *Uncertainty communication: issues and good practice*. Copernicus Institute for Sustainable Development and Innovation.

Knudby, A., Brenning, A., and LeDrew, E. (2010). New approaches to modelling fish–habitat relationships. *Ecological Modelling*, 221(3):503–511.

Knutti, R. and Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, 3(4):369–373.

Kocev, D., Vens, C., Struyf, J., and Džeroski, S. (2007). Ensembles of multi-objective decision trees. In *Machine Learning: ECML 2007*. Springer.

Kolding, J., Jacobsen, N. S., Andersen, K. H., and van Zwieten, P. A. (2015). Maximizing fisheries yields while maintaining community structure. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):644–655.

Kosara, R. and Mackinlay, J. (2013). Storytelling: the next step for visualization. *Computer*, 46(5):44–50.

Koziel, S. and Leifsson, L. (2013). *Surrogate-based modeling and optimization*. Springer.

Krupnick, A., Morgenstern, R., Batz, M., Nelson, P., Burtraw, D., Shih, J.-S., and McWilliams,

M. (2006). *Not a sure thing: making regulatory choices under uncertainty*. Resources for the Future.

Kucherenko, S. and Iooss, B. (2017). Derivative-based global sensitivity measures. In *Handbook of Uncertainty Quantification*. Springer.

Kucherenko, S., Rodriguez-Fernandez, M., Pantelides, C., and Shah, N. (2009). Monte Carlo evaluation of derivative-based global sensitivity measures. *Reliability Engineering and System Safety*, 94:1135–1148.

Kucherenko, S. and Song, S. (2016). Derivative-based global sensitivity measures and their link with Sobol' sensitivity indices. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 455–469. Springer.

Lall, S. P. and Tibbetts, S. M. (2009). Nutrition, feeding, and behavior of fish. *Veterinary Clinics of North America: Exotic Animal Practice*, 12(2):361–372.

Lamboni, M., Iooss, B., Popelin, A., and Gamboa, F. (2013). Derivative-based global sensitivity measures: general links with Sobol' indices and numerical tests. *Mathematics and Computers in Simulation*, 87:45–54.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Larrick, R. P., Burson, K. A., and Soll, J. B. (2007). Social comparison and confidence: when thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, 102(1):76–94.

Lazenby, M. J., Todd, M. C., and Wang, Y. (2016). Climate model simulation of the South Indian Ocean Convergence Zone: mean state and variability. *Climate Research*, 68(1):59–71.

Li, X. and Kong, J. (2014). Application of GA–SVM method with parameter optimization for landslide development prediction. *Natural Hazards and Earth System Sciences*, 14(3):525–533.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Lipkus, I. M. and Hollands, J. G. (1999). The visual communication of risk. *JNCI Monographs*, 1999(25):149–163.

Little, C. M., Horton, R. M., Kopp, R. E., Oppenheimer, M., and Yip, S. (2015). Uncertainty in twenty-first-century CMIP5 sea level projections. *Journal of Climate*, 28(2):838–852.

Liu, X. (2015). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling*. Sage Publications.

Livingston, P. A. (2013). *Incorporating fish food habits data into fish population assessment models*. Springer Science & Business Media.

Lohse, G. L. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6):297–308.

Longhurst, A. R. (2007). *Ecological geography of the sea*. Elsevier Inc.

Lorenz, S., Dessai, S., Paavola, J., and Forster, P. (2015). The communication of physical science uncertainty in European National Adaptation Strategies. *Climatic Change*, 132(1):143–155.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, 5-10 December*.

Lucas, D. D., Klein, R., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., and Zhang, Y. (2013). Failure analysis of parameter-induced simulation crashes in climate models. *Geoscientific Model Development*, 6(4):1157–1171.

Lynam, C. P., Llope, M., Möllmann, C., Helaouët, P., Bayliss-Brown, G. A., and Stenseth, N. C. (2017). Interaction between top-down and bottom-up control in marine food webs. *Proceedings of the National Academy of Sciences*, 114(8):1952–1957.

Lynam, C. P., Uusitalo, L., Patrício, J., Piroddi, C., Queirós, A. M., Teixeira, H., Rossberg, A. G., Sagarminaga, Y., Hyder, K., Niquil, N., et al. (2016). Uses of innovative modeling tools within the implementation of the Marine Strategy Framework Directive. *Frontiers in Marine Science*, 3(182):1–18.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. (2005). Visualizing geospatial information uncertainty: what we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160.

Maglogiannis, I. G., Karpousis, K., Wallace, M., and Soldatos, J. (2007). *Emerging artificial intelligence applications in computer engineering: real world AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. IOS Press.

Mahmood, A., Bennamoun, M., An, S., Sohel, F., Boussaid, F., Hovey, R., Kendrick, G., and Fisher, R. (2016). Automatic annotation of coral reefs using deep learning. In *OCEANS 2016 MTS/IEEE. Monterey, 19-23 September*.

Maier, H., Ascough Ii, J., Wattenbach, M., Renschler, C., Labiosa, W., and Ravalico, J. (2008). Uncertainty in environmental decision making: issues, challenges and future directions. *Developments in Integrated Environmental Assessment*, 3:69–85.

Mangin, T., Costello, C., Anderson, J., Arnason, R., Elliott, M., Gaines, S. D., Hilborn, R., Peterson, E., and Sumaila, R. (2018). Are fishery management upgrades worth the cost? *PloS ONE*, 13(9):e0204258.

Marshall, A. M. (2017). *Understanding shifts in body size distributions - a comparative study of the impacts of fishing and climate on North Sea demersal fishes*. PhD thesis, University of Sheffield.

Marshall, A. M., Bigg, G. R., Van Leeuwen, S. M., Pinnegar, J. K., Wei, H.-L., Webb, T. J., and Blanchard, J. L. (2016). Quantifying heterogeneous responses of fish community size structure using novel combined statistical techniques. *Global Change Biology*, 22(5):1755–1768.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., et al. (2010). Guidance note for lead authors of the IPCC Fifth Assessment Report on consistent treatment of uncertainties. *Intergovernmental Panel on Climate Change*.

McElhany, P., Steel, E. A., Avery, K., Yoder, N., Busack, C., and Thompson, B. (2010). Dealing with uncertainty in ecosystem models: lessons from a complex salmon model. *Ecological Applications*, 20(2):465–482.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

Meier, H., Hordoir, R., Andersson, H., Dieterich, C., Eilola, K., Gustafsson, B. G., Höglund, A., and Schimanke, S. (2012a). Modeling the combined impact of changing climate and changing nutrient loads on the Baltic Sea environment in an ensemble of transient simulations for 1961–2099. *Climate Dynamics*, 39(9-10):2421–2441.

Meier, H. M., Andersson, H. C., Arheimer, B., Blenckner, T., Chubarenko, B., Donnelly, C., Eilola, K., Gustafsson, B. G., Hansson, A., Havenhand, J., et al. (2012b). Comparing reconstructed past variations and future projections of the Baltic Sea ecosystem—first results from multi-model ensemble simulations. *Environmental Research Letters*, 7(3):034005.

Meier, M., Edman, M., Eilola, K., Placke, M., Neumann, T., Andersson, H., Brunnabend, S.-E., Dieterich, C., Frauen, C., Friedland, R., et al. (2019). Assessment of uncertainties in scenario simulations of biogeochemical cycles in the Baltic Sea. *Frontiers in Marine Science*, 6(46):1–29.

Merle, C. (2016). *Nouvelles méthodes d'inférence de l'histoire démographique à partir de données génétiques*. PhD thesis, Institut Montpelliérain Alexander Grothendieck, Denmark.

MERP (2017). *Report from the Annual Science Meeting, 10-12 October 2017, Sheffield, UK*. Marine Ecosystems Research Programme (MERP).

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.

Mol, L. (2011). The potential role for infographics in science communication. Master's thesis, Biomedical Sciences, Vrije University, Amsterdam, Netherlands.

Möllmann, C., Lindegren, M., Blenckner, T., Bergström, L., Casini, M., Diekmann, R., Flinkman, J., Müller-Karulis, B., Neuenfeldt, S., Schmidt, J. O., et al. (2013). Implementing ecosystem-based fisheries management: from single-species to integrated ecosystem assessment and advice for Baltic Sea fish stocks. *ICES Journal of Marine Science*, 71(5):1187–1197.

Momeni, A., Pincus, M., and Libien, J. (2018). *Introduction to statistical methods in pathology*. Springer.

Moreland, K. (2009). Diverging color maps for scientific visualization. In *Advances in Visual Computing*. Springer.

Morgan, M. G. (2009). *Best practice approaches for characterizing, communicating and incorporating scientific uncertainty in climate decision making*. DIANE Publishing.

Morissette, L. (2005). Addressing uncertainty in marine ecosystems modelling. In *Strategic Management of Marine Ecosystems*. Springer.

Morris, D. J., Speirs, D. C., Cameron, A. I., and Heath, M. R. (2014). Global sensitivity analysis of an end-to-end marine ecosystem model of the North Sea: factors affecting the biomass of fish and benthos. *Ecological Modelling*, 273:251–263.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Mulholland, D. P., Haines, K., Sparrow, S. N., and Wallom, D. (2017). Climate model forecast biases assessed with a perturbed physics ensemble. *Climate Dynamics*, 49(5-6):1729–1746.

Müller, A. C., Guido, S., et al. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001):768–772.

Murray, F., Needham, K., Gormley, K., Rouse, S., Coolen, J. W., Billett, D., Dannheim, J., Birchenough, S. N., Hyder, K., Heard, R., et al. (2018). Data challenges and opportunities for environmental management of North Sea oil and gas decommissioning in an era of blue growth. *Marine Policy*, 97:130–138.

NART (2013). *Workshop on data visualization to support ecosystem based management*. Gulf of Maine Research Institute, Portland, Maine.

Nelson, G. C., Bennett, E., Berhe, A. A., Cassman, K., DeFries, R., Dietz, T., Dobermann, A., Dobson, A., Janetos, A., Levy, M., et al. (2006). Anthropogenic drivers of ecosystem change: an overview. *Ecology and Society*, 11(2).

Neubauer, P. and Andersen, K. H. (2018). Thermal performance of fish is explained by an interplay between physiology, behaviour, and ecology. marxiv.org/gt9rn.

Nicholson, M. D. and Jennings, S. (2004). Testing candidate indicators to support ecosystem-based management: the power of monitoring surveys to detect temporal trends in fish community metrics. *ICES Journal of Marine Science*, 61(1):35–42.

Niiranen, S., Blenckner, T., Hjerne, O., and Tomczak, M. T. (2012). Uncertainties in a Baltic Sea food-web model reveal challenges for future projections. *Ambio*, 41:613–625.

Nikulin, G., Kjellström, E., Hansson, U., Strandberg, G., and Ullerstig, A. (2011). Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. *Tellus A: Dynamic Meteorology and Oceanography*, 63(1):41–55.

Nossent, J., Elsen, P., and Bauwens, W. (2011). Sobol' sensitivity analysis of a complex environmental model. *Environmental Modelling and Software*, 26(12):1515–1525.

Oppenlander, J. E. and Schaffer, P. (2017). *Data management and analysis using JMP: health care case studies*. SAS Institute Inc.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M. (2015). Does model performance improve with complexity? A case study with three hydrological models. *Journal of Hydrology*, 523:147 – 159.

O'Sullivan, D. and Perry, G. L. (2013). *Spatial simulation: exploring pattern and process*. John Wiley & Sons.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., et al. (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC.

Paine, P. J., Hooper, T., Spence, M. A., Heath, M. R., McPike, R., Bannister, H., Thorpe, R. B., and Blackwell, P. G. (in prep.). Modelling the effect of fishing levels on commercial fisheries revenue using bayesian belief networks.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., and Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, 129(3):257–299.

Palmer, T., Doblas-Reyes, F., Hagedorn, R., and Weisheimer, A. (2005). Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1463):1991–1998.

Patt, A. and Dessai, S. (2005). Communicating uncertainty: lessons learned and suggestions for climate change assessment. *Comptes Rendus Geoscience*, 337(4):425–441.

Patt, A. G. and Schrag, D. P. (2003). Using specific language to describe risk and probability. *Climatic Change*, 61(1-2):17–30.

Pauly, D., Christensen, V., and Walters, C. (2000). Ecopath, Ecosim, and Ecospace as tools for evaluating ecosystem impact of fisheries. *ICES Journal of Marine Science*, 57(3):697–706.

Payne, M. R., Barange, M., Cheung, W. W., MacKenzie, B. R., Batchelder, H. P., Cormon, X., Eddy, T. D., Fernandes, J. A., Hollowed, A. B., Jones, M. C., et al. (2015). Uncertainties in projecting climate-change impacts in marine ecosystems. *ICES Journal of Marine Science*, 73(5):1272–1282.

Peebles, D. and Ali, N. (2015). Expert interpretation of bar and line graphs: the role of graphicacy in reducing the effect of graph format. *Frontiers in Psychology*, 6:1673.

Peltier, J. (2013). *Excel charting dos and don'ts*. Peltier Technical Services.

Peters, R. H. (1983). *The ecological implications of body size*. Cambridge Studies in Ecology. Cambridge University Press.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., and Wagener,

T. (2016). Sensitivity analysis of environmental models: a systematic review with practical work flow. *Environmental Modelling and Software*, 79:214–232.

Pihur, V., Datta, S., and Datta, S. (2009). RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, 10(1):62.

Pihur, V., Datta, S., and Datta, S. (2018). *RankAggreg: weighted rank aggregation*. R package version 0.6.5.

Pinnegar, J. K., Goni, N., Trenkel, V., Arrizabalaga, H., Melle, W., Keating, J., and Óskarsson, G. (2015). A new compilation of stomach content data for commercially-important pelagic fish species in the Northeast Atlantic. *Earth System Science Data*, 7(1):19–28.

Pitsillou, M. and Fokianos, K. (2016). *dCovTS: Distance covariance and correlation for time series analysis*. R package version 1.1.

Plaganyi, E. E. (2007). Models for an ecosystem approach to fisheries. Technical report, FAO Fisheries Technical Paper 477, Food and Agriculture Organisation of the United Nations, Rome, Italy.

Pontes, G., Gupta, A. S., and Taschetto, A. (2016). Projected changes to South Atlantic boundary currents and confluence region in the CMIP5 models: the role of wind and deep ocean changes. *Environmental Research Letters*, 11(9):094013.

Power, M. and McCarty, L. S. (2006). Environmental risk management decision-making in a societal context. *Human and Ecological Risk Assessment*, 12(1):18–27.

Quetglas, A., Ordines, F., and Guijarro, B. (2011). The use of artificial neural networks (ANNs) in aquatic ecology. In *Artificial neural networks - application*. InTech.

Quispel, A., Maes, A., and Schilperoord, J. (2016). Graph and chart aesthetics for experts and laymen in design: the role of familiarity and perceived ease of use. *Information Visualization*, 15(3):238–252.

R Core Team (2018). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Räisänen, J. (2001). CO2-induced climate change in CMIP2 experiments: quantification of agreement and role of internal variability. *Journal of Climate*, 14(9):2088–2104.

Ratto, M., Castelletti, A., and Pagano, A. (2012). Emulation techniques for the reduction and sensitivity analysis of complex environmental models. *Environmental Modelling & Software*, 34:1–4.

Reed, J., Shannon, L., Velez, L., Akoglu, E., Bundy, A., Coll, M., Fu, C., Fulton, E. A., Grüss, A., Halouani, G., et al. (2016). Ecosystem indicators - accounting for variability in species' trophic levels. *ICES Journal of Marine Science*, 74(1):158–169.

Refsgaard, J. C., Arnbjerg-Nielsen, K., Drews, M., Halsnæs, K., Jeppesen, E., Madsen, H., Markandya, A., Olesen, J. E., Porter, J. R., and Christensen, J. H. (2013). The role of uncertainty in climate change adaptation strategies—A Danish water management example. *Mitigation and Adaptation Strategies for Global Change*, 18(3):337–359.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., and Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process - a framework and guidance. *Environmental Modelling & Software*, 22(11):1543–1556.

Regan, H. M., Colyvan, M., and Burgman, M. A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2):618–628.

Reinecke, K. and Gajos, K. Z. (2014). Quantifying visual preferences around the world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Toronto, Canada, 26 April - 1 May*.

Reintges, A., Martin, T., Latif, M., and Keenlyside, N. S. (2017). Uncertainty in twenty-first century projections of the Atlantic Meridional Overturning Circulation in CMIP3 and CMIP5 models. *Climate Dynamics*, 49(5-6):1495–1511.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, 11(5):559–623.

Rodriguez-Fernandez, M., Banga, J. R., and Doyle III, F. J. (2012). Novel global sensitivity analysis methodology accounting for the crucial role of the distribution of input parameters: application to systems biology models. *International Journal of Robust and Nonlinear Control*, 22(10):1082–1102.

Roeder, A. and Hill, J. (2009). *Recent advances in remote sensing and geoinformation processing for land degradation assessment*. CRC Press.

Roemmich, D., Johnson, G. C., Riser, S., Davis, R., Gilson, J., Owens, W. B., Garzoli, S. L., Schmid, C., and Ignaszewski, M. (2009). The Argo Program: Observing the global ocean with profiling floats. *Oceanography*, 22(2):34–43.

Roessig, J. M., Woodley, C. M., Cech, J. J., and Hansen, L. J. (2004). Effects of global climate change on marine and estuarine fishes and fisheries. *Reviews in Fish Biology and Fisheries*, 14(2):251–275.

Rose, M. R. and Harmsen, R. (1978). Using sensitivity analysis to simplify ecosystem models: a case study. *Simulation*, 31(1):15–26.

Rossberg, A. G. (2012). A complete analytic theory for structure and dynamics of populations and communities spanning wide ranges in body size. *Advances in Ecological Research*, 46:427–521.

Roughgarden, J. and Smith, F. (1996). Why fisheries collapse and what to do about it. *Proceedings of the National Academy of Sciences*, 93(10):5078–5083.

Rowan, T., Melbourne-Thomas, J., Constable, A., Wotherspoon, S., Hindell, M., McMahon, C., Lea, M.-A., Swadling, K., Kelly, P., and Blanchard, J. (2017). Size based models for understanding Southern Ocean food web structure and energy pathways. In *Book of Abstracts: XIIth SCAR Biology Symposium. Leuven, Belgium, 10-14 July*.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190.

Saary, M. J. (2008). Radar plots: a useful way for presenting multivariate health care data. *Journal of Clinical Epidemiology*, 61(4):311–317.

Saket, B., Endert, A., and Demiralp, C. (2018). *Task-based effectiveness of basic visualizations*. IEEE Transactions on Visualization and Computer Graphics.

Salkind, N., editor (2010). *Encyclopaedia of research design*. Sage Publications.

Salling, K. B. and Leleur, S. (2012). Modelling of transport project uncertainties: feasibility risk assessment and scenario analysis. *European Journal of Transport and Infrastructure Research*, 12(1):21–38.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280 – 297.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley and Sons.

Sarafanov, A. (2009). On the effect of the North Atlantic Oscillation on temperature and salinity of the subpolar North Atlantic intermediate and deep waters. *ICES Journal of Marine Science*, 66(7):1448–1454.

Sarrazin, F., Pianosi, F., and Wagener, T. (2016). Global sensitivity analysis of environmental models: convergence and validation. *Environmental Modelling and Software*, 79:135–152.

Savage, V. M., Gillooly, J. F., Brown, J. H., West, G. B., and Charnov, E. L. (2004). Effects of body size and temperature on population growth. *The American Naturalist*, 163(3):429–441.

Schmidt, J. (2012). Ordinal response mixed models: a case study. Master's thesis, Montana State University, Montana, USA.

Schmittner, A., Latif, M., and Schneider, B. (2005). Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations. *Geophysical Research Letters*, 32(23):L23710.

Schneider, S. H. and Moss, R. (1999). Uncertainties in the IPCC TAR: recommendations to lead authors for more consistent assessment and reporting. In Pachauri, R., Taniguchi, T., and Tanaka, K., editors, *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, pages 33–51. World Meteorological Organization, Geneva.

Scott, F., Blanchard, J., and Andersen, K. (2018). *mizer: multi-species size spectrum modelling in R*. R package version 1.0.

Scott, F., Blanchard, J. L., and Andersen, K. H. (2014). mizer: An R package for multispecies, trait-based and community size spectrum ecological modelling. *Methods in Ecology and Evolution*, 5(10):1121–1125.

Sedkaoui, S. (2018). *Data analytics and big data*. John Wiley & Sons.

Serpetti, N., Baudron, A. R., Burrows, M., Payne, B. L., Helaouet, P., Fernandes, P. G., and Heymans, J. (2017). Impact of ocean warming on sustainable fisheries management informs the Ecosystem Approach to Fisheries. *Scientific Reports*, 7(1):13438.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., and Khovanova, N. (2017). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, (in press).

Shannon, L. J., Cury, P. M., and Jarre, A. (2000). Modelling effects of fishing in the Southern Benguela ecosystem. *ICES Journal of Marine Science*, 57:720–722.

Shin, Y.-J., Houle, J. E., Akoglu, E., Blanchard, J. L., Bundy, A., Coll, M., Demarcq, H., Fu, C., Fulton, E. A., Heymans, J. J., et al. (2018). The specificity of marine ecological indicators to fishing in the face of environmental change: a multi-model evaluation. *Ecological Indicators*, 89:317–326.

Shin, Y.-J., Rochet, M.-J., Jennings, S., Field, J. G., and Gislason, H. (2005). Using size-based indicators to evaluate the ecosystem effects of fishing. *ICES Journal of Marine Science*, 62(3):384–396.

Shin, Y.-J. and Shannon, L. J. (2009). Using indicators for evaluating, comparing, and communicating the ecological status of exploited marine ecosystems. 1. The IndiSeas project. *ICES Journal of Marine Science*, 67(4):686–691.

Shrestha, D., Kayastha, N., and Solomatine, D. (2009). A novel approach to parameter uncertainty analysis of hydrological models using neural networks. *Hydrology and Earth System Sciences*, 13(7):1235–1248.

Simpson, S. D., Blanchard, J. L., and Genner, M. (2013). Impacts of climate change on fish. *Marine Climate Change Impacts Partnership: Science Review*.

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R., and Murphy, J. M. (2007). Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317(5839):796–799.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414.

Sobol', I. M. and Kucherenko, S. (2009). Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79(10):3009–3017.

Sobol', I. M. and Kucherenko, S. (2010). A new derivative based importance criterion for groups of variables and its link with the global sensitivity indices. *Computer Physics Communications*, 181(7):1212–1217.

Spangenberg, J. H. (2006). System complexity and scenario analysis. In *Ninth Biennial Conference of the International Society for Ecological Economics "Ecological sustainability and human well-being". New Delhi, India, 15-19 December*.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Speirs, D., Guirey, E., Gurney, W., and Heath, M. (2010). A length-structured partial ecosystem model for cod in the North Sea. *Fisheries Research*, 106(3):474–494.

Spence, M. A., Blackwell, P. G., and Blanchard, J. L. (2016). Parameter uncertainty of a dynamic multispecies size spectrum model. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):589–597.

Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., Serpetti, N., Speirs, D. C., Thorpe, R. B., and Blackwell, P. G. (2018). A general framework for combining ecosystem models. *Fish and Fisheries*, 19(6):1031–1042.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400.

Spiegelhalter, D. J. and Riesch, H. (2011). Don't know, can't know: embracing deeper uncertainties when analysing risks. *Philosophical Transactions of the Royal Society A*, 369(1956):4730–4750.

St-Louis, V., Clayton, M. K., Pidgeon, A. M., and Radeloff, V. C. (2012). An evaluation of prior influence on the predictive ability of Bayesian model averaging. *Oecologia*, 168(3):719–726.

Stäbler, M., Kempf, A., Smout, S., and Temming, A. (2019). Sensitivity of multispecies maximum sustainable yields to trends in the top (marine mammals) and bottom (primary production) compartments of the southern North Sea food-web. *PloS ONE*, 14(1):e0210882.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433(7024):403–406.

Steele, J. H., Thorpe, S. A., and Turekian, Karl, K., editors (2001). *Encyclopedia of ocean sciences*. Elsevier.

Sterba, Z. and Bláha, J. D. (2015). The influence of colour on the perception of cartographic visualizations. In *AIC Midterm Meeting. Tokyo, Japan, 19-22 May*.

Sudret, B. and Mai, C. V. (2015). Computing derivative-based global sensitivity measures using polynomial chaos expansions. *Reliability Engineering & System Safety*, 134:241–250.

Suessbrick, A., Schober, M. F., and Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association. Alexandria, Virginia, USA*.

Tan, A. C. and Gilbert, D. (2003). An empirical comparison of supervised machine learning techniques in bioinformatics. In *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics. Adelaide, Australia, 4-7 February*.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). Classification: basic concepts, decision trees, and model evaluation. In *Introduction to Data Mining*. Pearson/Addison-Wesley.

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540.

Terray, L., Corre, L., Cravatte, S., Delcroix, T., Reverdin, G., and Ribes, A. (2012). Near-surface salinity as nature's rain gauge to detect human influence on the tropical water cycle. *Journal of Climate*, 25(3):958–977.

Thomson, M., Doblas-Reyes, F., Mason, S., Hagedorn, R., Connor, S., Phindela, T., Morse, A., and Palmer, T. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, 439(7076):576–579.

Thorpe, R. B., Dolder, P. J., Reeves, S., Robinson, P., and Jennings, S. (2016). Assessing fishery and ecological consequences of alternate management options for multispecies fisheries. *ICES Journal of Marine Science*, 73(6):1503–1512.

Thorpe, R. B., Jennings, S., and Dolder, P. J. (2017). Risks and benefits of catching pretty good yield in multispecies mixed fisheries. *ICES Journal of Marine Science*, 74(8):2097–2106.

Thorpe, R. B., Le Quesne, W. J., Luxford, F., Collie, J. S., and Jennings, S. (2015). Evaluation and management implications of uncertainty in a multispecies size-structured model of population and community responses to fishing. *Methods in Ecology and Evolution*, 6(1):49–58.

Touzani, S. and Busby, D. (2014). Screening method using the derivative-based global sensitivity indices with application to reservoir simulator. *Oil & Gas Science and Technology*, 69(4):619–632.

Triola, M. F. (2006). *Elementary statistics*. Pearson/Addison-Wesley.

Turner, H. and Firth, D. (2012). Bradley-terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 48(9):1–21.

Tyre, A. J. and Michaels, S. (2011). Confronting socially generated uncertainty in adaptive management. *Journal of Environmental Management*, 92(5):1365–1370.

Ursin, E. (1973). On the prey size preferences of cod and dab. *Meddelelser fra Danmarks Fiskeri-og Havundersøgelser*, 7:85–98.

Uusitalo, L., Lehikoinen, A., Helle, I., and Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling and Software*, 63:24–31.

Van Asselt, M. B. and Rotmans, J. (2002). Uncertainty in integrated assessment modelling. *Climatic Change*, 54(1-2):75–105.

van der Sluijs, J., Janssen, P., Petersen, A., Kloprogge, P., Risbey, J., Tuinstra, W., and Ravetz, J. (2004). *RIVM/MNP guidance for uncertainty assessment and communication series: tool catalogue for uncertainty assessment*. Utrecht University, Utrecht, Netherlands.

van der Sluijs, J. P. (1997). *Anchoring amid uncertainty: on the management of uncertainties in risk assessment of anthropogenic climate change*. PhD thesis, Universiteit Utrecht, Utrecht, Netherlands.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K. (2011). The representative concentration pathways: an overview. *Climatic Change*, 109(1):5–31.

van Zwieten, P. A., Kolding, J., Plank, M. J., Hecky, R. E., Bridgeman, T. B., MacIntyre, S., Seehausen, O., and Silsbe, G. M. (2015). The Nile perch invasion in Lake Victoria: cause or consequence of the haplochromine decline? *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):622–643.

Vasantrao, K. V. (2011). Enhance accuracy in software cost and schedule estimation by using "Uncertainty Analysis and Assessment" in the system modeling process. *International Journal of Research and Innovation in Computer Engineering*, 1(1):6–18.

Vaughan, N. E. and Gough, C. (2016). Expert assessment concludes negative emissions scenarios may not deliver. *Environmental Research Letters*, 11(9):095003.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Villarini, G. and Vecchi, G. A. (2012). Twenty-first-century projections of North Atlantic tropical storms from CMIP5 models. *Nature Climate Change*, 2(8):604–607.

Walker, W., Harremoës, P., Rotmans, J., van der Sluijs, J., van Asselt, M., Janssen, P., and von Krauss, M. K. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4(1):5–17.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, 4(3):201–205.

Wang, W., Dungan, J., Hashimoto, H., Michaelis, A. R., Milesi, C., Ichii, K., and Nemani, R. R. (2011). Diagnosing and assessing uncertainties of terrestrial ecosystem models in a multi-model ensemble experiment: 1. Primary production. *Global Change Biology*, 17(3):1350–1366.

Ward, B. A. (2009). *Marine ecosystem model analysis using data assimilation*. PhD thesis, University of Southampton, Southampton, UK.

Watson, J. R., Stock, C. A., and Sarmiento, J. L. (2015). Exploring the role of movement in determining the global distribution of marine biomass using a coupled hydrodynamic–size-based ecosystem model. *Progress in Oceanography*, 138:521–532.

Wesselink, A., Challinor, A. J., Watson, J., Beven, K., Allen, I., Hanlon, H., Lopez, A., Lorenz, S., Otto, F., Morse, A., et al. (2015). Equipped to deal with uncertainty in climate and impacts predictions: lessons from internal peer review. *Climatic Change*, 132(1):1–14.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833.

Wilby, R. L. and Dessai, S. (2010). Robust adaptation to climate change. *Weather*, 65(7):180–185.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wong, B. and Candolin, U. (2015). Behavioral responses to changing environments. *Behavioral Ecology*, 26(3):665–673.

Yan, G., Wen-Jie, D., Fu-Min, R., Zong-Ci, Z., and Jian-Bin, H. (2013). Surface air temperature simulations over China with CMIP5 and CMIP3. *Advances in Climate Change Research*, 4(3):145 – 152.

Ye, N. (2013). *Data mining: theories, algorithms, and examples*. CRC Press.

Zhang, C., Chen, Y., and Ren, Y. (2015). Assessing uncertainty of a multispecies size-spectrum model resulting from process and observation errors. *ICES Journal of Marine Science*, 72(8):2223–2233.

Zhang, L., Thygesen, U. H., Knudsen, K., and Andersen, K. H. (2013). Trait diversity promotes stability of community dynamics. *Theoretical Ecology*, 6(1):57–69.

Zhang, Z. (2016). Decision tree modeling using R. *Annals of Translational Medicine*, 4(15):275.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series: a distance correlation approach. *Journal of Time Series Analysis*, 33(3):438–457.

Zhu, Q. and Zhuang, Q. (2013). Influences of calibration data length and data period on model parameterization and quantification of terrestrial ecosystem carbon dynamics. *Geoscientific Model Development Discussions*, 6(4):6835–6865.

Appendices

Appendix A

Chapter 4: Predicting species biomass using the random forest algorithm

The random forest algorithm in Chapter 4 tended to overestimate the biomass of each species when their simulated biomass was very low and underestimate the biomass of each species when their simulated biomass was high (see Figure A.1 for example). As the successful prediction of species survival and community coexistence requires the algorithm to accurately predict biomasses of less than 0.001g, the consistent overestimation of the algorithm at extremely low simulated biomasses may be the limiting factor that prevents the random forests from predicting coexistence with 100% accuracy.

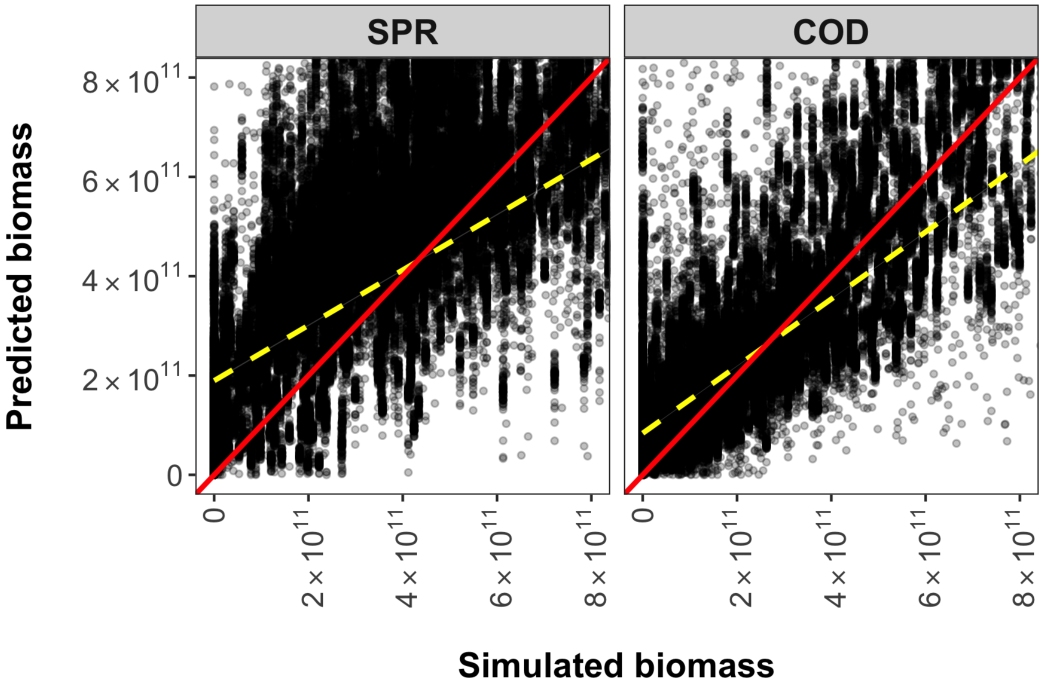


Figure A.1: The simulated biomass (g) of sprat (SPR; left) and Atlantic cod (COD; right) compared with the biomass of these species as predicted by the best performing random forests. The best performing random forests were defined as those with the lowest Root Mean Square Error (RMSE). The solid red line indicates where the data points should be if the observed and predicted biomass are equal (i.e. $y = x$). The dashed blue line represents a line of best fit through the predicted biomass of each species. Sprat and Atlantic cod were selected for plotting to provide examples from the smallest and largest fish species, but the patterns were consistent across all species. Please note that the results are presented for just one of the 100 testing datasets and the axes are limited to 8×10^{11} for plotting purposes.

Appendix B

Chapter 5: Climate models - sea surface temperature and sea surface salinity

The Coupled Model Intercomparison Project phase five (CMIP5) multi-model ensemble used in Chapter 5 consisted of projections from 14 different climate models for Sea Surface Temperature (SST) and 11 different models for Sea Surface Salinity (SSS). The models selected for both SST and SSS are listed in Table B.1.

Table B.1: The CMIP5 models that were selected for the analysis of spatio-temporal changes in the contributions of internal variability, model, and scenario uncertainty to the total variance of the projections of Sea Surface Temperature (SST) and Sea Surface Salinity (SSS). *Please note that projections of both SST and SSS were unavailable for the Red Sea and Persian Gulf province in the CanESM2 and MIROC5 models, whilst projections of SST were also unavailable for the Mediterranean and Black Sea province and the North East Atlantic Shelves province in the BNU-ESM model.

	SST	**SSS**
bcc-csm1-1	✓	
BNU-ESM*	✓	
CanESM2*	✓	✓
CCSM4	✓	
CESM1-CAM5	✓	
CNRM-CM5	✓	✓
CSIRO-Mk3-6-0		✓
EC-EARTH		✓
GFDL-CM3	✓	✓
GISS-E2-R	✓	✓
HadGEM2-ES	✓	✓
IPSL-CM5A-LR	✓	✓
MIROC5*	✓	✓
MPI-ESM-LR	✓	✓
MRI-CGCM3	✓	
NorESM1-M	✓	✓

Appendix C

Chapter 5: Longhurst biogeographical regions

To better understand the spatial variability in the relative contributions of internal variability, model, and scenario uncertainty to the total variance of the Coupled Model Intercomparison Project phase five (CMIP5) multi-model ensemble in Chapter 5, we divided the projections into distinct regions based on Longhurst's widely accepted partitioning of the global ocean into 54 biogeographical provinces (see Figure C.1; Longhurst (2007).

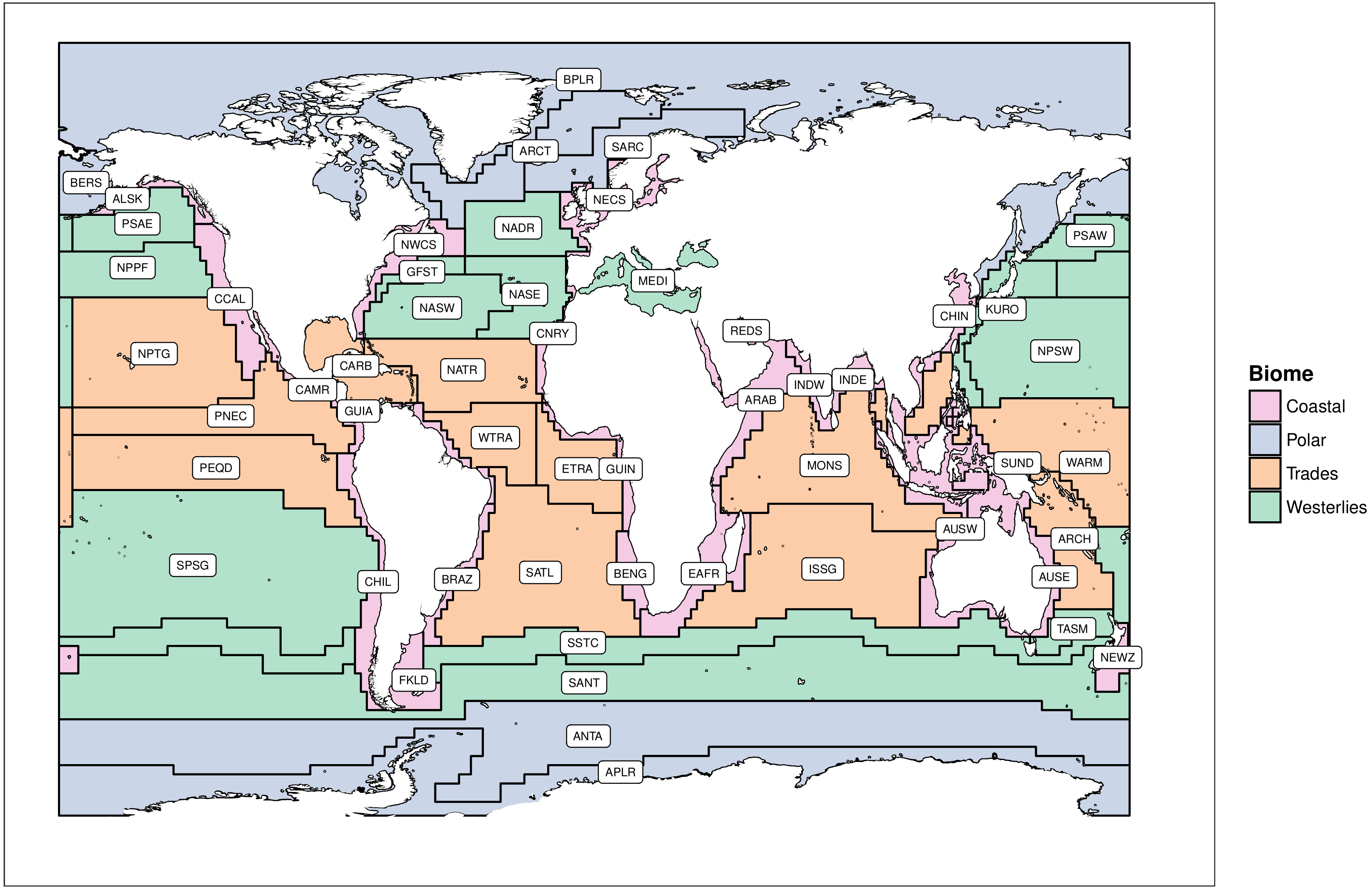


Figure C.1: The four biomes and 54 biogeographical provinces defined by Longhurst (2007). The full name of each province is provided overleaf.

Code	Description	Code	Description	Code	Description
ALSK	Alaska Downwelling Coastal	EAFR	E. Africa Coastal	NPPF	N. Pacific Polar Front
ANTA	Antarctic	ETRA	Eastern Tropical Atlantic	NPSW	N. Pacific Subtropical Gyre (West)
APLR	Austral Polar	FKLD	SW Atlantic Shelves	NPTG	N. Pacific Tropical Gyre
ARAB	NW Arabian Upwelling	GFST	Gulf Stream	NWCS	NW Atlantic Shelves
ARCH	Archipelagic Deep Basins	GUIA	Guianas Coastal	PEQD	Pacific Equatorial Divergence
ARCT	Atlantic Arctic	GUIN	Guinea Current Coastal	PNEC	N. Pacific Equatorial Countercurrent
AUSE	East Australian Coastal	INDE	E. India Coastal	PSAE	Pacific Subarctic Gyres (East)
AUSW	Australia-Indonesia Coastal	INDW	W. India Coastal	PSAW	Pacific Subarctic Gyres (West)
BENG	Benguela Current Coastal	ISSG	Indian S. Subtropical Gyre	REDS	Red Sea, Persian Gulf
BERS	N. Pacific Epicontinental	KURO	Kuroshio Current	SANT	Subantarctic
BPLR	Boreal Polar (POLR)	MEDI	Mediterranean Sea, Black Sea	SARC	Atlantic Subarctic
BRAZ	Brazil Current Coastal	MONS	Indian Monsoon Gyres	SATL	South Atlantic Gyral (SATG)
CAMR	Central American Coastal	NADR	N. Atlantic Drift (WWDR)	SPSG	S. Pacific Subtropical Gyre
CARB	Caribbean	NASE	N. Atlantic Subtropical Gyral (East) (STGE)	SSTC	S. Subtropical Convergence
CCAL	California Upwelling Coastal	NASW	N. Atlantic Subtropical Gyral (West) (STGW)	SUND	Sunda-Arafura Shelves
CHIL	Chile-Peru Current Coastal	NATR	N. Atlantic Tropical Gyral (TRPG)	TASM	Tasman Sea
CHIN	China Sea Coastal	NECS	NE Atlantic Shelves	WARM	W. Pacific Warm Pool
CNRY	Canary Coastal (EACB)	NEWZ	New Zealand Coastal	WTRA	Western Tropical Atlantic

Appendix D

Chapter 5: Signal-to-Noise Ratio of the projections

In Chapter 5, the Signal-to-Noise Ratio (SNR) of the projections of Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) was shown for the following decades: the 2010s, the 2050s, and the 2090s. In Figure D.1, the SNR of the projections is given for all available decades to further highlight the time periods in which we are most confident in the projections.

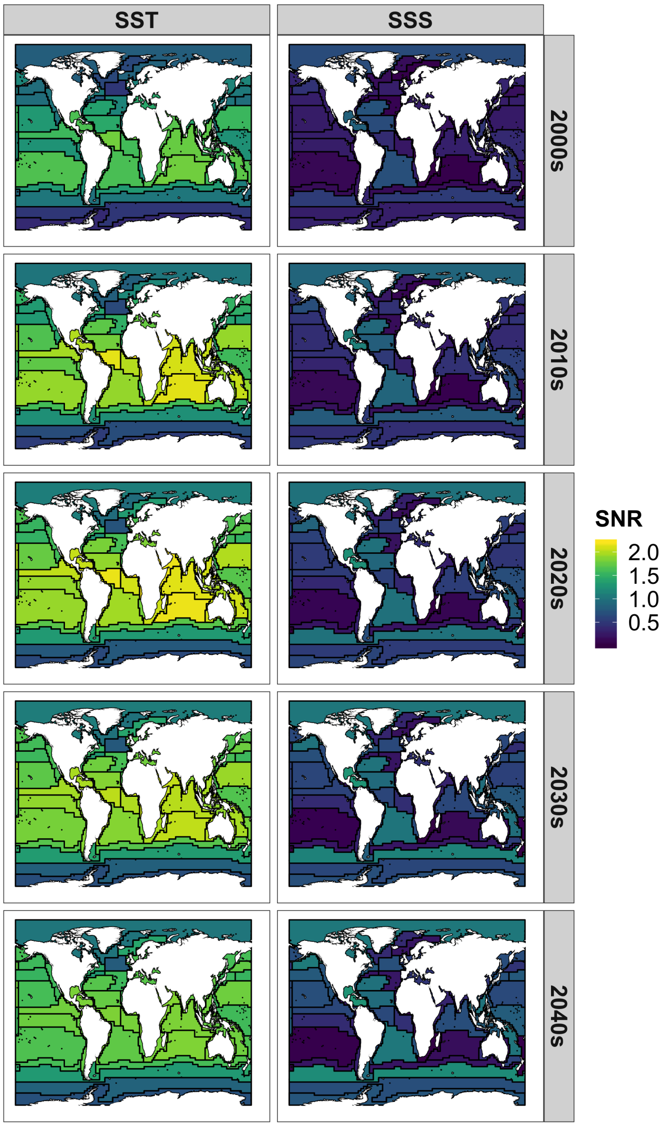


Figure continued overleaf.

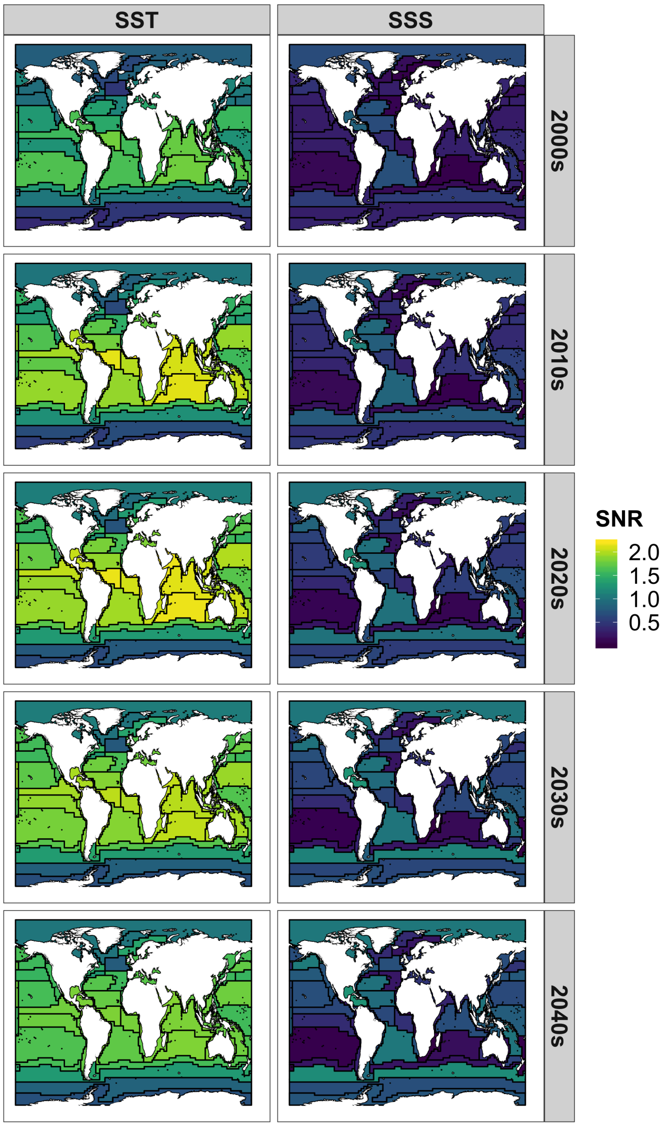


Figure D.1: The absolute Signal-to-Noise Ratio (SNR) for projections of future regional, decadal Sea Surface Temperature (SST; left) and Sea Surface Salinity (SSS; right) (90% confidence levels). All decades between the 2010s (top) and 2090s (bottom) are plotted to allow for comparisons to be made across the 21st century. The black lines delineate the 54 biogeographic provinces described by Longhurst (2007). A high SNR indicates high confidence in the projections for a given region, whilst a low SNR indicates low confidence in the projections for a given region.

Appendix E

Chapter 6: Climate models - surface air temperature

The Coupled Model Intercomparison Project phase five (CMIP5) multi-model ensemble used in Chapter 6 consisted of Surface Air Temperature (SAT) projections from 15 different climate models (see Table E.1 for a list of the selected models).

Table E.1: The 15 CMIP5 models that formed the multi-model ensemble used in the visualisation survey. The models were selected based solely on the availability of the appropriate Surface Air Temperature (SAT) data between 1850 and 2099.

	SAT
bcc-csm1-1	✓
BNU-ESM	✓
CanESM2	✓
CCSM4	✓
CESM1-CAM5	✓
CNRM-CM5	✓
CSIRO-Mk3-6-0	✓
EC-EARTH	✓
FIO-ESM	✓
GISS-E2-R	✓
IPSL-CM5A-LR	✓
MIROC-ESM	✓
MPI-ESM-LR	✓
MRI-CGCM3	✓
NorESM1-M	✓

Appendix F

Chapter 6: Survey visualisations

We developed ten different visualisations for the survey in Chapter 6, all of which depicted the same data from the Coupled Model Intercomparison Project phase five (CMIP5) multi-model ensemble. The visualisations included two versions of a line plot, two versions of a box plot, two versions of a dot plot, a radar plot, a cascade plot, a heat plot, and an infographic (see Figures F.1 to F.10). Please note that some of the visualisations were based on the work of Prof. Ed Hawkins and Dr. Rowan Sutton from the National Centre for Atmospheric Sciences (NCAS). The participants were provided with definitions of the median, standard deviation, percentiles, and the interquartile range alongside the appropriate visualisations to aid interpretation.

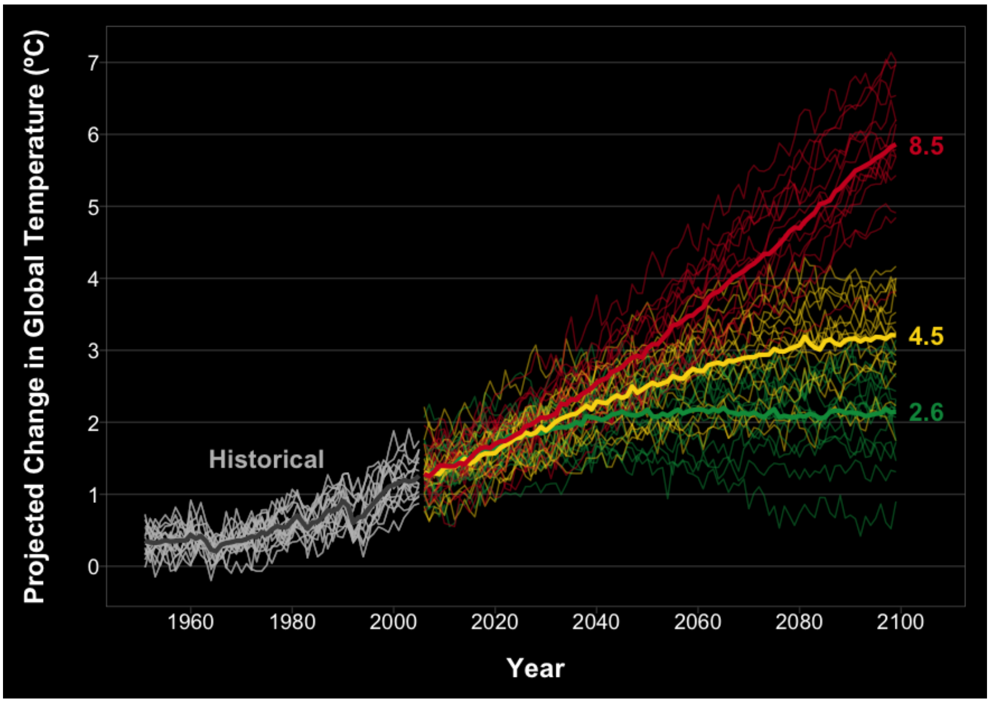


Figure F.1: Survey visualisation: line1. The thin lines represent the annual temperature change projected to occur by each of the 15 different climate models and the bold lines represent the multi-model average for each of the three scenarios.

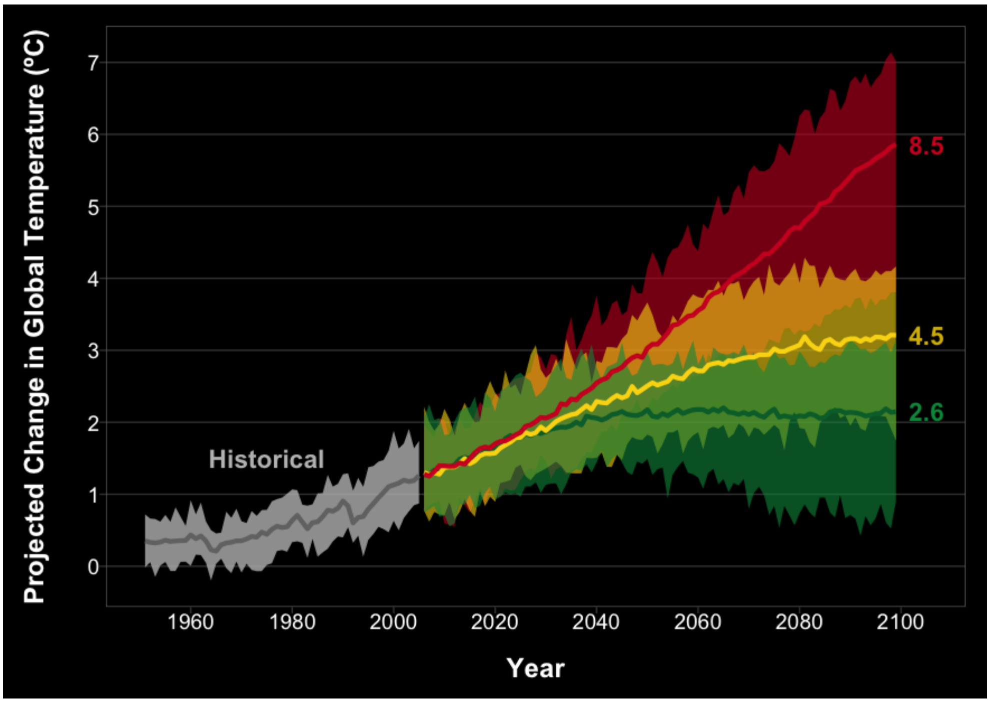


Figure F.2: Survey visualisation: line2. The bands represent the annual minimum and maximum temperature change projected to occur under each of the three scenarios and the bold lines represent the multi-model average for each scenario.

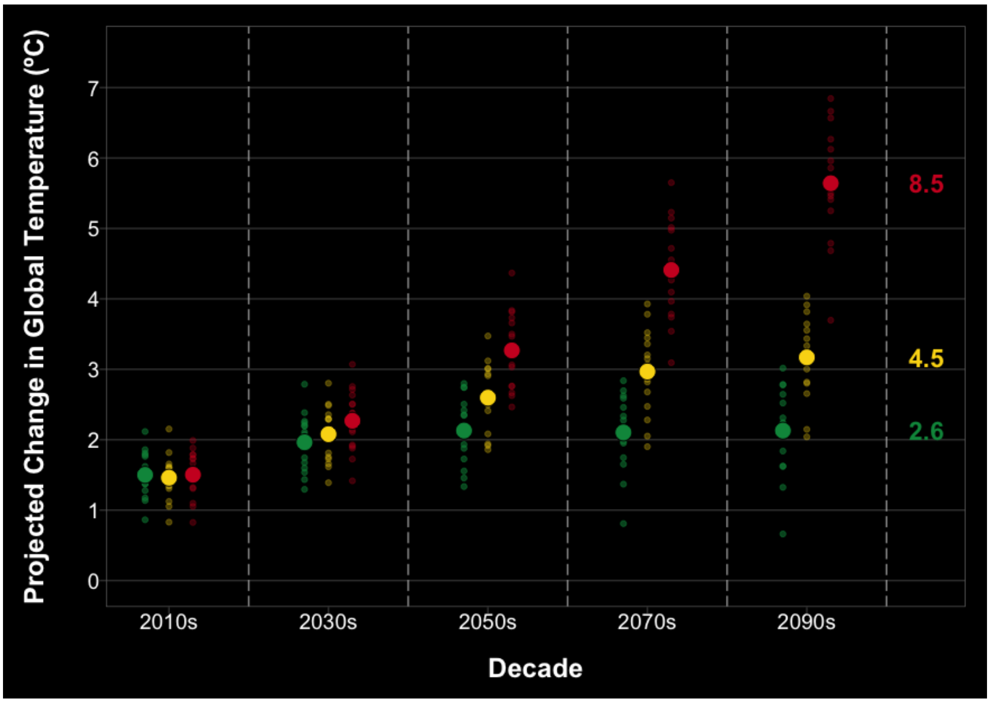


Figure F.3: Survey visualisation: dot1. The faint dots represent the average temperature change projected to occur by each of the 15 different climate models. The bold dots represent the multi-model average for each of the three scenarios. Please note that this visualisation depicts decadal averages and the data points have been slightly offset from one another to prevent overlap.

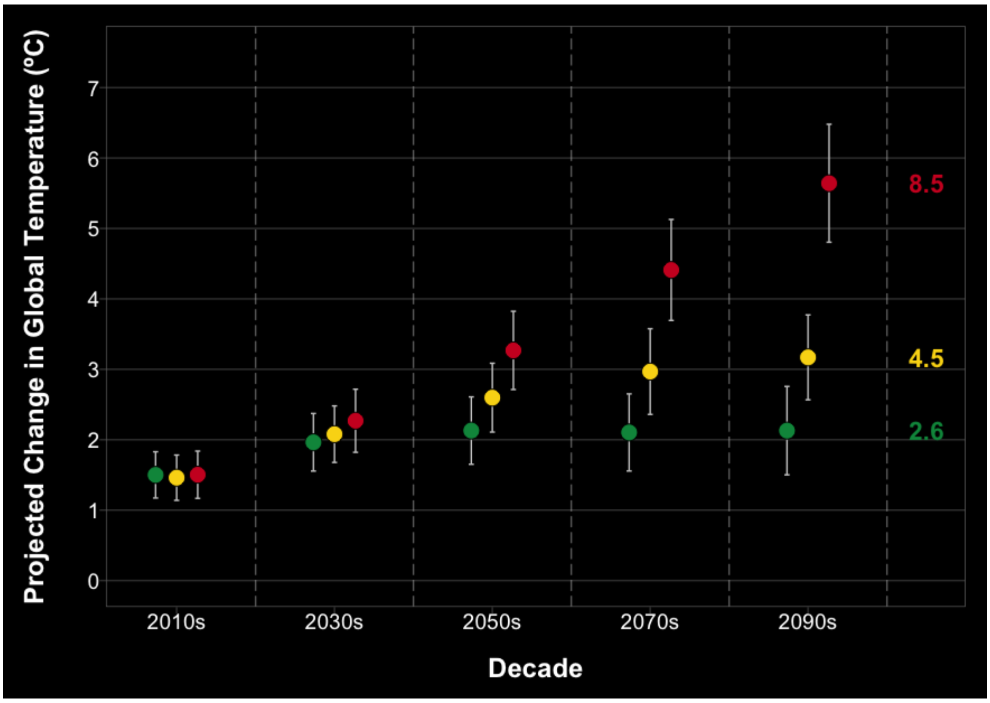


Figure F.4: Survey visualisation: dot2. The dots represent the multi-model average temperature change projected to occur under each of the three scenarios. The error bars depict the standard deviation of these projections. Please note that this visualisation depicts decadal averages and the data points have been slightly offset from one another to prevent overlap.

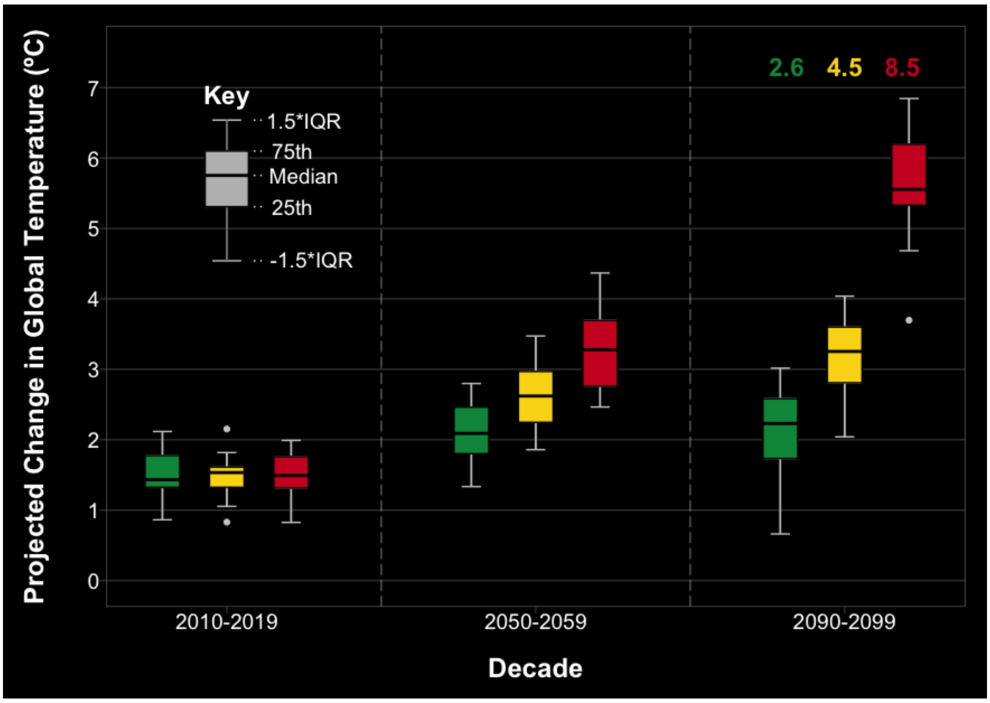


Figure F.5: Survey visualisation: box1. The boxes represent the 25th percentile, the median, and the 75th percentile of the temperature change projected to occur under each of the three scenarios. The error bars depict the minimum and maximum projected temperature change that is within 1.5 times the inter-quartile range. Any data point that falls outside the error bars is deemed an outlier and is shown as a dot (see bottom left). Please note that this visualisation depicts decadal averages, only three of which are shown in this visualisation to avoid over-crowding.

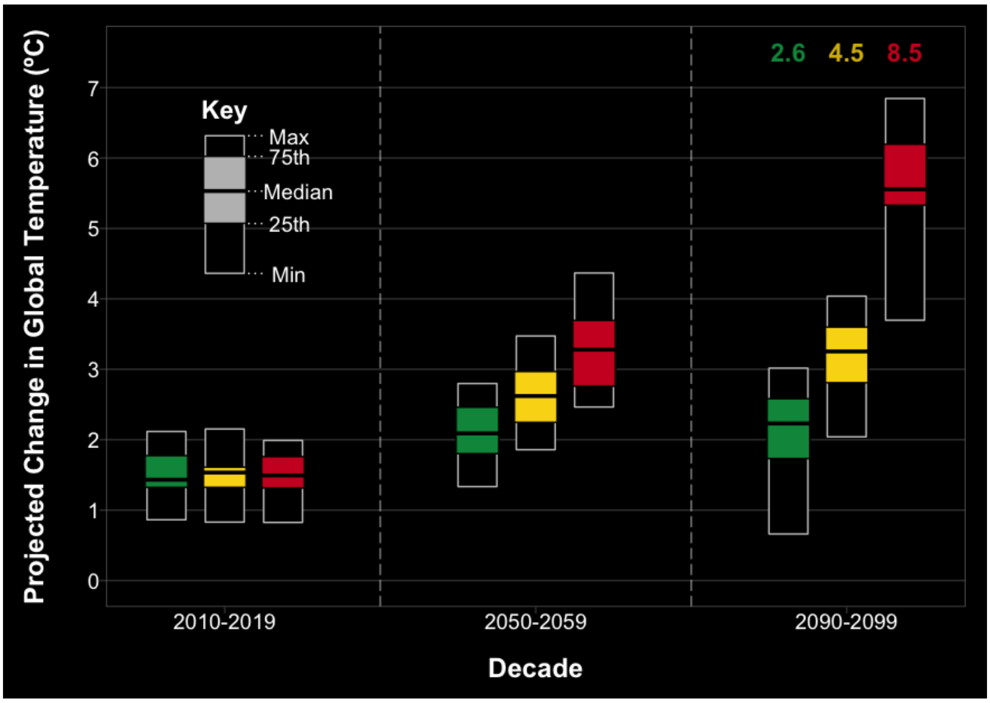


Figure F.6: Survey visualisation: box2. The boxes represent the minimum, 25th percentile, median, 75th percentile, and maximum temperature change projected to occur under each of the three scenarios. Please note that this visualisation depicts decadal averages, only three of which are shown in this visualisation to avoid over-crowding.

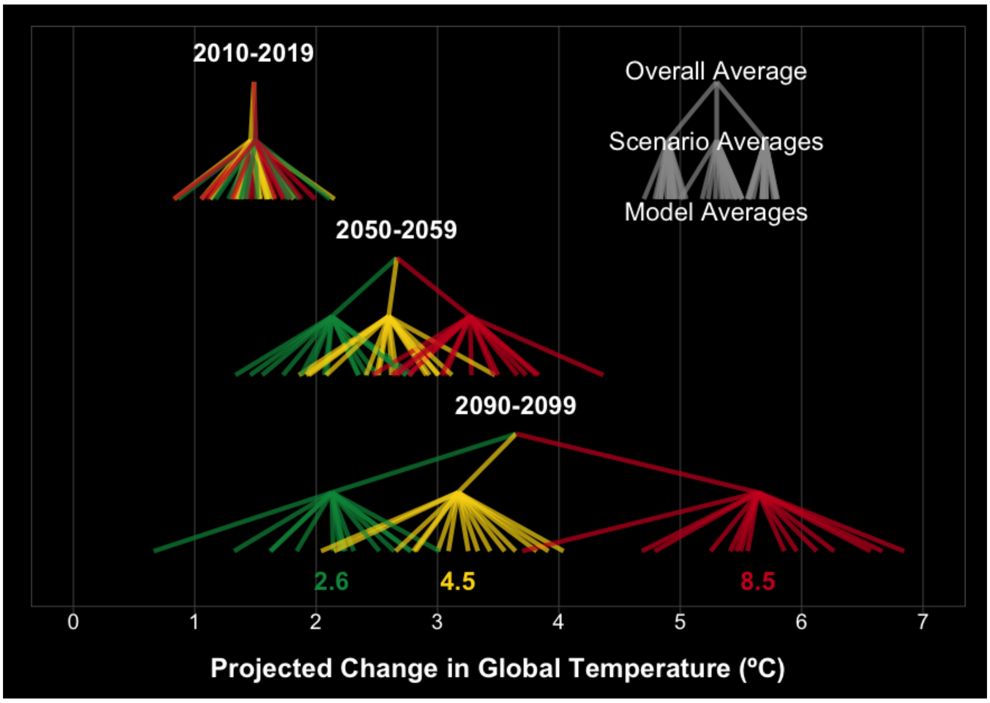


Figure F.7: Survey visualisation: cascade. For a given decade, the overall average of all of the projections is shown at the top as a single point. The three lines moving downwards from this point represent the multi-model average temperature change projected to occur under each of the three scenarios. Finally, the 15 lines moving downwards from each of three scenario averages represent the average temperature change projected to occur by each of the 15 different models. Please note that this visualisation depicts decadal averages, only three of which are shown in this visualisation to avoid over-crowding.

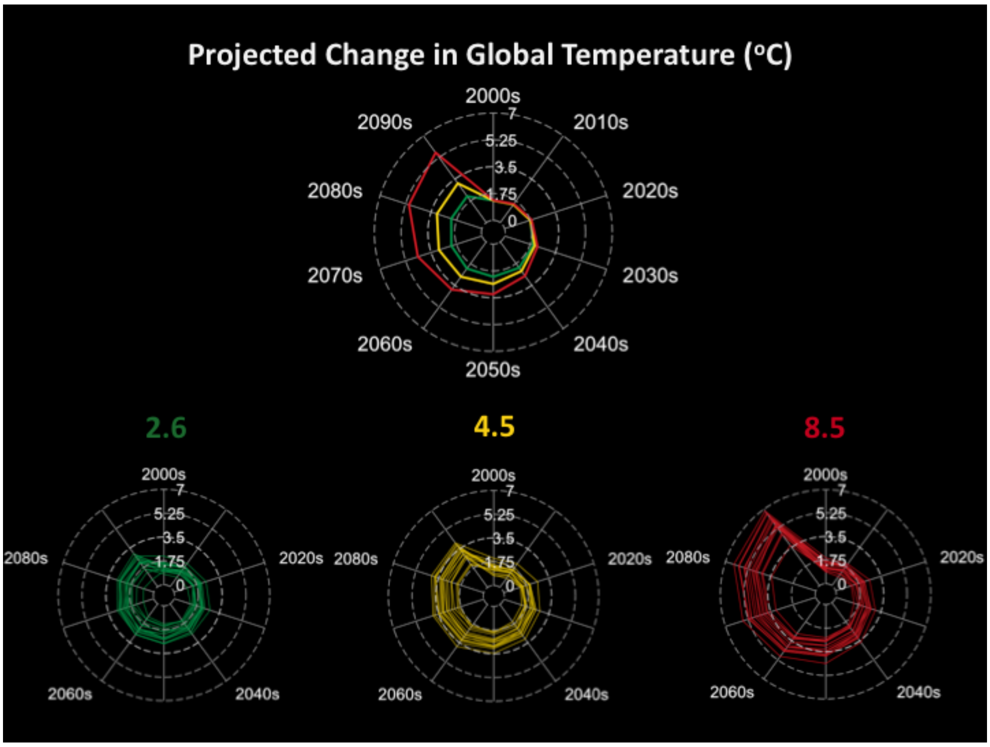


Figure F.8: Survey visualisation: radar. The bold lines (top) represent the multi-model average temperature change projected to occur under each of the three scenarios. The thin lines (bottom) represent the projections from each of the 15 different climate models under each scenario.

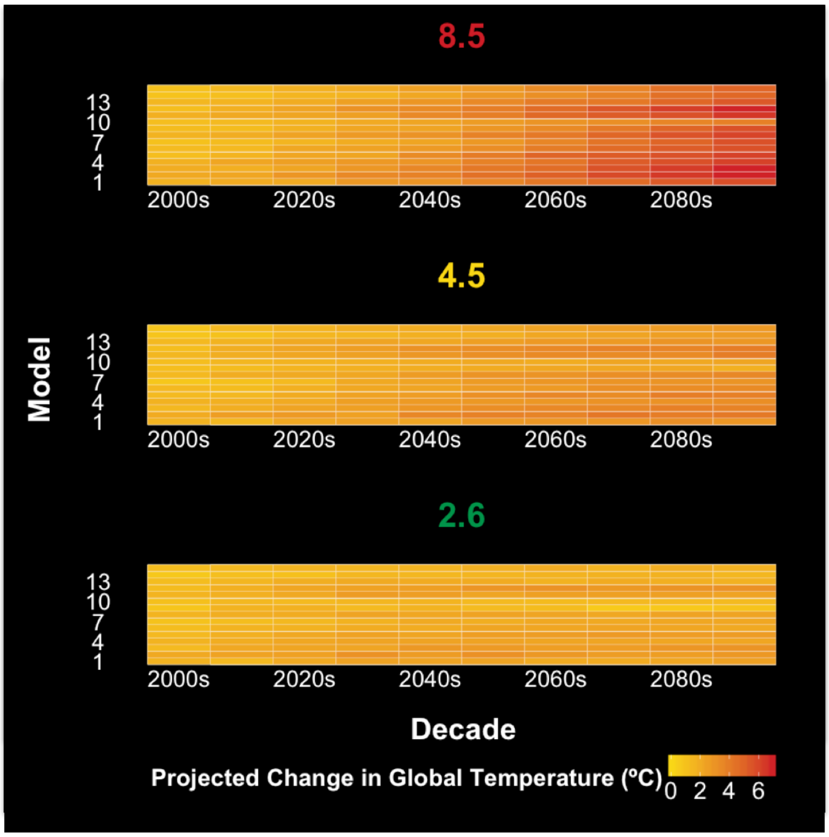


Figure F.9: Survey visualisation: heat. The decadal average temperature change projected to occur under each of the three scenarios (top, middle and bottom). Darker colours (red) represent a larger projected change in global temperature, whilst lighter (yellow) colours represent smaller projected changes in global temperature.

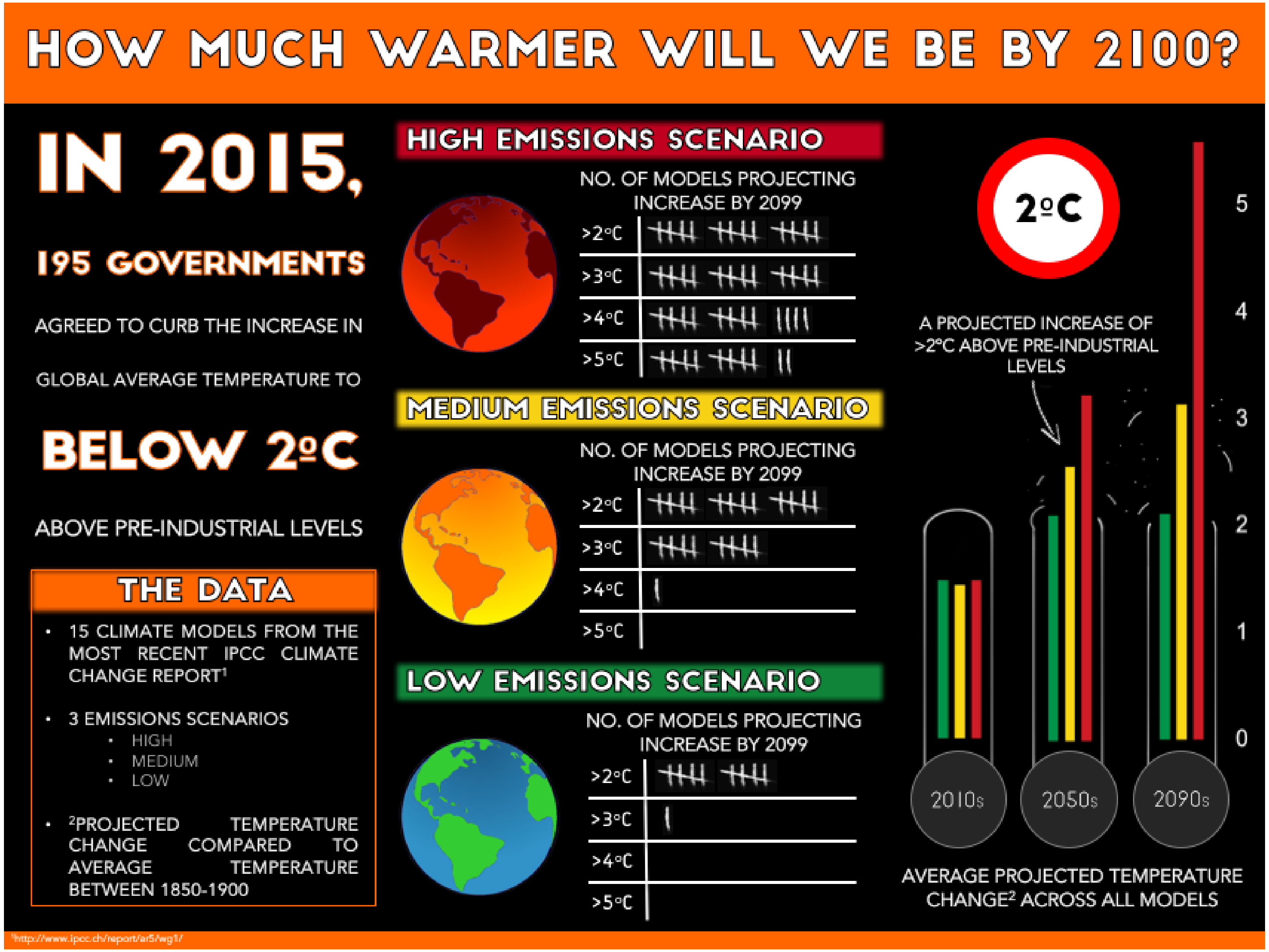


Figure F.10: Survey visualisation: infographic. Centre: The number of models projecting an average temperature increase of over 2, 3, 4 or 5°C by 2099 under each of the three scenarios. Right: The height of the bars represents the multi-model average temperature change projected to occur under each of the three scenarios. Please note that this visualisation depicts decadal averages, only three of which are shown in this visualisation to avoid over-crowding.

Appendix G

Chapter 6: Estimating the minimum and maximum projected temperature change

In the visualisation survey in Chapter 6, the participants were asked to estimate the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade using a randomly selected visualisation. However, it was not possible to correctly identify the minimum and maximum projected temperature change using the box1, dot2 and infographic plots (see Appendix F for all of the visualisations used in the survey). Because of this, we gave the participants the option of selecting ’not applicable’ when they believed they were unable to estimate the minimum and/or maximum. Over 40% of the participants selected ’not applicable’ when asked to estimate the minimum and/or maximum projected temperature change using the infographic, whilst just 4.1% and 6.4% of the participants selected ’not applicable’ when asked to estimate the minimum and/or maximum using the box1 and dot2 plots respectively (Figure G.1). 7.9% of the participants also felt that they were unable to estimate the minimum and/or maximum projected temperature change using the heat plot (Figure G.1). None of the participants were unable to estimate the minimum and/or maximum temperature change projected to occur in a given scenario and decade using the dot1 or line1 plots (Figure G.1).

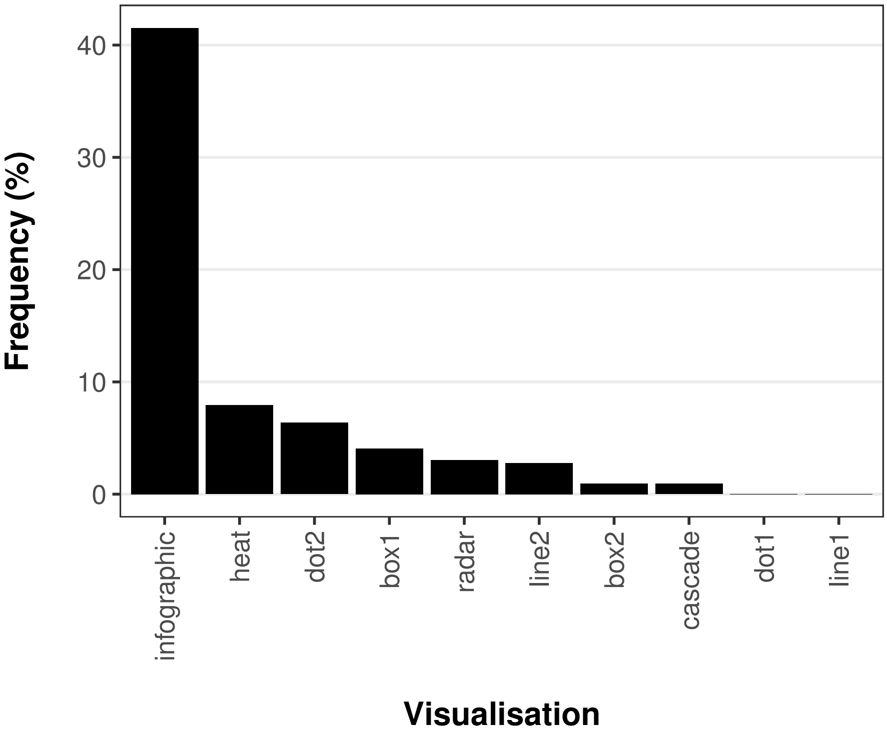


Figure G.1: The frequency (%) with which the survey participants selected the ’not applicable’ option when asked to estimate the minimum and maximum temperature change projected to occur in a given scenario and decade using the visualisation provided.

Appendix H

Chapter 6: Participant accuracy

In the visualisation survey in Chapter 6, the participants were asked to estimate the mean, minimum, and maximum temperature change projected to occur in a given scenario and decade using a randomly selected visualisation. The accuracy with which the participants were able to complete this task was analysed using three Generalised Linear Mixed Models (GLMMs). The predicted absolute difference (x10) (95% confidence interval) between the participants' estimates of the mean, minimum, and maximum projected temperature change and the true values given by the climate models are provided for all possible predictor variables in Figure H.1 to supplement the predictions given in Chapter 6, Section 6.4.2. The predictions shown here are very similar to the coefficients of the GLMMs described in Chapter 6, Section 6.4.2 but with slightly wider confidence intervals.

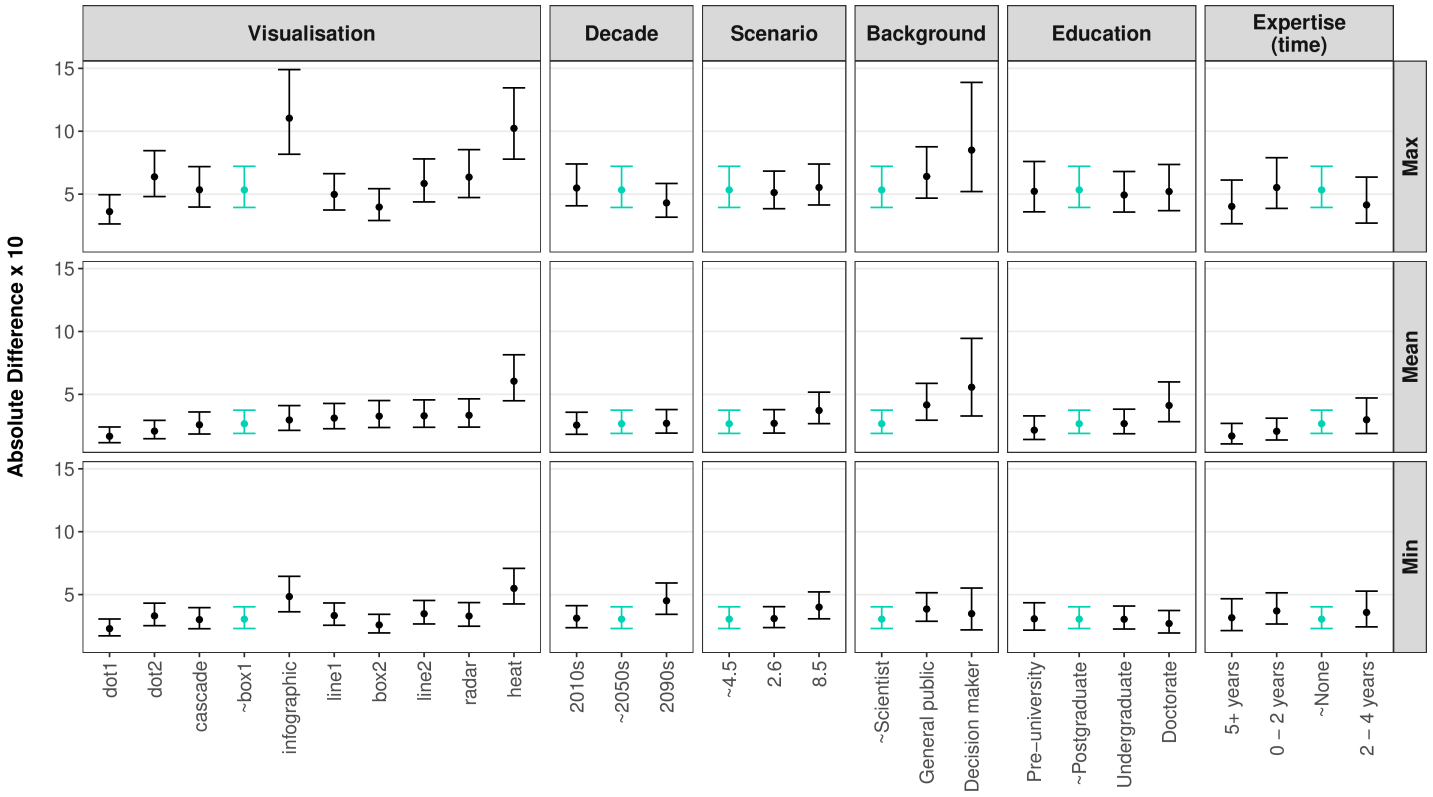


Figure H.1: The predictions (95% confidence interval) of the Generalised Linear Mixed Models that were used to analyse the absolute difference (x10) between the participants' estimates of the minimum (bottom), mean (middle), and maximum (top) temperature change projected to occur in a given scenario and decade and the true values given by the climate models. One level of each predictor variable is highlighted in teal and marked with a tilde to indicate the reference levels used to represent the 'typical' response (see Chapter 6, Section 6.3.5 for further details). Predictions are given for visualisation type, decade, scenario, background, level of education, and expertise in working with environmental models and/or their outputs.

Appendix I

Chapter 6: Visualisation preferences

In the visualisation survey in Chapter 6, Bradley-Terry (BT) models were used to analyse the participants' preferences for different visualisation types across five categories: the ability to view changes in temperature over time, the ability to view changes in uncertainty over time, the ability to retrieve specific values (such as the mean, minimum, and maximum), visual appeal, and ease of understanding. Preferences were measured in terms of predicted 'ability' (referred to as 'preference' from this point forth), where a greater preference score indicates the visualisation was preferred more often than a visualisation with a lower preference score. The preference scores are presented relative to the box1 plot, which was defined as the reference visualisation in the BT models (see Chapter 6, Section 6.3.5 for further details). The preference scores of the visualisations are also presented either with 95% 'comparison' intervals, which are based on quasi standard errors and allow for comparisons to be made across all of the visualisations, or with 95% confidence intervals that are based on (non-quasi) standard errors (see Chapter 6, Section 6.3.5 for further details).

In the main part of the analysis, decision-makers and environmental managers were removed from the data due to small sample sizes. For all five preference categories, the best-fitting models included either background or level of education, although there were no notable differences in the predicted preference scores of each visualisation between scientists and the general public (see Figure I.1 for example). There were only two apparent differences between individuals with different levels of education, both of which are discussed in Chapter 6, Section 6.4.

In an attempt to better understand the visualisation preferences of decision makers and environmental managers, we applied individual BT models to the survey data from the general public, scientists, and decision makers/environmental managers separately. The best-fitting models did not include any of the demographic information provided by the participants and as such the results are presented in terms of the overall preference scores of each visualisation only.

As expected, the 95% comparison intervals of the preference scores of each visualisation were much wider for decision makers and environmental managers than for the general public

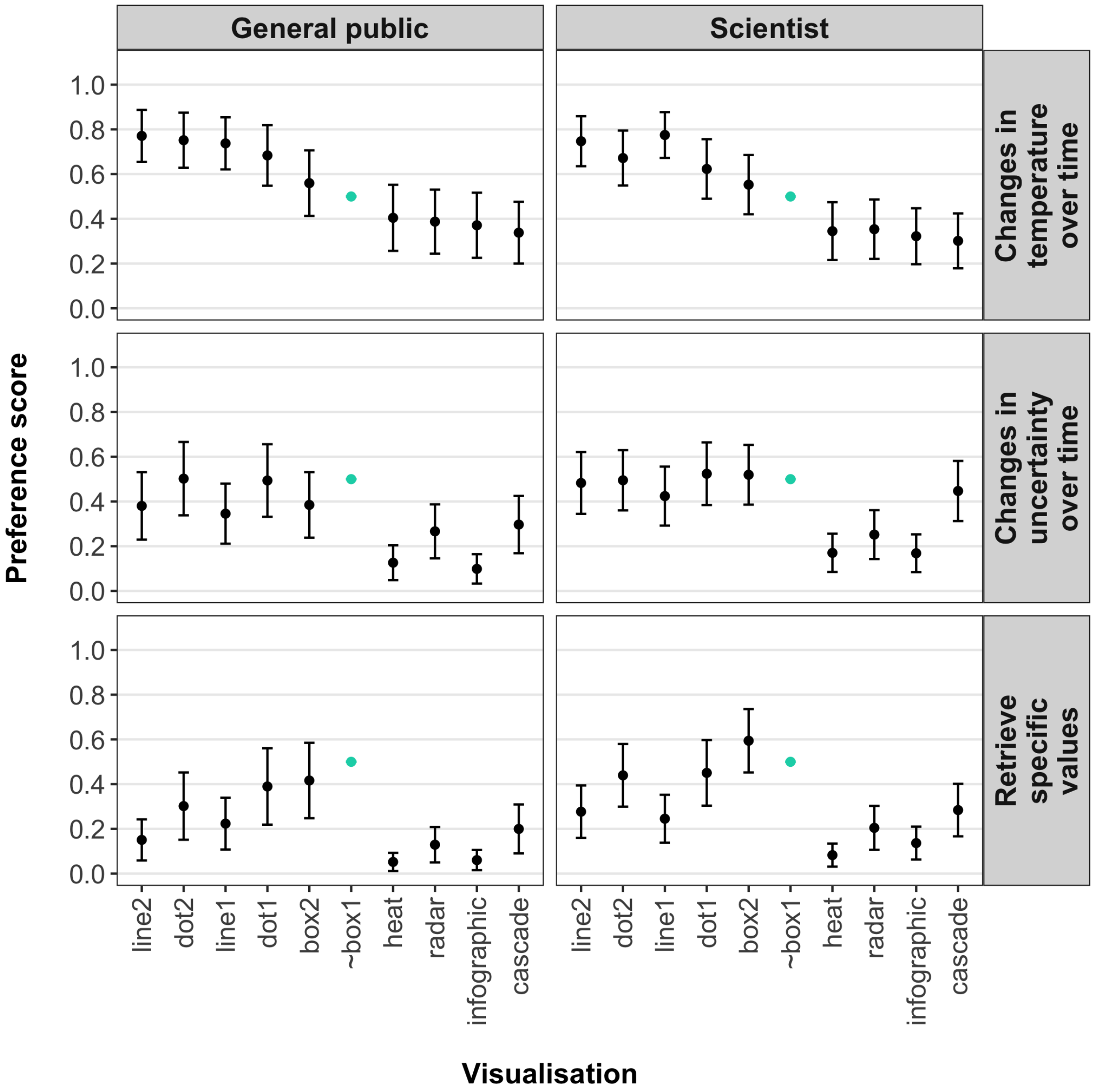


Figure I.1: A comparison of the predicted preference scores (95% confidence intervals) of each visualisation based on the preferences displayed by the general public (left) and scientists (right) in the following three categories: the ability to view changes in temperature over time, the ability to view changes in uncertainty over time, and the ability to retrieve specific values (such as the mean, minimum, and maximum) (see Chapter 6, Section 6.4.4 for the effect of participant background on visual appeal and overall ease of understanding). Comparison intervals are not used in this plot as we were unable to estimate the quasi standard errors of the visualisations across different levels of the predictor variables. Decision makers and environmental managers are not included the analysis due to small sample sizes. The preference scores of each visualisation are presented relative to the box1 plot (highlighted in teal and marked with a tilde), which was defined as the reference visualisation in the BT models.

and scientists (see Figure I.2). Such wide comparison intervals prevented any real differentiation between the visualisations based on the preferences displayed by the decision makers and environmental managers. However, the overall patterns in the preferences displayed by this group of individuals seem to broadly follow those of the scientists and the general public (Figure I.2), although a larger sample size would be required to confirm this theory.

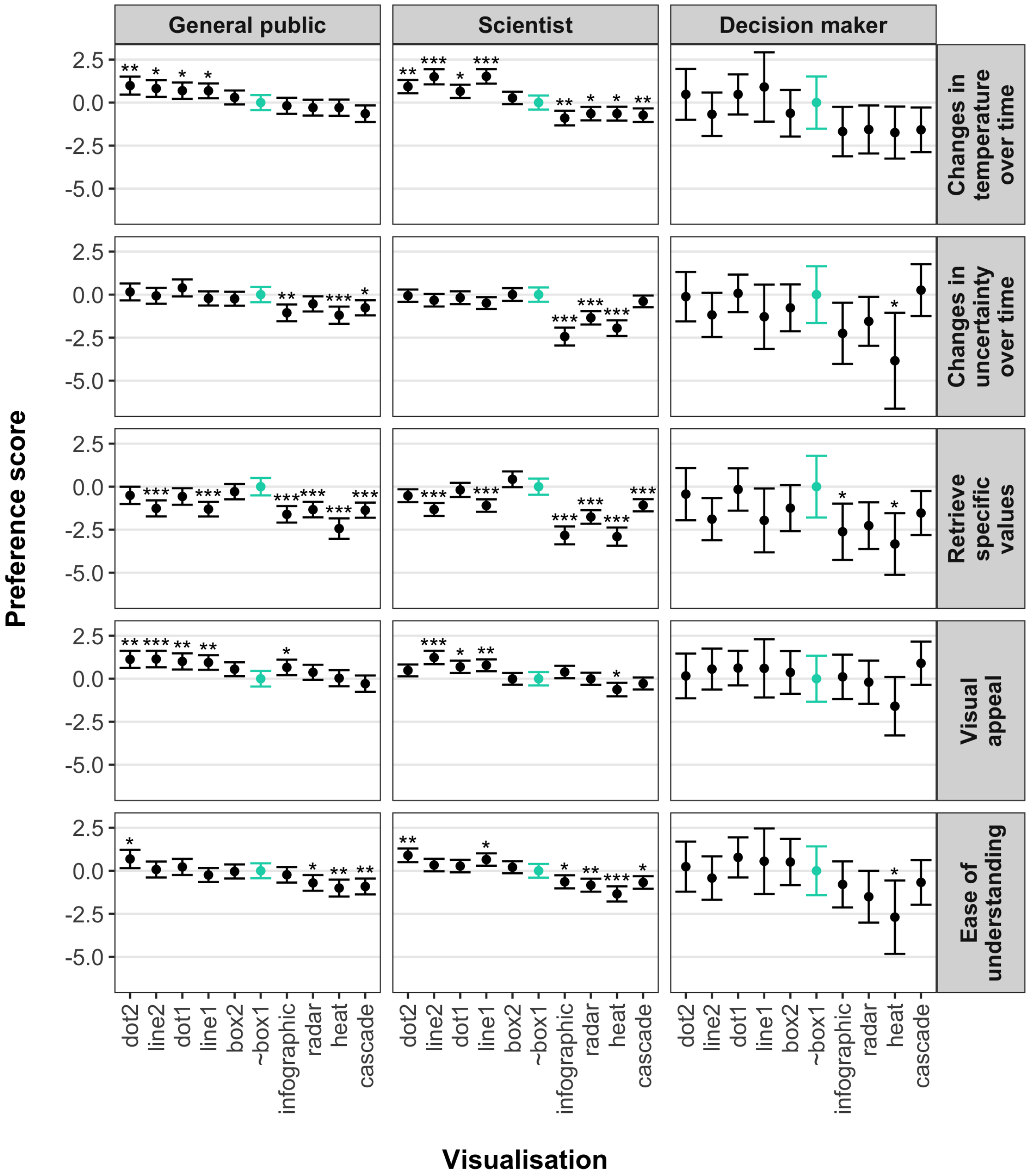


Figure I.2: A comparison of the predicted preference scores (95% ’comparison’ intervals) of each visualisation when a Bradley-Terry model was fit to the survey data from the general public (left), scientists (middle), and decision makers and environmental managers (right) separately. The preference scores of each visualisation are shown for all five preference categories: the ability to view changes in temperature over time, the ability to view changes in uncertainty over time, the ability to retrieve specific values (such as the mean, minimum and maximum), visual appeal, and overall ease of understanding. The 95% comparison intervals are estimated using quasi standard errors to allow for comparisons to be made across all of the visualisations. The preference scores of each visualisation are presented relative to the box1 plot (highlighted in teal and marked with a tilde), which was defined as the reference visualisation in the BT models. *p < 0.05, **p < 0.01, ***p < 0.001.

Appendix J

Chapter 6: Completion time

In the visualisation survey in Chapter 6, we found that the participants with doctoral training tended to be less accurate than those with pre-university, undergraduate, or postgraduate levels of education when they were asked to estimate the mean projected temperature change in a given decade and scenario using a randomly selected visualisation. One possible reason for this pattern is that those with doctoral training spent less time answering the questions, resulting in rushed responses that were slightly less accurate. However, this theory is difficult to prove as the only information we have on the time taken for each participant to complete the survey is relatively unreliable. This is because we only collected data on the time it took for the participants to submit each webpage. As each webpage included one section of the survey, there were multiple questions per page. Because of this, it is difficult to know how long each participant spent solely trying to identify the mean projected temperature change. It is also possible that some of the participants left the browser open whilst not actively taking part in the survey, thus skewing the results. For example, there were some instances where the time taken for a participant to complete section two, which included the interpretation of the mean (see Chapter 6, Section 6.3.3 for full details of section two), exceeded one week. Assuming that none of the participants spent more than one hour actively attempting to complete section two, it is not possible to differentiate between the completion time of the participants based on their level of education (Figure J.1). Further research is therefore required to determine whether the time taken for an individual to interpret a visualisation is affected by their level of education.

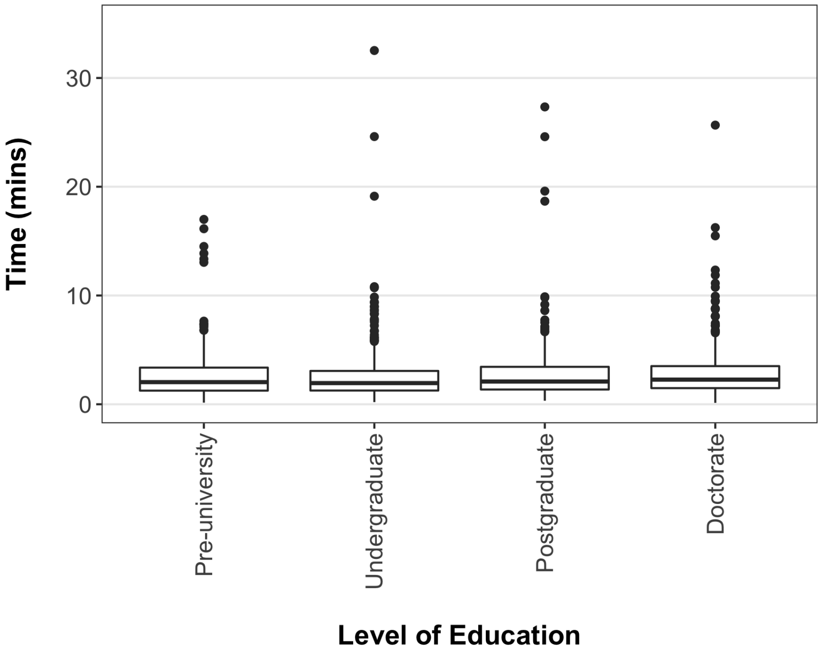


Figure J.1: The time (mins) taken for the participants to complete section two of the survey, which included the interpretation of the mean projected temperature change in a given scenario and decade using a randomly selected visualisation. All observations exceeding one hour were removed from the data prior to plotting.