Habitat Suitability Models for the analysis and prediction of macrobenthos in the North Sea

Habitatgeschiktheidsmodellen voor de analyse en voorspelling van macrobenthos in de Noordzee

Wouter Willems

Promotor: Prof. Dr. Magda Vincx
Co-promotors: Prof. Dr. Steven Degraer, Prof. Dr. Peter Goethals,
Prof. Dr. Vera Van Lancker

Academic year: 2009-2010

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor in Sciences (Biology)

Members of the reading committee:

Prof. Dr. Ann Vanreusel (Ghent University, Ghent, Belgium)

Prof. Dr. Karline Soetaert (NIOO-KNAW, Yerseke, The Netherlands)

Dr. Jan Vanaverbeke (Ghent University, Ghent, Belgium)

Members of the examination committee:

Prof. Dr. Dominique Adriaens, Chairman (Ghent University, Ghent, Belgium)

Prof. Dr. Magda Vincx, Promotor (Ghent University, Ghent, Belgium)

Prof. Dr. Steven Degraer, Copromotor (BMM, Brussels, Belgium)

Prof. Dr. Ing. Peter Goethals, Copromotor (Ghent University, Ghent, Belgium)

Prof. Dr. Vera Van Lancker, Copromotor (BMM, Brussels, Belgium)

Prof. Dr. Ann Vanreusel (Ghent University, Ghent, Belgium)

Prof. Dr. Karline Soetaert (NIOO-KNAW, Yerseke, The Netherlands)

Dr. Jan Vanaverbeke (Ghent University, Ghent, Belgium)

Dr. Tom Ysebaert (NIOO-KNAW, Yerseke, The Netherlands)

List of Abbreviations

AIC	Akaike Information Criterion
AICc	Akaike Information Criterion compensated for small sample sizes
ANN	Artificial Neural Network
AUC	Area Under the Curve
AW	Akaike Weight
BIC	Bayesian Information Criterion
BPI	Bathymetric Position Index
BPNS	Belgian Part of the North Sea
CAIC	Consistent Akaike Information Criterion
CART	Classification And Regression Trees
CCI	Correctly Classified Instances
CMOC	Combined Model Optimisation Criterion
DFA	Discriminant Function Analysis
ENFA	Ecological Niche Factor Analysis
FA	False Absences
FP	False Positives
GAM	General Additive Model
GLM	Generalised Linear Model
HSM	Habitat Suitability Model
LR	Logistic Regression
MAXENT	Maximum Entropy Modelling
MOC	Model Optimisation Criterion
MPA	Marine Protected Area
NMI	Normalised Mutual Information
NPP	Negative Predictive Power
p/a	presence/absence
PCA	Principle Components Analysis
PPP	Positive Predictive Power
pSCI	proposed Site of Community Interest
ROC	Receiver Operator Curve
SAC	Special Area of Conservation
SPA	Special Area of Protection
TA	True Absences
TP	True Presences

Table of Contents

Dankwoord

Wat hier nu voor je ligt is het resultaat van omzwervingen in de wereld van het modelleren en macrobenthos. Het onderwerp habitatmodelleren is niet vanzelf tot stand gekomen. Op uitwisseling in Canada leerde ik GIS kennen en wilde daarmee verderwerken. Na een passage via Magda kwam ik bij de mariene geologie bij Vera Van Lancker wegens de GIS expertise daar. Dan kwam Steven Degraer erbij als macrobenthosexpert. Uiteindelijk kwam Peter Goethals (op aanraden van Ann Vanreusel) en leidde me in, in de wereld van het habitatmodelleren.

Vooreerst wil ik mijn promotor Magda Vincx bedanken en mijn co-promotoren die ieder hun eigen expertise hebben aangebracht: Steven Degraer, Peter Goethals en Vera Van Lancker. Steven was de link tussen het modelleren en de biologische realiteit. Peter bracht me in contact met habitatmodelleren, en met interessante onderzoekers in zijn labo. Ook kon hij begrip opbrengen als mensen modelleer resultaten moeilijk konden vatten. Vera (samen met Els Verfaillie) zorgde voor de geologische ondersteuning in de vorm van GIS-lagen en ook bracht ze me incontact met andere mensen die mariene habitatmodellering doen via het MESH-project. Het IWT bedankt ik voor de doctoraatsbeurs.

Zonder het determineer- en opspoelwerk was er nooit de Macrodat databank die ik tijdens mijn doctoraat heb gebruikt. Alleen met een dergelijke grote databank kan je aan habitatmodelleren doen. Dus aan alle onderzoekers, laboranten en thesisstudenten merci. Ik hoop dat mijn werk de stalen een tweede leven heeft gegeven. Ook de sfeer op de Mariene was altijd wel de moeite tijdens de koffiepauzes, lunch en staalnames. Ook dank aan de ex-bureaugenoten: Sofie, Marijn, Guy, Sarah en Mikhael.

Tijdens mijn doctoraat kreeg ik hulp van twee thesisstudenten. Yves Salembier testte in habitatselectie-experimenten de modellen aan de realiteit. Tran Chinh Khuong, thanks for your interest in habitat modelling. You worked really independent, and managed modelling in R in a short time. I hope you can build a nice carreer in marine ecology.

Technische ondersteuning kreeg ik van verschillende mensen. Els Verfaillie, bedankt voor de vele samenwerking, zonder jouw GIS-lagen waren mijn resultaten niet half zo mooi en kon ik geen gebiedsdekkende kaarten maken. Ans Mouton en Andy Dedecker (Labo Aquatische Ecologie) gaven me een intro tot neurale netwerken en Matlab. Karel van Den Meersche, mede-assistent tijdens het practicum statistiek leidde me in in de wereld van R, een programma dat ik nu moeilijk zou kunnen missen. Stefan Vanaelst controleerde als statisticus mijn statistische creatie in de vorm van een model selectie criterium. Zac, thanks for correcting the English in the final version.

Ik kreeg de kans om mee te werken aan internationale projecten. I am greatful to the people in the ICES North Sea Benthos Project for the use of the North Sea macrobenthos data set and the interesting discussions. In the Mapping European Sea Habitats (MESH) project I met with some other researcher practicing habitat modelling for marine species. During the ecological modelling congresses I met some interesting people from all over the globe, and also the post-congres activities were worth it.

Als je bijna tien jaar in Gent blijft plakken, leer je al wat mensen kennen. Dat begon met studiegenoten zoals Bart J. en Davy die steevast bewonderend naar mijn grafieken keken. De Joris, de eeuwige laatste in de poolstatistiek. En dan zijn er natuurlijk nog Tanja, Katinka en nog zovelen meer. Door samenhuizen leerde ik heel wat volk kennen. Klaas, Bart, Pieter VO., Ellen en Sarah deelden met mij het huis. Bedankt voor de vele leuke momenten, het samen koken (en in het begin leren koken, bedankt Bart), de feestjes en uitstapjes. Vrienden van vrienden worden ook vrienden. Zo leerde ik via mijn huisgenoten nog wat volk kennen, zoals daar zijn: Martine, Griet, Mathias, Maureen, Jeroen, Pieter P., Mieke, Caroline, Heleentje, Bastiaan, An, Eric, Sarah Debaere, Jonas... Bedankt voor de gezellige feestjes, Blaarmeersen bezoekjes, spelletjesavonden waar al dan niet bloed vloeide tijdens Jungle Speed. En natuurlijk bedank ik ook de vrienden in de thuishaven Sint-Truiden: Ralf, Kris en Stijn. Bedankt voor de vele discussies in Op de Beeck of heel vroeger de Duplex. Het doet altijd weer deugd om te merken dat er nog niet veel veranderd is in Limburg ;)

Mijn passie voor fotografie en duiken kan ik uitleven samen met de mensen van de duikclub. Het zijn er teveel om op te noemen, maar bedankt voor de super sfeer op de weekendjes en clubreizen en de niet-duikactiviteiten. Dat we nog veel samen mogen duiken.

Mama en papa, zonder jullie was ik nooit aan mijn studies biologie kunnen beginnen. Merci om in mij te geloven en mij aan te moedigen om door te gaan. Ook dank aan mijn zus Marijke voor de hulp en steun als het wat moeilijker ging.

Tineke, in de laatste maanden was jij mijn tango-partner en luisterend oor. Je hebt van dichtbij meegemaakt wat een doctoraatstudent moet doorstaan om het boekje af te krijgen. Ik hoop je nog lang aan mijn zijde te hebben.

General Summary

The North Sea, characterised by densely populated and highly industrialised coastlines, is ranked among the marine regions most impacted by human pressures worldwide. Following impacts have been demonstrated as being problematic for the management of the North Sea ecosystem: 1) input of nutrients, chemicals and sewage, 2) shipping activities, 3) increase in invasive species, 3) heavy fishing activities and, 4) climate change effects (e.g. changes in temperature or current patterns).

Marine management is needed to halt further degradation and to evaluate and restore impacted sites. For this purpose, several policy instruments have been implemented in the EU: EU Habitat Directive (92/43/EC), EU Birds Directive (79/409/EC), EU Water Framework Directive (2000/60/EC) and more recently the EU Marine Strategy Framework Directive (2008/56/EC). Marine managers are often faced with limited and uncertain ecological information on which to base their decisions. However, the efficient implementation of marine decision support systems as a tool for marine spatial planning and management requires understanding of the processes that determine the observed distribution patterns of species in marine ecosystems. It is therefore necessary to gain insight in the temporal and spatial distribution of each ecosystem component (e.g. plankton, fish, seabirds, benthos). In practice, this means that good species distribution maps of important ecosystem components are required, that cover the marine region to be managed.

Currently the distribution of species is mostly known from point observations, and full cover species distribution maps are lacking. However, environmental variables are often available at a full cover scale (e.g. sediment grain size). Habitat suitability models (HSMs) relate the presence or abundance of a species in a location to a set of environmental variables, which then allows predicted distributions to be mapped across an entire region. Such full coverage species distribution maps are a good basis for marine management decisions; furthermore HSMs allow scenario simulations. HSMs can produce predictions of the species distribution, based on abiotic variables, at locations where information on species distribution was previously unavailable. These abiotic variables are often available on a full coverage basis, e.g. sediment grain size maps or satellite based temperature.

HSMs are a relatively recent modelling approach and their use is increasing. HSMs are also known under the synonyms “species distribution models” and “niche models”. HSMs originated in terrestrial ecology, but the discipline has entered into the field of marine ecology. HSM provide a cost-effective tool to integrate current data and knowledge, and to identify and prioritise locations for conservation. HSMs predict the suitability of the habitat in relation to the habitat preference of a species. The assumption is that, the more suitable the environment, the higher the probability that the species is present and will be found in high densities.

This Ph.D. thesis focused on the prediction of the spatial distribution of macrobenthos in the Belgian Part of the North Sea (BPNS) with HSMs. Macrobenthic species, defined as animals larger than

1 mm living in/upon the sea bottom, are often the main component in environmental monitoring programmes to evaluate the status of benthic ecosystems for marine management. There are several good reasons why macrobenthos is used in monitoring: 1) the species are macroscopic and thus more easy to handle and identify, 2) they are relatively immobile and thus strongly dependent on local conditions, 3) they are linked with biogeochemical processes in the sediment and perform important ecosystem functions (e.g. increasing habitat complexity, bioturbation, oxygenation) and, 4) benthic animals are the food source for many benthic fish, including the economically important species, such as plaice (*Pleuronectes platessa*) and cod (*Gadus morhua*) in the North Sea, and diving birds, e.g. the Common Scoter (*Melanitta nigra*).

Despite the increased number of applications of habitat suitability modelling in recent years, the methodology of these models can still be improved significantly. Therefore, **the general objective of this thesis** was the improvement of the existing methodology and more specifically the approach to model the distribution of macrobenthos species. In this thesis, HSMs have only been developed for a limited number of macrobenthos species, but with the modelling methodology proposed, models can be developed for other species in an efficient way.

In the introduction (**Chapter 1)** the challenges in marine management in the North Sea ecosystem are illustrated and the importance of macrobenthos species distributions to monitor the environmental status is highlighted. A technical introduction to HSMs for non-experts is presented as well, in order to give ecosystem managers a necessary background.

Based on the challenges identified in the current modelling methodology of HSMs, the objectives of this thesis were also laid out in Chapter 1. Three major challenges were identified in the modelling methodology; these will be treated in separate chapters: 1) choice of the modelling technique, 2) selection of the most optimal model (model selection) and, 3) assessment of the model reliability (model validation).

Challenge 1: Which modelling technique to use?

Several modelling techniques are being used to develop HSMs. The choice between different techniques is not obvious, as each technique claims to be the most optimal. In **Chapter 2**, the modelling techniques Logistic Regression (LR; a type of Generalised Linear Model, GLM) and Artificial Neural Networks (ANNs) were compared in their ability to predict the occurrence of *Lanice conchilega*, a common tube-building polychaete along the North-western European coastline. LR and ANNs were chosen, because they are the most frequently used statistical and machine learning techniques, respectively. ANNs are claimed to have a higher predictive performance, as they can model more complex species-environment relations. However, ANNs are also deemed to be black box models, as it

is difficult to gain insight in their inner workings. **Appendix I** provides an extensive technical background on ANNs, and **Appendix II** on GLMs.

Models were developed to predict the spatial distribution of the species *Lanice conchilega*. This species is known as a habitat engineer, increasing macrobenthic species diversity and abundance in soft sediments, through enhancement of the habitat complexity. *L. conchilega* is also an important food source for several demersal fish and, when occurring in high densities *L. conchilega* aggregations act as a refugium against predation for many organisms.

The models were developed with a data set collected in the Western Coastal Banks, a small region in the south west of the Belgian Part of the North Sea (BPNS). This data set was chosen because the samples were collected on a high resolution sampling grid (500m) and numerous environmental variables were measured. Some of those variables are not available in the data set of the complete BPNS region; as such the relevance of these variables for modelling could only be tested with the Western Coastal Banks data set.

Although several types of environmental variables were available in the data set (grain-size, currents, nutrients) only grain-size variables were used in the final models (median grain-size, mud and coarse fraction). ANN gave a higher predictive performance, based on a number of performance indicators (% correctly classified instances, area under the curve, specificity and sensitivity); however, there was a high correlation between the model output of LR and ANN.

ANNs can fit very complex species responses, but this can easily lead to overfitted models, that are not transferable to other regions or periods. In the *Lanice* example in this thesis, there was quite a high dissimilarity between ANNs produced for each of the three crossvalidation folds. The LRs, on the other hand, had a more similar predictive performance in each crossvalidation fold, pointing to a higher robustness of LRs. Presently, there is no established theory to determine the number of interneurons or the choice of the transfer functions for an ANN. Also, as no error distributions are assumed, no statistical tests are performed; hence the model output or model parameters of ANNs are not tested for significance. From a parsimonious point of view the LRs were superior in the modelling of the response of *Lanice*, as the model was simpler and the ANNs only performed slightly better.

Challenge 2: What is the most optimal combination of predictive variables?

The general aim of **Chapter 3** was the improvement of the model selection methodology for HSMs. HSMs use a combination of environmental variables that are assumed to determine the distribution of species. One of the challenges in habitat suitability modelling is the selection of an appropriate subset of predictive variables from all the variables in a data set. This variable selection and the choice of the

response modelled per variable (e.g. linear or unimodal) are together called model selection; this is considered a central step in the model development. Model selection methods aim at determining the most optimal model by adequately constraining the number of predictive variables used, and thus the model complexity. Stepwise model selection (e.g. forward or backward selection) is most often used in habitat suitability modelling. However, the stepwise model selection has a number of drawbacks: 1) due to the stepwise nature, the global optimal can be missed; 2) the stepwise selection is sensitive to multicollinearity of the predictive variables; and 3) stepwise selection precludes multimodel inference and prediction as only one model is selected.

In this chapter, a new model selection approach was proposed: the Combined Model Optimisation Criterion (CMOC). The CMOC approach has been set up to deal with the shortcomings in the model selection approach that is currently most often used in regression based HSMs: a stepwise model selection based on a single data set and without paying attention to the species prevalence in this data set. The CMOC is based on the general model optimisation criterion (MOC) framework incorporating both model fit and model complexity. CMOC values are calculated based on five MOCs: the Akaike Information Criterion (AIC), an adapted version of the AIC (AICc), the Bayesian Information Criterion (BIC), the Consistent Akaike Information Criterion (CAIC) and the F-statistic. The MOCs of the model calibration and test set are combined in the CMOC; in this way the generalisation ability of the models on an independent test is incorporated in the model selection. Good generalisation is necessary to use the model in other regions or periods.

The proposed CMOC methodology was first tested on artificial data of a virtual species, because this provides full control over the data set. Overall, the proposed model selection approach managed to find the true models that were used to generate the virtual species presences. Logistic regression was used as a modelling technique, because this technique is often used in HSMs, is relatively simple and is well established statistically. In a second step, the model selection methodology was applied to field observations of *Abra alba*, a marine bivalve species. This species is an indicator for the *A. alba* macrobenthic community in the Southern North Sea. The CMOCs, based on each of the five MOCs, selected slightly different variable combinations, but the variables median grain size, depth and bathymetric position index were selected with each of the five CMOCs.

The proposed CMOC model selection approach has a number of advantages over the current methodology. The exhaustive testing of all variable combinations increases the chance of finding an optimal model, and also causes the CMOC approach to deal better with multicollinearity of predictive variables. The CMOC is not based on a contingency table and therefore does not require the arbitrary choice of a cut-off for presence. Bootstrap resampling is used in the CMOC approach to increase the reliability of the model selection. This resampling produced model replicas, which increases the

robustness of the model performance estimate. During the resampling the prevalence of the species was always kept at 50% to avoid influences of the species prevalence on the parameter estimates. The CMOC model selection can be applied to any modelling technique for which a likelihood can be calculated (e.g. GLM, general additive models, ANNs).

Challenge 3: Are the model predictions reliable?

There is a disproportionally large effort in the development of HSMs, compared to the model validation. Traditionally, HSMs are only validated by comparing the observations with the model predictions. The general goal of **Chapter 4** is to plea for an integrated validation of marine HSMs which also validates the ecological soundness of the models. Such an integrated validation considers: 1) the classical model validation of observations vs. predictions; 2) a conceptual scheme bringing together ecological knowledge from literature and allowing inference about the causality of predictive variables; 3) habitat preference experiments allowing distinguishing the fundamental and realised niche; and 4) an assessment of the sample distribution over the range of the predictive variables.

To illustrate the suggested model validation improvements, HSMs were developed for the marine bivalve species *Donax vittatus* (Da Costa 1778). This species was chosen because: 1) extensive ecological literature is available on the species, 2) habitat preference experiment results are available and, 3) the species is a food source for juvenile plaice (*Pleuronectes platessa*). The Combined Model Optimisation Criterion (CMOC) was used for model selection. Logistic regression was used as a modelling technique. Four models were retained after the model selection, and a multimodel prediction was performed, with the CMOC as a weighing factor. Only depth and median grain size were retained as predictive variables in the selected models.

Ecological knowledge from literature on the species was combined in a conceptual scheme. Such a scheme, as well as the habitat preference experiments, was useful to determine if the relation of a variable with the species distribution was causal or rather correlative (i.e. a proxy variable). Knowledge on the causality of a variable is crucial when models are transferred to other regions or periods. Experimental validation confirmed the modelled sediment grain size response. Habitat preference experiments allowed identifying the fundamental niche, while field observations can only provide insight in the realised niche. If habitat preference experiments are available, they should be used to assess the causality of variables in the prediction of the species distribution. If no experiments are available, the feasibility of experiments should be considered. The distribution of samples over the range of each variable was assessed. This allowed to determine if the sampled range is sufficiently sampled and if a variable is correlated to the presence of the species. Sometimes ecologically useful

variables will not be chosen in the model selection as the sampled range is insufficient. For each variable in the data set an integrated discussion was provided why this variable was (not) chosen in the final models; this was based on the conceptual scheme, the experimental results and the sample distribution along the range of the variable.

General discussion

In **Chapter 5** a general discussion was provided. The improvements to the modelling methodology as identified in the three challenges, were discussed in a broader sense. A future outlook on habitat suitability modelling was provided. This included an overview of the most important sources of error and bias in HSMs. Next, an overview of some of the ecological traits of species influencing the model development was provided. Another source of uncertainty that was discussed related to HSMs, is the spatiotemporal mismatch of environmental variables and species observations. As a final part of the discussion on the future of HSMs, some developments in modelling techniques are discussed. An overview of the applications and advantages of spatially explicit and mechanistic models is provided.

A separate discussion on the macrobenthos models developed in this thesis is provided. Finally some applications of HSMs for macrobenthos species in the BPNS are proposed: HSM-based sampling, scenario simulations and further research in community HSMs for macrobenthos.

The **general conclusions** of this research are:

- When model parsimony is considered important, logistic regression models are superior; these models are simpler and the predictive performance is only slightly lower than the ANNs.
- The CMOC model selection approach has several advantages compared to the widely used stepwise model selection. The main advantage is that the model validation is incorporated in the model selection process.
- An integrated validation of HSMs is necessary to develop reliable and ecologically sound models. Such an integrated validation considers: 1) model validation with field observations, 2) ecological knowledge combined in a conceptual scheme, 3) habitat preference experiments; and, 4) influence of the sampled range of each variable.
- Habitat suitability modelling is a very useful tool to model the spatial distribution of macrobenthos species in the North Sea. This was demonstrated for several species in this research. More specific applications of HSMs for macrobenthos species are to be expected in the near future.

Algemene samenvatting

De Noordzee wordt algemeen beschouwd als een mariene regio die onder sterke invloed van menselijke activiteiten staat, omwille van de dichtbevolkte en geïndustrialiseerde landen die er rond liggen. Verschillende negatieve impacten op het Noordzee-ecosysteem zijn veroorzaakt door activiteiten op het vasteland: aanvoer van nutriënten, chemicaliën en rioolwater. Enkele van 's werelds drukste havens liggen aan de Noordzee, wat deze zee één van de meest bevaren mariene gebieden wereldwijd maakt. Zo een intensief marien verkeer verhoogt de kans op olielekken en de introductie van invasieve soorten via het ballastwater. De vissersvloot bestaat hoofdzakelijk uit boomkorvaartuigen, die netten met zware wekkerkettingen die diep in de bodem ploegen en benthische soorten verstoren. Klimaatverandering is een andere bedreiging voor de mariene biodiversiteit in de Noordzee. Enkele van de verwachte effecten zijn veranderingen in de wind en stromingspatronen en verzuring door oplossen van CO_2.

Marien beheer is nodig om een verdere degradatie van de Noordzee te voorkomen en om de natuurlijke waarde te beschermen. Voor dit doel zijn door de Europese Gemeenschap verscheidene beheersinstrumenten gelanceerd: EU Habitat Richtlijn (92/43/EC), EU Vogel Richtlijn (79/409/EC), EU Water Framework Directive (2000/60/EC) en meer recent de EU kaderrichtlijn Mariene Strategie (2008/56/EC). Mariene beheerders worden vaak geconfronteerd met beperkte ecologische informatie, waarop ze hun beslissingen moet baseren. Nochtans vereist de efficiënte implementatie van beslissingsondersteunende systemen in marien beheer kennis van de processen die de geobserveerde ruimtelijke verspreiding van ecosysteem componenten bepalen. Daarom is het noodzakelijk om inzicht te verwerven in de temporele en ruimtelijke verspreiding van iedere ecosysteemcomponent: vissen, zeevogels, macrobenthos, meiobenthos, etc. In de praktijk betekent dit dat goede verspreidingskaarten van soorten nodig zijn voor het mariene gebied dat beheerd wordt.

Op dit moment is de verspreiding van soorten vooral gekend van puntwaarnemingen en gebiedsdekkende soortverspreidingskaarten ontbreken vaak. Omgevingsvariabelen (bv. korrelgrootte), zijn echter vaak beschikbaar op een gebiedsdekkende schaal. Habitatgeschiktheidsmodellen (HGM's) relateren het voorkomen of de densiteit van een soort op een bepaalde plaats aan de omgevingsvariabelen, wat dan toelaat om de verspreiding van de soort te voorspellen over een heel gebied. Zulke gebiedsdekkende verspreidingskaarten van soorten vormen een goede basis voor mariene beheersbeslissingen, HGM's laten ook toe simulaties te doen. HGM's kunnen de verspreiding van een soort voorspellen, op locaties waar enkel omgevingsvariabelen beschikbaar zijn. Deze variabelen zijn meestal beschikbaar op een gebiedsdekkende schaal. Het alsmaar toenemende aantal mariene karteringsinitiatieven zorgt ervoor dat er veel gebiedsdekkende variabelen ter beschikking komen.

HGM's zijn een relatief recente modelleertechniek en het aantal toepassingen neemt nog steeds toe. HGM's zijn ook gekend onder de synoniemen "soortverspreidingsmodellen" (species distribution models) en "niche modellen" (niche models). HGM's kunnen gebruikt worden als een kostenefficiënte methode om de huidige kennis en data te integreren. HGM's voospellen de geschiktheid van het habitat in relatie tot de habitatpreferentie van een soort. De assumptie is dat, hoe geschikter het habitat hoe hoger de kans dat een soort aanwezig is of voorkomt in hoge densiteiten.

Dit doctoraat focuste op de voorspelling van de ruimtelijke verspreiding van macrobenthos in het Belgisch Deel van de Noordzee (BDNZ) aan de hand van HGM's. Macrobenthossoorten, gedefinieerd als dieren groter dan 1 mm die in of op de bodem leven, zijn vaak het belangrijkste studieobject in milieumonitoring programma's die als doel hebben de status van benthische ecosystemen te evalueren. Er zijn verschillende goede redenen waarom macrobenthos gebruikt wordt in monitoringsprogramma's: 1) de soorten zijn macroscopisch en dus makkelijker te identificeren, 2) ze zijn relatief immobiel en dus sterk afhankelijk van lokale omstandigheden, 3) ze zijn sterk gekoppeld aan de biogeochemische processen in het sediment en voeren belangrijke ecosysteemfuncties uit (verhogen habitatcomplexiteit, bioturbatie, oxygenatie, etc.) en 4) benthische soorten zijn een voedselbron voor bodemvissen zoals (*Pleuronectes platessa*) schol en kabeljauw (*Gadus morhua*) en sommige zeevogels, bv. de zwarte zeeëend (*Melanitta nigra*).

Ondanks het stijgend aantal toepassingen van HGM's de laatste jaren, kan de methodologie van deze modellen nog significant verbeterd worden. De algemene doelstelling van dit doctoraat was dan ook de verbetering van de bestaande modelleermethodologie, en meer specifiek de benadering om macrobenthossoorten te voorspellen. HGM's zijn enkel ontwikkeld voor een beperkt aantal soorten, maar de voorgestelde methodologie kan voor andere soorten op een efficiënte wijze toegepast worden.

Hoofdstuk 1 van dit doctoraat is een introductie die de lezer de nodige achtergrond verschaft in de huidige uitdagingen in marien beheer in de Noordzee en hoe macrobenthossoorten gebruikt kunnen worden om de milieustatus te monitoren. Een tweede deel van de introductie is een technische handleiding tot HGM's voor niet-experts. Verschillende aspecten van HGM's werden besproken: de theoretische achtergrond van HGM's, data voor modelleren, modelleertechnieken, modelselectie en model validatie. Vervolgens werd een overzicht gegeven van de huidige toepassingen van HGM's in marien beheer. Gebaseerd op de uitdagingen in de huidige modelleer methodologie van HGM's, werden op het einde van hoofdstuk 1 de doelstellingen van dit doctoraat voorgesteld. Drie hoofddoelstellingen werden geïdentificeerd in het verbeteren van de modelleermethodologie: 1) keuze van de modelleertechniek, 2) selectie van meest optimale model (model selectie) en, 3) nagaan van de

modelbetrouwbaarheid (model validatie). Ieder van deze drie topics werd als een uitdaging beschouwd in het verbeteren van de huidige modelleermethodologie.

Uitdaging 1: welke modelleertechniek te gebruiken?

Verschillende modelleertechnieken worden gebruikt om HGM's te ontwikkelen. De keuze tussen verschillende technieken is vaak niet eenvoudig, omdat vele technieken claimen de beste te zijn. In **hoofdstuk 2** werden de modelleertechnieken Logistische Regressie (LR; een type Gegeneraliseerde Lineaire Model, GLM) en Artificiële Neurale Netwerken (ANN's) vergeleken. Deze technieken werden gekozen omdat ze respectievelijk de meest gebruikte statistische en artificiële intelligentie modelleer technieken zijn. Over ANN's wordt beweerd dat ze een hogere voorspellingskracht hebben omdat ze meer complexe relaties kunnen modelleren tussen een soort en de omgevingsvariabelen. Maar ANN's worden ook vaak als "black box" modellen beschouwd, omdat het moeilijk is inzicht te krijgen in de interne werking van de modellen. In **Appendix I** werd een uitgebreide technische achtergrond over ANN's gegeven, en in **Appendix II** over GLM's.

HGM's werden ontwikkeld met beide technieken voor de soort *Lanice conchilega*, een veel voorkomende soort polycheet langs de Noordwest-Europese kustlijnen. Deze soort is gekend als een habitatingenieur, die lokaal de diversiteit en densiteit van geassocieerde soorten verhoogt in zachte substraten, door de habitatcomplexiteit te verhogen. *L. conchilega* is een belangrijke voedselbron voor verschillende demersale vissen, en wanneer hoge densiteiten van *L. conchilega* voorkomen kunnen de aggregaties als refugium tegen predatie fungeren.

De modellen werden ontwikkeld met een dataset die verzameld was in het gebied van de westelijke kustbanken, een klein gebied in het zuidwesten van het BDNZ. Hoewel er verschillende types omgevingsvariabelen in de dataset beschikbaar waren (granulometrische, nutriënten- en stromingsvariabelen), werden alleen granulometrische variabelen gebruikt in de finale modellen (mediane korrelgrootte, slib % en grof zand %). ANN's hadden een hogere voorspellende kracht op basis van een aantal performantieindicatoren (% correct voorspelde stalen, Area Under the Curve, sensitiviteit en specificiteit), maar er was een hoge correlatie tussen de modelvoorspellingen van LR's en ANN's.

ANN's kunnen complexere relaties modelleren, maar dit kan vaak leiden tot modellen die moeilijk transfereerbaar zijn naar andere regio's of periodes. In het *Lanice* voorbeeld in dit hoofdstuk was er een vrij groot verschil tussen de modelperformantie van de kruisvalidatie folds bij de ANN's. Bij de LR's, was er een kleiner verschil tussen de folds, wat wijst op een grotere robuustheid van de LR's voor kleine verschillende in de dataset. Op dit moment is er geen onderbouwde theorie om het aantal

interneuronen of de keuze van de transferfuncties bij ANN's te bepalen. Omdat er bij ANN's geen distributie van de modelresiduelen wordt verondersteld, worden er bij ANN's geen significantietesten gedaan. Vanuit het standpunt van modelparsimonie, waren LR's meer geschikt om het voorkomen van de soort *Lanice conchilega* te modelleren, aangezien de LR's eenvoudiger waren en ANN's slechts een licht hogere modelperformantie hadden. Het voordeel van LR's is dat dit een wijdverspreide modelleertechniek is, met een goed onderbouwde statistische methodologie. De significantie van het complete model en van individuele modelparameters kan getest worden met LR's en de model-parameters zijn direct interpreteerbaar.

Uitdaging 2: wat is de meest optimale variabelencombinatie?

De algemene doelstelling van **hoofdstuk 3** was de verbetering van de modelselectiemethodologie voor HGM's. HGM's gebruiken een combinatie van variabelen om de verspreiding van soorten te voorspellen. Deze variabelen worden verondersteld de verspreiding te bepalen. Eén van de uitdagingen van habitatgeschiktheidsmodelleren is de selectie van een optimale subset van variabelen. Deze variabelenselectie en de keuze van de gemodelleerde respons per variabele (bv. een lineaire of unimodale respons) worden samen modelselectie genoemd, en modelselectie wordt als een centrale stap in de modelontwikkeling beschouwd. Meestal wordt er een stapsgewijze modelselectie gebruikt in HGM's (bv. voorwaartse of achterwaartse modelselectie). Deze stapsgewijze methode heeft echter wel een aantal nadelen: 1) door de benadering in stappen kan het globaal optimale model gemist worden, 2) de stapsgewijze methode is gevoelig voor multicollineariteit van variabelen en 3) multimodel inferentie en predictie is uitgesloten omdat er slechts één optimaal model wordt gekozen.

In dit hoofdstuk werd een nieuwe modelselectiemethode voorgesteld: het Gecombineerde Model Optimalisatie Criterium (GMOC). De GMOC-methode werd voorgesteld om de nadelen met de stapsgewijze modelselectie methode aan te pakken. De GMOC is gebaseerd op het Model Optimalisatie Criterium (MOC) framework, dat de modelperformantie en complexiteit tegelijk in rekening brengt. GMOC-waardes werden berekend aan de hand van vijf veel gebruikte MOC's: het Akaike Informatie Criterium (AIC), een aangepaste versie hiervan (AICc), het Bayesiaanse Informatie Criterium (BIC), het Consistente Akaike Informatie Criterium (CAIC) en de F-statistiek. De MOC's van de model calibratie- en testset werden gecombineerd in de GMOC, en op deze manier werd de modelgeneralisatie op basis van onafhankelijke data geïncorporeerd in de modelselectie. Goede modelgeneralisatie is belangrijk wanneer een model gebruikt wordt in een ander gebied of in een andere periode.

De voorgestelde GMOC-methodologie werd eerst getest op artificiële data voor een artificiële soort, omdat dit een volledige controle van de dataset toelaat. De modelselectiemethodologie liet toe om de datageneratie modellen, die gebruikt werden om de artificiële data te genereren, te reconstrueren. Logistische regressie werd gebruikt als modelleertechniek omdat deze techniek vaak wordt gebruikt voor HGM's, relatief eenvoudig is en theoretisch goed onderbouwd. De voorgestelde modelselectiemethode werd ook toegepast op observaties van *Abra alba*, een mariene bivalve. Deze soort is een indicator van de *A. alba* macrobenthos gemeenschap in de zuidelijke Noordzee. De GMOC's, gebaseerd op ieder van de vijf MOC's, selecteerden licht verschillende variabelencombinaties, maar de variabelen mediane korrelgrootte, diepte en bathymetrische positie index werden in alle modellen gekozen.

De GMOC-methode heeft een aantal voordelen tov. de huidige stapsgewijze methodologie. Omdat alle combinaties van variabelen worden vergeleken verhoogt de kans om het optimale model te vinden. Dit zorgt er ook voor dat deze methode beter om kan met multicollineariteit van variabelen. De GMOC-methode is niet gebaseerd op een contingentie tabel en daarom is het niet nodig om een arbitraire cutoff waarde te kiezen voor de aanwezigheid van een soort. Bootstrap resampelen werd gebruikt in de GMOC-methode om de betrouwbaarheid van de modelselectie te verhogen. Deze resampling produceerde modelreplica's, die de robuustheid van de schatting van de modelperformantie verhogen. Tijdens het resamplen werd de prevalentie van de soort (ratio aanwezigheden/aantal stalen), steeds op 50% gehouden om een beïnvloeding te voorkomen van de modelparameterschatting door extreme prevalentiewaardes. De GMOC-modelselectie kan worden toegepast voor iedere modelleertechniek waarvoor een *likelihood* kan worden berekend (bv. GLM's, gegeneraliseerde additieve modellen, ANN's).

Uitdaging 3: zijn de modellen betrouwbaar?

Er is een disproportionele verhouding in de tijd en moeite die gebruikt wordt om HGM's te ontwikkelen, in verhouding tot het valideren van de modellen. Traditioneel worden HGM's gevalideerd door de observaties te vergelijken met de modelpredicties. De algemene doelstelling van **hoofdstuk 4** is te pleiten voor een geïntegreerde validatie van mariene HGM's, die ook de ecologische relevantie van de modellen beschouwt. Zo een geïntegreerde validatie bestaat uit: 1) klassieke modelvalidatie door observaties en predictie te vergelijken, 2) een conceptueel schema dat ecologische kennis uit de literatuur samenbrengt, 3) habitatpreferentie-experimenten en 3) een analyse van de verdeling van de stalen over het bereik van iedere variabele.

Om de gesuggereerde verbeteringen in de modelvalidatiemethodologie te illustreren, werden HGM's ontwikkeld voor de mariene bivalve *Donax vittatus* (Da Costa 1778). Deze soort werd gekozen omdat: 1) uitgebreide literatuur beschikbaar was over deze soort, 2) resultaten van habitatpreferentie-experimenten beschikbaar waren en 3) deze soort een voedselbron is voor juveniele schol (*Pleuronectes platessa*). De GMOC werd gebruikt als modelselectiemethode en logistische regressie als modelleertechniek. Na de modelselectie werden vier modellen weerhouden en een multimodelpredictie werd berekend met de GMOC van ieder model als wegingsfactor. Alleen diepte en mediane korrelgrootte werden weerhouden als voorspellende variabelen in de vier finale modellen.

Ecologische kennis uit de literatuur werd gecombineerd in een conceptueel schema. Zo een schema was nuttig om een eerste beeld te vormen of een variabele eerder een causale variabele is of eerder een proxy variabele, die een correlatie heeft met het voorkomen van de soort maar geen causale relatie. Wanneer modellen gebruikt zullen worden in andere regio's of periodes is het cruciaal om te weten of een variabele een causale relatie heeft met het voorkomen van een soort. Experimentele validatie bevestigde de gemodelleerde respons voor de variabele mediane korrelgrootte. Habitatpreferentie-experimenten lieten toe de fundamentele niche te identificeren, terwijl veldobservaties alleen de gerealiseerde niche konden afbakenen. Wanneer experimentele data beschikbaar zijn, moeten deze data gebruikt worden om de causaliteit van variabelen na te gaan. Als dit soort gegevens niet beschikbaar is, kan de haalbaarheid van dergelijke experimenten overwogen worden.

De verdeling van stalen over het bereik van een variabele werd onderzocht voor iedere variabele in de data set. Dit liet toe om na te gaan of het bemonsterde bereik voldoende was en of de variabele een relatie vertoonde met het voorkomen van de soort. Soms werden ecologisch relevante variabelen niet gekozen in de modelselectie omdat het bemonsterde bereik niet voldoende was. Voor iedere variabele in de dataset was er een geïntegreerde discussie waarom een variabele wel of niet gekozen was in de variabele selectie. Deze discussie was gebaseerd op het conceptuele schema, de experimentele resultaten en de verdeling van de stalen over het bereik van iedere variabele.

Algemene discussie

In **hoofdstuk 5** werden de bevindingen van dit doctoraat in een breder kader bediscussieerd. De discussie volgde de drie uitdagingen die geïdentificeerd werden in de modelleermethodologie werden in dit doctoraat. Een toekomstperspectief voor het gebruik van HGM's werd gegeven, met een discussie van de voornaamste bronnen van bias en lage modelbetrouwbaarheid. Vervolgens werd een overzicht

gegeven van ecologische soortkenmerken die de modelontwikkeling kunnen beïnvloeden. Een andere bron van onzekerheid die in deze context werd besproken was het ontbreken van een spatiotemporele overlap tussen voorspellende variabelen en soortsobservaties. Als laatste deel van de algemene discussie werden sommige ontwikkelingen op gebied van modelleertechnieken besproken en werden de voordelen van ruimtelijk expliciete modellen en mechanistische HGM's uitgelegd.

De HGM's voor macrobenthossoorten die in dit doctoraat werden ontwikkeld werden apart besproken. En finaal werden sommige toepassingen van HGM's voor macrobenthossoorten voorgesteld: bv. HGM-gebaseerde staalnames, simulatie van beheerscenario's.

De algemene conclusies van dit onderzoek zijn:

- Logistische regressie modellen zijn te verkiezen boven ANN's, vooral wanneer model parsimonie gewenst is. HGM's waren complexer en hadden slechts een licht hogere modelperformantie.
- De GMOC-modelselectiemethode heeft verschillende voordelen in vergelijking met de vaak gebruikte stapsgewijze methode. Het belangrijkste voordeel is dat de modelvalidatie in de modelselectieprocedure wordt geïncorporeerd.
- Een geïntegreerde validatie van HGM's is nodig om betrouwbare en ecologische onderbouwde modellen te ontwikkelen. Zo een validatie integreert:: 1) modelvalidatie met veldobservaties, 2) ecologische kennis, samengebracht in een conceptueel schema, 3) habitatpreferrentie-experimenten en 4) invloed van het bemonsterde bereik van iedere variabele.
- HGM's zijn een waardevolle techniek om de ruimtelijke verspreiding van macrobenthossoorten in de Noordzee te voorspellen. Dit werd aangetoond voor verschillende soorten. Meer specifieke toepassingen van HGM's voor macrobenthossoorten zullen in de toekomst te verwachten zijn.

Preface

In a quantitative analysis of human impacts on marine ecosystems (Halpern *et al.*, 2008), the North Sea was ranked among the regions most impacted by human pressures worldwide. The OSPAR Quality Status Report of the North Sea (OSPAR, 2000) identified no less than 33 specific human pressures in the North Sea. Marine management is thus needed to stop further degradation of the North Sea and to restore impacted sites. For this purpose, several policy instruments have been introduced: EU Habitat Directive (92/43/EC), EU Birds Directive (79/409/EC), EU Water Framework Directive (2000/60/EC) and more recently the EU Marine Strategy Framework Directive (2008/56/EC). The latter directive targets good environmental status of all EU marine waters by 2021. A sound marine management requires detailed knowledge on the spatial distribution of species and habitats (Pittman *et al.*, 2007). Currently the distribution of species is known only from point observations, and full cover species distribution maps are mostly lacking. However, environmental variables are often available at a full cover scale (e.g. sediment grain size). Habitat suitability models (HSMs) relate the presence or abundance of a species in a location to a set of environmental predictors, which then allows predicted distributions to be mapped across an entire region (Barry and Elith, 2006; Elith and Leathwick, 2009).

This thesis focuses on the prediction of the spatial distribution of macrobenthos in the Belgian Part of the North Sea (BPNS) with HSMs. Macrobenthic species, defined as animals larger than 1mm which live in/upon the sea bottom, are often the main component in environmental monitoring programmes to evaluate the status of the benthic ecosystems for marine management. These species have limited mobility and thus are good indicators of the local environmental status (Rees *et al.*, 2007). Sampling and treatment of macrobenthos is however time-consuming and expensive. It involves expensive ship time and species identifications by trained personnel. Therefore, taking extra samples to obtain knowledge on the spatial distribution of species is not always feasible. The development of HSMs, based on the available environmental data, is thus a cost-effective way to develop species distribution maps for macrobenthic species.

HSMs are a relatively recent modelling approach and their use is increasing, especially in the last years. HSMs are also known under the synonyms “species distribution models” and “niche models”. The rapid growth of HSMs in recent years is clearly evident by the rise in the number of publications applying HSMs. The scholarly article “Predictive habitat distribution models in ecology” by Guisan and Zimmerman (2000) has been cited around 1200 times (August 2010). This also means that during the course of this PhD research the use of habitat suitability models has increased exponentially and modelling techniques and theory have evolved. Chapter 2 was written at the beginning of the research, and uses the modelling techniques and insight available at the time of writing. But weaknesses identified in this chapter (i.e. model selection and model validation approach) are dealt with in later chapters.

A standard modelling methodology for HSMs is however lacking at the moment (Araujo *et al.* 2006). Numerous modelling techniques are available (Segurado and Araujo, 2004; Elith and Leathwick, 2009), but there is no agreement on which technique should be used in which situation. The selection of appropriate predictor variables is a central step in most modelling efforts (Guisan and Zimmermann, 2000; Heikkinen *et al.*, 2006), but little discussion is found on the choice of the model selection methodology in most HSM papers. Furthermore, there is a disproportionally large effort in developing HSM models, compared to the essential need for a proper validation of the models (Eastwood *et al.*, 2003). Because the HSM methodology is not yet optimally developed, this thesis will mainly focus on improving the HSMs modelling methodology with emphasis on the application towards macrobenthos species in marine environments. HSMs have only been developed for a limited number of macrobenthos species, but with the modelling methodology proposed in this research, models can be developed for other species in an efficient way.

The first chapter of this thesis is an introduction that aims at providing the reader with the necessary background on challenges for marine management in the North Sea is facing and how macrobenthos species can be used to monitor the environmental status. A second part of the introduction provides a technical introduction to HSMs for non-experts, which will aid in the understanding of the technical aspects of HSMs. Following, an overview of the current applications of HSMs in marine management is provided. Based on the challenges identified in the introductory chapter, the objectives of this thesis are laid out in a final part of the first chapter.

Three major challenges are identified in the HSM methodology; these will be treated in separate chapters in this thesis:

Challenge 1: Which modelling technique to use? (Chapter 2)

Numerous alternative modelling techniques are available to model the distribution of species. Two commonly used modelling techniques are compared: Artificial Neural Networks (ANNs) and Generalised Linear Models (GLMs). As this chapter was written early in the course of the PhD, later chapters sometimes contradict the findings of this chapter. But as this chapter is already published, the changes to the original published manuscript were kept to the strict minimum.

Challenge 2: What is the most optimal combination of predictive variables? (Chapter 3)

Most data sets available for HSM development contain a set of potential predictive variables. The challenge is finding the most optimal combination of variables to model the distribution of species. An

improved model selection approach is proposed. This chapter is a response to the need for an objective model selection methodology, a need identified in chapter 2.

Challenge 3: Are the model predictions reliable? (Chapter 4)

Classical model validation of observations vs. predictions is often the only way models are validated. An integrated model validation is proposed that also validates the ecological soundness of the models.

Chapter 1. General introduction

1.1. Marine management in the North Sea and the role of macrobenthos

1.1.1. The worldwide oceans: an ecosystem under pressure

Presently, there is an enormous anthropogenic pressure on the oceans, which offer a whole range of indispensable goods and services to mankind (Beaumont *et al.*, 2007). It is estimated that two-thirds of the total economic value provided by global ecosystems is generated by marine ecosystems (Costanza *et al.*, 1998). The idea of the unlimited oceans with undepletable fish stocks is shattered as scientific evidence of the human impacts on the oceans builds up (Worm *et al.*, 2009). Globally more than half of the world population lives within 60 kilometres of the coast (UNEP, 2004). Land-based pressures on the marine ecosystem include runoff of pollutants and nutrients into coastal waters (Halpern *et al.*, 2008). The current intensive agricultural practice causes eutrophication and algal blooms, which lead to anoxic conditions or even "dead zones" at the mound of large rivers in some areas (Diaz and Rosenberg, 2008).

Statements such as "only 10% of big ocean fish remain" (Myers and Worm, 2003) and new concepts as "fishing down marine food webs" (Pauly *et al.*, 1998) illustrate the urgency for a sustainable fisheries management. As the higher trophic level species, such as cod and tuna, are removed, the top-down control on their prey is removed and the whole ecosystem is affected.

There is a broad consensus that the increased global output of greenhouse gasses causes climate change (IPCC, 2007), for which marine systems are particularly vulnerable (Cheung *et al.*, 2009). Local invasions and extinction (collectively called species turnover) due to climate change, will affect biodiversity, community structure and ecosystem functions (Cheung *et al.*, 2009). The worldwide oceanic circulation is driven by density differences: slightly cooler water sinks and wells up thousands of kilometres further. If global warming would interfere with this mechanism, the global circulation pattern could change (IPCC, 2007). In the tropics, coral reefs are threatened by global warming as they can only persist within a narrow temperature range. Increased water temperatures will cause coral bleaching and influence the calcification ratio of hermatypic corals (Howe *et al.*, 2002). The increased CO_2 levels can cause ocean acidification, which influences the geochemistry of the ocean, and affect the calcification rate of marine organisms.

As an example to study the current and future pressures on marine systems, the North Sea is chosen as study area in this thesis, and in the next paragraphs some general information and particular characteristics of this system are provided.

1.1.2. The North Sea

The North Sea has a surface area of 750000 km². It embraces the entire English Channel, bordered by the UK, Belgium and France, up to the waters of the Skagerrak and Kattegat in the east, bounded by Denmark, Norway, and Sweden. Since the borders are not closed, water exchange occurs through the influx of Atlantic water to the north, to a lesser extent via the Channel, and also from the Baltic to the east, along with a northward efflux. The mean flushing time of the North Sea water is estimated to be one year (OSPAR Commission, 2000). The influx of oceanic waters links the North Sea to the general oceanic circulation and allows the Gulf Stream to transport warmer water into the North Sea. The North Sea is situated on the continental shelf, with depths not exceeding 50m in the southern North Sea. Only in the northern North Sea and the Norwegian Trench depths up to 700m can be reached. Due to the shallow bathymetry, most of the North Sea water is mixed in winter and only the deeper offshore central and northern regions become stratified (Ducrotoy *et al.*, 2000).

Fig.1.1. Overview map of the countries bordering the North Sea. The black line indicates the limits of the drainage basin of the rivers mounding in the North Sea. Source: www.ospar.org

In a quantitative analysis of the human impact on marine ecosystems, performed by Halpern *et al.* (2008), the North Sea was ranked among the regions most impacted by human pressures worldwide. The drainage basin of the rivers mounding in the North Sea (Fig. 1.1) consists of densely populated, highly industrialised countries (OSPAR, 2000). The North Sea is therefore prone to land based inputs of nutrients, chemicals and sewage. The coast hosts some of the largest ports in the world, making the

North Sea one of the most traversed marine areas of the world (Ducrotoy *et al.*, 2000). Increased shipping traffic therefore increases the chance of accidental pollution. Additionally, the ballast water of all these ocean sailing ships can be a transport medium for invasive species, which can harm the ecological system and wipe out native species.

The OSPAR quality status report of the North Sea (OSPAR, 2000) identified no less than 33 human pressures, which were ranked in priority classes taking into account severity, spatial scale and recovery time. OSPAR is a commission of fifteen Governments of the western coasts and catchments of Europe, together with the European Community, that cooperate to protect the marine environment of the North-East Atlantic. In addition to land based organic contaminants and nutrient input, fisheries' impacts make up the top category of human pressures. To target benthic fish, such as plaice or cod, highly destructive beam trawling is used in the sandy parts of the North Sea. This fishing gear uses heavy tickler chains that effectively plough the bottom and causes long term changes to the macrobenthic communities (Frid *et al.*, 2000; Piet *et al.*, 2000). This has changed the species composition from larger, more long-lived species to smaller, more opportunistic species (OSPAR, 2000).

The high coastal human population densities lead to high intensities of activities. Each summer millions of tourists occupy the beaches and disturb coastal birds and mammals (Ducrotoy *et al.*, 2000). Coastal defence constructions often stop the dynamics in dune or estuarine ecosystems (Ducrotoy *et al.*, 2000). The maintenance of shipping lanes and ports requires frequent dredging (OSPAR, 2008) and this dredged material is often contaminated with traces of anti-fouling paint or sewage.

1.1.2.1. Belgian part of the North Sea

The Belgian Part of the North Sea (BPNS) has a surface area of 3600 km², which encompasses only 0.5% of the whole North Sea (Fig. 1.2). It is situated in the southernmost part of the North Sea, with a maximum depth of 46m. The main bottom features found in the BPNS are sandbanks alternating with gullies (Van Hoey *et al.*, 2004). The sandbank systems are dynamic, as strong tidal currents and wave action can rework the top layer of the sediment.

Nearby the coast, flood-dominated currents head northeast, while more offshore ebb-dominated currents head towards the southwest (Luyten *et al.*, 2003). The strong tidal currents, in combination with the shallow depth, cause the water column to be fully mixed throughout the whole BPNS (Luyten *et al.*, 2003). The coastal waters have a high turbidity due to nearby river mouths, e.g. Rhine, Meuse, Scheldt, Yzer, Authie, Canche (Lacroix *et al.*, 2004). More offshore, the turbidity and nutrient levels decrease, together with reduced primary production. The north-eastern coastal area is influenced by the turbidity plume of the Scheldt and Rhine/Meuse. The sediment grain size in the BPNS becomes coarser further offshore and is the result of the interaction between seabed morphology and currents (Verfaillie *et al.*,

2006). The BPNS has a soft-bottom substratum, ranging from mud, close to the shore, to coarse sand in the north (Verfaillie *et al.*, 2006), with patches of gravel on the Hinderbanks region (Houziaux *et al.*, 2007).

1.1.2.2. Marine management of the North Sea

The North Sea is one of the most exploited marine areas in the world (Maes *et al.*, 2005; Douvere *et al.*, 2007). Therefore, marine spatial planning is urgently needed for the management of human activities in the North Sea region (Douvere *et al.*, 2007). A number of European policy instruments have been implemented to protect the marine environment of the North Sea. The EU Birds Directive (79/409/EC), an outcome of the Ramsar Convention on Wetlands (1971), proposes protection measures for the sea and coastal waters where birds are living. Member States have designated special protection areas (SPAs) for important bird areas. In the BPNS, for example, 3 SPAs are designated by the federal Act on the protection of the marine environment (Act of 20 January 1999, amended by Act of 17 September 2005; Fig. 1.2). SPA1 is an important site for the Sandwich Tern (*Sterna sandvicensis*) and the Great-crested Grebe (*Podiceps cristatus*), SPA2 is a crucial site for the life and reproduction of the Common Scoter (*Melanitta nigra)*, the Great-crested Grebe, the Common Tern (*Sterna hirundo*), the Sandwich Tern and the Little Gull (*Hydrocoloeus minutus*) and SPA3 is important for Common Tern, Little Gull and Little Tern (Haelters *et al.*, 2004).

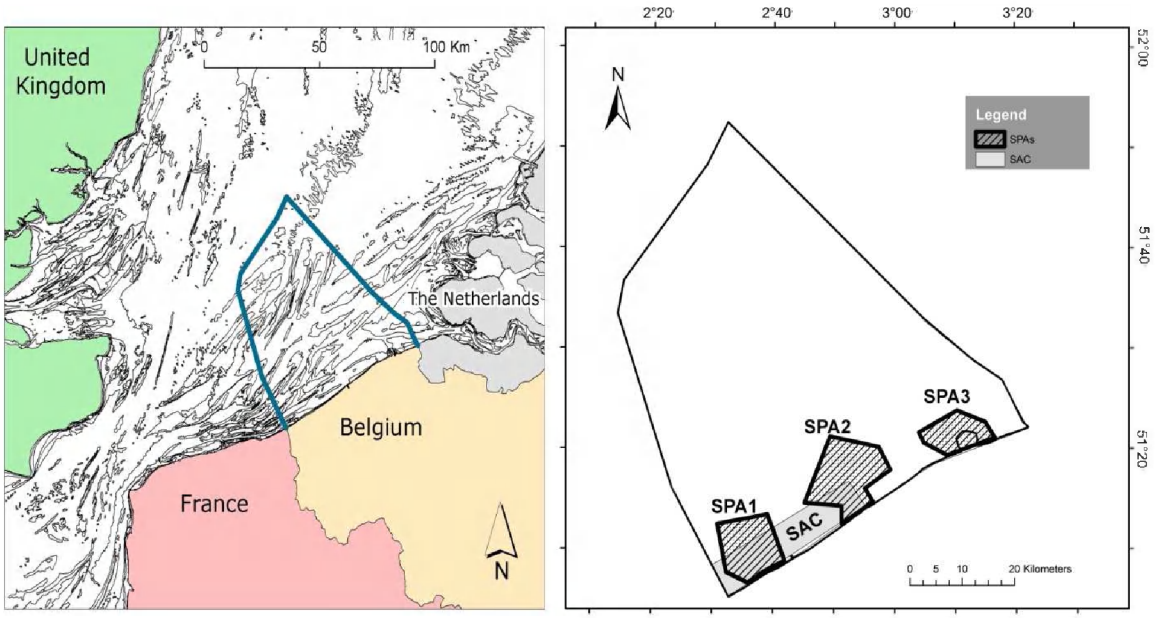

Fig. 1.2. Map of the Belgian part of the North Sea (BPNS). SPA: Special Area of Protection; SAC: Special Area of Conservation. SACs and SPAs designated by the federal Act on the protection of the marine environment (Act of 20 January 1999, amended by Act of 17 September 2005).

The EU Habitat Directive (92/43/EC) aims at the maintenance of a minimum level of biodiversity in Europe. Special Areas of Conservation (SACs) are an integral part of the directive and required the first listing of proposed Sites of Community Importance (pSCIs) by June 1998. Protection measures have to be taken in the SACs to ensure that the quality of the habitats does not deteriorate and that no negative impacts occur on the species for which the SACs are designated. According to the EU Birds Directive (79/409/EEC), Member States are required to designate Special Protection Areas (SPAs) for the conservation of a specific list of bird species. The network of both SPAs and SACs is the Natura2000 network. The EU Water Framework Directive (2000/60/EC) is another highly relevant policy instrument. It came into force in 2000 and requires that all coastal waters must reach at least a good environmental status by 2015 and defines how this should be achieved through the establishment of environmental objectives and ecological targets. As the sum of all existing measures and efforts, whether taken at international, EU or national level, was clearly not sufficient to protect Europe's marine environment, the European Marine Strategy Framework Directive (2008/56/EC) was adopted in 2008. This directive aims to ensure that all EU marine waters have a "good environmental status" by 2021. Human activities impacting the maritime environment have so far been addressed on a sectoral basis rather than holistically. The European Marine Strategy Framework Directive takes a new, ecosystem-based approach addressing all the pressures and impacts on the marine environment.

1.1.3. The need for species distribution information

The ever present conflict between several human activities and marine ecosystems requires the development of decision support systems for an integrated marine management and spatial planning. Marine spatial planning and decision support systems are only recently introduced, with examples in Taiwan (Chang *et al.*, 2008), the Great Barrier Reef in Australia (Olsson *et al.*, 2008) and the Eastern Scotian Shelf in Canada (Walmsley *et al.*, 2007). In the US, the Magnuson-Stevenson Act (NOAA, 1996) amends the protection of "essential fish habitat" to decrease anthropogenic disturbance (Le Pape *et al.*, 2003). And in the EU several directives have been implemented, as mentioned earlier, which aim at an ecosystem based marine management and spatial planning.

Marine managers are often faced with limited and uncertain ecological information on which to base their decisions (Pittman *et al.*, 2007). However, the efficient implementation of marine decision support systems as a tool for marine spatial planning and management requires understanding of the processes that determine the observed distribution patterns of species in marine ecosystems (Heglund, 2002). It is therefore necessary to gain insight in the temporal and spatial distribution of each ecosystem

component: fish, seabirds, macrobenthos, meiobenthos, etc. In practice, this means that good species distribution maps are required which cover the marine region to be managed. Only with good quality, full cover maps of the distribution of the ecosystem components under concern, it will be possible to manage them well and to mitigate the impact of human pressures on the system. For example, if one wants to set up an offshore wind farm or plan a dredging operation, good knowledge of the species that might potentially be impacted is necessary for a reliable impact assessment. Another example where good species distribution maps are necessary is the delimitation of Marine Protected Areas (MPAs). An efficient MPA should maximally engulf the species requiring protection. Marine biological valuation (Derous *et al.*, 2007) requires knowledge of the spatial distributions of indicator species.

Distribution maps of marine species are usually maps with point observations based on biological samples (e.g. grab samples of macrobenthos species). But often the samples are not evenly distributed over the surface of an area. Marine areas with a high sampling density are mostly closer to shore and, on a global scale, closer to richer countries. Consequently, there are a lot of blank spaces in most species distribution maps as the presence of the species is unknown in between point observations. One way to obtain full cover species distribution maps, is to interpolate the density of species, based on the observed densities (Degraer *et al.*, 2008). This would however require a good spatial distribution of the biological observations, which is rarely the case. Even for the well studied BPNS, the sampling density is very low in some regions (Fig. 1.3).

Habitat Suitability Models (HSMs) allow to generate full-cover species distribution maps, based on the available set of biological point observations and environmental variables. These models enable the prediction of the suitability of the habitat for a certain (group of) species (Guisan and Zimmermann, 2000; Elith and Leathwick, 2009). The main idea is that if the habitat conditions, e.g. sediment grain size, temperature and salinity, fall within the range preferred by the species, there is a high probability that the species will be present. As such, it becomes possible to produce a species distribution map, based on knowledge of physical habitat variables. A great advantage is that these physical variables are mostly available on a full coverage basis, e.g. sediment grain size maps or satellite based temperature. Full cover in this context means that there is a raster data set, and for each pixel in the raster there is a value. Full cover data can be obtained by satellite images, which have a full cover output straight away, or by interpolating point-based observations. Thus, if these full cover layers are fed into a HSM full cover species prediction maps can be generated.

The application of HSMs therefore requires a relation between the physical habitat and the species observed. Such a link has been observed for many ecosystem components. For macrobenthos a relation with several environmental variables (e.g. median grain size and mud %) has been observed in the North Sea (Van Hoey *et al.*, 2004; Degraer *et al.*, 2008; Pesch *et al.*, 2008; Willems *et al.*, 2008).

Davies *et al.* (2008) modelled the distribution of the cold-water coral *Lophelia pertusa* based on a set oceanographic variables. The distribution of fish species (Le Pape *et al.*, 2003) and marine mammals (Redfern *et al.*, 2006) has been linked to sea water temperature and bathymetry. Similar relations of the species distribution with physical habitat variables have been observed for birds (Ballance *et al.*, 2006) and algae (Sandman *et al.*, 2007). Thus, it is possible to use HSMs to predict the occurrence of marine species by using the information of the spatial variance of the physical variables. Such physical habitat variables are more practical and cheaper to measure, compared to biological observations.

The ever increasing effort of marine habitat mapping of several countries provides full cover maps of environmental variables that can potentially be used for habitat suitability modelling. In the context of habitat mapping several initiatives have collated existing maps of environmental variables and several surveys have filled the gaps in the data coverage. In European waters two large habitat mapping initiatives have recently taken place. The European project MESH (Mapping European Sea Habitats, 2004-2007, www.searchmesh.net) developed integrated habitat maps of the North Sea and made them available in a catalogue and a webGIS application. Similarly, the European Balance-project (2005-2007, www.balance-eu.org) made similar habitat maps for the Baltic Sea. Additionally, several national habitat mapping projects (e.g. Norway: www.mareano.no; Irish Sea: www.habmap.org) are producing valuable maps of the environmental variables which can be used in future habitat suitability modelling exercises.

1.1.4. Macrobenthos in environmental monitoring

Macrobenthic species, defined as animals larger than 1mm which live in/upon the sea bottom, are often the main component in environmental monitoring programs to evaluate the status of the benthic ecosystems and to support an ecologically sustainable marine management (Rees *et al.*, 2002; Degraer *et al.*, 2008). The distribution pattern of macrobenthos is often used to support an ecologically sustainable marine management. For example, Borja *et al.* (2003) developed a biotic index based on the benthic species, which goes from 0 (unpolluted) to 7 (extremely polluted). They based this index on the paradigm of Pearson and Rosenberg (1978) on the influence of the enrichment of organic matter on benthic communities. But their biotic index proved also to be useful for the assessment of anthropogenic impacts, such as heavy metal inputs or physical habitat alterations. If habitat suitability maps which give the probability of presence of species are available, they can be used to calculated biotic indices, and possibly even at a full cover scale.

There are several good reasons why macrobenthos is used in monitoring: 1) the species are macroscopic and thus more easy to handle and identify, 2) they are relatively immobile and thus

strongly dependent on local conditions, 3) they are linked with the biogeochemical processes in the sediment (Snelgrove and Butman, 1994) and perform important ecosystem functions (e.g. Rabaut *et al.*, 2007), 4) benthic animals are the food source for many benthic fish, including the economically important species, such as plaice (*Pleuronectes platessa;* Burrows and Gibson, 1995) and cod (Amstrong , 1982; *Gadus morhua*) in the North Sea, and diving birds, e.g. the Common Scoter (*Melanitta nigra*, Degraer *et al.*, 1999b). The fact that macrobenthic organisms remain more or less immobile is very relevant in biological monitoring programs worldwide (Rees *et al.*, 2002). Their local presence/absence or densities can provide information about local changes in the marine environment, as opposed to one-shot physical measurements of human impacts. The limited mobility of macrobenthic species is also a very useful property of the species for the development of HSMs. As a result, the species' occurrence is more strongly linked to local conditions, as compared to fish and marine mammals that can travel thousands of kilometres, crossing different habitats. The relation between the physical environment and the spatial distribution of macrobenthic species, a prerequisite for HSM, has been frequently observed in ecological research (e.g. Gray, 1981; Snelgrove and Butman, 1994). The relation of species distributions with specific environmental variables is demonstrated by a number of researchers. For example, several authors demonstrated the relation between sediment grain size and the distribution of macrobenthos (Van Hoey *et al.*, 2004; Willems *et al.*, 2007; Degraer *et al.*, 2008; Pesch *et al.*, 2008; Willems *et al.*, 2008).

Macrobenthic animals mainly feed on the organic matter coming down from the productive pelagic system, where most or all primary production is happening. This spatial decoupling of production and consumption makes marine benthic environments fundamentally different from terrestrial systems (Snelgrove, 1999). Macrobenthic species play a crucial role in the degradation and remineralisation of organic matter (Borja *et al.*, 2000). Some macrobenthic species are known as habitat engineers, e.g. the polychaete *Lanice conchilega* (Rabaut *et al.*, 2007; see Chapter 2), which structures the environment by building tubes or burrows. Such structures increase the habitat complexity and provide a habitat suitable for other species. As such *Lanice conchilega* increases the diversity locally (Rabaut *et al.*, 2007). Additionally, the bio-irrigation activities of habitat engineers bring organic matter, as well as oxygen, to the deeper sediment layers, which would otherwise be anoxic (Jones and Jago, 1993; Foster and Graf, 1995). According to their feeding type, the species can be classified as predators, deposit feeders, selective deposit feeders and filter feeders (Hartmann-Schröder, 1996, Le Pape *et al.*, 2007).

In the BPNS the macrobenthos consists mainly of polychaetes (43% of the species), crustaceans (34% of the species) and molluscs (16% of the species; Van Hoey *et al.*, 2004). Van Hoey *et al.* (2004) identified four main macrobenthos communities on the BPNS using a multivariate analysis based on species composition of collected samples. The four communities differ drastically, both in

habitat and species composition: 1) the muddy-fine sand *Abra alba-Mysella bidentata* community has high densities and diversity; 2) the *Nephtys cirrosa* community occurs in well-sorted sandy sediments and is characterised by low densities and diversity; 3) very low densities and diversity typify the *Ophelia limacina-Glycera lapidum* community, which is found in coarse sandy sediments and 4) the *Eurydice pulchrae-Scolelepis squamata* community which is typical for the upper intertidal zone of sandy beaches. Degraer *et al.* (2008) developed a HSM that predicts the spatial distribution of the four macrobenthos communities on the BPNS based on the different habitat preference of each community.

1.1.5. Data availability for macrobenthos modelling

Macrobenthos samples of the Belgian part of the North Sea are compiled in the Macrodat database of the Marine Biology section (Ghent University). This database is continuously updated as new monitoring programs collect macrobenthos samples (Fig. 1.3). From the period 1977-1983, 500 samples are available, the more recent samples were collected from 1994 on. Since monitoring programs often focused on sandbanks or coastal regions, the spatial sample distribution is uneven. Based on this database a distribution atlas of the macrobenthos on the BPNS was produced (Degraer *et al.*, 2006).

Full cover maps with physical variables are available for the BPNS and make this region an ideal case study for the development and application of HSMs. The Renard Centre for Marine Geology (RCMG, Ghent University) has developed interpolated, full cover maps for several sedimentological parameters: median grain size, % mud, sand and gravel (Verfaillie *et al.*, 2006; Van Lancker *et al.*, 2007). Based on a database with 9000 grab samples, full coverage sedimentological maps have been created by interpolation techniques which take also secondary variables into account (kriging with external drift; Verfaillie *et al.*, 2006). Additionally, the RCMG has developed a series of full cover maps for topographical variables, derived from a bathymetry grid of the BPNS. Such variables are slope (first spatial derivative of depth), Bathymetric Position Index (BPI, Lundblad *et al.*, 2006; Wilson *et al.*, 2007) and bottom rugosity (Verfaillie *et al.*, 2009a). The BPI is a measure of where a location is, in reference to the surrounding locations. It is calculated by comparing the depth of a pixel in a bathymetry grid, with the depth of surrounding pixels. In the BPNS the BPI was calculated based on an 80m pixel size bathymetry grid. A broad scale BPI used 20 surrounding pixels for the BPI calculation, and the small scale BPI 8 pixels. Rugosity is the ratio of the surface area to the planar area and is a measure for terrain complexity or "roughness". The Management Unit of the North Sea Mathematical Model (MUMM) provides full cover bottom current speed and bottom shear stress maps of the Southern North Sea, obtained from the 3D baroclinic hydrodynamic COHERENS model (Luyten *et al.*, 2003). This model has a horizontal resolution of about 250x250 m and a vertical resolution of ten layers. North Sea wide data

on median and maximum chlorophyll-a concentration in the surface water were obtained from MERIS satellite images from the REVAMP-project (Peters *et al.*, 2005).

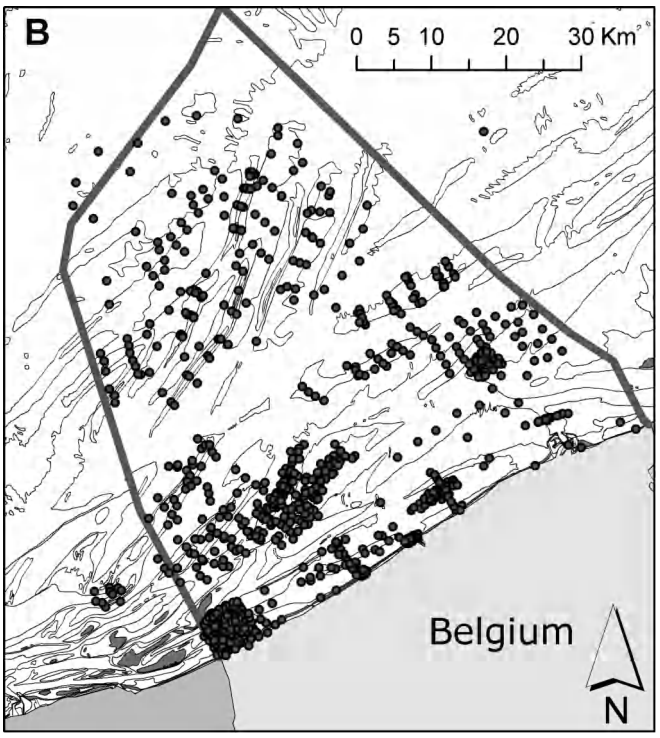

Fig. 1.3. Overview of the MACRODAT macrobenthos data set in the Southern North Sea. The bold line indicates the borders of the Belgian part of the North Sea, Projection WGS84 UTM31N.

1.2. Habitat suitability modelling principles

In this section, a brief introduction to habitat suitability modelling will be provided. Such an introduction is necessary as HSMs are not well known by most benthic ecologists and marine biologists in general. After discussing the history and fundamentals of HSMs, the theoretical background and the data requirements for HSM are discussed. Next, a brief overview of the model development and the model validation is provided. Finally an overview of the applications of HSMs in marine management is given. More extensive introductions to HSMs are available in the literature (e.g. Guisan and Zimmermann, 2000; Stauffer, 2002; Guisan and Thuiller, 2005; Pearson, 2007; Elith and Leathwick, 2009; Franklin and Miller, 2009). This overview of HSMs will focus on models to predict the spatial distribution of species, rather than models to predict communities or species diversity.

The most logical strategy for estimating the actual or potential spatial distribution of a species is to first characterise the environmental conditions that are preferred by a species, and next to identify how these conditions are distributed in a region (Pearson, 2007). If, for example, a polychaete species is observed

within the 200-300 µm sediment grain size range in a region, locations in the region with such sediment would have a high probability that the species is present.

To quantify the probability that a species is present or can occur with a certain density in a more objective and practical way, HSMs are used. These models can predict the expected density or the probability of presence of a species for a combination of environmental variables. This goes further than identifying that a species will be present or not in the 200-300 µm sediment range. For each grain size observed, the HSM will provide a probability of presence of the species. HSMs are mostly multivariate, and allow to model the probability of presence for a combination of several predictive variables. Hence their name, HSMs model the suitability of the local environment for the species. The assumption is that, the more suitable the environment, the higher the probability that the species is present and will be found in high densities. The emphasis must be put on probability of presence. In fact not the presence of a species is predicted directly, but the suitability of the local habitat for the species.

The majority of the HSMs are correlative models (i.e. data-driven models) that estimate the species response by associating species' occurrence records with suites of environmental variables which are expected to affect the species' probability of persistence (Pearson, 2007; Elith and Leathwick, 2009). Very recently some HSMs have been developed in a mechanistic or knowledge-based way (Kearney and Porter, 2009). Mechanistic HSMs require knowledge on the ecophysiological processes that determine the spatial distribution of species (Kearney and Porter, 2009), while correlative models only require a set of species observations and environmental variables (Pearson *et al.*, 2007). Examples of mechanistic HSMs are still rare and this thesis will only deal with correlative modelling approaches. Kearney and Porter (2009) argue that a strategy involving both correlative and mechanistic approaches may provide very robust predictions of species potential ranges in the future (see Chapter 5 General Discussion).

In the last two decades the use of HSMs for the prediction of plant and animal distributions has grown rapidly (Guisan and Thuiller, 2005). Applications of HSMs in marine and freshwater environments are relatively recent but the use of HSMs is also increasing rapidly in these environments (Elith and Leathwick, 2009). Contemporary HSMs combine concepts from ecological and natural history traditions which originally observed the species-environment relations, with more recent developments in statistics and information technology (Elith and Leathwick, 2009). Generally, three phases can be identified in the history of HSMs (Guisan and Thuiller, 2005): 1) non-spatial statistical quantification of species–environment relationship based on empirical data, 2) expert-based (non-statistical, non-empirical) spatial modelling of species distribution (e.g. Habitat Suitability Index; Rubec *et al.*, 1998) and 3) spatially, statistical modelling of species distributions.

In the literature HSMs are sometimes termed resource selection functions (e.g. Norcross *et al.*, 1999), Species Distribution Models (SDMs; e.g. Austin, 2002) and Ecological Niche Models (e.g. Hirzel and Arlettaz, 2003). Modern HSMs strongly rely on the research developments in physical geography, spatial interpolation (e.g. Verfaillie *et al.*, 2009b), remote sensing (e.g. Peters *et al.*, 2005) and Geographical Information Systems (GIS) in general. These research fields provided the widely available full cover data layers which are used as predictive variables in HSMs (Elith and Leathwick, 2009).

1.2.1. Ecological theory habitat suitability modelling

Theoretical justification of HSMs has always relied on the niche theory (Guisan and Thuiller, 2005; Hirzel and Le Lay, 2008): the habitat preference of a species is modelled, and as such the niche of the species. To link HSMs with the niche theory first the concepts geographic and environmental space need to be introduced. Species are observed at certain locations in the geographic space, which is the world around us. The observations in the geographic space can be visualised on maps (Pearson, 2007). Locations in the geographic space are referenced with their coordinates (latitude and longitude). At each location in the geographic space a number of environmental variables are also observed (e.g. temperature or sediment grain size). Each of these variables makes up one dimension of the n-dimensional environmental space (Elith and Leathwick, 2009). The observations collected in the geographic space can thus be plotted in the environmental space by using the environmental variables as coordinates.

Both the niche theory and HSMs operate in the environmental space. In the ecological theory of HSMs, a distinction is made between the fundamental niche *sensu* Hutchinson (1957) and the realised niche (Pearson, 2007). The fundamental niche is an n-dimensional hypervolume, and each point in this hypervolume presents a combination of environmental variables that allows the species to survive and persist (Pearson, 2007). The fundamental niche of a species assumed in the context of HSMs is only limited by physiological constraints. The realised niche in the environmental space is a subspace of the fundamental niche considering both physical dispersal limitations as well as biotic interactions (e.g. competition, predation; Rodder and Lotters, 2009). Dispersion limitation can leave suitable habitat patches unoccupied and biotic interactions can cause low densities of a species in an optimal habitat.

The realised niche is what is observed in field observations, and thus reflects all constraints imposed on the actual distribution of a species, both physiological, biological and geographical (Pearson *et al.*, 2007). As HSMs are calibrated with field observations, it is mostly assumed that HSMs model the realised niche (Heglund, 2002; Guisan and Thuiller, 2005). The ecological niche theory thus relates a set of environmental variables to the fitness of species, while HSMs relate environmental variables to

the probability of presence of species (Hirzel and Le Lay, 2008). In a recent review on the link between both theories, Hirzel and Le Lay (2008) concluded that in spite of the relationship between both theories, the concepts are weakly linked in the literature and there is a strong need for better integration.

To determine the fundamental niche, and develop HSMs that predict the probability of presence of species in the absence of biotic interactions and dispersion limitations, knowledge on the species' physiology is needed. Such knowledge can be obtained in habitat preference experiments (e.g. Alexander *et al.*, 1993; Wright *et al.*, 2000).

1.2.2. Data requirements for habitat suitability modelling

1.2.2.1. Species observation data

HSMs are developed by estimating the suitability of a combination of environmental variables based on a set of species and variable observations. Species distribution data may be densities (individuals/surface unit), presence-only (i.e. records of localities where the species has been observed) or presence/absence observations (i.e. records of presence and absence of the species at sampled localities). Each type of species observation demands a specific modelling approach and the HSMs developed will also have a specific model output (see further).

When only presence observations are available, it is not possible to determine if a sufficient part of the variable range was sampled (see Chapter 4), while this is well possible with presence/absence observations (Phillips *et al.*, 2009). Presence observations can only provide positive proof that a species is present, when an observation is done. Presence only observations have no information on the spatial distribution of the sampling effort.

In this thesis the HSMs will only be developed with presence/absence observations of macrobenthos species. Variables can only be used to predict the distribution of species if a sufficient part of the variable range is sampled. There is thus a need for more research on the effect of the sampled variable range on the model parameters fitted.

The ratio of the number of presence observations on the total number of samples is termed the prevalence (Jimenez-Valverde and Lobo, 2006). The prevalence of a species in a data set is often related to the overall rareness of a species in a region. The sampling design will determine if rare species will have a low prevalence in a data set or not. If more samples are collected in suitable environments the prevalence will be higher in the data set than the overall rareness in the region. As most modelling techniques model the probability of presence and thus the expected proportion of presence samples for a given combination of predictor variables, the prevalence in a data set can introduce a major bias in the modelled response (Jiménez-Valverde *et al.*, 2009). For example, species

with a low prevalence will more likely be predicted to be absent, as this maximises the fit of the model to the data set, where the species is mostly absent. There is thus a need for compensation of the effect of species prevalence in the HSM modelling methodology.

1.2.2.2. Predictive variables

The predictive variables used in HSMs can be point observations or full cover layers (e.g. satellite imagery). Full cover predictive variable layers are needed when the aim is to produce full cover species distribution maps. Because modelling a large area does not necessarily imply considering a coarse resolution, the scale of predictive variables is often expressed separately as resolution (or pixels size) and extent (e.g. the North Sea; Guisan and Thuiller, 2005).

Predictive variables should ideally be causal variables that determine the distribution of the species because they directly influence the species' physiology (Guisan and Thuiller, 2005). The opposite of causal variables are proxy or indirect variables, which are easy and cheap measurable approximations for other variables, e.g. suspended particulate matter from satellite images can be used as a proxy for the food available to filter feeders. Proxy variables often have a more indirect and correlative relation with the prediction of the species. The use of proxy variables may therefore lower the generalisation ability of HSM when used in other regions because local correlative relations are used in the model (Randin *et al.*, 2006). In the assessment of the model reliability and to determine if a model can be used in other regions, it is therefore crucial to know if the predictive variables are causal or rather proxy variables (Luoto *et al.*, 2002). At the moment such analyses are mostly lacking in the modelling methodology.

1.2.3. Model development

Several modelling techniques are used in the field of habitat suitability modelling to model the species-environment relation. A number of comparative reviews on the techniques are available (Olden and Jackson, 2002a; Segurado and Araujo, 2004; Elith *et al.*, 2006; Meynard and Quinn, 2007). The modelling techniques can be classified in several ways. First the possible model outputs of HSMs are discussed, later the modelling techniques used in this thesis are discussed in more detail.

1.2.3.1. Calibration data determine the model output

The model output is determined by the type of model calibration data available, and in turn the model output determines the possible modelling techniques and performance indicators that can be used. The calibration data can be continuous species densities or discrete presence/absence observations.

Models calibrated with density data predict densities in a fairly straightforward manner. Examples of modelling techniques predicting densities are Generalised Linear Models (GLMs, Attrill *et al.*, 1999; Maes *et al.*, 2004), Generalised Additive Models (GAMs, Maravelias and Papaconstantinou, 2003) and Artificial Neural Networks (ANNs, Maravelias *et al.*, 2003). Some authors have also developed coupled two stage models predicting presences as well as densities (Jensen *et al.*, 2005; Koubbi *et al.*, 2006). A first model predicts the presence of the species and a second model predicts the density, conditional on the species being present. Because species absence observations could mask the underlining dependence on environmental variables, the density model is calibrated using samples where the density does not equal zero.

HSMs can be developed from presence-only observations, in addition to a set of observations of environmental variables (Brotons *et al.*, 2004; Barry and Elith, 2006; Elith *et al.*, 2006; Olivier and Wotherspoon, 2006). Only the locations where the species is present are available, but there is no information on locations where the species is absent. Despite their limitations, use of such data is often justified by the lack of systematic survey data, e.g. models based on museum collections (Elith and Leathwick, 2009). Examples of modelling techniques are Ecological Niche Factor Analysis (ENFA, Hirzel and Arlettaz, 2003) and maximum entropy modelling (MAXENT, Phillips and Dudik, 2008). These modelling techniques use background environmental data of the whole study region from environmental data grids. The idea behind these techniques is that the environmental variables at places where the species is occurring are compared to the overall distribution of these environmental variables (Pearson, 2007). As such, a discrimination can be made between suitable and less suitable variable combinations.

Most models predicting the probability of presence of species use both species presence/absence observations to develop the model. As such the proportion of presences observations expected for a certain combination of predictive variables is modelled. Modelling densities of species rather than presence/absence would be more sensitive (Thrush *et al.*, 2003), but the use of presence/absence data to develop models is expected to avoid bias from seasonality, long term fluctuations and different sampling methods (Ysebaert *et al.*, 2002). A possible disadvantage to the use of presence/absence observations for model calibration is the occurrence of false absence observations in the data set: the habitat is suitable, but the species is absent. Especially rare, clumped, large and mobile species are expected to generate more false absences observations (Ysebaert *et al.*, 2002). An advantage of presence/absence data is that these data convey valuable information about surveyed locations (enabling analyses of biases and the sampled range per variable) and prevalence (Phillips *et al.* 2009). Several modelling techniques can be used with species presence/absence observations and in the next section some modelling techniques will be discussed in more detail. In this thesis only presence/absence observations will be used to develop HSMs.

When no absence observations are available, some modellers have generated pseudo-absences in the studied region, based on grids of the environmental variables (Chefaoui and Lobo, 2008; Václavík and Meentemeyer, 2009). These pseudo-absences can then further be treated as if they were normal absence observations. The pseudo-absences may be selected randomly (e.g. Stockwell and Peters, 1999) or according to a set of weighting criteria (e.g. Engler *et al.*, 2004). An important difference between the pseudo-absence approach and the background approach of presence-only techniques, is that pseudo-absence models do not include presence localities within the set of pseudo-absences (Pearson, 2007).

1.2.3.2. Modelling techniques

The modelling techniques used in this thesis are correlative algorithms, which essentially use the correlative relation between the species observations and the environmental variables. The causal relation between the species distribution and variables is not necessarily known. These algorithms are also called empiric (Guisan and Zimmermann, 2000) or data-driven, because they estimate the model parameters from a field data set. Elith *et al*. (2006) found out differences between predictions by 16 correlative modelling techniques. The models with the highest predictive performance were those that were able to model complex relations in the data, including interactions among predictor variables (Elith *et al.*, 2006).

The correlative modelling techniques can be divided in two classes: statistical and machine learning techniques. Statistical methods are based on error distributions, testable null hypotheses, generate *p*-values, etc. Examples of techniques used in habitat suitability modelling are Discriminant Function Analysis (DFA; e.g. Stevens and Boness, 2003; Degraer *et al.*, 2008), GLMs (e.g. Le Pape *et al.*, 2007; McBreen *et al.*, 2008), GAMs (e.g. de Segura *et al.*, 2007; Bekkby *et al.*, 2008). DFA is used to discriminate between discrete categories, for example presence or absence sites, based on a set of discriminant functions applied to the predictive variables (Stevens and Boness, 2003) or between species communities (Degraer *et al.*, 2008). DFA creates linear combinations of variables with normal errors that best discriminate between sites defined *a priori* by the presence or absence of a species or a community. GAMs use semi-parametric, data-defined smoothers to fit non-linear functions, therefore GAMs have a higher flexibility regarding the shape of the response (Elith *et al.*, 2006; Schroder, 2008)

In this thesis GLMs will be used and therefore an extensive introduction to this modelling technique is available in Appendix II. GLMs allow choosing the error distribution and the link function, depending on the data used for modelling. GLMs are a generalisation of general linear models because 1) other distributions can be assumed besides the normal distribution, 2) both the response variable and

the predictive variables can be categorical variables and 3) GLMs allow to model also nonlinear functions of the mean (Kutner *et al.*, 2005). Logistic regression (LR) is a widely used type of GLM in habitat suitability modelling (Schroder, 2008), where the model is calibrated with categorical presence/absence species observations, and therefore the error distribution is binomial and the link function logistic (Agresti, 2002). GLMs with a logistic link function are used to predict the probability of presence of species based on a data set with presence/absence observations.

Several machine learning algorithms are used in HSMs: Artificial Neural Networks (ANNs), Classification And Regression Trees (CART) and recently the maximum entropy approach has been introduced (MAXENT; Phillips and Dudik, 2008). The CART algorithm splits all the samples based on the value of one predictive variable (e.g. depth is more or less then 25m). CART thus partition the range of each predictive variable and provide a probability of species presence for each variable combination (i.e. classification trees; e.g. MacLeod *et al.*, 2007; Pesch *et al.*, 2008) or provide a regression formula to be applied within a certain range of the predictive variables (i.e. regression trees; Dzeroski and Drumm, 2003). MAXENT finds the species distribution that is closest to uniform (thus highest entropy), with the constraint that the expected value of each environmental variable (or its transform and/or interactions) under this estimated distribution matches its empirical average (Phillips *et al.*, 2006).

As ANNs will be used in this thesis an in-depth discussion of this technique can be found in Appendix I. ANNs consist of interconnected neurons divided into layers: an input layer, output layer and one or several hidden interlayers (Lek and Guegan, 1999). Each neuron receives a number of inputs from the neuron on the previous layer that are multiplied by an interconnection weight and these are being summed. Next this sum, plus a bias term, is fed into a transfer function, of which the output is then passed onto the next neuron. Similar to the β-terms in regression, interconnection weights thus determine the relation and relative influence of each predictive variable to the ANN model output. If sufficient interneurons are used, ANNs can approximate any function (Lek and Guegan, 1999). But ANNs are deemed to be "black-box" models, as it is difficult to gain insight in the meaning and mechanistic relevance of the equations that relate the inputs and outputs (Olden and Jackson, 2002b). The choice of the number of interneurons and which transfer functions to use remains an issue to be solved; at present researchers most often determine this by trial and error (Dedecker *et al.*, 2004).

1.2.3.3. Model selection

HSMs use a combination of environmental variables, which are assumed to determine the distribution of species (Barry and Elith, 2006). One of the challenges in habitat suitability modelling is the selection of an appropriate subset of predictor variables from all the variables in a data set. This variable selection

and the choice of the response modelled per variable (e.g. linear or unimodal) are together called model selection, and model selection is considered a central step in the model development (Guisan and Zimmermann, 2000; Heikkinen *et al.*, 2006; Franklin and Miller, 2009). Model selection methods aim at determining the most optimal model by adequately constraining the number of predictive variables used, and thus the model complexity (Reineking and Schroder, 2006).

Model selection algorithms are needed to generate alternative models and then select the most optimal model for a given data set. Stepwise model selection is most often used in habitat suitability modelling (e.g. Attrill *et al.*, 1999; McBreen *et al.*, 2008). During a stepwise model selection approach, predictive variables are sequentially added to a model (i.e. forward selection) or removed from a model (i.e. backward selection) until an optimal model has been found (Franklin and Miller, 2009). However, the stepwise model selection has a number of drawbacks: 1) due to the stepwise nature the global optimal can be missed (Reineking and Schroder, 2006), 2) the stepwise selection is sensitive to collinearity of the predictive variables (Hirzel and Guisan, 2002) and 3) stepwise selection precludes multimodel inference and prediction; as only one model is selected. An improved model selection approach is thus needed to find the globally most optimal model for a species in case of multicollinearity. This improved model selection approach should also allow multimodel inference and prediction by quantifying how optimal each model is.

1.2.4. Model Validation

1.2.4.1. Types of model validation

Model validation is necessary to determine if a model is appropriate for the intended goal and to compare different modelling methods (Pearce and Ferrier, 2000; Redfern *et al.*, 2006). Model validation also tests for problems in model fitting: overfitting or underfitting of data sets (Guisan and Thuiller, 2005). The model validation strategy used will mainly be determined by the available data and the predicted model output (Pearson, 2007).

Models are usually validated by comparing predictions with the original observations. A distinction can be made between internal and external model validation (Randin *et al.*, 2006) based on whether an independent data set is used for model validation. External validation uses an independent data set that is collected at another point in time (e.g. Iampietro *et al.*, 2005), at another location (e.g. Clark *et al.*, 2004) or both (e.g. Francis *et al.*, 2005). Araujo and Guisan (2006) state that external model validation is to be preferred form internal model validation, as the use of an independent data set tests the model generalisation. Generalisation is defined by Cheng and Titterington (1994) as a models' ability to perform well on data that were not used to train it. The generalisation ability of the model also

determines the model transferability to other regions or periods. By full transferability, Randin *et al.* (2006) mean that: 1) the internal validation of models fitted in region 1 and 2 must be similar; 2) a model fitted in region 1 must at least retain a comparable external validation when applied in region 2, and vice-versa; and that 3) internal and external predictions have to match in both regions.

As independent data sets are not always available, the single data set available is often split in a part for model calibration (the calibration or training set) and a part for model testing (the test set). Ideally the calibration and test set performance should be similar. If the training set performance is higher, the model has overfitted ("learned by heart") the training data and cannot perform well with new data. Internal validation can however lead to an overestimation of the predictive performance as the test set is only pseudo-independent, because the calibration and test set are part of the same original data set. This will cause the model performance to be lower on independent data.

The original data set can be split in a number of ways during the internal model validation. K-fold crossvalidation is most often used (e.g. Francis *et al.*, 2005; Haputhantri and Jayawardane, 2006). In the case of k-fold crossvalidation the original data set is iteratively split up in k parts of which k-1 parts make up the calibration set, and the remaining part is used as test set. Model performance statistics are then reported as the mean over the k model replicas (Fielding and Bell, 1997). Choosing the number of crossvalidation folds is always a trade off between using more samples for model calibration or for model validation. The extreme case when k equals the number of samples in the original data set is called leave-one-out, as only one sample is left apart for model testing (Pearson *et al.*, 2007). This approach is used when the number of samples is very low (e.g. <20, Pearson, 2007). Alternatively to k-fold partitioning, bootstrap resampling can be used to create replicas of the original data set. In contrast to data splitting, bootstrapping methods resample the original set of data randomly with replacement (i.e. the same occurrence record could be included in the test data more than once).

1.2.4.2. Model performance indicators

To compare and assess HSMs during the model validation, a quantification of the models' predictive performance is necessary. Therefore, several performance indicators have been developed. The basic idea behind each indicator is that the model predictions and the field observations are compared. The type of model performance indicator is dependent on the model output. The discussion here will be limited to performance indicators for models that are calibrated with a data set containing presence/absence observations of a species, as only this kind of HSMs will be constructed in this thesis.

The predictive performance of models that predict the probability of presence of a species is mostly assessed by first converting the continuous model output, [0 – 1], to presence/absence coding

[0, 1]. For this conversion a cut-off for presence needs to be chosen. Liu *et al.* (2005) reviewed different methods to determine the optimal cut-off level. No single cut-off for presence was optimal for all models (Liu *et al.*, 2005), making the choice of a specific cut-off subjective.

The agreement of the predictions with the observations is tabulated in a contingency table also called confusion matrix (Table 1.1). The resulting contingency table shows the matches between the observations and predictions (true presence and true absence). The mismatches, false presence and false absence, are also termed commission and omission errors respectively.

Table 1.1. A contingency table or confusion matrix. a = True Presence (observed present and predicted present); b = False Presence (observed absent and predicted present); c = False Absence (observed present and predicted absent); d = True Absence (observed absent and predicted absent). A-C: examples of possible contingency tables.

		Observed	
		Present	*Absent*
Predicted	*Present*	a (TP)	b (FP)
	Absent	c (FA)	d (TA)

A

		Observed	
		Present	*Absent*
Predicted	*Present*	0	0
	Absent	5	95

B

		Observed	
		Present	*Absent*
Predicted	*Present*	25	25
	Absent	25	25

C

		Observed	
		Present	*Absent*
Predicted	*Present*	40	20
	Absent	10	30

From the contingency table (Table 1.1) a whole series of model performance indicators can be calculated (Table 1.2). For a number of possible contingency tables (Table 1.1), the values for each of the model performance indicators are calculated to illustrate how each indicator responds. Contingency table A (Table 1.1A) is an illustration of a model result that predicts a very rare species (prevalence 5%), to be absent everywhere, which is a nonsense prediction. Contingency table B (Table 1.1B) is an example of the result of model that predicts the probability of presence to be 50%, regardless of the actual presence of the species. Contingency table C (Table 1.1C) is from a model with a quite good prediction.

The % Correctly Classified Instances (CCI, e.g. Iampietro *et al.*, 2005) is commonly used, but very sensitive to the prevalence of the species (Manel *et al.*, 2001). If a HSM predicts a species with very low or very high prevalence to be, respectively, absent or present in all the samples, this model will have a very high CCI. Contingency table A has a CCI of 95%, while the model predictions are in fact nonsense (Table 1.2). Similarly, contingency table B has a CCI of 50%, while the model is predicting a

probability of 50%, which is no better than guessing. Sensitivity is the ratio of true presences over the number of presence observations and the specificity is the ratio of true absences over the number of absence observations (Fielding and Bell, 1997). Two other indicators of the models' predictive performance are the Positive Predictive Power (PPP; Table 1.2) and the Negative Predictive Power (NPP; Table 1.2; Fielding and Bell, 1997). The contingency tables A and B, result in high values for the specificity and NPP, but have very low values for the sensitivity and PPP. Only a combination of the sensitivity and specificity or the NPP and PPP indicators, can thus unmask the nonsense model predictions.

Table. 1.2. Contingency table derived measures of model predictive performance. Adapted from Fielding and Bell (1997). a = True Presence (observed present and predicted present); b = False Presence (observed absent and predicted present); c = False Absence (observed present and predicted absent); d = True Absence (observed absent and predicted absent). N is the total number of samples in the data set (= a + b + c + d). The example contingency tables are provided in Table 1.1. NA: not possible to calculate.

		Example		
Performance Indicator	Formula	*A*	*B*	*C*
Species prevalence	$\frac{a+c}{N}\cdot 100$	5%	50%	50%
% Correctly Classified Instances (CCI)	$\frac{a+d}{N}\cdot 100$	95%	50%	70%
Sensitivity	$\frac{a}{a+c}$	0	0.5	0.57
Specificity	$\frac{d}{b+d}$	1	0.5	0.6
Positive Predictive Power (PPP)	$\frac{a}{a+b}$	0	0.5	0.67
Negative Predictive Power (NPP)	$\frac{d}{c+d}$	0.95	0.5	0.75
Cohen's Kappa	$\frac{[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/N)]}{[N-(((a+c)(a+b)+(b+d)(c+d))/N)]}$	0	0	0.4
Normalised Mutual Information index (NMI)	$\frac{[-a.\ln(a)-b.\ln(b)-c.\ln(c)-d.\ln(d)+(a+b).\ln(a+b)+(c+d).\ln(c+d)]}{[N.\ln(N)-((a+c).\ln(a+c)+(b+d).\ln(b+d))]}$	*NA*	0	0.12

The Cohen's Kappa (e.g. Dedecker *et al.*, 2004) and the Normalised Mutual Information (NMI, Forbes, 1995) are performance indicators that are compensating for the prevalence of the species (Fielding and Bell, 1997). For theoretical reasons the NMI is the most optimal performance indicator, but is rarely used (Forbes, 1995). The NMI also compensates for the prevalence of the species, and additionally takes into account the special status of the presence observations, which are more reliable than the absence observations, since the latter observation has a higher chance to be false due to sampling uncertainty (Forbes, 1995). In the example contingency tables (Table 1.2), the Kappa and NMI values for the two nonsense predictions A and B, are zero or not possible to calculate (NA). A χ^2-test based on the contingency table is also used to assess whether the model predictions are significantly different from a null-model (Norcross *et al.*, 1999; Ysebaert *et al.*, 2002).

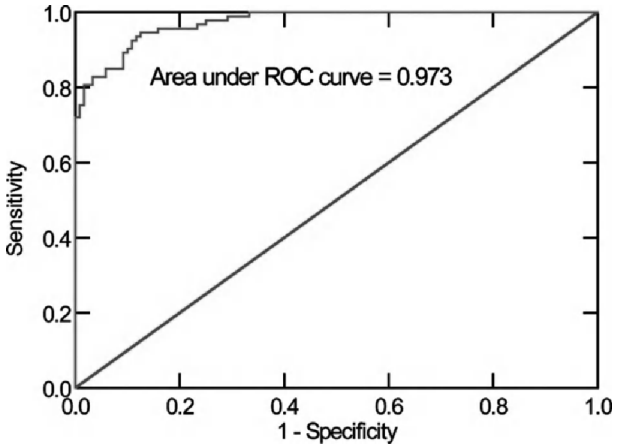

Fig. 1.4. Example of a Receiver Operator Curve for a very reliable model. The diagonal line indicates the ROC of a null model, which randomly predicts presences and absences.

As the choice of one cut-off value to convert a continuous model output to presence/absence is subjective, some authors use the Receiver Operator Curve (ROC, Swets, 1988). First the sensitivity and the specificity are calculated for a range of cut-off values for presence [0,1]. Each cut-off for presence results in a contingency table, which is used to calculate a specificity and sensitivity value. Next all these sensitivity values are plotted against the 1-specificity values (Fig. 1.4). When all these points are connected, the ROC-curve is obtained (Fig. 1.4). The ROC-curve also allows calculating the Area Under the Curve (AUC, Swets, 1988), in order to summarize the ROC in one value. A perfect model would have a sensitivity and specificity of one for each cut-off for presence value, which would result in an AUC of one. A nonsense model which predicts the probability of presence to be 50% everywhere, results in sensitivity and specificity values of 0.5 for each cut-off values. This generates an AUC of 0.5 for a nonsense model, illustrated by the diagonal line in Fig. 1.4. Although the AUC does not require a cut-off for presence, this performance indicator has some drawbacks: 1) it ignores the goodness-of-fit of

the model (Lobo *et al.*, 2008), 2) the performance of the model in regions that are not practically used is incorporated in the AUC (Lobo *et al.*, 2008), and 3) the AUC is not independent of the prevalence of the species, contrary to common belief (Maggini *et al.*, 2006; Lobo *et al.*, 2008).

Another group of performance indicators are based on the likelihood of the model given the observations. Likelihood-based Model Optimisation Criteria (MOCs) are often used in model validation. Commonly used examples are the Akaike Information Criterion (AIC, Akaike, 1973) and the Bayesian Information Criterion (BIC, Burnham and Anderson, 2004). The AIC was used by Kupschus (2003) and Haputhantri and Jayawardane (2006) in the context of habitat suitability modelling.
The MOC is defined as:

$$MOC = goodness\text{-}of\text{-}fit + \lambda\, model\ complexity \quad (1.1)$$

A general MOC consists of three parts: 1) a measure of the model fit to the observations, 2) a measure of the model complexity and 3) a regularisation parameter λ (Equation 1.1, Reineking and Schroder, 2006). The goodness of the model fit is quantified in the MOC as likelihood of a model given the data. The model complexity term equals the number of model parameters p. The regularisation parameter λ determines the relative weight of the model complexity p in the MOC formula (Equation 1.1; Reineking and Schroder, 2006). The likelihood of the model given the observations can be calculated for different model outputs (density or presence/absence). By choosing different λ values, a different trade-off between model fit and model complexity can be obtained. With increasing model complexity, the model fit will increase monotonously, but also the risk of overfitting the model. The MOCs thus have the benefit that they allow to penalise for increasing model complexity. As such they allow selecting which models have the highest parsimony, rather than the model with the highest fit to the data.

1.3. Applications of habitat suitability models

1.3.1. Habitat suitability model-based sampling

Sampling marine environments is expensive and time consuming. HSM-based sampling allows to increase the sampling efficiency (Guisan *et al.*, 2006a), because fewer samples can be collected, while maintaining the amount of ecological information. The general idea is that an initial HSM is developed to direct future sampling. This initial model can be calibrated with samples that are already available

(previous monitoring or even museum collections) or that are collected in a first exploratory survey (e.g. during the planning phase of a monitoring program). Graham *et al.* (2004), for example, successfully used museum collections to calibrate an initial HSM to direct the sampling of frog species.

If samples are collected in an exploratory survey the samples should ideally be randomly stratified per habitat type, to ensure that rarer habitat types receive equal sampling effort. Verfaillie *et al.* (2009a) proposed a method to identify habitat types based on solely physical and/or chemical data layers which is very useful in this context. In a later stage the initial HSM will be used to allocate the sampling effort for new samples, to each habitat type in a region. The allocation of the sampling effort can be based on the model residuals per habitat type, while also taking into account the local habitat heterogeneity. In regions with a bad fit of the initial model more samples should be collected, while fewer samples can be taken in a region with good model fit and low spatial heterogeneity.

To improve the sampling of rare species, Guisan *et al.* (2006a) used HSM-based sampling. They used an initial HSM to discover new populations of the rare species in locations where the predicted habitat suitability is high, but no samples were yet collected. Next, they used new samples collected in these locations to improve the model. In a simulation, Guisan *et al.* (2006a) found that HSM-based sampling reduced the sampling effort to find rare species with 70%, in comparison with random sampling. Graham *et al.* (2007) discovered tropical deep-water refugia of kelp species thanks to the use of HSM-based sampling. Similarly, Raxworthy *et al.* (2003) used HSMs to point out locations where species new to science are expected to be found. In regions that were spatially separated from the native region and had high predicted presences for a species, they found not the predicted species, but a related species new to science

1.3.2. Establishment of Marine Protected Areas (MPAs)

To protect species from extinction and marine habitats from further deterioration, marine managers are setting up Marine Protected Areas (MPAs, Kaiser, 2005). In comparison with terrestrial reserves, MPAs have a higher openness and connectivity with the surrounding region because the rates of dispersal of nutrients, planktonic organisms and larvae are higher (Stergiou and Browman, 2005). Carr *et al.* (2003) argue that this openness has an influence on the delimitation of marine reserves, and thus the suitability of the habitat around the MPA should also be considered.

The current MPAs are often established based on a limited number of species observations or based solely on abiotic variables, but mostly not based on full cover species distributions maps (Stevens and Connolly, 2004; Wilson *et al.*, 2005). If HSMs are used to develop full cover species distribution maps, these maps can be combined via reserve selection algorithms (Guisan and Thuiller, 2005; Wilson

et al., 2005) to estimate the most optimal MPA location and borders (Cabeza *et al.*, 2004). Such algorithms are always a trade-off between the risk of inadequately conserving biodiversity and the amount of sampling precision that can be sacrificed (Wilson *et al.*, 2005).

A critique on MPAs is that their utility for highly mobile species is questionable (Kaiser, 2005). The borders of MPAs should thus be dynamic and based on the momentary suitability of a location for the species. A solution would be an integrated framework of HSMs that generate species distribution maps based on dynamic information of the environmental variables, combined with a reserve selection algorithm that dynamically estimates the most optimal MPA borders. As such the temporal dynamics of migratory species can be captured and specific phenomena (e.g. seasonal spawning depending on oceanographic trigger conditions) can be protected. An integration of HSMs and reserve selection algorithms also allows the MPA borders to be refined and updated to incorporate both new biological observations and environmental observations (Canadas *et al.*, 2005).

All the environmental variables in the HSMs should have a full spatial cover of the region. Ideally, these variables should have a high spatial resolution, otherwise microhabitats with specific species to protect can be overseen in the MPA delimitation process. Also the problem of "released matching" can happen, when several habitat requirements for a species are met in a single grid cell based on the aggregated variable grids, but inside the cell the suitable conditions do not overlap on a microscale (Guisan and Thuiller, 2005).

Attention should be given to minimise false presence-errors in the HSMs because locations where the target species is absent should not be included in the MPA (Loiselle *et al.*, 2003). As the spatial extent of an MPA will always be limited, these errors could prevent the protection of locations that do contain the species. When setting up an MPA, source-sink effects (Pulliam, 2000) should also be considered. These effects can cause species to be in a suboptimal habitat which is suitable enough to survive, but not suitable enough to support reproduction. Sink habitats should not be included in the MPA unless the source population is also included (Guisan and Thuiller, 2005), as the species can still go extinct if the source habitat receives no protection. In reality however, it's hard to determine if the protected habitat is a source or sink habitat (Guisan and Thuiller, 2005).

In the Mediterranean, Canadas *et al.* (2005) proposed an MPA for Cetacean species based on the modelled presence and group size of dolphin species. Louzao *et al.*, (2006) estimated first the foraging range of seabirds from presence/absence data, and then the foraging grounds where high bird densities were predicted. They proposed an MPA with a high protection core zone where the bird densities predictions were highest. A buffer zone with diffuse protective measures, delimited by the foraging range of the birds, was proposed around the core zone.

1.3.3. Environmental impact assessment and habitat loss

The current practice in most marine impact studies (e.g. fishing, dredging impact or pollution) is to compare the species composition in a reference area where the impact is absent, with an impacted area (e.g. Van Dalfsen *et al.*, 2000). Differences found in the species compositions between the reference and impacted region are then attributed to the impact. However, this comparison only holds under the assumption that the habitat suitability is exactly similar in the two regions, otherwise the effect of the impact cannot be separated from the effect of the different habitats on the species composition. It is not likely that two locations are identical in their physical habitat. If the variance in the species distribution due to variance in the local habitat suitability can be controlled for by HSMs, the effect of a human impact can be quantified more objectively (Redfern *et al.*, 2006).

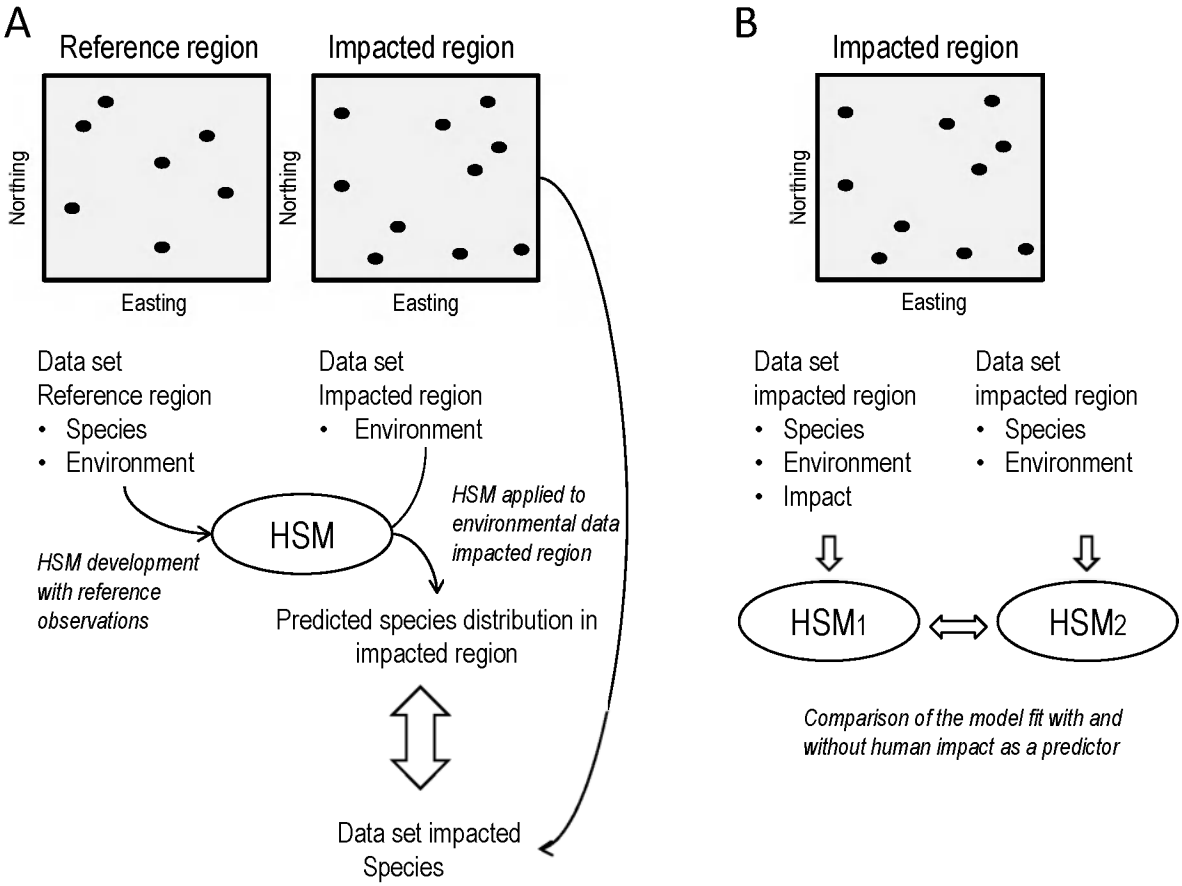

Fig. 1.5. Two approaches to use Habitat Suitability Models (HSMs) to estimate the effect of human impacts. A. Approach when a reference and impacted region are available, B. Approach when the strength of the human impact is known.

Two possible approaches in the use of HSMs to quantify the effect of human impacts can be followed, based on whether the strength and distribution of the human impact is known or not (Fig. 1.5). In the most common situation, the intensity of the human impact is not measured directly but a reference region is used where the impact is absent (Fig. 1.5A). Based on a data set collected in the reference region, the habitat preference of the species in the absence of the impact is modelled. Next, this HSM is applied to a data set with only physical variables from the impacted region. The effect of the human impact is quantified by comparing the observed species in the impacted region, with the species predicted to be present based on the HSM that was developed with observations in the reference region. The difference in species composition is then assumed to be due to human impact.

In case the strength and spatial distribution of the human impact is known (e.g. fishing effort maps or distribution sediment dredging), the impact can be used as a predictive variable along with the environmental variables (HSM1; Fig. 1.5B). We define in this context the effect of a human impact as an alteration of the habitat suitability, which can be observed as a change in the species distribution. Using the human impact as a predictor variable should improve the fit of the model HSM1 compared to HSM2, the model without the human impact as a predictor. This is because locations where the habitat is suitable but the impact is strong will have a different predicted probability of presence. If the inclusion of the human impact variable significantly improves the HSM model fit, an effect of the human impact can be assumed. A major advantage of this approach is that a reference region is not needed, as the local effect of the impact is considered in the HSM. Another advantage is that, because the human impact is a predictive variable in the HSM, scenario simulations can easily be performed. For example: will the total suitability for a species in a region decrease or not as a function of a human impact? This can be used as a management tool to calculate the amount of suitable habitat that needs to be compensated after being lost due to human impacts. The total equivalent of suitable habitat in this context can be calculated as the sum of the suitability in all grid cells, times the surface area of a grid cell. In the ideal situation, also biological interactions should be considered, to truly distinguish the effect of the human impact on the species distribution, apart from other effects.

Stevens and Boness (2003) used HSMs to model the effect of human disturbance on seal breeding sites by using the human presence as a variable together with other physical predictive variables. They concluded that human disturbance decreased the chance that a breeding site was used, although all the other variables were in the suitable range. Avissar (2006) modelled the effect of different beach replenishment scenarios on horseshoe crabs. Based on a simulation using realistic forcing factors, she was able to conclude that the fill sediment should be very similar to the natural sediment for maximum habitat suitability.

1.3.4. Spatially explicit stock assessment and migration modelling

Stock assessments are traditionally used in fisheries science to estimate the population size of exploited species. The current population models of commercial species mostly ignore the local and temporal variance in habitat suitability, as these models only consider the number of spawning fish, the natural mortality and the fishing mortality (Cadrin and Secor, 2009). By coupling the assessment models with HSMs it is possible to capture some of the spatiotemporal variance of the habitat which is often not included in the current models (Sundermeyer *et al.*, 2005). At the moment the larval recruitment is often only related to the biomass of the adult population, but relations with environmental variables have been pointed out (see Rice, 2005 for an overview). For example, water temperature can influence the spawning and survival rate of the larvae, and thus can greatly influence the recruitment (Gibson, 1994).

A step ahead is to make stock assessment models spatially explicit (Cadrin and Secor, 2009) by introducing a model grid where the properties of the grid cells are based on measured values for the habitat variables (e.g. depth, sediment grain size,...). In a spatially explicit modelling approach, Cheung *et al.* (2009) combined a fish stock population model, an advective-diffusive model for passive larval migration as well as a habitat suitability model. They assumed the carrying capacity per grid cell to be proportional to the habitat suitability. Adults in such models are assumed to migrate to neighbouring suitable grid cells if the density is higher than the carrying capacity of the grid cell. Migration caused by crowding is thus related to population growth or by a decrease in the habitat suitability, which lowers the carrying capacity (Reyes *et al.*, 1994; Sundermeyer *et al.*, 2005). To model active migration in time, dynamic information on the environmental variables should be available. For example, Dreyfus-Leon and Kleiber (2001) modelled the behaviour of individual fish, searching for their optimal habitat and fishing vessels aiming for the fish in a spatially explicit model. Another example of HSM and population model integration is provided by Reyes *et al.* (1994), who included model terms for both the population biology (birth rate, mortality), as well as a term for the local habitat suitability of a model grid cell. Sundermeyer *et al.* (2005) used temperature, bottom sediment type, and bottom depth, as environmental variables in a spatially explicit stock assessment model for cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*). Based on such distribution estimates, simulations are possible for different fisheries management schemes. For example, the effect of fisheries exclosure zones or local fishing effort reductions can be simulated. This approach will also allow combining stock assessment models and HSMs, with socio-economic models of the fishing industry, which should ultimately lead to a more sustainable fisheries management.

In case no population model and stock assessment program are available yet, HSM can estimate the total stock size based solely on the suitability of the local habitat. All variables for the HSM

need to be full cover, and are stacked to form a variable grid. Next, a HSM is developed that predicts the density of individuals per grid cell. The stock estimate is then performed by summing the predicted density in each grid cell and multiplying this by the surface of the grid cell. The result is an estimate of the total number of individuals in the region, and of the spatial distribution of the density of the species over the region. Later on, this habitat-based stock estimate can be combined with population based stock assessment models if these models are developed. Bello *et al.* (2005) modelled the density of spiny lobsters per habitat type and performed a spatially explicit stock assessment over the whole region. Clark *et al.* (2004) performed a stock estimate for shrimp based on the local habitat suitability. This model provided an estimate of the total stock, as well as a spatial distribution of the species in the region.

1.3.5. Invasive species modelling

As worldwide boating traffic increases and non-native species are grown in aquaculture, the risk for invasive species in marine systems increases. Invasive species threaten biodiversity, marine industries (including fishing and tourism) and human health (Bax *et al.*, 2003). Invasive species are successful because they are often released from competitors, pathogens and predators (Rodder and Lotters, 2009). Only if invasive species populations are still localised and small, they can be eradicated (Inglis *et al.*, 2006; Capinha *et al.*, *accepted*).

HSMs can be used as an early warning system to point out which locations have the highest risk of invasion, which allows concentrating the monitoring. HSMs can efficiently assess the potential for invasion for a large number of species, even before their introduction (Peterson and Vieglais, 2001). Habitat suitability models have a limited accuracy in providing predictions of the actual timing of future invasions as they do not explicitly incorporate the demographic or human-induced processes that cause invasions (e.g. ballast water discharges, new aquaculture species, etc.; Gallien *et al.*, *in press*). At the moment HSMs are often reasonable alternatives when more process based modelling tools are missing (Gallien *et al.*, *in press*).

As niche conservatism is assumed, the niche characteristics of species are expected not to evolve fast enough to adapt to quick environmental changes, thus species must either track their suitable environment in space or die (Wiens and Graham, 2005; Hirzel and Le Lay, 2008). Also due to niche conservatism, the habitat preference of an invasive species is thus assumed to be equal in its original and new environment, at least over short time scales. However, the habitat preference of a species may shift after invasion.

The development of HSMs for invasive species needs extra care, as these species are not in equilibrium with their environment (Václavík and Meentemeyer, 2009; Capinha *et al.*, *accepted*), i.e. the

species is not present in all suitable habitat patches, although equilibrium is an assumption of HSMs. Species presence observations can thus be trusted, but the observation of species absence cannot be trusted. A species can be absent due to the unsuitability of the habitat, or because the species has not colonised this suitable habitat patch.

There are two approaches to model the habitat preference of invasive species of which no reliable absence observation are available: the use of presence-only modelling techniques (e.g. Ecological Niche factor Analysis, Hirzel *et al.*, 2002; or maximum entropy MAXENT, Phillips *et al.*, 2006) or alternatively the generation of artificial absence observations, termed pseudo-absences (Václavík and Meentemeyer, 2009). But presence-only models tend to overpredict the actual range of invasions because dispersion limitation is not included in simple HSMs (Václavík and Meentemeyer, 2009), so the predictions should be considered as potential habitat for colonisation. On the other hand HSMs models can be very sensitive to the choice of pseudo-absences generation method (Capinha *et al.*, *accepted*). Because an invaded region can have variable combinations not present in the native region, Capinha *et al.* (*accepted*) suggest to use all species observations, thus from both native and the invaded region (if available), to have the highest coverage of the environmental variables. Native range observations remain potentially useful in the absence of invasion data even if environmental conditions in native and invasive ranges differ slightly (Capinha *et al.*, *accepted*).

Few examples of HSMs for marine invasive species can be found in the literature. Le Pape *et al.* (2004) modelled the effect of an invasive mollusc on the habitat suitability of a habitat for fish. Particle dispersion models and HSM were used by Inglis *et al.* (2006) in the early detection of invasive bivalve species.

1.3.6. Climate change impact modelling

Climate change can impact the patterns of marine biodiversity through changes in species distributions. The combined effect of species migrations and extinctions in a region (collectively called species turnover) will affect the biodiversity, community structure and ecosystem functioning, especially in sub-polar region, the tropics and semi-enclosed seas (Cheung *et al.*, 2009). To identify effects of climate change, HSMs are used to distinguish the variance in species distributions due to the local habitat suitability from the variance due to climate change. Given the assumption of niche conservatism on the time scale of climate change (Rodder and Lotters, 2009), species will either track their suitable environment in space or go extinct (Hirzel and Le Lay, 2008). When different climate change scenarios are used together with other environmental variables as inputs for the HSMs, the species distributions and extinctions can be simulated. When the climate changes, the suitable climate range of a species

might have shifted to a region where other environmental conditions are unsuitable, or some climate conditions might disappear locally. The latter effect makes marine biodiversity in the high latitude regions the most sensitive to climate change (Cheung *et al.*, 2009).

HSMs allow modelling the necessary movement of species to follow their preferred climate range, while taking into account the local habitat suitability in the regions the species move to. HSMs can thus model the location and overall quantity of suitable habitat in case of different climate change scenarios. Using HSMs to model the climate envelope of species can provide a useful first approximation of the impact of climate change if the spatial scale is well considered (Pearson and Dawson, 2003). On a large scale the climate is expected to determine the distribution of species, while biotic interactions are expected to play on smaller scales (Pearson and Dawson, 2003; Gallien *et al.*, *in press*).

The challenge is to develop HSMs that are reliable when used beyond the temperature range of the original observations used to calibrate the model. Using HSMs under future climate change is a case of model extrapolation which requires specific model validation approaches. Mostly, the validation performed is postdiction or hindcasting: the model is calibrated with current species-climate relations and then tested in the reconstruction of past species distributions, sometimes even based on fossils (Araujo and Rahbek, 2006). Another validation approach is to apply the HSMs to a data set from a region with temperatures in the range of the temperature predicted in the future. The drawback is that often species in these regions are not present in the original model training data from the first region. A last validation option is to use experimental data, but in this case the modelling of the realised niche from the field observations, is compared to the broader fundamental niche as observed in the experiments. HSM can also be based completely on experimental habitat and temperature preference observations (Guisan and Thuiller, 2005).

HSMs based on field observations are likely to model the realised niche, which is the result of biological interactions limiting the fundamental niche. The extent to which biological interactions limit the fundamental niche of a species depends on the strength of the biological interactions, the competing species densities and the spatial scale of the observations. The assumption that all species will respond similarly to environmental change is not realistic, thus it might be possible that HSMs will provide erroneous predictions of the realised niche due to different biological interactions in case of climate change that limit the fundamental niche in a different way (Pearson and Dawson, 2003; Hirzel and Le Lay, 2008). Under climate change, the correlative relation between species distributions and proxy variables (e.g. depth) can change, because there was only an indirect relation between the variable and the species response. This can be avoided by using more causal variables that are directly determining the physiological limits of the species in HSM.

Dispersion limitation needs to be considered when modelling species distributions and extinction risk under climate change, otherwise predicted and observed future distributions might differ considerably due to barriers that prevent species from migrating to their most suitable habitat (Pearson and Dawson, 2003). Correlating the current climate with the observed species distribution will not always identify the full climatic range of the species, because the species was unable to reach all suitable patches (Pearson and Dawson, 2003). When making predictions with HSMs, dispersion limitation should be incorporated to model the risk for species extinctions. If a species is unable to track its suitable habitat it will have a high risk of extinction.

Shifts in species distributions can be simulated by evaluating changes in the species distributions under climate change scenarios (Cheung *et al.*, 2009). As each meteorological climate model has its own limitations and variance, a greater understanding of climate scenarios and how these can be used as input for HSMs is needed (Beaumont *et al.*, 2008). There is no single best climate model at this moment, so it is more reliable to consider the whole range of climate models available and compare the range of outcomes of the subsequent HSMs for a species (Beaumont *et al.*, 2008).

1.3.7. Assessment of the strength of biotic interactions

HSMs mostly model the distribution of species based on abiotic variables only (Elith and Leathwick, 2009). The effect of biotic interactions can be modelled by using the density or presence of interacting species as predictive variables. When biotic interactions are included as predictive variables, the variance in the species distribution explained by the HSM will increase (Guisan and Thuiller, 2005). As HSMs cannot model feedback interactions as in population models, this approach should be regarded as a coarse method of exploring and modelling species interactions. If the physical habitat is suitable, but the species is absent this can be due to biotic interactions. Including those interactions in the HSM, can thus improve the predictive performance of HSMs. In the ideal situation human impacts should also be considered in the HSM, if they are assumed to influence the species distribution.

The relative contribution of biotic interactions can be determined by quantifying the extra variance explained when the densities of other species are used as a predictive variable. If this inclusion increases the model fit considerably, both species are assumed to interact in a negative or positive way, depending on the sign of the estimated model parameter. The different interactions that can be included are predation (density predator and/or prey), competition (density competitor), parasitism (density host or parasite) and mutualism (density mutualistic species). The main advantage of including other species as predictors, such as the predators or prey of the modelled species, is that these species densities are

often more direct and causal predictors. As such, trophic interactions can be included in a simple way. A drawback is that the density of one species is needed to predict another, which limits the use of the models for extrapolations.

Species lower in the food chain are expected to have a more direct link with environmental variables. Higher up in the food chain, species are mainly determined by the densities of their prey, but environmental variables still limit their distribution at larger spatial scales. A predator is thus responding indirectly to the environmental variables that determine the density of its prey. A competitively dominant species will be influenced less by the presence of competing species and will therefore face weaker negative biotic interactions (Guisan and Thuiller, 2005; Hirzel and Le Lay, 2008). As a result, the fundamental and occupied niches of dominant species are expected to be more similar compared to subordinate species (Guisan and Thuiller, 2005).

The interpretation of the partitioning of the explained variance between abiotic and biotic variables is not straightforward. Environmental variables might have a strong relation with the species distribution, but when an interacting species is introduced as a variable, the contribution of environmental variables might become marginal (Elith and Leathwick, 2009). Sometimes it thus remains unclear if improvement of model fit after inclusion of one species as predictor truly reflects biological interactions or reflects the absence of an important environmental predictor (Guisan and Thuiller, 2005).

When including biotic interactions in the HSM, scale aspects come into play, as each ecological process and interaction has its proper scale (Guisan and Thuiller, 2005). A mismatch between the spatial distribution of predator and prey should be avoided by using fine resolution observations, otherwise the problem of "released matching" (Guinet *et al.*, 2001) will happen: the prey-predator relation will change when observed at another spatial resolution. At fine resolution the prey avoids the predator, at lower resolution the predator seemed to attract the prey because they both share the same physiological range limits. In general, biotic interactions are mostly expected to have an effect on species distributions on small scales (Hirzel and Le Lay, 2008), while on macro-scale physiological limits are expected to determine the species distributions (Pearson and Dawson, 2003).

The most widely used biotic variable in the marine HSM literature is prey abundance, which has a very direct and causal link with the presence and density of predators. Le Pape *et al.* (2007) used densities of benthic epifauna prey to model the density of flatfish. Bradshaw *et al.* (2002) found squid and prey fish densities to be important predictors for the modelling of fur seal colonies. Rooper *et al.* (2005) used the density of prey invertebrates as a predictive variable to predict the flathead sole distribution. Van Tomme *et al.* (*submitted*) iteratively used the density of one species as a predictive variable in the HSM to predict another species in a macrofauna data set. Based on the extra variance explained by including species densities as predictors, a scheme with hypothised positive and negative

interactions between all species was drawn. This scheme can be a start for future experiments or food chain modelling with dynamic models.

1.3.8. HSMs in biogeography and phylogeny

As evolution is usually too slow to lend itself to experimentation, analytical techniques such as HSMs appear to be a valuable addition to biogeography and phylogenetic research (Hirzel and Le Lay, 2008). New insights in biogeographical and biodiversity patterns can be gained from investigating the evolutionary patterns of niche diversification. Species adapt over evolutionary time to new environmental conditions due to natural selection, but only when the new conditions are not too different from the ancestral niche of the species. Thus, over short time scales niche conservatism (Wiens and Graham, 2005) is assumed, while over a long time scale niche evolution and the emergence of new species is expected.

Conservatism of the preferred range of an environmental variable can be inferred through ancestral character state estimation in a phylogenetic tree (Verbruggen *et al.*, 2009). By plotting the habitat preference of species living today on a phylogenetic tree, the preference of extinct ancestors can be inferred. Then it becomes possible to infer if the preference for an environmental variable has been conserved over evolutionary times (Wiens and Graham, 2005; Verbruggen *et al.*, 2009). For example, Verbruggen *et al.* (2009) plotted the temperature preference, as modelled in a HSM, on a phylogenetic tree of the algal genus *Halimeda* and observed that the genus is characterised by a conservatism of the temperature preference for tropical temperatures, but one section of the genus managed multiple times to invade colder habitats independently.

HSMs can also be applied to investigate speciation patterns. Allopatric speciation is generally caused by a geographic barrier between two subpopulations that consists of unsuitable environmental conditions or a migration barrier (e.g. land, deep ocean; Wiens and Graham, 2005). These barriers result in a dispersion limitation between the two subpopulations which can then evolve to separate species over time because the genetic exchange has stopped. To infer if allopatric speciation has happened in the formation of two related species, HSMs can be used to project the modelled habitat preference of a species living at one side of the barrier, on the habitat of the whole region. In case of dispersion limitation, and thus allopatric speciation, the region of the second species might be suitable for the first species, but due to the migration barrier, the species cannot disperse there and allopatric speciation has taken place. Niche conservatism is important in allopatric speciation because it will limit adaptation to the unsuitable habitat in the geographic barrier, and two new species will evolve (Wiens and Graham, 2005). Verbruggen *et al.* (2009) applied the habitat preference of some algae species to

the whole ocean and were able to conclude that suitable, but distant, habitats were often not occupied by the species, strengthening the case for dispersal limitation.

Parapatric speciation is when species evolve in neighbouring regions, which are often separated by potential contact zones that have a lower habitat suitability. The lower the suitability of this contact area, the lower the chance of genetic exchange between subpopulations at both sides of the zone (Rissler and Apodaca, 2007). An application of HSMs here, is to model the suitability of the contact zone (Rissler and Apodaca, 2007). The local suitability of the contact zone can be used as a weighing factor in the calculation of the effective distance between the two neighbouring regions. The effective distance is here defined as the geographic distance times a factor determined by the local suitability of the contact zone (Adriaensen *et al.*, 2003), with less suitable habitat increasing the effective distance.

Such a least cost migration model (Adriaensen *et al.*, 2003) based on HSMs would allow to compare the genetic distance with the effective distance. A high cost, thus high effective distance, to migrate between subpopulations will increase the chance of parapatric speciation. Not only species, but also metapopulation structures can be studied in this way, by comparing the connectivity based on the habitat suitability, with the genetic distance between metapopulations.

HSM can also be used for species delimitation, by assessing if there are two or more distinct ecological niches within one species. Cryptic diversity or taxonomic errors can thus be suggested by comparing the habitat preferences for subspecies that are expected to be separate species. Even more proof can be generated when combining HSM with mitochondrial DNA and morphological data to re-evaluate species limits (Raxworthy *et al.*, 2007).

1.4. General objectives of the thesis

Despite the increased number of applications of habitat suitability modelling in recent years, the methodology of these models can still be improved significantly. Therefore, the general objective of this thesis is the improvement of the existing HSM methodology and more specifically the approach to model the distribution of macrobenthos species. In the introduction of this thesis several shortcomings of the methodology were pointed out. Three specific topics of the model development of HSMs are chosen for further research: 1) choice of the modelling technique, 2) model selection and 3) model validation. Each topic can be seen as a challenge in the current HSM methodology that this thesis aims to deal with. In this section each challenge will be discussed, as well as the data sets used and the species for which the models will be developed.

Challenge 1: Which modelling technique to use? (Chapter 2)

Numerous alternative modelling techniques are available to model the distribution of species. In this chapter, two commonly used correlative modelling techniques for absence/presence data will be compared: artificial neural networks and logistic regression, a type of GLM. Each modelling technique has several advantages and disadvantages (see Chapter 1 and Appendix I-II), and these will be compared against each other.

Models will be developed to predict the spatial distribution of the species *Lanice conchilega,* a common tube-building polychaete along the North-western European coastline. Marine management of the BPNS would greatly benefit from knowledge of the spatial distribution of this species. This species is known as a habitat engineer, increasing macrobenthic species diversity and abundance in soft sediments that lack any structure, through enhancement of the habitat complexity (Zühlke *et al.*, 1998; Zühlke, 2001; Rabaut *et al.*, 2007). *L. conchilega* is also an important food source for several demersal fish (Rijnsdorp and Vingerhoed, 2001) and, when occurring in high densities *L. conchilega* aggregations act as a refugium against predation for many organisms (Woodin, 1978).

The models will be developed with a data set collected in the Western Coastal Banks, a small region in the south west of the BPNS (Fig.1.3). This data set was chosen because the samples were collected on a high resolution sampling grid (500m) and numerous environmental variables were measured.

Challenge 2: What is the most optimal combination of predictive variables? (Chapter 3)

The general aim of Chapter 3 is the improvement of the model selection methodology for HSMs. Model selection is considered a central step in the model development (Guisan and Zimmermann, 2000; Heikkinen *et al.*, 2006; Franklin and Miller, 2009). Stepwise model selection is most often used in habitat suitability modelling (e.g. Attrill *et al.*, 1999; McBreen *et al.*, 2008), but this approach has a number of disadvantages (see Chapter 1). Therefore, the Combined Model Optimisation Criterion (CMOC) is proposed in this chapter which combines the predictive performance of the model as well as the model complexity for both the model calibration data and for independent data. The proposed modelling approach will be able to deal better with the properties of most data sets which are opportunistically collated to calibrate HSMs (e.g. low/high prevalence, multicollinearity, etc). The CMOC is not based on a contingency table and therefore does not require the arbitrary choice of a cut-off for presence.

The newly developed model selection approach will be applied to determine the most optimal logistic regression models. Logistic regression is chosen because this technique is often used in HSMs (e.g. Ysebaert *et al.*, 2002; Le Pape *et al.*, 2003), is relatively simple and well established statistically. GLMs have a transparent model structure and a comparatively low number of parameters (Reineking and Schroder, 2006). The conclusions and the proposed methodology can be applied universally to other HSM modelling techniques.

The proposed methodology will first be tested on artificial data of a virtual species (Hirzel *et al.*, 2001), because this provides full control over the data set. In a second step, the model selection methodology will be applied to field observations of *Abra alba*, a marine bivalve species. This species is an indicator for the *A. alba* macrobenthic community in the Southern North Sea, which is one of the ecologically most important soft-sediment macrobenthic communities along the coastal areas of the English Channel and Southern Bight of the North Sea (Van Hoey *et al.*, 2005).

The Macrodat data base with samples of the whole BPNS will be used to calibrate the models. This data set is representative for most marine data sets used in HSM development: the samples are a collation of different research projects over several years and only a limited number of variables is available for all the samples.

Challenge 3: Are the model predictions reliable? (Chapter 4)

There is a disproportionally large effort in developing HSM models, compared to the validation of the models (Eastwood *et al.*, 2003). Traditionally, HSMs are only validated by comparing the observations with the model predictions. The general goal of Chapter 4 is to plea for an integrated validation of marine HSMs which also validates the ecological soundness of the models. Such an integrated validation will consider: 1) the validation of the model with species observations (internal or external model validation), 2) ecological insights from the literature, 3) habitat preference experiments and, 4) assessment of the distribution of the model calibration data over the range of the predictive variables.

Ecological knowledge from the literature on the modelled species will be combined in a conceptual scheme. Such a scheme, as well as the habitat preference experiments, will be useful to determine if the relation of a variable with the species distribution is causal or rather correlative (i.e. a proxy variable). Knowledge on the causality of a variable is crucial when models will be transferred to other regions or periods (Luoto *et al.*, 2002; Randin *et al.*, 2006). The habitat experiments performed will also allow exploring the extent of the fundamental niche, as the preference for one variable is tested while other effects are excluded. For each variable in the data set it will be determined if the sampled variable range is sufficient to discriminate species absence and presence. Sometimes ecologically

useful variables will not be chosen in the model selection as the sampled range is insufficient. The model development is based on the CMOC approach proposed in the previous chapter. Logistic regression will be used as a modelling technique.

To illustrate the suggested model validation improvements, a HSM based on GLMs will be developed for the bivalve species *Donax vittatus* (Da Costa 1778). This species was chosen because extensive ecological literature is available on the species, habitat preference experiment results are available and the species is a food source for juvenile plaice (*Pleuronectes platessa*; Burrows and Gibson, 1995).

Chapter 2. Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*

Adapted from:

Willems, W., P. Goethals, D. Van den Eynde, G. Van Hoey, V. Van Lancker, E. Verfaillie, M. Vincx, and S. Degraer. 2008. Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*. Ecological Modelling 212:74-79.

Abstract

Grab samples to monitor the distribution of marine macrobenthic species (animals >1mm, living in the sediment) are time consuming and give only point based information. If the habitat preference of a species can be modelled, the spatial distribution can be predicted on a full coverage scale from the environmental variables. The modelling techniques Generalised Linear Models (GLM) and Artificial Neural Networks (ANNs) were compared in their ability to predict the occurrence of *Lanice conchilega*, a common tube-building polychaete along the North-western European coastline. Although several types of environmental variables were in the data set (granulometric, currents, nutrients) only three granulometric variables were used in the final models (median grain size, mud% and coarse sediment fraction). ANNs slightly outperformed GLM for a number of performance indicators (% correct predictions, specificity and sensitivity), but the GLM were more robust in the crossvalidation procedure.

Key words:
Lanice conchilega, polychaete, habitat preference, Generalised Linear Models (GLMs), Artificial Neural Networks (ANNs)

2.1. Introduction

This research will focus on *Lanice conchilega*, a common tube-building polychaete along the North-western European coastline. This species was chosen because of its role as habitat engineer, increasing macrobenthic species diversity and abundance in soft sediments that lack any 3D structure, through enhancement of the habitat complexity (Zühlke *et al.*, 1998; Zühlke, 2001; Rabaut *et al.*, 2007). *Lanice conchilega* is also an important food source for several demersal fish (Rijnsdorp and Vingerhoed, 2001; Rabaut *et al.*, 2007) and, when occurring in high densities *Lanice* acts as a refugium against predation for many organisms (Woodin, 1978).

The aims of this chapter are: 1) to identify the environmental variables determining the distribution of *L. conchilega*, 2) to search for the most optimal model describing the habitat preferences of *L. conchilega* and, 3) to compare the modelling performance of Generalised Linear Models (GLMs, Hastie *et al.*, 2001) and Artificial Neural Networks (ANNs, Lek and Guegan, 1999) when applied to a marine data set.

2.2. Material and methods

2.2.1. Data availability

All samples used were collected in the near shore part of the Belgian Part of the North Sea (BPNS; Fig. 2.1) within the framework of the HABITAT-project (Degraer *et al.*, 2002; Degraer *et al.*, 2003) in October 1999, March 2000 and November 2000. The major part of the samples (265) were collected in the area of the western Coastal Banks (WCB), a small complex of sandbanks and swales covering a wide range of soft sediment habitats (Degraer *et al.*, 1999a). Outside of the WCB, 38 additional samples were collected in November 2000, along four transects perpendicular to the coastline. The samples were collected with a Van Veen grab (sampling surface area: 0.1 m²) and sieved alive over a 1 mm sieve. In each sample the presence of adult *L. conchilega* individuals was assessed.

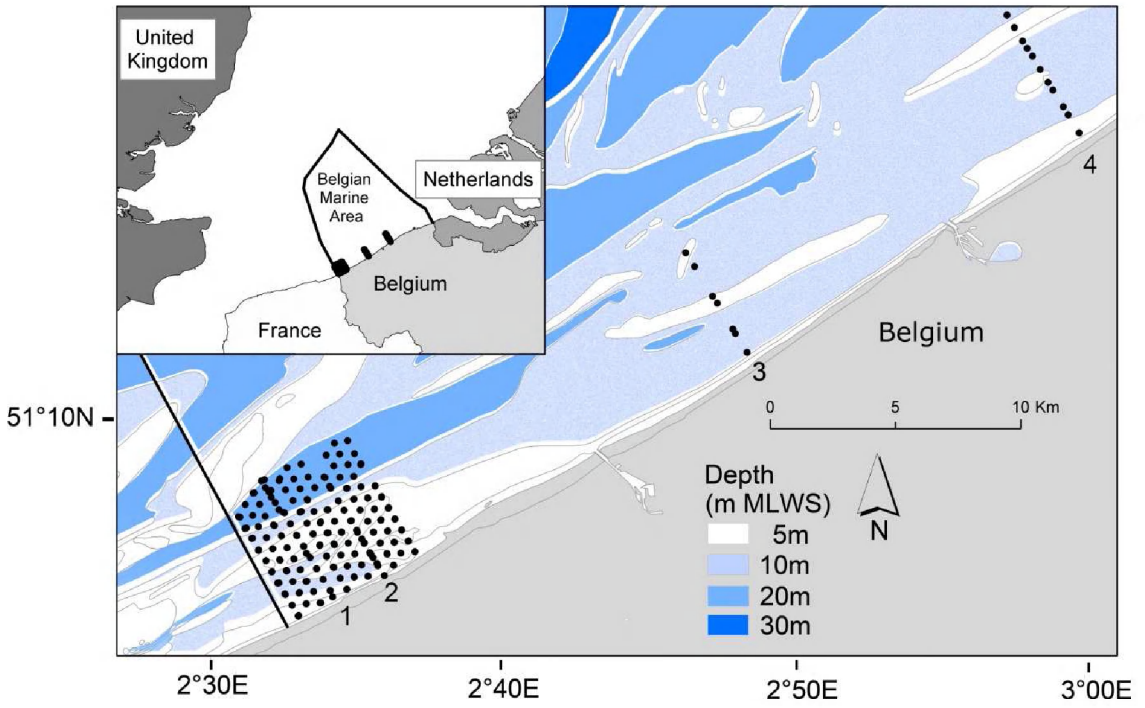

Fig. 2.1. Overview map of the sampling locations in the Belgian Marine Area. Numbers indicate the four transects. Projection UTM 31N WGS84.

In total 29 environmental variables were available in the data set. The range of each variable is provided in Table 2.1. A sediment subsample was taken of each Van Veen grab with a 3.6 cm diameter core to measure nutrient concentrations in the interstitial water: nitrate and nitrite (NO_3+NO_2), ammonium (NH_4), orthophosphate (PO_4) and silica (Si). Sediment granulometry was determined: the sediment fraction < 850 µm was analysed with a LS Coulter laser counter (volume %), while the sediment fraction >850 µm

was weighted (mass %). The following variables were calculated for the <850 µm size fraction: median grain size, mean grain size, mean/median grain size ratio (M/M ratio), the 10% and 90% percentiles (d10 and d90), mode, standard deviation of the grain size distribution, skewness, kurtosis, the volume percentages of the 0-63 µm (hereafter: mud%), 63-125 µm, 125-250 µm, 250-500 µm and 500-800 µm fractions, as well as the mass percentage of the > 850µm fraction (hereafter % coarse fraction).

Table 2.1. Range of the variables in the data set. d10 and d90: 10% and 90% quantile of the grain size distribution. U and Umax: median and maximum bottom current. BSRTM and BSRTX: median and maximum bottom shear stress.

Variable	Min	Mean	Max	Variable	Min	Mean	Max
L. Conchilega density m^{-1}	0.00	56.32	2300.20	800-850 µm	0.00	0.43	9.56
Mean grain size	8.61	238.48	617.00	% coarse fraction	0.00	5.11	86.02
d10	0.98	148.17	431.00	NO_2 (µg/l)	0.00	18.65	275.00
Median grain size	7.98	256.14	655.60	NO_3 (µg/l)	6.00	725.90	4738.00
d90	63.18	407.79	849.40	NO_3+NO_2 (µg/l)	3.00	552.21	4753.00
Mean/Median ratio	0.50	0.91	1.08	NH_4 (µg/l)	328.00	6785.73	66463.00
Mode grain size	4.97	267.30	853.00	PO_4 (µg/l)	77.00	1848.14	13489.00
Stdev. grain size	1.26	2.08	6.80	Si (µg/l)	157.00	1336.64	4887.00
Skewness	1.58	5.46	46.18	U (m/s)	0.15	0.24	0.34
Kurtosis	-5.76	-1.83	0.65	UMAX (m/s)	0.32	0.51	0.80
0-63µm	-1.12	9.53	52.99	BSTRM (m/s)	0.26	0.47	0.92
63-125 µm	0.00	3.20	20.20	BSTRX (m/s)	0.67	1.68	4.06
125-250 µm	1.58	41.93	82.40	Chla max (mg m^{-3})	34.58	83.58	100.26
250-500 µm	0.00	39.87	78.40	Chla median (mg m^{-3})	8.16	12.58	17.03
500-800 µm	0.00	6.83	61.50	Depth (m MLSW)	-16.65	-7.17	0.00

Bottom current speed and bottom shear stress were obtained from the 3D baroclinic hydrodynamic COHERENS model (Luyten *et al.*, 2003). This model has a horizontal resolution of about 250x250 m and a vertical resolution of ten layers. U and Umax are the median and maximum bottom current, and BSRTM and BSRTX are the median and maximum bottom shear stress. Median and maximum chlorophyll-a concentration in the surface water were obtained from MERIS satellite images of 2003 from the REVAMP-project (Peters *et al.*, 2005).

The distribution of the samples over the range of each predictive variable can have undersampled, oversampled and unsampled regions (see 4.2.4) In Fig 2.2. A visualisation of the sample distribution is provided for the variables in the final models. The depth range is also provided, because this variable is important when setting use limits for the obtained HSM.

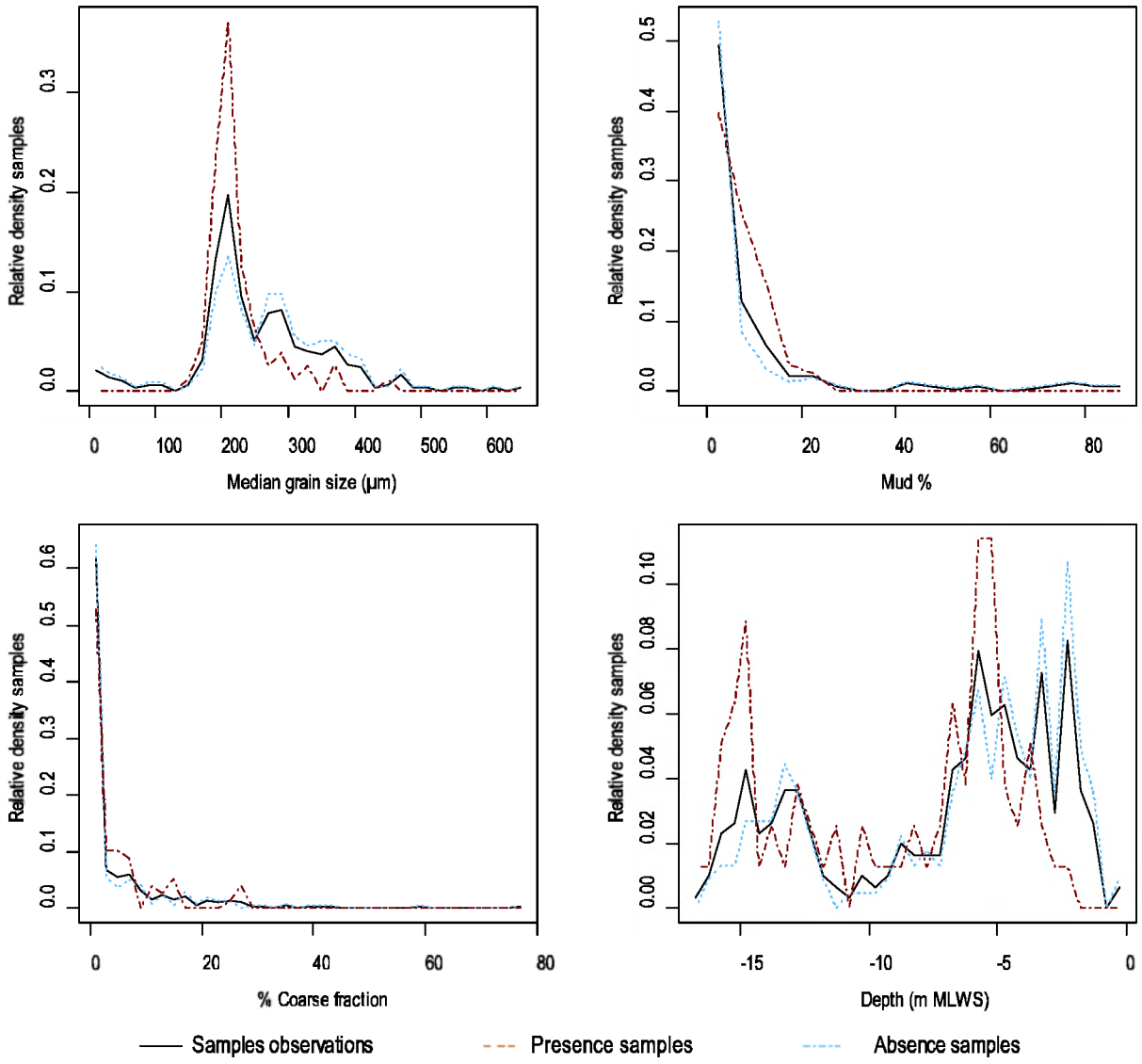

Fig. 2.2. Distribution of the observations over the range of the variables in the final models and additionally the depth range of the samples.

2.2.2. Modelling techniques

2.2.2.1. Variable selection

Since related variables (e.g. all granulometric variables) were expected to be highly correlated and thus redundant, Principal Component Analysis (PCA, Smith, 2002) was used to analyse the relationships among the variables before inclusion into the models. A varimax rotation was performed to maximise the independence of the Principal Components (PCs). After an exploration of the variable relations in

the data set, a forward stepwise selection was performed to select a final GLM. In a final step an ANN was developed with the same variable combination as obtained with the GLM stepwise selection.

2.2.2.2. GLM: logistic regression

To predict the absence or presence of *L. conchilega,* logistic regression (Trexler and Travis, 1993), a type of GLM, was used. Logistic regression has been widely used in ecology (Paruelo and Tomasel, 1997; Ysebaert *et al.*, 2002; Verween *et al.*, 2007) and predicts the probability (between 0 and 1) that a species will occur, based on the environmental conditions. In appendix II in this thesis an extensive introduction to GLMs and logistic regression is provided. Since the sample distribution was binary (present or absent), the logit link function was used. The forward stepwise likelihood-ratio method was used to select the best set of variables. Interaction terms and non-linear terms (i.e. quadratic) of each variable were also included in the set of predictive variables. The Wald-test (Kutner *et al.,* 2005) was used to test the significance of the model parameters β. The Wald-test tests if H_0: $\beta = 0$ is true by using the large-sample normality approximation of maximum-likelihood estimates. The Wald-statistic z^2 is obtained by dividing a parameter estimate by its standard error and then squaring it. In case H_0 holds, z^2 has an χ^2 distribution with one df. It was necessary to convert the continuous logistic regression output [0 – 1], to discrete absences and presences to create a contingency table. The cut-off value for species presence was based on the prevalence of the species *L. conchilega* in the data set (present in 26% of the samples, cut-off of 0.26) as proposed by Ysebaert *et al.* (2002). The analysis was performed with SPSS version 12.0 (SPSS, Inc., Chicago IL).

Threefold crossvalidation was used to test the robustness of the models. The complete data set was randomly split in three parts and two parts were iteratively used to construct a model, and the third part to test the model. Next, a final model was constructed with all the data.

2.2.2.3. Artificial Neural Networks

Artificial Neural Networks (ANNs) are a technique from the field of machine learning (Lek and Guegan, 1999). They have a similar structure as the human brain: a network of connected neurons. In appendix I of this thesis an extensive introduction to ANNs is provided. The neurons are the building blocks of the ANN. Data enters a neuron from several other neurons, is summed and then fed into an activation function, which generates the output of the neuron. Neurons can pass on information because they are connected. The importance of a connection is expressed as an interconnection weight. The adjustment of these weights will influence the model output (Lek and Guegan, 1999). Through a learning algorithm,

the weights will be adjusted iteratively, increasing the agreement between the observed and predicted presence of the species (Lek and Guegan, 1999).

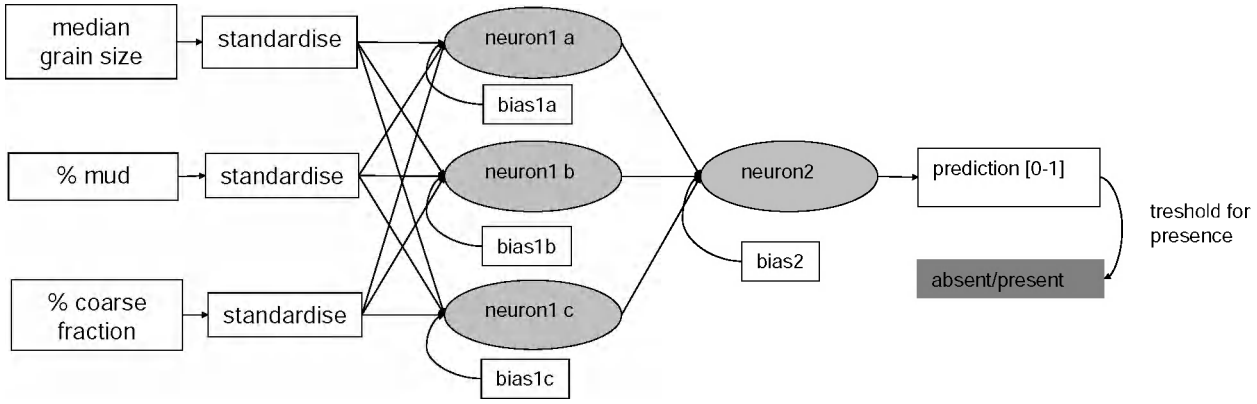

Fig. 2.3. The Artificial Neural Network model, constructed with all data, which predicts the presence of *Lanice conchilega*.

The ANNs in this research have their neurons organised in three layers: environmental variables are presented at the input layer, are passed on to the hidden layer which processes the information and an output layer which generates the prediction of the probability of presence of *L. conchilega*. The interneurons use a logistic transfer function, while the output neuron uses a linear transfer function. Prior to the model calibration, the environmental variable values were standardised to the interval [-1, 1]. The species was predicted to be present if the ANN output was larger than 0.5 (i.e. the cut-off for presence). As for GLMs, threefold crossvalidation - used to test the robustness of the models. The ANNs were constructed in MATLAB 6.1 using the neural networks toolbox.

2.2.3. Model performance and variable contribution

In order to assess and compare the predictive power of GLMs and ANNs several performance indicators were calculated. Most indicators were based on a two by two contingency table containing the number of True Positive (TP), False Negative (FN), True Negative (TN) and False Positive (FP) predictions (see 1.2.4.2. Model performance indicators). The overall percent of correct predictions was expressed as the % of Correctly Classified Instances (CCI, Fielding and Bell, 1997). The ratio of the number of correctly classified species absences over the total number of predicted species absences was calculated as the Negative Predictive Value. Similarly, the Positive Predictive Value was calculated. The model specificity and sensitivity were calculated. Cohen's Kappa (Dedecker *et al.*, 2004) was calculated as well, because this indicator is expected to be compensated for the prevalence of the

species in the original data set. An alternative statistical parameter used to express the performance of a model is the Area Under the Curve (AUC; Fielding and Bell, 1997). This parameter expresses the area under the Receiver Operator Curve (ROC), which is 1 for a perfect model and 0.5 for a null model without parameters. The AUC is not dependent on a single cut-off for presence at which the species is present, but a series of cut-off values are used (each resulting in a sensitivity and specificity value). In this way the AUC evaluates the model output in a continuous, instead of a discrete manner. Although the AUC does not require the choice of single a cut-off for presence, this performance indicator has some drawbacks 1) it ignores the goodness-of-fit of the model (Lobo *et al.*, 2008), 2) the performance of the model in regions that are not practically used is incorporated in the AUC (Lobo *et al.*, 2008), and 3) the AUC is not independent of the prevalence of the species, contrary to common believe (Maggini *et al.*, 2006; Lobo *et al.*, 2008). Finally, the Pearson correlation between the output of the ANNs and GLM final models was calculated to assess the similarity of the predictions.

The relative contribution of an environmental variable in the prediction of the probability of presence of *L. conchilega* was assessed. The Wald statistic (see Appendix II) for each variable in the GLM was calculated (Ysebaert *et al.*, 2002). Equivalently, the partial derivatives (PaD) method was used for assessing variable contribution for the ANN (Dedecker *et al.*, 2004). This method calculates the partial derivatives of the ANN output with respect to the input are calculated (Dimopoulos et al., 1995). Per variable the sum of squared partial derivatives is calculated and averaged over all the neurons. The input variable that has the highest sum, influences the output most.

2.3. Results

2.3.1. Principle components analysis

The first five PCs explain 78 % of the variance in the data and were clearly associated with groups of related environmental variables (Table 2.1). PC1 was most correlated with the sorting of the sediment and the fine sediment fraction, PC2 with the coarser fraction and with general sediment variables, PC3 only with current characteristics, PC4 with the shape of the grain size distribution and PC5 with the nutrient concentrations in the interstitial water.

Table 2.2. Principal Component Analysis: rotated scores (Varimax rotation) of the variables for the first five principal components (PC). The percentage of variance explained is shown in the upper right for each PC. Only scores with an absolute value above 0.50 are shown.

PC1	38%	PC2	16%	PC3	10%	PC4	9%	PC5	5%
St dev grain size	0.96	500-800 µm	0.94	U	0.93	Skewness	0.96	NH_4	0.73
38-63 µm	0.92	Mode	0.91	UMAX	0.93	Kurtosis	-0.93	PO_4	0.65
% mud	0.87	d90	0.89	BSTRM	0.92			NO_3+NO_2	-0.56
63-125 µm	0.78	Median grain size	0.84	BSTRX	0.76				
M/M ratio	-0.78	125-250 µm	-0.79						
d10	-0.77	Mean grain size	0.75						
Mean grain size	-0.62	% coarse fr.	0.71						

2.3.2. Comparison techniques and variable contribution

The GLM forward selection algorithm allowed to select the most optimal set of environmental variables for the GLMs (Table 2.2). The selected set contained only three granulometric variables: median grain size, % mud and % coarse fraction, along with the quadratic terms mud%^{2} and % coarse fraction^2 (all interaction terms were rejected).

Table 2.3. The model parameters of the GLM (logistic regression) model with all data for prediction of the presence of *Lanice conchilega*. The parameter estimates β, the standard error (S.E.) on the estimates and the Wald statistic are provided. An asterisk indicates the Wald statistic was significant.

	β	S.E.	Wald
Median grain size	-0.02	0	14.95*
Mud %	0.26	0.09	9.05*
Mud²	-0.01	0	9.02*
% coarse	0.28	0.08	11.33*
% coarse²	-0.01	0	8.68*
Constant	2.19	1.1	3.97*

The same variables that were selected in the stepwise model selection of the GLM model (median grain size, mud% and % coarse fraction) were used in the ANNs. ANNs with one hidden layer containing three neurons were constructed (Fig. 2.3). The ANN interconnection weights and bias values are provided in Table 2.3. Direct interpretation of these parameters is however not straightforward.

Table 2.4. Parameters of the Artificial Neural Network model, constructed with all data, for the prediction of the probability of presence of *Lanice conchilega*. Each connection in the ANN is described by its start and end. For each connection the interconnection weight is provided, as well as the bias for each neuron.

	Connection start			Bias
Connection end	Median grain size	% mud	% coarse	bias1
neuron1a	1.41	-4.83	-6.08	-3.36
neuron1b	5.28	0.45	1.18	-1.32
neuron1c	2.44	6.43	0.99	-4.74
	neuron1_a	neuron1_b	neuron1_c	bias2
neuron2	-2.88	-2.62	-5.56	1.05

Table 2.5. Model performance indicators for the models developed in each of the three crossvalidation folds, and for the models developed with all data. CCI: Correctly Classified Instances, NPV: Negative Predictive Value, PPV: Positive Predictive Value, spec.: specificity; sens.: sensitivity, AUC: Area Under the Curve. Model performance of the models developed with all data, was calculated on all data as no samples were kept in a separate test set.

	Generalised Linear Models				Artificial Neural Networks			
	Calibration set			All data	Calibration set			All data
	fold 1	fold 2	fold 3		fold 1	fold 2	fold 3	
CCI	78.1	74.5	73.5	78.0	82.1	85.7	82.0	80.6
NPV	0.90	0.90	0.91	0.92	0.88	0.90	0.86	0.89
PPV	0.56	0.51	0.50	0.55	0.66	0.74	0.69	0.56
specificity	0.79	0.74	0.71	0.77	0.88	0.91	0.91	0.85
sensitivity	0.77	0.77	0.81	0.81	0.67	0.71	0.58	0.66
Kappa	0.50	0.44	0.43	0.49	0.54	0.63	0.52	0.48
AUC	0.83	0.81	0.83	0.83	0.85	0.85	0.88	0.85
	Test set				Test set			
CCI	74.5	71.4	73.5	/	78.6	72.4	81.6	/
NPV	0.51	0.47	0.50	/	0.85	0.81	0.92	/
PPV	0.91	0.87	0.88	/	0.62	0.50	0.54	/
spec.	0.72	0.72	0.74	/	0.86	0.82	0.85	/
sens.	0.81	0.69	0.73	/	0.59	0.48	0.70	/
Kappa	0.45	0.36	0.41	/	0.46	0.30	0.49	/
AUC	0.82	0.77	0.77	/	0.81	0.77	0.80	/

The modelled response of the final models was relatively similar between the GLM and ANN models (Fig. 2.4). After conversion of the predicted outcomes to a binary coding (using the cut-off for presence values), a number of performance indicators were calculated (Table 2.4). For each performance indicator, a t-test was performed between the results of the GLM and ANN. The values were pairwise compared for each fold. The CCI, specificity and sensitivity of the ANNs were significantly different from their GLM counterpart (paired t-test, 6 df; $p< 0.05$), but this is also related to the different cut-off for presence used for both modelling techniques. The NPV and PPV and the Cohen's Kappa were not significantly higher for the ANNs (NPV and PPV: 6 df, $p>0.05$; Cohen's Kappa: 3 df; $p> 0.05$). The AUC was significantly larger for the ANNs (paired t-test; 3 df; $p<0.05$). The Pearson correlation between the ANN and GLM final model output was high: $r=0.89$ ($p<0.01$).

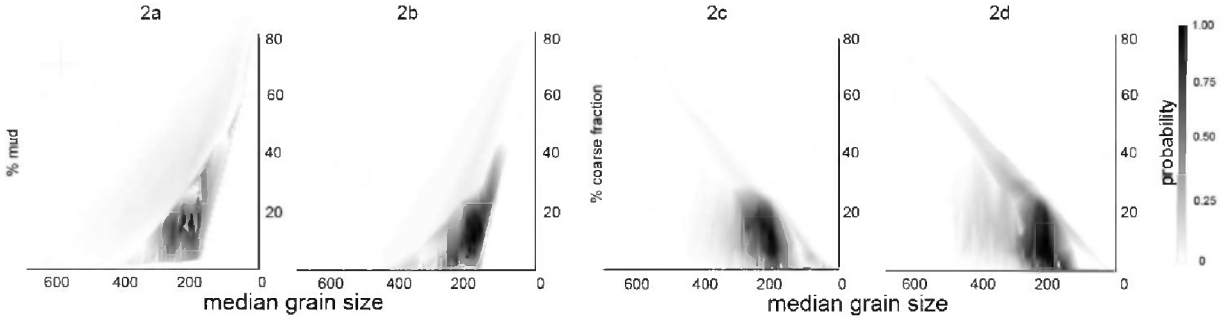

Fig. 2.4. Predicted probability of presence of *L. conchilega* for the final GLM (2.4a and 2.4c) and ANNs (2.4b and 2.4d) calibrated with all samples.

Table 2.6. The relative contribution (%) of the variables in the prediction of the presence of *L. conchilega*. For the GLM the percentages are based on the Wald statistic, for the ANNs on the partial derivatives method. The relative contribution of the variables is provided for the models developed in the threefold crossvalidation. "All data" indicates the relative variable contribution in the final model created with all samples.

		Fold			
	Variable	1	2	3	All data
GLM	Median grain size	27.5	25.5	37.8	26.2
	% mud	20.3	14	10.8	15.9
	% mud^2	16.3	13	14.9	15.8
	% coarse fraction	19.2	23.6	15.9	19.9
	% coarse fraction^2	11.6	17	5.6	15.2
	Constant	5	6.9	15	7
ANN	Median grain size	21.2	5.1	2.6	18.7
	% mud	49.8	6.3	5.9	37.2
	% coarse fraction	29	88.6	91.6	44.1

For the GLMs, the relative contribution of the variables in predicting the presence of *L. conchilega* was similar between the three folds and the final model with all data (Table 2.5). Overall the median grain size was most important in the GLMs, the other two variables and their quadratic terms had a similar contribution to the prediction, the constant was less important (Table 2.5). The ANNs showed a high variability in variable contribution between the folds and the final model. The order of variable contribution in the final model was reversed compared to the GLM: from % coarse fraction over mud% to median grain size.

2.4. Discussion

2.4.1. Selection of environmental variables

PCA allowed only to assess the mutual relations of the variables, but did not allow to distinguish if the variables were important in the prediction of *L. conchilega*. In a later stage, the forward selection algorithm was very important in selecting the final set of appropriate environmental variables. A good selection of environmental variables is crucial to obtain reliable HSMs. The data set contained numerous predictive variables. Using all variables would increase the number of model parameters to be estimated. Also, the inclusion of redundant variables (e.g. all sediment grain size variables) would complicate the assessment of the relative variable contribution, an assessment which is very helpful to derive ecological insights from a model. The relative contribution of two highly redundant variables will be "shared" between them, underestimating their importance.

The stepwise selection used in this chapter has some major drawbacks (see 3.1.2.1.): 1) the selection can get stuck in local optima, 2) is sensitive to multicollinearity and 3) only provides one final model. Therefore the need for a more reliable model selection methodology was identified in this chapter. This need is answered in chapter 3 with the proposed Combined Model Optmisation Criterion.

2.4.2. Modelled habitat preference

Although a data set with various types of environmental variables was available (granulometric, bottom currents, nutrients and chlorophyll), only three granulometric variables were selected during the modelling exercise. The relative order of the contribution of these variables in the GLMs was median grain size, % coarse fraction (sum of relative importance of linear and quadratic term) and % mud (idem % coarse). For the ANNs the same predictive variables were used as for the GLMs, but the order of the variable contribution observed was not consistent between the folds and the final model with all samples. The range per variable where the species is expected to be present (predicted probability >

cut-off for presence) is provided here. The GLM predicts the species to be present between a median grain size between 150 and 325 µm, a mud percentage between 0 and 23, and a coarse fraction between 0 and 30 percent. The ANN predicts the species to be present between a median grain size between 150 and 290 µm, a mud percentage between 0 and 23, and a coarse fraction between 0 and 30 percent. Based on the limited depth range sampled, a use limit is set on the application of the models. The models should only be used within the 0 to -17m depth range.

Previous studies did point out the importance of granulometric variables (Gray, 1981; Snelgrove and Butman, 1994), but Buhr and Winter (1977) indicated that currents also have an important effect on a much smaller scale for *L. conchilega*. In the models presented here, currents were not selected as predictive variable. However this could be due to the spatial resolution of the oceanographic model (Luyten *et al.*, 2003) which was only 250x250 m. Chlorophyll-a was expected to be important in predicting the occurrence of *L. conchilega*, as it is a proxy for local food input. However chlorophyll was not selected for the models, probably because it showed little variance on such a small scale and the fact that only data from the year 2003 were available. The reason that nutrients were not chosen in the models could be due to a seasonal effect. There could thus be a mismatch in the temporal scale of observation: on one hand nutrients that change on short time scale were measured at one point in time and on the other hand species distributions that are relatively constant in time.

The models did not perform equally well throughout the whole range of the variables. This is due to the fact that there were only very little samples in some parts of the variable range. The model should therefore only be used within the variable range of the original data set used to construct the models (Fig. 2.2). The majority of the samples were found in the 150-400 µm range for the median grain size. The mud % of most samples ranged between 0 and 20%, while the % coarse fraction was zero for most samples (Fig. 2.2.). Another important use limit for the model is the depth range of the samples used to develop the model. The depth distribution is bimodal with peaks around -12m and -5m, no samples above the mean low water spring level, so no samples on the beach. This sets the use limits of the model from -17m to 0m water depth. Inclusion of samples at shallower depths in the data set should be considered to broaden the application of the model to the beach.

The maximum values of the CCI are 80-85%. There appears to be a limit on the maximum predictive performance that can be achieved when the species distribution is predicted from the available variables. This could be explained by potentially useful predictive variables which were not measured and the patchy distribution of *L. conchilega* (Heuers *et al.*, 1998). This distribution, due to biological interactions and recruitment fluctuations, introduces noise in the data set: a percentage of the samples in suitable environments would have no *L. conchilega.*

2.4.3. Generalised linear models vs. artificial neural networks

Although there was a very strong correlation between the model outputs of both modelling techniques, ANNs significantly outperformed GLM for a number of performance indicators (i.e. CCI, specificity and sensitivity). Also the AUC was significantly higher for the ANNs, indicating that the difference in performance is not due to the difference in the cut-off for presence of the species between both techniques. For the GLM each quadratic term and each interaction term had to be explicitly presented to the selection algorithm. The superior performance of ANNs could be explained by the fact that non-linear functions and variable interaction are inherent to the architecture of the ANN, because of the connections between the neurons. The higher number of model parameters of the ANN (16 parameters vs. 6 in the GLM) allowed to fit the species-environment relation more precisely, but at the same time made the ANNs less parsimonious in comparison with the GLM counterpart. The interpretation of the numerous ANN parameters was not straightforward and also in the literature ANNs are often called "black-box" models (Olden and Jackson, 2002b).

The effect of only two variables could be visualised simultaneously (Fig. 2.4), the graphs were only a simplification of the model predictions, which are in a multivariate space. However, performance and the relative variable contribution showed a higher dissimilarity between the folds for the ANNs. This could be due to the higher internal complexity of the ANNs or a high dependence on the initial conditions during the training of the network. From a parsimonious point of view the final GLM was superior, as the number of model parameters was lower in comparison with the ANNs, while the predictive performance was similar to the ANNs.

2.5. Conclusions

***Which environmental variables in the data set determine best the spatial distribution of* L. conchilega?**

Several types of environmental variables were in the data set (granulometric, currents, chlorophyll a concentration and nutrient concentrations), but only granulometric variables were selected as predictive variables. In a stepwise variable selection the variables median grain size, mud% and coarse fraction % were selected as the variables that were best suited to predict the spatial distribution of *Lanice conchilega*.

What is the most optimal model to describe the habitat preferences of* L. conchilega*?

The most optimal GLM to predict the spatial distribution of *Lanice conchilega:*

$$logit(probability\ of\ presence) = median\ grain\ size + mud\% + mud\%^2 + coarse\ fraction + coarse\ fraction^2 + error$$

The most optimal ANN used the environmental variables median grain size, mud% and coarse fraction %. There were 3 interneurons on a single hidden layer and all these neurons used the sigmoid transfer function.

Which modelling technique was most optimal to model the distribution of* L. conchilega*?

The ANNs had a slightly higher predictive performance compared to the GLM. But the ANNs had a higher variability in the predictive performance in the crossvalidation. When model parsimony is considered important, GLMs were superior, as the models were simpler and the predictive performance was only slightly lower than the ANNs.

Acknowledgements

The data was obtained from the project "Intensive evaluation of the evolution of a protected benthic habitat (HABITAT) funded by OSTC (project number: MN/02/89) and AWK (Coastal Waterways Division; file numbers: 99380 and 200.455). Chlorophyll data were obtained from the REVAMP-project: (Peters *et al.*, 2005).

Chapter 3. Improved model selection in habitat suitability modelling

Abstract

Habitat suitability models (HSMs) are being used increasingly in a conservation context to model the spatial distribution of species. Based on the relation of the species with the environment, the distribution of the species can be predicted where only physical habitat variables are available.

Model selection is necessary to find the most optimal variable combination to predict species distributions. In this research, a new model selection approach was proposed based on the general Model Optimisation Criterion (MOC) framework which incorporates both model fit and model complexity. The MOC of the model calibration and test set were combined, and in this way the generalisation ability of the models on an independent data set was incorporated in the model selection. Good generalisation was necessary to use the model in other regions or periods.

All variable combinations were tested in an exhaustive model selection to find the most optimal model. Bootstrapping was used to create replicas of the calibration and test set for each variable combination. This resampling produced model replicas per variable combination, which increased the reliability of the model performance estimate. During the resampling, the prevalence of the species was always kept at 50% to avoid influences on the parameter estimates. The model selection performance was also compared with the widely used Cohen's Kappa, Normalised Mutual Information (NMI) and the Area Under the Curve (AUC). The CMOC model selection, which combines the MOCs of the calibration set and the test or validation set, can be applied to any modelling technique for which a likelihood can be calculated.

The proposed approach was first applied to a virtual species to test its performance and later to field observations of *Abra alba*, a marine bivalve. Overall the proposed model selection approach managed to find the true models that were used to generate the virtual species presences. When applied to real species observations, the selection approach allowed choosing the most optimal model, by making a trade-off between model complexity and model fit, while at the same time testing the model generalisation ability.

Key words:

Habitat suitability modelling, species distribution modelling, model selection, variable selection, model optimisation criteria, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), *Abra alba*

3.1. Introduction

3.1.1. Habitat Suitability Models

Because natural habitats are globally under threat, conservation management needs to take measures to halt the deterioration of pristine areas and the loss of biodiversity. Such management decisions demand for a good knowledge on the ecosystems and habitats, in particular of the spatial distribution of plants and animals. To observe this distribution, sampling schemes have been set up to collect species distribution data at sampling locations. When these observations are plotted onto maps, they mostly provide only point-based estimates of the species distributions. To attain the level of full cover species distribution maps, researchers have been using Habitat Suitability Models (HSMs; Guisan and Zimmermann, 2000; Austin, 2007; Hirzel and Le Lay, 2008). HSMs can be used to create full cover distribution maps based on widely available maps of physical habitat variables. HSMs can produce predictions of the species distribution at locations where no species observations are available (Guisan and Zimmermann, 2000).

HSMs allow to analyse and predict the distribution of species based on the local suitability of the habitat for the species. The general assumption is that more suitable habitats have a higher probability of species presence and will support higher species densities (Barry and Elith, 2006). Other assumptions of HSMs include that the species distribution is in equilibrium with the environment (Guisan and Thuiller, 2005; Václavík and Meentemeyer, 2009). Besides the development of species distribution maps, more sophisticated HSM applications exist, e.g. simulation of the ecological effect of management scenarios, guidance of new sampling efforts (Raxworthy *et al.*, 2003) or the prediction and risk assessment of species invasions (Le Pape *et al.*, 2004; see Chapter 1.3.). To model the species-environment response in HSMs, several modelling techniques have been used. Most commonly Generalised Linear Models (GLMs) and General Additive Models (GAMs) have been used. Segurado and Araujo (2004) and Elith *et al.* (2006) provide comprehensive overviews of modelling techniques used in habitat suitability modelling. In this research, a model selection methodology for GLMs is proposed, but the methodology is generally applicable.

3.1.2. Finding optimal models: model selection

HSMs use a combination of environmental variables to predict the spatial distribution of suitable habitats for a species. These variables are assumed to determine the distribution of species (Barry and Elith, 2006). Therefore, the selection of the most optimal combination of predictive environmental variables is a central step in the HSM development (Guisan and Zimmermann, 2000; Heikkinen *et al.*, 2006). Model selection methods aim at determining the most optimal model by adequately constraining the number of predictive variables used, and thus the model complexity (Reineking and Schroder, 2006). Model selection is necessary to select the most parsimonious models and avoid models with highly correlated

predictive variables. Including variables which share information, on the distribution of the species, will lead to multicollinearity which causes inflated variances of the model parameter estimates and hence too large confidence intervals (Miller, 2002).

Model complexity, the number of variables in a model, is often considered relative to the complexity of an unknown true model (Anderson *et al.*, 1998). Models with fewer parameters are said to underfit, whilst models with more parameters than the true model are said to overfit. The problem of model under- and overfitting is avoided when only the necessary variables to predict the species are in the model. If all available variables would be included, the complexity of the model and thus the number of model parameters would raise quickly.

For biological data based on field observations of species, the concept of a true model which completely captures the species-environment relations seems inappropriate, and the model complexity of the most optimal model is expected to depend on the sample size (Anderson *et al.*, 1998) as well as the complexity of the modelled ecosystem. This is because smaller, more subtle effects that are modelled by including more environmental variables or higher order terms, can often only be revealed as the sample size increases (Anderson *et al.*, 1998; Burnham and Anderson, 2004). In the context of biological field data, there is thus no valid concept of model underfitting (Anderson *et al.*, 1998), as the true model is infinitely complex, and the chosen model will always be too simple by definition. Thus for biological observations, the concept of a single true model does not hold, as the complexity of such true model is fixed by definition, and thus cannot change as a function of the sample size *n* (Burnham and Anderson, 2004). When no true model is assumed to exist as with biological field observations, under- and overfitting of a specific model are relative to the complexity of the parsimonious model, which is named hereafter the optimal model (Anderson *et al.*, 1998).

3.1.2.1. Model selection methodologies

To perform model selection, a methodology is needed to generate alternative variable combinations and compare the models with these combinations. If the number of variables is low, it is feasible to manually develop and compare HSMs with alternative variable combinations. For a higher number of variables, two alternative methodologies are used: stepwise model selection and the exhaustive model selection algorithm.

Stepwise model selection

Stepwise model selection is an automatic procedure often used in regression-type models, where predictive variables are sequentially added to a model (i.e. forward selection) or removed from a model (i.e. backward selection) until an optimal model has been found. The stepwise method can be

understood as a steepest descent optimisation method (Reineking and Schroder, 2006). At each step, a neighbourhood of models around the current candidate model is investigated, where this neighbourhood consists of all models that differ by one variable (Reineking and Schroder, 2006). Mostly an F-test is used to compare models that differ by one variable. If the significance of the F-test is below a preset value (e.g. α = 0.05), the variable is kept in the model and the stepwise selection continues. Other procedures exist that have also a stepwise nature. Genetic algorithms, a technique sometimes used in machine learning, use a population of chromosomes to represent alternative models (D'heygere *et al.*, 2003). On the chromosome the combination of variables included in the model is coded binary (e.g. variable1: included; variable2: not included). The chromosomes evolve and in each step, called a generation, there is evolution due to selection and mutation. The selection is done in each step, the fitness of the alternative models is evaluated, less fit models are removed and mutation cause new alternative models to appear. As such a set of models are evaluated in each step, but all possible models are not compared.

Stepwise model selection methods are usually fast and are used very often (e.g. Maes *et al.*, 2004), however the stepwise procedure has three major shortcomings. A first drawback is that it is prone to get stuck in local optima and hence might miss the overall most optimal model (Reineking and Schroder, 2006). This is due to the stepwise nature of the procedure that can only compare models which differ by one variable only. The choice of a forward or backward procedure is arbitrary (Anderson *et al.*, 1998), as researchers do not know where the true model or optimal model is in the sequence of competing models.

A second drawback of stepwise approaches is their sensitivity to multicollinearity. Variables are thus sometimes omitted from the model based on spurious correlations with other variables, rather than for ecological reasons (Hirzel and Guisan, 2002). If the model selection procedure is repeated with a different order of variables, one of the two correlated variables will be in the model alternately, which leads to instability in the model selection (Prost *et al.*, 2008).

A last drawback is the fact that stepwise approaches such as backward selection and forward selection, only provide one final model, while an exhaustive model selection approach provides a quantitative value for the model fit and model complexity for each possible model, which is useful for multimodel predictions. The use of one single model is not always realistic, especially not for biological observations, where the true model is never known.

Exhaustive model selection

An alternative to stepwise selection is the exhaustive or "all subsets" model selection approach where all possible variable combinations are considered to produce a series of models (Hosmer *et al.*, 1989).

Exhaustive model selection is time consuming, but will find the most optimal variable combination in a given set of environmental variables. In this research, exhaustive model selection will be used because it guarantees to find the most optimal model and it provides model fit values for each variable combination. These model fit values, together with the model complexity, can be used to calculate Model Optimisation Criteria (see 3.2.).

Because the number of possible models rises quickly with the number of variables in an exhaustive approach, it is necessary to first select a limited set of variables from the data set. This first selection should ideally be done by ecological experts of the species or by using the habitat preference of related species. With this limited set, an exhaustive approach can then find the most optimal variable combination.

3.1.2.2. Model validation and generalisation

In the current model selection practice, mostly one data set is used to select one optimal model. This same data set is thus used to calibrate the different models and later to test the fit of these models. Sometimes a second, independent data set is used to validate the predictive performance of the selected model for data that have not been used to calibrate the model. In the current practice, the selected model is only applied to the independent set after the model selection has taken place, and thus only the performance of the single selected model with the independent data is assessed. The predictive performance of the model on the independent data is thus not considered in the model selection when forward or backward selection approaches are used. The incorporation of the model performance on independent data in the model selection process would greatly improve the current model selection methodology. It would be possible to select models that have a good trade-off between the model fit for the calibration set and for the test set.

The test set can contain truly independent data or the test and calibration set can be part of the same data set. If independent data are used, these can be truly spatiotemporally independent, thus from another period and/or region. When only one data set is available, this data set is split in a model calibration data set to calibrate the models, and a test set used only to test the models. The splitting of the data set can be: 1) a one-time split, 2) a k-fold split (k-fold crossvalidation) or 3) a bootstrap resampling which generates calibration and test sets iteratively.

The test set obtained by splitting one data set is only pseudo-independent from the calibration set. Validation with pseudo-independent data is called internal validation, while validation with truly independent data is called external validation. External validation can assess whether a model can generalise well on unseen data and can thus more reliably be transferred to other regions or periods.

For full transferability, Randin *et al.* (2006) require that: 1) the internal validation of models fitted in region 1 and 2 must be similar; 2) a model fitted in region 1 must at least retain a comparable external validation when applied in region 2, and *vice versa*; and that 3) internal and external spatial predictions have to match in both regions.

3.1.3. Aims

The general aim of this paper is the improvement of the model selection methodology for HSMs. Improvements to the common practice of model selection are proposed that should lead to the selection of globally optimal models.

The most important shortcomings identified in the current model selection are: 1) Independent data are not optimally used in the current model selection procedure, only post hoc validation of a single chosen model; 2) the stepwise approach can miss the globally optimal model, is sensitive to multicollinearity and does not allow multimodel inference; 3) the data set is used once without replication and; 4) the prevalence of the species in the calibration set greatly influences the model selection. The Combined Model Optimisation Criterion (CMOC) approach proposed in this chapter aims to improve on these shortcomings. The proposed methodology will first be tested on artificial data of a virtual species (Hirzel *et al.*, 2001), because this provides full control over the data set. In a second step, the model selection methodology will be applied to field observations of *Abra alba*, a marine bivalve species.

3.2. Material and methods

3.2.1. Modelling technique: logistic regression

Logistic regression (LR; Agresti, 2002, Appendix II) will be used as modelling technique in this research to model the species-environment relations in the HSMs. LRs are a type of generalised linear models (GLMs) which have the flexibility to choose a link function between the random and systematic component, and the possibility to assume different distribution functions of the response variable. This link function allows to have predictions within the range of the observed responses and to use a linear combination of predictors (Guisan and Zimmermann, 2000). LRs are used to model the relation between the binomial (present or absent) species occurrence observations and a set of predictive environmental variables. Because the model output of the LR should be constrained to the interval [0 – 1], the logistic link function is applied in the LR (see Appendix II). The model parameters $\hat{\theta}$ in a LR are estimated by

means of maximum likelihood estimation (Agresti, 2002), which maximises the likelihood of the estimated parameters $L(\hat{\theta})$ for the given data set.

LR is chosen because this technique is often used in HSM (e.g. Ysebaert *et al.*, 2002; Le Pape *et al.*, 2003), is relatively simple and well established statistically. LR has a transparent model structure and a comparatively low number of parameters (Reineking and Schroder, 2006). The conclusions and the proposed methodology can be applied to other HSM modelling techniques.

3.2.2. Model Optimisation Criteria (MOCs)

For the purpose of model selection it is necessary to have an objective measure to quantify the trade-off between the fit of the model to the observations and the complexity of the model, in order to select the most optimal model. A Model Optimisation Criterion (MOC) allows ranking models from most optimal to worst model. The general MOC framework meets the requirements of a good model performance assessment index laid out by Burnham and Anderson (2004): 1) the MOC is established from the data for each fitted model, 2) it fits into a general statistical inference framework (maximum likelihood is used), 3) the MOC reduces to a number for each fitted model.

A general MOC consists of three parts: 1) a measure of the model fit to the observations, 2) a measure of the model complexity p and 3) a regularisation parameter λ (Reineking *et al.*, 2006) (Equation 3.1). The goodness of the model fit is quantified in the MOC as the natural logarithm of the likelihood of a model with the parameter vector $\hat{\theta}$ (Equation 3.2). The model complexity term equals the number of model parameters p. The regularisation parameter λ determines the relative weight of the model complexity p in the MOC formula (Equation 3.1; Reineking and Schroder, 2006). The regularisation parameter λ times the model complexity p equals the penalisation term of the MOC (Reineking and Schroder, 2006). The lower the MOC value, the more optimal a model is. A more complex model will thus probably have a better model fit, but the elevated number of model parameters will also results in a higher penalty term when the MOC is calculated. The smaller the relative weight of the penalty term, the more complicated the selected models, given that the model fit remains the same (Shono, 2005). As such, the model with the lowest MOC, will be a trade-off between maximal model fit (correct predictions) and minimal model complexity (number of model parameters). A model with minimal MOC has maximal parsimony and is assumed to have little or no over- or underfitting of the species environment relations.

$$MOC = -goodness\text{-}of\text{-}fit + \lambda \cdot model\ complexity \quad (3.1)$$

$$MOC = -2\ln(L(\hat{\theta})) + \lambda \cdot p \quad (3.2)$$

In this research, five MOCs will be used that fit in the framework of the general MOC (Equation 3.1). Historically, the first MOC proposed, is the Akaike Information Criterion (AIC, Table 3.1; Akaike, 1973; Akaike, 1974). Akaike found a formal relationship between Kulback-Leiber information (a dominant paradigm in information theory) and the likelihood theory (the dominant paradigm in statistics; Burnham and Anderson, 2004). This finding made it possible to combine the model parameter estimation (i.e. maximum likelihood estimation) and model selection under a unified model optimisation framework (Burnham and Anderson, 2004). The goodness-of-fit is estimated by the expected, relative Kulback-Leiber information which is measured as the maximised log-likelihood function *L* of the estimated set of model parameters $\hat{\theta}$ (Shono, 2005).

The Kulback-Leiber information can be understood as a distance between a given model and the true model (Anderson *et al.*, 1998; Burnham and Anderson, 2004). The model with the lowest AIC looses the least information in comparison with the true model. There is a penalisation term in the AIC, $\lambda \cdot p$, which is the number of parameters in the model *p* times the regularisation parameter λ, a constant which equals 2 for the AIC. Hurvich and Tsai (1989) adapted the original AIC formula by adding an additional penalty term that is a function of the sample size *n* (AICc, Table 3.1). This second order correction is an improvement of the AIC for small samples size, where the AIC would otherwise select models that are too complex (Anderson *et al.*, 1998). With increasing sample size, the AIC and AICc become asymptotically identical.

Other MOCs are based on the Bayesian theorem. The Bayesian Information Criterion (BIC) is a criterion derived from a Bayesian context with equal priors for each competing model and uniform priors for its parameters (Shono, 2005). In the BIC formula (Table 3.1), the regularisation parameter λ equals the logarithm of the number of samples *n*. The Consistent Akaike Information Criterion (CAIC; Table 3.1; White, 1998) has a slightly different regularisation parameter $\lambda = ln(n + 1)$. So the CAIC penalty term for the model complexity will be slightly larger, and the CAIC will select slightly simpler models than the BIC (Shono, 2005).

Although the BIC and CAIC formulae resemble the AIC formula, the model selection concept differs. The BIC and CAIC focus on the selection of the true subset of variables (Shono, 2005). If the true model is among the candidate models, the probability of selecting this model with the BIC approaches one as the sample size grows infinitely large(Hastie *et al.*, 2001). The BIC and CAIC thus are said to be consistent MOCs that theoretically select the true model in case the number of samples *n* reaches infinity. This is because high sample numbers lead to a very high penalty for the inclusion of

extra variables as λ is depending on the number of samples n. Only the true model has a log-likelihood of zero, because it fits the data perfectly and will thus have a minimal MOC value when compared to all other models. Consistency is desirable when MOCs are applied to a large data set (Shono, 2005). Consistent criteria provide an asymptotically unbiased estimate of complexity of the true model (Anderson *et al.*, 1998). Neither the AIC, AICc nor the F-statistic (see further) are consistent MOCs (Reineking and Schroder, 2006), because more complex models are selected as the number of samples n grows.

The F-statistic, the fifth MOC that will be considered, is the most commonly used test statistic in model selection. This statistic is calculated in the F-test, a test that compares models (mostly during stepwise model selection), that differ by one variable only. The F-statistic is obtained by taking the ratio of the log-likelihoods of the two nested models. This likelihood ratio is asymptotically χ^2 distributed with one degree of freedom, under the null hypothesis that the estimated value of the model parameter by which the models differ, is zero (Reineking and Schroder, 2006).

Table 3.1. Equations of the five model optimisation criteria used in this chapter. For each criterion the regularisation term λ and the significance α of each variable in the model are provided, as well as the fact whether a criterion is consistent (will select the true model if the number of samples n goes to infinity). AIC = Akaike Information Criterion; AICc = AIC with small sample correction; CAIC = Consistent Akaike Criterion; BIC = Bayesian Information Criterion. *f(λ)*: the significance α is a function of λ. *L( )* = likelihood. n = nr. of samples, p = nr. of model variables, $\hat{\theta}$ = vector estimated model parameters.

Criterion	λ	α	Consistent	Formula
AIC	2	0.157	no	$AIC = -2\ln(L(\hat{\theta})) + 2p$
AICc	$2+\frac{2(p+1)}{n-p-1}$	$f(\lambda)$	no	$AICc = -2\ln(L(\hat{\theta})) + (2+\frac{2(p+1)}{n-p-1})p$
CAIC	$ln(n)+1$	$f(\lambda)$	yes	$CAIC = -2\ln(L(\hat{\theta})) + (\ln(n)+1)p$
BIC	$ln(n)$	$f(\lambda)$	yes	$BIC = -2\ln(L(\hat{\theta})) + \ln(n)p$
F-statistic	3.841	0.05	no	$F-statistic = -2\ln(L(\hat{\theta})) + 3.841 \cdot p$

Reineking *et al.* (2006) explain how the F-statistic fits in the general MOC framework as it also calculates a model goodness-of-fit and has a penalty term. Via the χ^2 distribution of the log-likelihood ratio, the λ value corresponding to a certain significance level α in the F-test can be calculated (Reineking and Schroder, 2006). As such, the most commonly used significance level $\alpha = 0.05$ for the F-test, corresponds with a regularisation term $\lambda = 3.841$. In the general MOC framework, the addition of one extra model parameter (because an extra variable is added) will thus increase the MOC with an extra penalty of 3.841 ($\lambda \cdot 1$ extra parameter), which is compensated by a decrease in the model fit term $-2\ln\left(L(\hat{\theta})\right)$ of 3.841 or more, in case this variable is significant at $\alpha = 0.05$ (Reineking and Schroder, 2006). If the model fit term decrease is less than 3.841, the model with one variable extra will have a higher MOC value than the model with one variable less, and the simpler model will be the more optimal one.

Similarly, the significance level α corresponding to the regularisation term λ can be calculated for all other MOCs as shown in Table 3.1. For the widely used AIC, this would mean that a variable added to the model would need to be significant at α of 0.157 in a likelihood ratio test in order to be included in the model. In case of the AICc, BIC and CAIC, the corresponding significance of the variables is dependent on the number of samples *n*, and for the AICc also on the number of predictive variables *p*. The BIC applied to a model with 100 samples, for example, would result in a regularisation value of $\lambda = \ln(100)$, which corresponds with $\alpha = 0.0319$. For BIC and CAIC, a higher number of observations would thus require variables to be more significant to enter the model.

3.2.3. Combined Model Optimisation Criterion (CMOC)

To maximise the generalisation and transferability of the models chosen in the model selection here we propose the Combined Model Optimisation Criterion. The CMOC is maximal for the most optimal model, in contrast with the MOC which is minimal in that case. The CMOC combines the MOCs of the calibration set and a test or validation set. To combine both MOCs, it is necessary to make them both relative. Because the log-likelihood will increase for a given model when the number of samples *n* rises, independent of how good the species-environment relation is fit, MOC values for a model can only be compared when they have been calculated from the data sets with the same size (Turkheimer *et al.*, 2003).

As a first step to calculate the CMOC, the MOC values of the calibration set and the test or validation set are averaged over the 1000 model replica created for each variable combination (Equation 3.3). To make the $\overline{MOC_i}$ value of the model *i,* relative and independent of the number of samples *n*, the Akaike Weight (AW; Wagenmakers and Farrell, 2004) of the $\overline{MOC_i}$ values is calculated.

For each alternative model *i,* first the $\Delta\overline{MOC_i}$ is calculated by subtracting the minimum over all alternative models of the average MOC per model created in the exhaustive approach (Equation 3.4). The AW is then further calculated as a ratio (Equation 3.5). The AW of a model is a value between zero and one, and the sum of the AW over all models is one.

$$\overline{MOC_i} = \frac{1}{1000}\sum_{j=1}^{1000} MOC_{ij} \qquad i = 1,...,k \text{ (k = nr. alternative models)} \quad (3.3)$$

$$\Delta\overline{MOC_i} = \overline{MOC}_i - \min(\overline{MOC}) \quad (3.4)$$

$$AW(\overline{MOC_i}) = \frac{e^{-\frac{\Delta\overline{MOCi}}{2}}}{\sum_1^k e^{-\frac{\Delta\overline{MOCr}}{2}}} \quad (3.5)$$

k = nr models

$$CMOC_i = m\, AW(\overline{MOC}_{i\,calibration}) + (1-m) AW(\overline{MOC}_{i\,test}) \quad (3.6)$$

In the $CMOC_i$, the Akaike weights of the calibration set and the test or validation set MOC_i's for model i are then combined (Equation 3.6). The relative contribution of the calibration set and the test/validation set MOC Akaike weights is determined by the weight *m,* which can be chosen in the interval [0, 1]. In the following part of this chapter, *m* = 0.5, which means both data sets have equal contribution to the CMOC. The choice of *m* will shift the emphasis in the model selection from accurate predictions on the calibration data (*m* = 1), to maximal generalisation and thus transferability of the model (*m* = 0). The CMOC meets the requirements of a good model performance assessment index (Burnham and Anderson, 2004): 1) the log-likelihood, number of samples *n* and model parameters *p* used in the CMOC formula, is dependent on the data set, 2) the CMOC is based on commonly established principles: maximum likelihood, Kulback-Leiber information and parsimony, 3) all the terms in the formula add up to one number, the CMOC, which allows to classify the alternative models.

3.2.4. Application of the CMOC methodology: virtual species

The major difficulty while evaluating new model selection methods with biological field data is the fact that the true model or even the most optimal model is unknown (Austin, 2007). Therefore, the newly proposed CMOC in this research will first be applied in the model selection for a virtual species (Hirzel *et al.*, 2001). Virtual species occurrences are artificially generated based on a true model specified by us, which generates the species-environment responses (Hirzel *et al.*, 2001). Later, the CMOC model selection methodology will be applied to field observations of the marine bivalve species *A. alba*. This will allow testing the methodology on real species observations, which have multiple sources of error and bias. All analyses are performed in R (R core development team, 2009).

In the virtual species approach, the species-environment relations are known exactly because the true model that generates the species occurrence data is specified by the modeller (Reineking and Schroder, 2006). As the true model is known, the assessment of the model selection methodology is straightforward and certain (Hirzel *et al.*, 2001). The ability to manipulate the virtual species, allows to isolate, and thus better understand, problems encountered when dealing with real species (e.g. the relative importance of predictive variables) which will lead to better modelling methodologies (Hirzel *et al.*, 2001). One objection to artificial data generation is that the current theory used to generate species response curves is simple and unrealistic (Austin, 2007). However, Hirzel *et al.* (2001) argue that a model selection methodology which fails to find the true model structure even for virtual species data constructed based on a simple theory is unlikely to find the most optimal model for real field observations.

The three steps in the virtual species approach to assess the model selection abilities of the CMOC are: 1) generation of virtual species data, 2) model development and 3) model evaluation (Fig. 3.1).

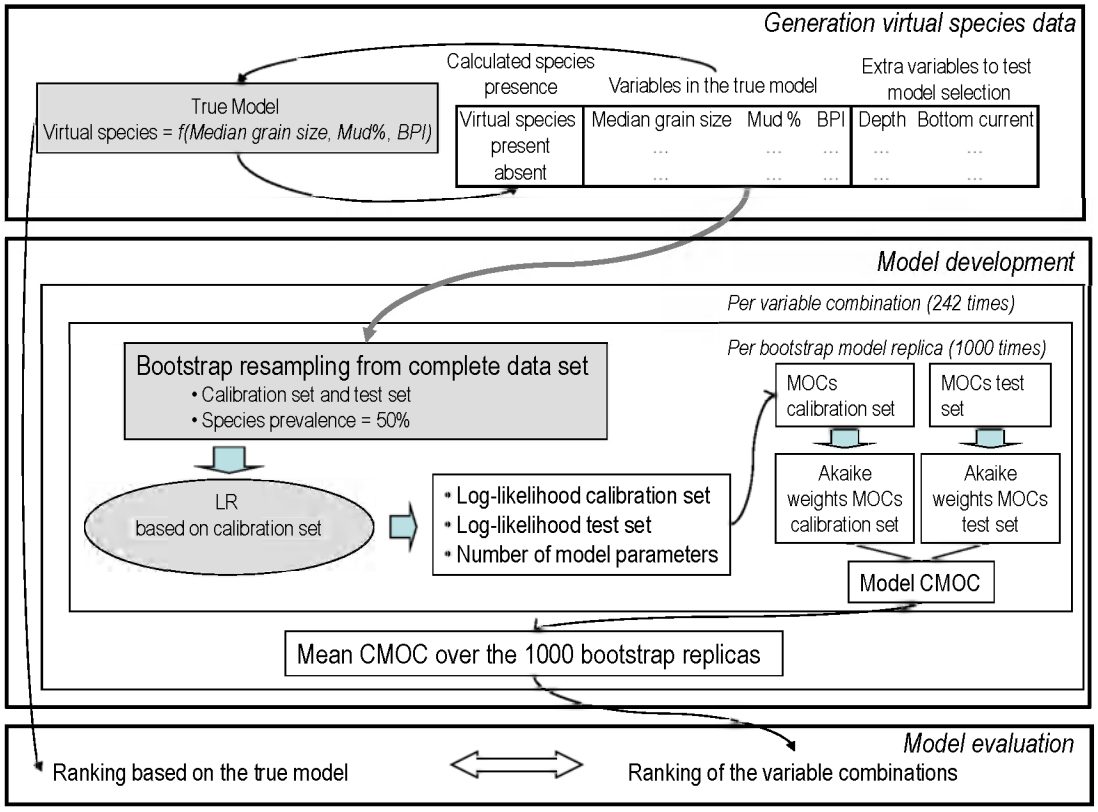

Fig. 3.1. Overview of the steps in the virtual species approach to assess the model selection abilities of the Combined Model Optimisation Criterion (CMOC).

3.2.4.1. Generation of virtual species data

The first step in the model selection methodology for the virtual species is the generation of artificial occurrence data for each sample in a data set based on environmental variables (Fig. 3.1). These environmental variables are taken from a data set with real field observations (Hirzel *et al.*, 2001), collected between 1977 and 2004 on the Belgian Part of the North Sea (BPNS). Five environmental variables are chosen from all available environmental variables in the data set (Table 3.2). Because based on previous modelling exercises (Degraer *et al.*, 2008; Willems *et al.*, 2008), they have been shown to be useful to model the habitat preference of macrobenthic species, and these variables are available at a full cover scale for the BPNS, so full cover distribution maps can be created if needed.

Table. 3.2. Overview of the data sets. BPI = Bathymetric Position Index (Lundblad *et al.*, 2006)

	Training/test set			Validation set		
Nr. samples	786 samples			204 samples		
Year	1977-2001			2002-2003		
	min	mean	max	min	mean	max
A. Alba (dens./m²)	0.0	112.0	7582.8	0.0	10.1	1325.5
Median grain size (µm)	8.0	277.8	658.0	9.8	228.7	490.2
Mud %	0.00	3.12	99.04	0.00	8.69	92.94
Depth (m)	-35.90	-10.03	5.41	-20.43	-4.88	4.35
BPI	-573.0	-8.3	329.0	-229.0	-30.1	200.0
Bottom current (ms^{-1})	0.035	0.517	0.884	0.117	0.493	0.836

The sediment median grain size, the mud% and the depth are actual measurements at the time of sampling, while the Bathymetric Position Index (BPI; Lundblad *et al.*, 2006) is derived from a high resolution bathymetry raster data set (80m pixel size). The depths are standardised to the Mean Low Water Spring level (MLWS), which cause some depths to be > 0 m (Table 3.2.). The BPI indicates whether a sample is on a ridge (BPI > 0), or in a trough (BPI < 0) (Lundblad *et al.*, 2006). The maximum bottom current speed (m/s) is derived from the COHERENS 3D baroclinic model (Luyten *et al.*, 2003). To explore the multicollinearity in the data set, a Kendalls Tau correlation analysis was performed. The environmental variables depth and bottom current have a moderate correlation (r = -0.61; Table 3.3.). Other variables have a low or almost no correlation (Table 3.3.).

Table. 3.3. Correlation coefficients of the environmental variables in the complete data set (Kendall Tau correlation).

	Median grain size	Mud %	Depth (m)	BPI	Bottom current (ms^{-1})
Median grain size		-0.29	-0.24	-0.03	0.31
Mud %	-0.29		-0.16	0.31	0.10
Depth (m)	-0.24	-0.16		-0.29	-0.61
BPI	-0.03	0.31	-0.29		0.24
Bottom current (ms^{-1})	0.31	0.10	-0.61	0.24	

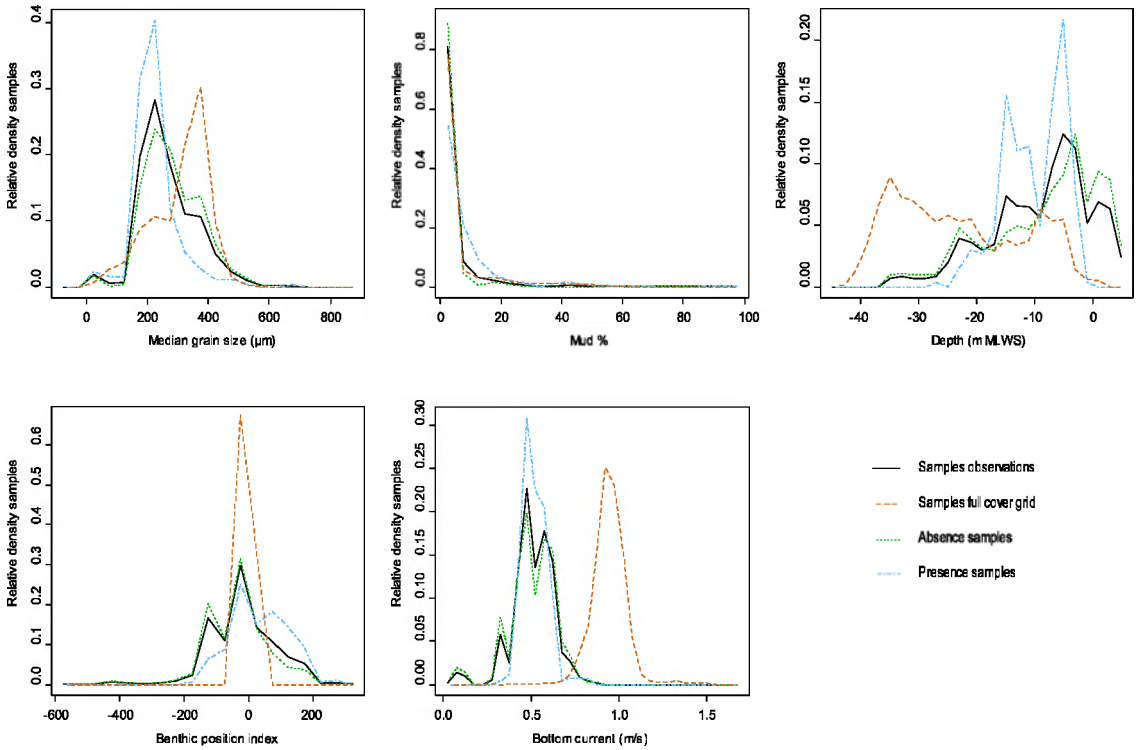

Fig. 3.2. Distribution over the range of the variables in the complete data set used to generate virtual species data and to model the distribution of *Abra alba.*

To obtain artificial species occurrences based on a set of predictive variables, the niche coefficient (Hirzel *et al.*, 2001) is calculated for each sample in the data set. The niche coefficient expresses the probability that the virtual species is present.

In a first step the shape of the species-environment response h_i *is* stated for each environmental variable in the data set (Fig. 3.3). The response shapes chosen are based on visual inspection of real

species-environment relations in the data set. Three types of species-environment responses are used in this research. With the Gaussian response the highest chance of finding the virtual species is at intermediate values of the variable (median grain size, depth and maximum bottom current; Table 3.2). In a negative exponential response curve the chance of finding the virtual species decreases exponentially as the environmental variable increases (mud content; Table 3.2). The simplest response is the linear response, which is used for the BPI (Table 3.2).

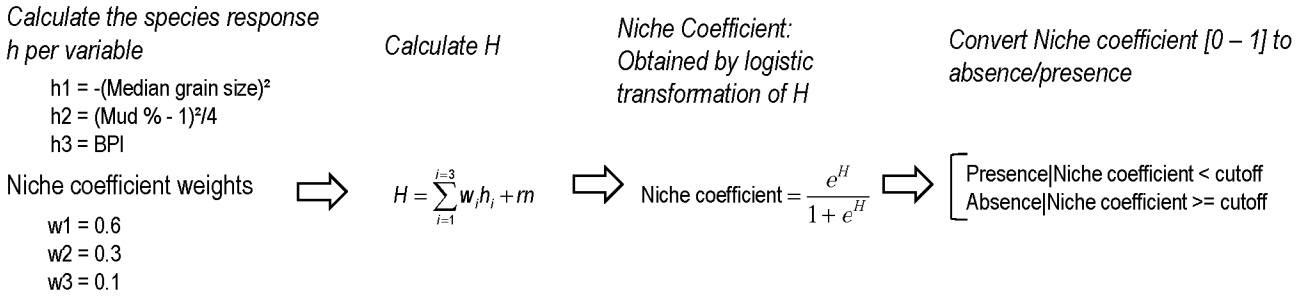

Fig. 3.3. Example of the calculation of the virtual species absence/presence for the true model with three environmental variables (TM3).

The species response *H* is calculated as a weighted sum of the calculated response h_i per environmental variable, times the niche coefficient weight w_i, plus a random noise term *rn* (Fig. 3.3; Hirzel *et al.*, 2001). The niche coefficient weights give the relative importance of variables in the final niche coefficient *H*. In order to use niche coefficient weights w_i that are simple to interpret, the environmental variables are first transformed to the interval [-1, 1] (Reineking and Schroder, 2006). The niche coefficient weights w_i can then be chosen in the interval [0 – 1], with a sum of one for all weights. To account for stochastic effects and random error, a random noise term *rn* (mean = 0, sd = 0.05) is added to the niche coefficient (Fig. 3.3; Hirzel *et al.*, 2001).

The niche coefficient is obtained by a logistic transformation of *H* to the interval [0 – 1]. As a next step, this niche coefficient, is converted to discrete absences/presences [0, 1], with the use of a cut-off for presence. For each of the three true models (see further), the cut-off was chosen as to obtain a prevalence of presence samples of 50% in the resulting data set. As LR models the proportion of presence samples, this prevalence in the calibration set can greatly influence the modelled response (Liu *et al.*, 2005).

Table 3.4. The predictive variables in the data set used to generate virtual species data. For each variable the response type and equation of the niche coefficient response is stated (Hirzel *et al.*, 2001). The niche coefficient weight indicates the relative importance of the variable. In case the niche coefficient weight is not stated, the variable is not included in the model.

Predictive variable	Species response h_i	Response type	Niche coefficient weight w_i TM1	TM3	TM5
Median grain size (µm)	$-x^2$	Gaussian	1.0	0.6	0.2
Mud %	$(x - 1)^2/4$	exponential	-	0.3	0.2
Benthic Position Index	x	linear	-	0.1	0.2
Depth (m MLWS)	$-x^2$	Gaussian	-	-	0.2
Maximum bottom current (m/s)	$-x^2$	Gaussian	-	-	0.2

In this research three alternative true models are created, which correspond with three possible situations in model selection (Table 3.2): the true model is very simple (one variable; TM1), the true model is very complex comprising all variables (five variables; TM5) or intermediate (three variables; TM3). Different niche coefficient weights per variable are chosen for each of the three true models (Table 3.2). TM1 has no weighing as there is only one variable, TM3 uses different weights for the three variables and TM5 has an equal weight for all five variables. During the model selection, the goal will be to retrieve which variables were in the true model for each of the three true models.

3.2.4.2. Model development

The exhaustive or all subsets model selection approach is used in this model selection approach. To find the most optimal model, models are developed with all possible combinations of the five variables in the data set (Table 3.2) and the second order terms of these variables (e.g. $depth^2$). Higher order terms and interaction terms are not considered because the emphasis is on the model selection methodology. A hierarchical model selection approach (Kutner *et al.*, 2005) is followed. Second order terms are thus only allowed if the first order term is already present in the model. This gives a total of 242 variable combinations.

For each of the 242 variable combinations, the bootstrap algorithm samples 1000 times a calibration and test set from the complete data set with 990 samples. The calibration and test sets have an equal size of 495 samples, thus half the size of the complete data set. Each bootstrap calibration set is used to create one model for each variable combination, which results in 242000 models. For each of the three true models (TM1, TM3, TM5) this bootstrap process is repeated. The bootstrapping increases the reliability of the model selection because 1000 replica models are created for each variable combination. The bootstrap resampling is stratified by species presence, to obtain a species prevalence

of 50%. Such a stratified bootstrap resampling is performed by separately sampling an equal number of samples from the presence and the absence samples, and then rejoining the samples in one data set. Liu *et al.* (2005) advice to use a prevalence of 50%, because it is more robust.

After development of all the LRs, the five MOCs (AIC, AICc, CAIC, BIC and the F-statistic) are calculated per true model for the calibration and test set for each of the 240000 models. The average of the MOC over the 1000 replica models per alternative variable combinations is calculated for the calibration and test set (Equation 3.3). The CMOC (Equation 3.6) is calculated as a weighted average of the averaged MOCs of the calibration and test set, after these have been calculated to Akaike weights.

For the sake of comparison with other HSM literature, the Cohen's Kappa and the Normalised Mutual Information (NMI) are provided (Fielding and Bell, 1997). Kappa and NMI are measures of the correct classification of observations that are compensated for the prevalence of the species in the calibration set, although the prevalence is kept at 50% in this research. They are based on contingency tables which require the continuous prediction of the LR [0 – 1], to be converted to discrete absences and presences [0, 1] by using a cut-off for presence. A contingency table is obtained by tabulating the virtual species presence against the predicted species presence. Selection of the cut-off for presence is critical for contingency based model performance indicators (Liu *et al.*, 2005).

Another criterion frequently used in HSM, the Area Under the Curve (AUC; Swets, 1988), is provided as well. The AUC is the surface under the receiver-operator curve, which is constructed by plotting the sensitivity values against 1-specificity for a series of cut-off for presence-values (Swets, 1988). The AUC is 1 for a perfect model and 0.5 for a nonsense model. The AUC statistic appears to be independent of prevalence only in its middle range (Maggini *et al.*, 2006), and thus benefits as well from the prevalence which is kept at 50%.

3.2.4.3. Model evaluation

Comparison CMOC model ranking and true model ranking

To compare the model ranking based on the CMOCs with the true model ranking, a measure is needed to quantify the dissimilarity between the true model and a given model. This dissimilarity is quantified by the number of variables a model is over- and underfitting in comparison with the true model used to generate the species response. Therefore the number of over- and underfitting variables is combined in the Euclidean distance to the true model. This can be conceptualised by plotting for each alternative model the number of overfitting variables versus the number of underfitting variables (Fig. 3.4). In this plot, the true model used to generate the virtual species response, is in the origin (0,0), as this model has no over- or underfitting. Now the Euclidean distance can be calculated from a given model to the

true model in the origin. This Euclidean distance ranking can be compared with the CMOC ranking where more optimal models have a lower CMOC.

	Median grain size	Mud %	BPI	Depth	Max. current	Median grain siz²	Mud %²	BPI²	Depth²	Max. current²	
True model 3	1	0	1	0	1	0	0	1	0	0	4 variables
Model *i*	0	0	1	0	0	0	1	1	0	0	3 variables
Underfitting	1	0	0	0	1	0	0	0	0	0	underfitting 2 variables
Overfitting	0	0	0	0	0	0	1	0	0	0	overfitting 1 variable

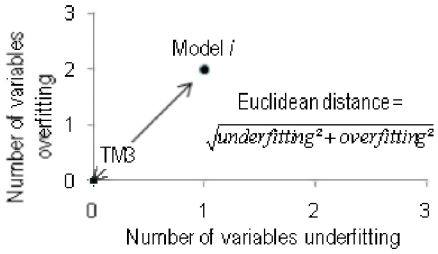

Fig. 3.4. Example of the calculation of the Euclidean distance from a model *i* to the true model TM3.

Are models most similar to the true model selected?

An alternative assessment of the model selection abilities of the proposed methodology is to determine whether the models that differ from the least from the true model have the highest CMOCs. Therefore, the sum of the CMOC is calculated for the models that have an Euclidean distance <2 from the true model (Fig. 3.5). If this sum approaches one than the models most similar to the true model, have the highest CMOC and thus are selected as the most optimal models.

How often do optimal models contain a given variable?

To assess how important a variable is in the prediction of the species presence, it is necessary to study the CMOC of the models that include this variable. Variables that are found in a more optimal model with a higher CMOC will have a higher relative contribution in a multimodel inference, and thus the summed CMOCs will be closer to one. The relative contribution of the variable *a* is thus calculated as the sum of the CMOC of the models that contain variable *a*:

$$\sum_{i=1}^{242} (CMOC_i \mid \text{variable a} \in \text{model i})$$

Multimodel weighted average of the number of variables

The average number of variables in the models for each CMOC is a means to determine if a CMOC is rather preferring complex or more simple models. To obtain a multimodel weighted average of the number of variables, for each model the number of variables is multiplied by the CMOC of this model, and these values are summed over all the models:

$$\sum_{i=1}^{242} (CMOC_i \cdot p)$$

The average number of variables can be plotted against the regularisation parameter λ of each of the five MOC used to calculate the CMOCs (Fig. 3.7.). As such, the impact of the choice of one of the five MOCs with its specific regularisation parameter λ on the complexity of the selected models can be assessed.

3.2.5. The species *Abra alba*

The model selection methodology is applied to *Abra alba,* a deep burrowing, small bivalve. *Abra alba* is a characteristic inhabitant of shallow inshore muddy fine sand or mud (Van Hoey *et al.*, 2005). Van Hoey *et al.* (2004) describe the complete sampling methodology. This species is an indicator for the *A. alba* macrobenthic community in the Southern North Sea (Van Hoey *et al.*, 2004).

The data set containing the species observations is the same as used for the virtual species, but now also the field observations of a real species are used. To have a second data set for model validation that is temporally independent, the complete data set is split based on the years. The pre-2001 data set contains 786 samples and the post-2001 data set contains 204 samples. The pre-2001 data set will be resampled during the bootstrap resampling to generate calibration sets, while the post-2001 data set is resampled to generate validation sets. As such an external validation of the model on unseen data from a different period is performed to assess the generalisation and transferability of the model. The number of samples is 79 in the calibration set and the prevalence of the species is kept at 50%.

The proposed model selection methodology based on the CMOC is applied to real species observations. Compared to the steps in the virtual species model selection, only the second step in the scheme (Fig. 3.3), model development, is relevant for real species. Field observations are used to model the species-environment response in an exhaustive model selection, where 1000 replica models are created for each variable combination. Evaluation against a true model is not possible as the true model is unknown for the species. Therefore only a ranking of the models based on the CMOC is obtained. Similarly as with the virtual species, an assessment of how often variables are included in the most optimal models is performed (Fig. 3.8). Also, a weighted average of the number of variables versus the regularisation term λ is plotted (Fig. 3.9).

3.3. Results

3.3.1. Virtual species

3.3.1.1. Comparison CMOC model ranking and true model ranking

The properties of the models that were two variables or less different from the true model (Euclidean distance ≤ 2) are shown in Appendix III of this thesis. The models are ordered in the table according to the similarity to the true model, i.e. the Euclidean distance.

TM1

The overall pattern in the CMOC values and the rankings based on these variables, shows a good model selection performance of the CMOC. Models most similar to the true model have high CMOC values. The very simple TM1 with only two variables, is very sensitive to the omission of one of these variables. The models that are underfitting because they lack the variable median grain size², have a very low CMOC, and have ranks between 120-160 out of 242 models based on the CMOC (see Appendix III). The Kappa, NMI and AUC are not low for these models though. CMOC based on the CAIC and BIC criteria had a different spread compared to the other three criteria. Models most similar to the true model had very high CAIC and BIC CMOCs, while the other CMOCs values (based on AIC, AICc and F-statistic) have more spread in the distribution over all the 242 models.

TM3

The spread of the CMOC is high for the AIC, AICc and the F-statistic, with lots of models receiving a low CMOC while the BIC and CAIC are more selective and give very high CMOCs to only a few models. The most optimal model according to all five CMOC is lacking the variable mud² in comparison with TM3. None of the CMOCs thus selects the true model, but they select a model that underfits with one variable. The models lacking mud or median grain size have a CMOC of zero, and also the series of models that lacks the BPI and/or the median grain size² also have a CMOC of zero, which indicates that these variables cannot be omitted. The ranking based on the Kappa, NMI and AUC did not match with the true model ranking or the CMOC ranking.

TM5

TM5 is a very complex model, so relative to TM5, a model can only overfit by adding the term BPI^2. The model chosen as the most optimal model by all five CMOCs and the Kappa, NMI and AUC, is the model that equals TM5 without the mud^2 term. TM5 itself is ranked as the second most optimal model. For this true model, the CMOCs and the Kappa, NMI and AUC agreed in the selection of the most optimal model.

3.3.1.2. Models most similar to true model selected?

In the selection of the optimal models for the three true models, the BIC and CAIC have a good model selection performance (the summed CMOC approaches one), while the AIC and AICc perform badly at model selection of the simpler TM1 and TM3 (Fig. 3.5). The F-test has an intermediate performance in the model selection, and follows the pattern of the AIC. For the most complex true model, TM5, the model selection is near perfect for all the CMOCs and the Kappa, NMI and the AUC, as the sum of the CMOCs of the models most similar to TM5 (Euclidean distance ≤ 2), is almost one (Fig. 3.5).

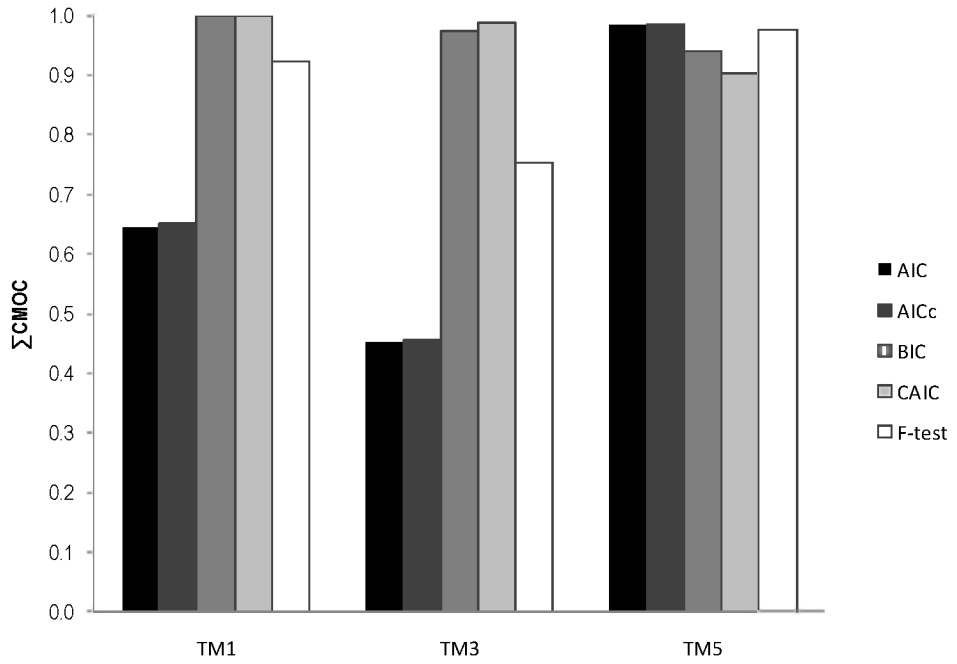

Fig. 3.5. Sum of the Combined Model Optimisation Criteria of the models that differ by two or less variables from the true models TM1, TM3 or TM5 (Euclidean distance ≤ 2).

3.3.1.3. How often do optimal models contain a given variable?

TM1

The model terms Median grain size and Median grain size² have a summed CMOC of almost one for each of the five CMOCs. This means that all the models selected as optimal models, included these variables and thus from a multimodel inference point of view, these variables were strongly selected. Overall the AIC and AICc gave higher summed CMOCs, especially for the calibration set. The CAIC and the BIC selected variables more correctly, and the contribution of variables not in the TM1 is very low to zero.

TM3

The variable Mud² is in TM3, but has very low summed CMOCs (Fig. 3.6) and is thus missed by the model selection approach. The other four variables that are in the TM3, each have a summed CMOC of almost one for all five CMOCs. The variables that are not in TM3, Depth and Currents, still received a moderate (0.6) summed CMOC of the AIC and AICc criteria that clearly select rather complex models.

TM5

Median grain size² has only an intermediate summed CMOC for the CIAC (0.49) and BIC (0.62), although this variable is in TM5. The Mud² has even lower summed CMOC values, and is included as well in TM5. All the CMOCs excluded thus Mud² in most of the optimal models. The BPI² is not in the TM5, but has very similar variable contribution values as the Mud². For the most complex true model TM5, the AIC and AICc perform the best, as these criteria prefer more complex models.

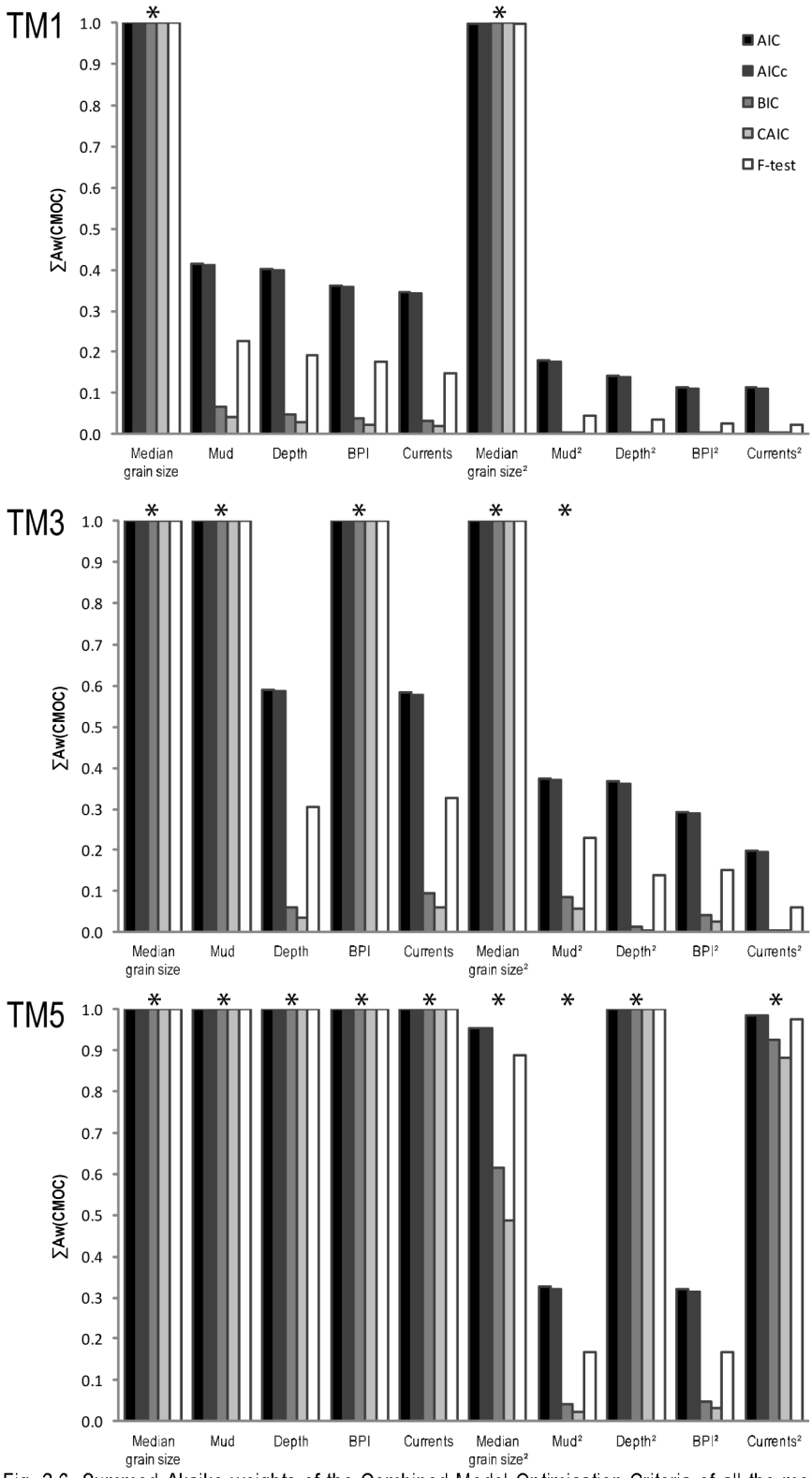

Fig. 3.6. Summed Akaike weights of the Combined Model Optimisation Criteria of all the models that included the specific variable. An asterisk indicates that the variable was included in the true data generating model. TM1, TM3 and TM5: true models for the virtual species. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion. BPI: Bathymetric Position Index.

3.3.1.4. Average number of variables, weighted by the CMOCs

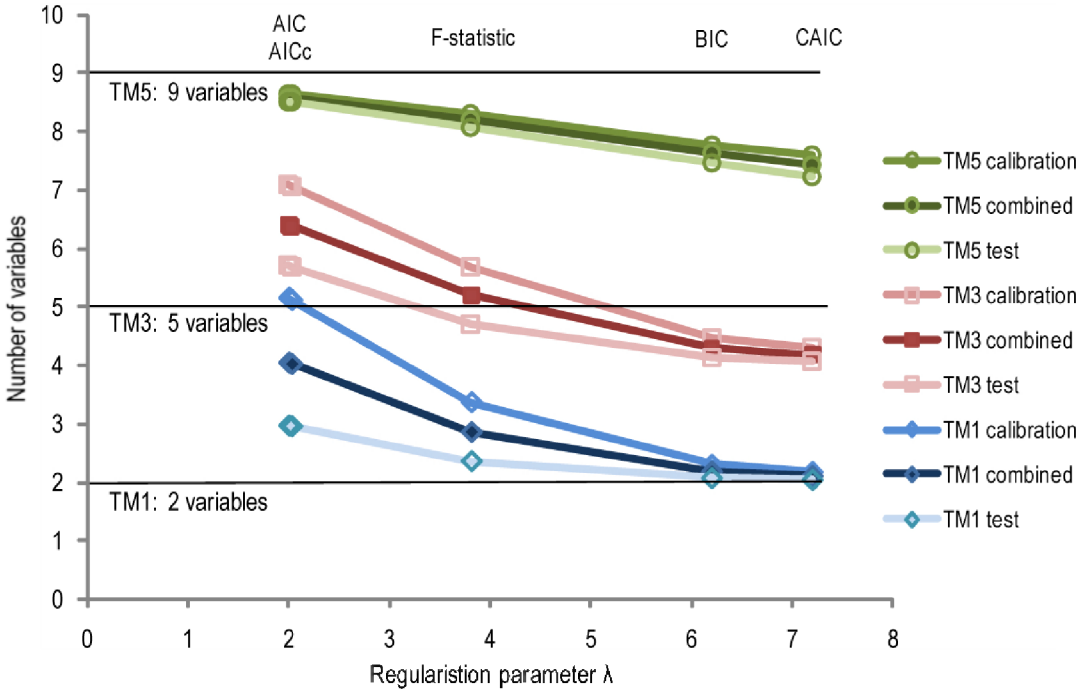

Fig. 3.7. Plot of the weighted average of the number of model variables for each CMOC versus the regularisation parameter λ of the CMOC for each of the three true models TM1, TM3 and TM5. TM1, TM3 and TM5: true models for the virtual species. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion

There is a clear relation between the average number of variables and the regularisation parameter λ of the MOCs. Generally the number of variables decreases with increasing λ (Fig. 3.7). TM and TM3 show a non-linear descent, while TM5 shows a linear descent. For TM1 the number of variables in the selected models converges to the number of variables in the true model TM1 when the regularisation parameter λ is maximal (with the CAIC CMOC). Lower values of λ (e.g. AIC and AICc) lead to models that overfit with two variables relative to TM1. During the model selection to find TM3, high λ values generated underfitting models, while low λ values generated overfitting models. The complexity of the true model (five variables) would have required a λ value of 4.5. For both the TM1 and TM3 model selection, the variance of the number of variables decreases with increasing λ, while the TM5 shows an opposite effect. The selection of a model with a correct number of variables for TM5 (nine variables) would have required a λ of around one, one unit below the λ of the AIC and AICc.

3.3.2. The species *Abra alba*

3.3.2.1. Model selection results *Abra alba*

The model selected as the most optimal model differs for the five different CMOCs. The CMOC based on the F-statistic selects the most complex model with 9 parameters (*A. alba ~ median grain size + mud + depth + bpi + currents + median grain size² + mud² + depth² + ε*). The AIC and AICc select a slightly simpler model with 8 parameters (*A. alba ~ median grain size + mud + depth + bpi + currents + mud² + depth² + ε*). The CIAC and BIC selected the simplest model with 5 parameters (*A. alba ~ median grain size + depth + bpi + depth² + ε*). An alternative to using one single model is multimodel prediction, with the CMOC per model as a weighing factor to determine the contribution of each model in the model prediction (see Chapter 4). The calibration set MOC AWs are more evenly spread over the possible models, while the test set MOCs have high values for the most optimal models, but decrease rapidly thereafter.

Table 3.5. Results of the model selection for the species *Abra alba.* The fifteen most optimal models based on the Akaike weights of the Combined Model Optimisation Criteria (CMOC) are provided. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion. NMI: Normalised Mutual Information. AUC: Area Under the Curve. The total number of model parameters includes the model intercept. The CMOC values in bold indicate that this model was the most optimal based on the specific CMOC. BPI: Bathymetric Position Index.

	CMOCs (expressed as Akaike Weights*10²)															Contingency table based indicators														
	Calibration set					Test set					CMOC					Calibration set			Test set												
Nr. Parameters	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	Kappa	NMI	AUC	Kappa	NMI	AUC	Median grain size	Mud	Depth	BPI	Currents	Median grain size²	Mud²	Depth²	BPI²	Currents²
5	0.0	0.0	1.2	0.3	0.0	1.6	1.9	54.9	41.4	14.8	0.8	0.9	**28.1**	**20.8**	7.4	0.58	0.26	0.79	0.56	0.25	0.78	1	0	1	1	0	0	0	1	0	0
6	0.0	0.0	2.5	0.9	0.1	8.0	9.1	28.7	35.6	30.3	4.0	4.6	15.6	18.3	15.2	0.59	0.28	0.80	0.57	0.25	0.78	1	0	1	1	1	0	0	1	0	0
9	19.9	21.3	14.2	22.9	30.5	9.7	9.1	0.0	0.2	2.3	14.8	15.2	7.1	11.6	**16.4**	0.64	0.33	0.82	0.61	0.30	0.81	1	1	1	1	1	1	1	1	0	0
8	2.4	2.8	17.0	16.6	9.3	36.4	36.9	1.4	4.6	22.0	**19.4**	**19.8**	9.2	10.6	15.6	0.63	0.31	0.82	0.60	0.29	0.80	1	1	1	1	1	0	1	1	0	0
8	2.5	2.9	17.6	17.2	9.7	0.8	0.8	0.0	0.1	0.5	1.7	1.9	8.8	8.7	5.1	0.63	0.32	0.82	0.60	0.29	0.80	1	1	1	1	0	1	1	1	0	0
7	0.2	0.2	11.0	6.5	1.5	0.6	0.6	0.2	0.4	0.9	0.4	0.4	5.6	3.5	1.2	0.62	0.30	0.81	0.60	0.28	0.80	1	1	1	1	0	0	1	1	0	0
7	0.1	0.1	6.9	4.1	1.0	2.3	2.5	0.9	1.7	3.5	1.2	1.3	3.9	2.9	2.2	0.61	0.29	0.81	0.59	0.27	0.79	0	1	1	1	1	0	1	1	0	0
10	27.0	26.7	2.0	5.2	16.5	6.9	6.0	0.0	0.0	0.7	16.9	16.3	1.0	2.6	8.6	0.65	0.34	0.83	0.61	0.30	0.81	1	1	1	1	1	1	1	1	0	1
9	4.1	4.4	2.9	4.7	6.3	6.3	5.9	0.0	0.1	1.5	5.2	5.1	1.5	2.4	3.9	0.64	0.32	0.82	0.60	0.29	0.80	1	1	1	1	1	0	1	1	0	1
8	0.5	0.6	3.5	3.4	1.9	3.9	4.0	0.1	0.5	2.4	2.2	2.3	1.8	2.0	2.2	0.63	0.31	0.81	0.60	0.28	0.80	0	1	1	1	1	0	1	1	0	1
6	0.0	0.0	0.2	0.1	0.0	0.8	0.9	2.9	3.7	3.1	0.4	0.5	1.6	1.9	1.6	0.58	0.27	0.79	0.56	0.25	0.78	1	0	1	1	0	0	0	1	1	0
7	0.0	0.0	0.7	0.4	0.1	4.0	4.3	1.5	3.0	6.0	2.0	2.1	1.1	1.7	3.1	0.60	0.28	0.80	0.58	0.26	0.79	1	0	1	1	1	0	0	1	0	1
10	11.7	11.6	0.9	2.3	7.2	1.9	1.6	0.0	0.0	0.2	6.8	6.6	0.4	1.1	3.7	0.65	0.33	0.82	0.61	0.29	0.80	1	1	1	1	1	1	1	1	1	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.6	2.1	0.3	0.0	0.0	2.3	1.1	0.2	0.56	0.25	0.78	0.55	0.24	0.77	1	0	1	0	0	0	0	1	0	0
8	0.3	0.3	2.1	2.1	1.2	0.2	0.2	0.0	0.0	0.1	0.3	0.3	1.1	1.0	0.6	0.61	0.30	0.81	0.59	0.27	0.79	1	1	1	0	1	1	1	1	0	0

3.3.2.2. How often do optimal models contain a given variable?

The variables *depth* and *depth*² are included in all the models with a high CMOC, followed by BPI and median grain size. AIC and AICc give considerably higher CMOCs for all the variables, because these MOCs prefer more complex models.

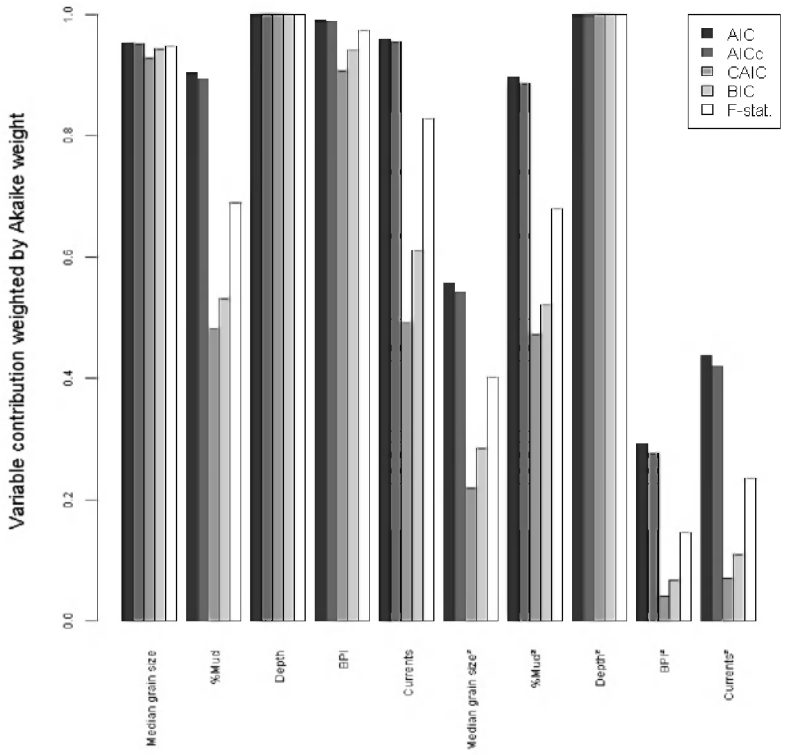

Fig. 3.8. Summed Akaike weights of the Combined Model Optimisation Criteria (CMOCs) of all the models predicting *A. alba* that included the specific variable. The higher the value, the higher the CMOC of the models that included the variable. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion. NMI: Normalised Mutual Information. AUC: Area Under the Curve.

3.3.2.3. Average number of variables, weighted by the CMOCs

The relation between the regularisation parameter λ and the weighted average number of variables is not as straightforward as with the virtual species. Although the λ is maximal for the CAIC, the BIC selects on average the simplest model. There is no monotonously decreasing trend as with the virtual species (Fig. 3.7).

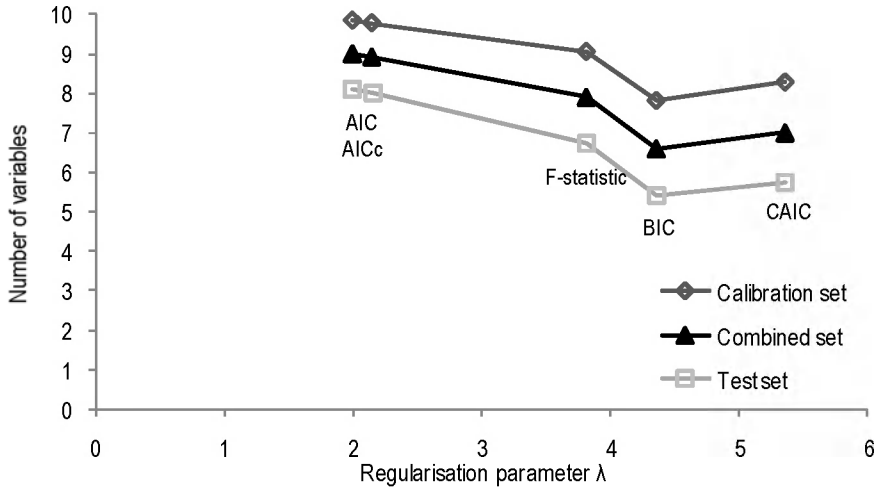

Fig. 3.9. Plot of the average number of variables for each CMOC versus the regularisation parameter λ of the CMOC for each of the models to model the response of *Abra alba*. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion.

3.4. Discussion

3.4.1. Combined Model Optimisation Criterion model selection approach

The CMOC model selection approach has been set up to deal with the shortcomings in the model selection approach which is currently most often used in regression based HSMs: a stepwise model selection based on a single data set and without paying attention to the species prevalence in this data set is used. The shortcomings of the stepwise model selection approach are listed below, together with the improvements that are made by using the CMOC approach.

Independent data not optimally used in the current model selection, only post hoc validation of a single chosen model

The CMOC approach effectively combines the MOC for a calibration set and a test or validation set. This second data set can be completely independent or can be constructed by splitting the original data set. As such the predictive performance of each model for independent or pseudo-independent data is incorporated in the model selection process. This incorporation causes the selected model(s) to have a high generalisation ability and thus transferability to other regions and periods. The weighing parameter

m in the CMOC formula determines the trade-off between choosing an optimal model for the calibration or test set, thus choosing a model with less or more model generalisation ability, respectively.

A stepwise approach can miss the global optimal model, is sensitive to multicollinearity and does not allow multimodel inference

All models have a CMOC value and can be compared, while in the stepwise approach only a one-by-one comparison is done during each step in the model selection, which allows the optimal model to be found in the given set of models.

Multicollinearity is not a problem for CMOC, because all variable combinations are calculated separately. The stepwise approach becomes instable when multicollinearity is present, because one of the two variables correlated is chosen alternatingly (Prost *et al.*, 2008). The data generating model (or a very similar model) was found in the virtual species approach, even if correlated variables were used to generate the virtual species data (e.g. bottom current and depth: $r = -0.61$).

As the CMOC provides a value for each model, it is possible to do a multimodel prediction with the CMOC as a weighing factor for the contribution of each model to the final prediction. The stepwise approach just provides one chosen model. Another advantage is that the CMOC approach uses for each variable combination the maximal number of complete observations, while a backward stepwise selection can only use the observations that are complete for every variable in the data set.

The data set is used once without replication

Real data are only a snapshot of a dynamical situation and can only give a partial and instantaneous observation of the species-environment relations (Hirzel *et al.,* 2001). Bootstrap resampling increases the reliability of the model selection by creating replica calibration and test sets (Araujo and Guisan, 2006; Prost *et al.*, 2008), which are used to create replica models. The model selection is thus based on numerous model replicas, and not on one model calibrated with one data set.

The species prevalence greatly influences the model selection

Logistic regression models model the expected proportion of samples where the species is present [0-1], i.e. the probability of presence, for a given value of the environmental variables. The estimation of the model parameters of a logistic regression is thus sensitive to the prevalence of the species in the

calibration set. In the proposed CMOC approach, the species prevalence is kept at 50% during bootstrap resampling because this provides the best trade-off between omission (false absence) and commission (false presence) errors for the LR models (Liu *et al.*, 2005). A prevalence of 50% also allows a more objective comparison of the contingency based model performance criteria kappa and NMI.

3.4.2. Cohen's Kappa, NMI and AUC

The Cohen's Kappa, NMI and AUC are model performance indicators that are often used to compare models. These three indicators showed little variance for the different models compared to the CMOC, which had very high values for the most optimal models and very low values for the rest. The CMOC thus had a much higher discriminatory power between models that were similar to the true model and other models. For the virtual species, the CMOCs clearly selected a few models which were most similar to the true model. Also, the calculated Akaike weights of the CMOCs are easier to interpret, as they are relative values that add up to one. Another disadvantage of Kappa and NMI, compared to the CMOCs, is the requirement to choose an arbitrary cut-off for presence to construct the contingency table. Liu *et al.* (2005) compared 12 different methods to determine this cut-off, but no single best method was found.

In the virtual species analyses there was only an agreement between the CMOC model ranking and the Kappa, NMI and AUC model ranking when the true model was very complex (TM5). The good model selection performance of these three indicators, can be explained by the absence of a penalisation term for model complexity. The Kappa, NMI and AUC only look at the model fit and thus select more complex models, as these models always have a better model fit to the data.

3.4.3. Which MOC is the best for model selection?

The five MOCs used in this research fit in the general framework of the model optimisation criterion. The MOCs differ only by the value of their regularisation parameter λ (Table 3.1). As such, the choice of a particular MOC comes down to the choice of a fixed λ value (AIC, F-statistic), an equation that relates λ to the number of samples (CAIC and BIC) or to the number of variables in the model (AICc). The parameter λ determines the trade-off between good model fit (low λ) or low model complexity (high λ).

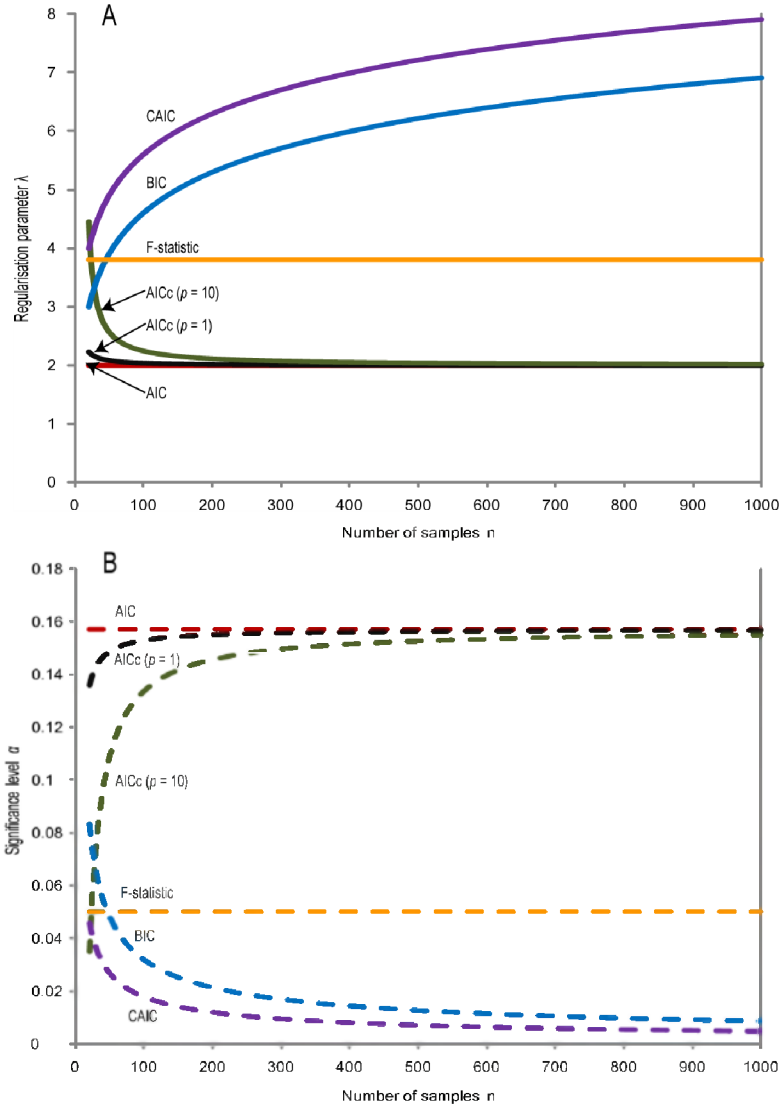

Fig. 3.10.A. Relation between the number of samples *n* in a data set and the regularisation parameter *λ* in the model optimisation framework, 4.9.B. Relation between the number of samples *n* in a data set and the significance *α* in the likelihood ratio framework. AIC = Akaike Information Criterion; AICc = AIC with small sample correction; CAIC = Consistent Akaike Criterion; BIC = Bayesian Information Criterion.

The *λ* can also be interpreted in the framework of the F-test (Reineking and Schroder, 2006), which tests whether the model with one parameter extra has a significantly higher likelihood. As such, the *λ* indicates how much the model fit has to rise to allow the model to include one extra variable in the model. The obtained ratio in the F-test of two models that differ only by one variable is equal to the *λ* in the MOC framework. Via the χ^2 distribution of this log-likelihood ratio, a significance level *α* in the F- test can be calculated for any *λ* value that is used in a MOC (Table 3.1; Reineking and Schroder, 2006). Thus, the addition of one extra variable to the model will thus increase the MOC with an extra penalty of

λ, which is compensated by an increase in the model fit (log-likelihood) of λ or more, in case this variable is significant at a value α which is linked to the specific λ value used (Reineking and Schroder, 2006).

In Fig. 3.10 the regularisation parameter λ and the matching significance α are provided for different numbers of samples n in a data set. The AIC and the F-test have a fixed λ and thus also a fixed significance α. The λ of the AICc goes asymptotically to 2 if the number of variables raises, the λ value of the AIC. The λ values of the CAIC and BIC keep on increasing logarithmically with an increasing number of samples n in the data set. In the virtual species analysis the number of samples n = 494 resulted in λ values of the CAIC (7.20) and BIC (6.20) that were more than three times larger than the AIC λ of 2. This higher λ turned out to be beneficial in the model selection, as the CAIC and BIC were able to find the true model or a very similar model independent of the complexity of this true model. The AIC and AICc performed bad at model selection for the simpler TM1 and TM3 (Fig. 3.5). The F-statistic had an intermediate model selection performance, and followed the pattern of the AIC. Similarly, in the model selection for the species *Abra alba* the summed CMOC AWs of the AIC and AICc were considerably higher for all the variables which again indicates that these CMOCs prefer more complex models.

For the virtual species there was a very clear relation between the λ and the weighted average of the number of variables over all models (Fig. 3.7). For the species *Abra alba* the relation between the regularisation parameter λ and the weighted average number of variables was not as straightforward as with the virtual species (Fig. 3.9). It can be concluded that overall the choice of the λ directly influences the complexity of the models chosen in the model selection.

The CAIC and BIC are consistent criteria which will find the true model in case the number of samples n goes to infinity (Anderson *et al.*, 1998). The main characteristic of consistent criteria is that λ is a function of the number of samples. Consistency is desirable when MOCs are applied to large data sets (Shono, 2005). This is because with rising sample numbers, the model fit will increase monotonously, which causes the log-likelihood to decrease. If the λ value is fixed, as with non-consistent MOCs, the complexity of the selected models will increase as the log-likelihood will keep on decreasing while the penalisation stays constant. Consistent criteria have an increasing λ when the number of samples rises to compensate for the ever decreasing log-likelihood when the number of samples rises. With increasing sample numbers the model complexity is thus penalised more and more, which causes consistent criteria to choose models with a trade-off between complexity and model fit, independent of the number of samples. In the virtual species analysis, both consistent MOCs (CAIC and BIC) did a good model selection for all complexities of the true model, while the other non-consistent MOCs only performed well for the most complex model TM5.

Consistent criteria allow the model to get slightly more complex when more samples are available. The justification to allow more complex models is that greater sample sizes bring on more information on the species-environment relation. More time and area effects may thus be revealed by the data when the number of samples increases (Anderson *et al.*, 1998). For example, with increasing sample size age-specific, sex-specific or year-specific species-environment relations might be found (Anderson *et al.*, 1998). Inclusion of these factors can lower the MOC in case the drop in log-likelihood is larger than the λ for the given number of samples. Consistent MOCs are thus to be preferred for the selection of models developed based on biological observations, because the true model has an infinite complexity and therefore the models chosen by the MOC should also get more complex, if the number of samples goes to infinity.

3.4.4. Virtual species approach to assess model selection

The virtual species approach proved to be very useful to assess the model selection abilities of the CMOC. Contrary to an analysis with field observations, the true model was known (Austin, 2007). The Euclidean distance allowed to quantify the similarity between the true model and the selected models. The virtual species approach was realistic because: 1) real environmental observations were used with a realistic correlation structure (Table 3.3.) and measurement errors, 2) random noise was introduced in the species-environment relation to introduce stochasticity and 3) the relative contribution of the variables was not equal, the model selection was challenged to find variables with a small contribution to the species-environment response.

Comparing the CMOC values between the virtual species and *Abra alba*, the values are less spread out for the virtual species with a few models having high values and the rest very low values. While with the virtual species only a few models were very good at modelling the response, and so the rest received very low CMOCs. For the virtual species there appeared to be a smaller selection of models that were good at modelling the species presences and absences. The difference in model selection results can be attributed to the fact that there is only a single data generating model for the virtual species and only a limited amount of noise is introduced in the species-environment relation. In the prediction of real species field observations the true model is not known and there are several sources of error. Also there is no guarantee that all the necessary environmental variables are in the data set.

3.4.5. Model selection for the species *Abra alba*

All the models selected as the most optimal by all of the five CMOCs for the species *Abra alba* used the variables median grain size, depth and BPI. The models selected by the CMOC based on the F-statistic also used the variables mud and bottom current speed. The importance of the sedimentological variables median grain size and mud was already observed by Degraer *et al.* (2008), who predicted the spatial distribution of the *A. alba* community with these variables as predictors. In sediment preference experiments, Alexander *et al.* (1993) observed that *Abra alba* is a specialist species with a narrow sediment preference range and the species has a preference for silt and mud. The importance of the variable depth can be explained by the relation of this variable, with numerous other variables on the BPNS (e.g. distance to coast, amount of suspended matter). The BPI indicates whether a sampling location is on a crest or in a through. The importance of this variable could be due to its relation with small scale current patterns or the fact that organic matter accumulates in troughs. Gogina *et al.*, 2010 modelled the probability of presence of *Abra alba* based on the organic matter concentration in the German Bight. For the BPNS, however, this variable was not measured in most of the samples in the data set used.

3.4.6. Practical application of the CMOCs approach

The proposed CMOC approach is a tool to select the most optimal variable combination to model the species-environment relation. Because this is an exhaustive approach, the number of variable combinations rises exponentially with the number of variables considered for inclusion in the model. In this context, Araujo and Guisan (2006) argue that automatic model selection should not be seen as a substitution for pre-selecting relevant ecophysiological variables based on deep knowledge of the biogeography and ecology of species. Therefore it is advisable to perform a first manual variable selection based on visual exploration of the species-environment relations, previous ecological knowledge by experts or from models of related species. Although the exhaustive approach does not suffer from instability when highly correlated variables are considered, prior omission of these variables reduces the number of possible models and reduces the error on the estimated model parameters.

The traditional idea that only the variables in the selected best model are important, while excluded variables are not important is too simplistic (Burnham *et al.*, 2004). Multimodel approaches consider all possible models and thus require a weighing term that expresses the relative importance of each model. The CMOC model selection approach is suitable for multimodel prediction, inference and

reliability estimates. At a higher level, these mean predictions over all the bootstrap replicas can be used to produce a weighted mean over all variable combinations with the CMOC as weight.

The CMOC is based on the model optimisation criterion framework which incorporates the likelihood and the number of model parameters. As such, the CMOC can be calculated for each modelling technique that can have a likelihood calculated, for example all GLMs and GAMs. Also for neural networks or other techniques a likelihood can be calculated. The CMOC can also be used for modelling techniques that predict species densities and abundance classes, as long as a likelihood can be calculated. GLMs, for example, can predict both densities, probabilities of presence and densities classes. For all these model outputs a likelihood, and thus A CMOC, can be calculated for the GLMs.

3.4.7. Future Research

Future research on the application of the CMOC approach should focus on the interaction between the model regularisation parameter λ, the number of variables in the data set, the weighing parameter m in the CMOC and the number of variables in the chosen model. This would however require a very large number of models to be calibrated to investigate possible interaction effects between the variables and was beyond the scope of this chapter.

A clear relation between the λ value and the number of variables in the chosen model has been found in the virtual species analysis and a similar relation also for the species *Abra alba*. However, can a universal relation be found between on one side the number of samples which determines the λ value and on the other side the number of variables in the models selected? In the consistent MOCs the number of samples determines λ, and thus the complexity of the chosen models. Ideally, this complexity should not be determined by the number of samples in the data set, but by an estimator of the true information content of the samples. If the samples contain more information on the species-environment relation, the model is allowed to get more complex.

Using the number of samples to estimate the model complexity works only under the assumption that each observation is totally independent, and thus brings on one full degree of freedom. Most HSM data sets are opportunistic compilations of samples, which often have autocorrelation because the samples are all collected in a small area or replica samples are taken. Another source of autocorrelation is oversampling of some environmental variable combinations and undersampling of other combinations. A HSM-based sampling design (see 1.3.1.) can guarantee a well stratified sampling over all habitat types. Autocorrelated samples bring on less information per sample as they are not completely independent (Legendre, 1993). As a result, the effective sample number is lower than the

actual number of samples in the data set. The λ based on the real number of samples is larger than the λ based on the effective number of samples. Thus the penalisation for model complexity in case of autocorrelated samples is higher, leading to the selection of simpler models. In reality, the number of samples often cannot be chosen. For consistent MOCs, the fixed sample number determines the λ value which in term determines the significance α at which variables are included in the model. In practice thus only the MOC can be chosen and the weighing parameter m in the CMOC.

3.5. Conclusions

The Combined Model Optimisation Criterion (CMOC) approach is proposed as a model selection method for HSMs which is superior to stepwise model selection. The CMOC approach is an exhaustive, robust model selection approach based on information theoretic measures, which chooses parsimonious models to maximise the transferability of the models to other regions or periods. Some advantages of the COMC are:

- The CMOC approach incorporates model validation into the model selection to increase the generalisation ability of the selected models on spatially or temporally independent data.
- The exhaustive testing of all variable combinations increases the chance of finding an optimal model, and also causes the CMOC approach to deal better with multicollinearity of predictive variables.
- Bootstrap resampling is used in the CMOC approach to increase the reliability of the model selection. During the resampling the species prevalence is kept at 50%, to minimise bias in the model parameter estimation.
- The CMOC approach can be used for multimodel inference and prediction, and has parameters to shift the emphasis in the model selection from high predictive performance on the calibration data or on the test data.
- The CMOC approach is not based on contingency table-based model performance indicators, so an arbitrary cut-off for presence does not need to be chosen.
- The CMOC approach was successfully tested with artificial species data, and applied to a real species *Abra alba*.

Possible disadvantages of the CMOC model selection approach are the slower model selection in comparison with stepwise approaches and fact that the CMOC approach is not included in commonly available statistical software.

Acknowledgements

Special thanks goes out to all the contributors to the Macrodat data base. The Management Unit of the North Sea Mathematical Models (MUMM) provided the current speed data. The Renard Centre for Marine Geology (RCMG, UGent) provided the full cover median grain size, depth and bathymetric position index grids.

Chapter 4. Integrated validation of marine habitat suitability models

Abstract

A sound management of species and habitats requires good knowledge on the spatial distribution of these entities. However, mostly a limited set of point observations is available, instead of full cover surveys. Habitat Suitability Models (HSMs) can predict the distribution of suitable habitats for a species on a full cover scale by using the species-environment relationship. As the number of HSM applications rises the need for improvement of the modelling methodology increases. In general the ecological relevance of the HSMs is rarely assessed. Therefore, the general goal of this chapter is a plea for an integrated validation of HSMs to test the ecological relevance of the models produced. Such an integrated validation considers: 1) model validation of observations vs. predictions, 2) a conceptual scheme bringing together ecological knowledge from the literature and allowing inference about the causality of predictive variables, 3) habitat preference experiments which allow distinguishing the fundamental and realised niche and, 4) an assessment of the sample distribution over the range of the predictive variables.

HSMs were developed for the bivalve species *Donax vittatus* with a data set from the Belgian part of the North Sea. The Combined Model Optimisation Criterion (CMOC) was used for model selection, as this approach combines the model performance on a test set and an independent validation set. Logistic regression, a type of generalised linear model, was used as a modelling technique. Four models were retained after the model selection, and a multimodel prediction was done, with the CMOC as a weighing factor. Only depth and median grain size were retained as predictive variables. Habitat preference experiments in the ecological literature confirmed the modelled sediment grain size response. For each variable in the data set an integrated discussion was provided why this variable is (not) chosen in the final models based on the conceptual scheme, habitat preference experiments and the sample distribution along the range of the variable.

Key words:
Habitat suitability modelling, species distribution modelling, model validation, habitat preference experiments, burrowing speed, *Donax vittatus*, model validation

4.1. Introduction

4.1.1. Habitat suitability models

Habitat Suitability Models (HSMs, Guisan and Zimmermann, 2000; Austin, 2007; Hirzel and Le Lay, 2008) are used increasingly in ecosystem management to get insight in the spatial distribution of

species. Sound management decisions need to be supported by knowledge of the current and future distributions of vulnerable species and their preferred habitat. HSMs can predict the distribution of suitable habitats for a species on a full cover scale by using the species-environment relationships. If the habitat matches with the preferred habitat of the species, the habitat is suitable for the species to occur. Without HSMs, the knowledge of the species distribution is often limited to a sparse set of point observations (e.g. sediment grab samples). Environmental variables are often available at a full cover scale and these variables can be used as inputs for HSMs.

The general assumption is that more suitable habitats have a higher probability of species presence and will support higher species densities (Barry and Elith, 2006). Another assumption for HSMs is that the species is in steady-state equilibrium with its environment (Guisan and Thuiller, 2005; Václavík and Meentemeyer, 2009). This means that the species is present where the habitat is suitable and absent in case the habitat is unsuitable (Barry and Elith, 2006). The equilibrium assumption is important because at one point in time, both the species, as well as the physical habitat are observed, and HSMs assume that the habitat is determining the species distribution observed at a given point in time (Guisan and Thuiller, 2005).

The theoretical justification of HSMs leans heavily on the niche theory (Guisan and Thuiller, 2005; Hirzel and Le Lay, 2008). The ecological niche is the subspace of the environmental space where the combination of environmental variables is most suitable for the species. Modelling the habitat preference of a species thus equals the delimitation of its niche. The fundamental niche of a species is only limited by physiological constraints. The realised niche in the environmental space is a subspace of the fundamental niche, and is limited by both physiological and dispersal limitations, as well as biotic interactions (e.g. competition, predation; Rodder and Lotters, 2009).

The fundamental niche can be delimited by setting up habitat preference experiments (e.g. Alexander *et al.*, 1993; Wright *et al.*, 2000; de la Huz *et al.*, 2002) where the preferred range for each environmental variable is tested in the absence of the influence of other variables, biotic interactions or dispersion limitation. The realised niche is what is observed in field observations, and thus reflects all constraints imposed on the actual distribution of a species, both physiological, biological, anthropological and geographical (Pearson *et al.*, 2007). HSMs based on field observations thus model the realised niche (Guisan and Thuiller, 2005).

4.1.2. Need for an integrated model validation

As HSMs are a relatively recent ecological modelling technique, methodological research is rather limited (Huettmann and Diamond, 2001). A general overview of the methodological aspects is provided

by several authors (e.g. Guisan and Zimmermann, 2000; Redfern *et al.*, 2006; Elith and Leathwick, 2009; Franklin and Miller, 2009). Some methodological research focused on the choice of the modelling technique (Segurado and Araujo, 2004; Elith *et al.*, 2006), the impact of the resolution of environmental variables (Guisan *et al.*, 2007), the transferability of models (Randin *et al.*, 2006), or which model output to choose (Lutolf *et al.*, 2006).

There is a disproportionally large effort in developing HSM models and maximising the predictive performance, compared to the validation of HSMs (Rykiel, 1996; Eastwood *et al.*, 2003; Araujo and Guisan, 2006). Assessment of only the predictive success by comparing predictions and observations is insufficient and a test of ecological realism of the models is also needed (Austin, 2007). When models are only validated with observations an assessment of the causality of each variable is not possible and a distinction between the realised and fundamental niche per variable cannot be made. Model validations by comparing the predictions with field observations can be done with a part of the original data set not used for model calibration (test set), this is called internal validation (Randin *et al.*, 2006). However, this test set is only pseudo-independent of the model calibration set and the model performance on truly independent data may be significantly lower (Beutel *et al.*, 1999).

In this chapter an extension to the current practice of model validation with field observations is proposed. The validation will be improved by incorporating knowledge on the species' ecology into the validation, as well as experimental model validation. An assessment of the influence of the sample distribution over the range of the predictive variables, on the model selection and on the models' predictive performance is provided.

4.1.3. Integration of species ecology

The ecology of the modelled species is rarely considered in HSM research (Austin, 2007). Often an automated variable selection algorithm is applied to a set of readily available variables, without much further considerations of the species ecology. However, the combination of ecological knowledge and methodological modelling skills is more important than the precise statistical method used (Austin, 2007). Therefore a better integration of HSMs with ecological theory is needed (Guisan and Thuiller, 2005). The most efficient and comprehensive method is to combine the current knowledge on the species ecology in a conceptual scheme. Such schemes have been used as a basis for some marine HSMs for fish (Gibson and Robb, 2000), crabs (Avissar, 2006) and seagrasses (van Katwijk *et al.*, 2000). In a conceptual scheme variables that are relevant in the determination of the species distribution according to ecological literature should be shown. Additionally, variables that are in the data set but are not considered in the literature can be included (e.g. BPI). Creating a conceptual scheme will: 1) force

researchers to consider the possible causal relation and the direction of the relation between variables and/or the species, 2) help to determine which variables need to be measured (Gibson, 1994) and included in an optimal model (Austin, 2007) and 3) help to determine which important variables are missing in the data set. A model can be validated by comparing the variables selected in the model selection, with variables that are determining the species distribution according to the ecological literature.

Some variables determine the species distribution in a direct and causal way, e.g. the food availability. The opposite of causal variables are proxy variables that are widely available and cheaply measurable approximations for other variables (Guisan and Zimmermann, 2000). Based on the ecological knowledge combined into the conceptual scheme, it is possible to assume if a variable is causal in the prediction of the species distribution. Causal variables have little or no intermediate variables, between the variable and the species distribution. Proxy variables often have a more indirect and correlative relation with the prediction of the species distribution, because there are several intermediate variables between the proxy variable and the species. Proxy variables cannot explain the causal nature of the relationship, but are practically useful in forecasting change over the range of the environmental conditions sampled (Thrush *et al.*, 2003). The identification of causal variables is crucial as it is desirable to predict the distribution of species based on ecological parameters that are believed to be causal driving forces of the species distribution (Guisan and Zimmermann, 2000). When proxy variables with local correlative relations are used in the model, the transferability of the model to other regions and periods is limited (Guisan and Zimmermann, 2000).

4.1.4. Experimental model validation

Habitat preference experiments are a valuable extra model validation approach (Huxham and Richards, 2003). Experiments serve to test if the relation of a variable with the distribution of the species is causal or merely correlative and, to assess the fundamental niche of the species for the variable in the absence of biotic interactions and dispersion limitation. The experimental validation is truly independent of field observations, while ecological literature often is based on field observations. HSMs can quantify the contribution each variable has in the prediction of the distribution of the species, but this does not allow distinguishing between correlative and causal relations. Assessing true cause-and-effect relationships requires controlled experiments on the effect of one variable, while controlling for the others (Thrush *et al.*, 2003). Ideally, a HSM should be tested with habitat preference experiments, however very few such examples can be found in the marine literature. Wright *et al.* (2000), for example, developed a HSM for

sandeel and assessed the sediment preference in controlled experiments to validate the modelled species response.

Especially when a model is to be applied in another region or period, model validation by means of habitat preference experiments is very valuable. The more causal the variables in the transferred model are, the more reliable the model predictions will be in other regions or periods. In some HSM applications it can be mandatory to identify also the fundamental niche for certain variables (e.g. to identify potential colonisable areas for invasive species; Wiens and Graham, 2005).

4.1.5. Assessment of the sample distribution

Ideally, data sets to develop HSMs should be obtained in planned sampling campaigns, but in reality well designed sampling campaigns for modelling are rare and mostly the data are an opportunistic post-hoc data compilation from existing sources (Austin, 2007). Such a collation can create a good spatial cover and a large sample database. The drawback is that the spatial distribution of the samples is likely to be biased for the purpose of modelling, because the samples are heavily clustered and some habitat types are under- or oversampled. This will mostly result in an uneven coverage of the range of the environmental variables which can lead to a poor fit in an undersampled part of the variable range. The distribution of the samples over the variable range can also affect if a variable is useful to determine the distribution of a species. The model can thus also be validated by interpreting the modelled response, together with the distribution of the data used to calibrate the model.

4.1.6. Aims

The general aim of this chapter is to plea for an integrated validation of HSMs. Such an integrated validation would consider the validation of the model with 1) species observations (internal or external model validation), 2) ecological insights from the literature, 3) habitat preference experiments and 4) an assessment of the distribution of the model calibration data over the range of the predictive variables.

To illustrate the suggested model validation improvements, a HSM is developed for the bivalve species *Donax vittatus* (Da Costa 1778). This species was chosen because extensive ecological literature is available on the species, habitat preference experimental results are available and, the species is a food source for juvenile plaice (*Pleuronectes platessa*; Burrows and Gibson, 1995), an abundant fish species in the North Sea. An overview of the ecological knowledge on the species will be combined in a conceptual scheme. In a multimodel approach, HSMs were developed based on the

CMOC model selection approach (see Chapter 3). The selected models were validated with ecological knowledge from the literature. Habitat preference experiments from the ecological literature were used to validate the modelled species response for the variable median grain size.

4.2. Material and methods

4.2.1. Habitat suitability models

4.2.1.1. Modelling technique

To predict the presence of species, different modelling techniques exist (Elith *et al.*, 2006). In this research, logistic regression (LR; Agresti, 2002; Appendix II) a type of Generalised Linear Model (GLM) is chosen as modelling technique because: 1) this technique is often used in HSM (e.g. Ysebaert *et al.*, 2002; Le Pape *et al.*, 2003), 2) is relatively simple and, 3) statistically well established. LR has a transparent model structure and a comparatively low number of parameters (Reineking and Schroder, 2006). More complex techniques (e.g. Artificial Neural Networks) would require more in-depth technical discussion. The proposed methodological improvements (incorporation species ecology, experimental model validation and improved sampling design) can later be applied to any HSM modelling technique.

GLMs have the possibility to assume different distribution functions of the response variable and the flexibility to choose a link function between the random and systematic component. This link function allows to have predictive variables within the range of the observed responses and, to use a linear combination of predictors (Guisan and Zimmermann, 2000). LRs are used to model the relation between the binomial (present or absent) species occurrence observations and a set of predictive environmental variables. Because the model output of the LR should be constrained to the interval [0 – 1], the logistic link function is applied in the LR (see Appendix II). This continuous model prediction can then be converted to discrete presence/absence information by using a cut-off for presence (Liu *et al.*, 2005). To minimise the bias in the estimation of the model parameters due to the low prevalence of the modelled species in the data set (prevalence = 0.10), the prevalence will be kept at 50% during the bootstrap resampling. A cut-off for presence of 0.5 will be used.

4.2.1.2. Data preparation: study area and data set

The BPNS has a surface of 3600 km² (±0.5% of the total area of the North Sea), and is situated in the Southern North Sea. The BPNS is relatively shallow with a depth of maximum -45 m. The seabed surface is characterised by a highly variable topography, with a series of sandbanks and swales. A broad spectrum of sediments is found, ranging from clay to coarse sands (Van Hoey *et al.*, 2004).

A data set with 990 grab samples collected between 1977 and 2003 is used to develop HSMs (Fig. 4.1). Animals were sieved on a 1 mm sieve and counted. This complete data set is split based on the collection period to obtain a validation set that is temporally independent. The pre-2001 data set contains 786 samples and is resampled during the bootstrap resampling to generate a calibration set and a test set for each bootstrap replica (See 4.3.1.3. Model development). The post-2001 data set contains 204 samples and is resampled in a bootstrap procedure to generate a validation set for each bootstrap replica. As such, an external validation of the model with data from a different period is performed to assess the temporal generalisation and transferability of the model.

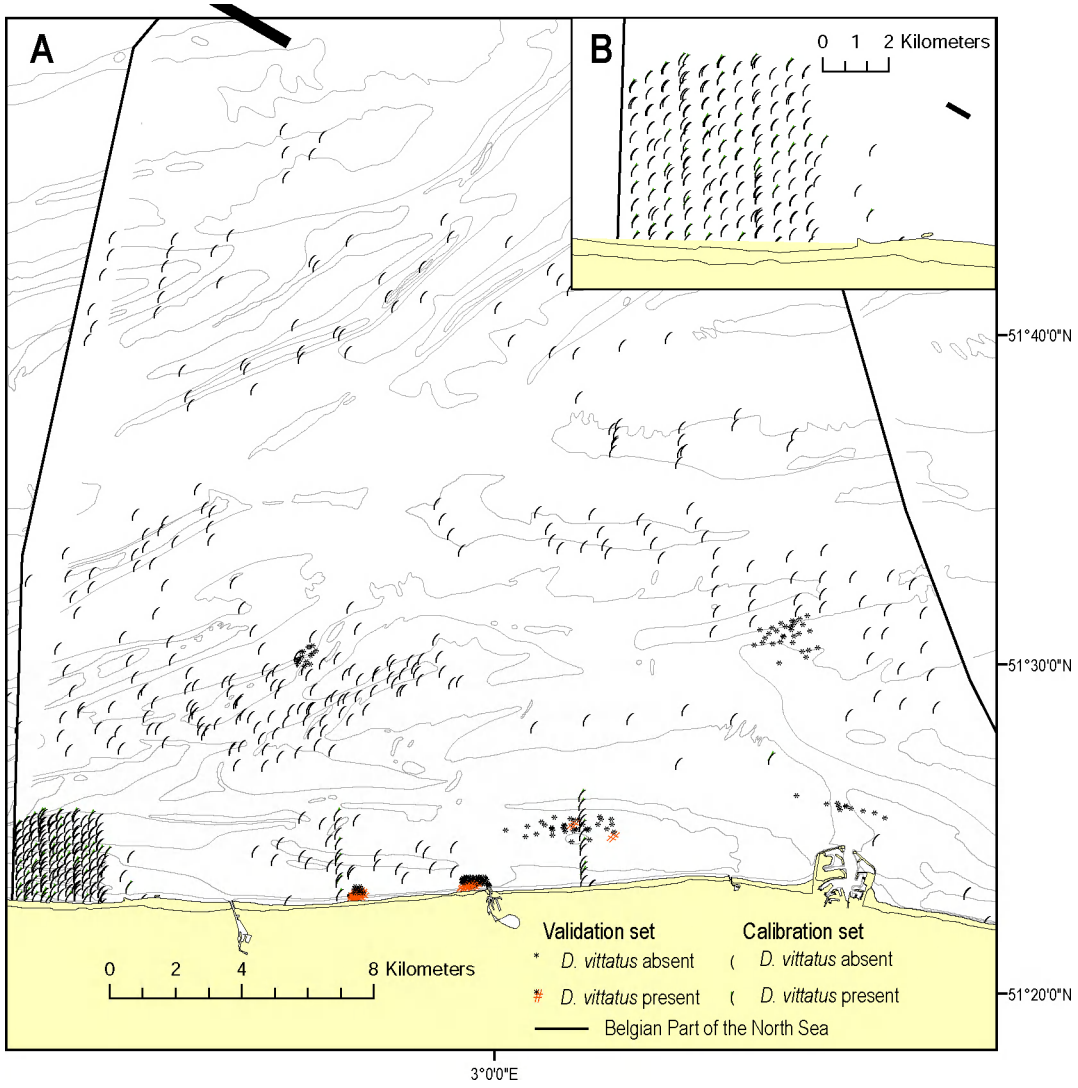

Fig. 4.1. Distribution of the training/test and the validation set samples in the Belgian part of the North Sea. Projection UTM 31N WGS84. A. Overview of the Belgian Part of the North Sea (BPNS). B. Detail view of the Western Coastal Banks.

Five environmental variables are considered as potential variables to predict the distribution of *D. vittatus* (Table 4.1). These variables are used because: 1) they are available as full cover grids, 2) they represent several aspects of the habitat (sedimentology, topology and currents) and, 3) they have been shown to be useful in the prediction of the habitat preference of macrobenthic species in previous research (Degraer *et al.*, 2008; Willems *et al.*, 2008). Other variables were available, some also at a full cover scale, but were not considered for the model. Salinity, for example, is a good predictor of macrobenthic species distributions in estuarine environments where there is a clear gradient (Ysebaert *et al.*, 2002), but on the scale of the BPNS there is no variance in the salinity which would be useful to model species distributions. The water temperature is expected to influence the densities and growth of macrobenthic species (Reiss *et al.*, 2006), but also displays little spatial variance on the BPNS.

Table. 4.1. Overview of the data sets. BPI = Bathymetric Position Index (Lundbald *et al.*, 2006).

	Training/test set			Validation set			Full cover layers		
Nr. samples	786 samples			204 samples			55726 grid cells		
Year	1977-2001			2002-2003			/		
	min	mean	max	min	mean	max	min	mean	max
D. vittatus (dens./m²)	0.0	3.4	194.9	0.0	5.3	282.7	/	/	/
Median grain size (µm)	8.0	277.8	658.0	9.8	228.7	490.2	0.0	307.3	891.1
Mud %	0.00	3.12	99.04	0.00	8.69	92.94	0.00	5.10	75.68
Depth (m MLWS)	-35.90	-10.03	5.41	-20.43	-4.88	4.35	-45.72	-22.99	1.74
BPI	-573.0	-8.3	329.0	-229.0	-30.1	200.0	-659.0	-0.3	415.0
Bottom current (ms^{-1})	0.035	0.517	0.884	0.117	0.493	0.836	0.049	0.943	1.659

The sediment median grain size, the mud % (4-63µm) and the depth are actual measurements at the time of sampling, while the Bathymetric Position Index (BPI; Lundblad *et al.*, 2006) is derived from the high resolution bathymetry raster (80m pixel size). The depths are standardised to the Mean Low Water Spring level (MLWS), which cause some depths to be > 0 m. The BPI indicates the relative elevation of a location in relation to the surroundings. Thus whether a sample is on a ridge (BPI > 0), or in a trough (BPI < 0; Verfaillie *et al.*, 2006). The maximum bottom current (m/s) is derived from the COHERENS 3D baroclinic model (Luyten *et al.*, 2003).

Table. 4.2. Correlation coefficients of the environmental variables in the complete data set (Kendall Tau correlation).

	Median grain size	Mud %	Depth (m)	BPI	Bottom current (ms^{-1})
Median grain size		-0.29	-0.24	-0.03	0.31
Mud %	-0.29		-0.16	0.31	0.10
Depth (m)	-0.24	-0.16		-0.29	-0.61
BPI	-0.03	0.31	-0.29		0.24
Bottom current (ms^{-1})	0.31	0.10	-0.61	0.24	

Full cover data grids of each of the predictive variables are available for each of the five environmental variables and will be used to produce full cover prediction of the probability of presence for *D. vittatus* . Each grid cell has a size of 250x250m. To create full coverage median grain size maps, kriging with an external drift was used, taking into account bathymetry as a secondary variable to assist in the interpolation (see Verfaillie *et al.*, 2006). The map of the mud % was created, using ordinary kriging with directional variograms for the anisotropy of the data (Van Lancker *et al.*, 2007). The BPI was derived from the bathymetry grid (Flemish Authorities, Agency for Maritime and Coastal Services, Flemish Hydrography).

4.2.1.3. Model development

The central step in the development of HSMs is the model selection, thus the selection of the most optimal combination of predictive variables (Guisan *et al.* 2000, Heikkinen *et al.* 2006). If all available variables would be included, the complexity of the model and thus the number of model parameters would increase quickly. Therefore, automated stepwise variable selection (see Chapter 2) is most often used with LR. But the stepwise procedure has some major shortcomings (see Chapter 3): 1) it is prone to find local optima and miss the overall most optimal model (Reineking and Schroder, 2006), 2) the procedure is sensitive to multicollinearity of the predictive variables, 3) only one final model is provided in the stepwise approach.

In this research an exhaustive model selection approach is applied to find the most optimal combination of environmental variables (Fig. 4.2). This approach was proposed in Chapter 3. Exhaustive model selection tests and compares combinations of predictive variables and aims to find a globally optimal model. As a measure of how optimal a variable combination is for species modelling, the Combined Model Optimisation Criterion (CMOC, Chapter 3) is calculated. The CMOC can be used

for multimodel inference and prediction when the CMOC per variable combination is used as a weighing factor to calculate a multimodel average.

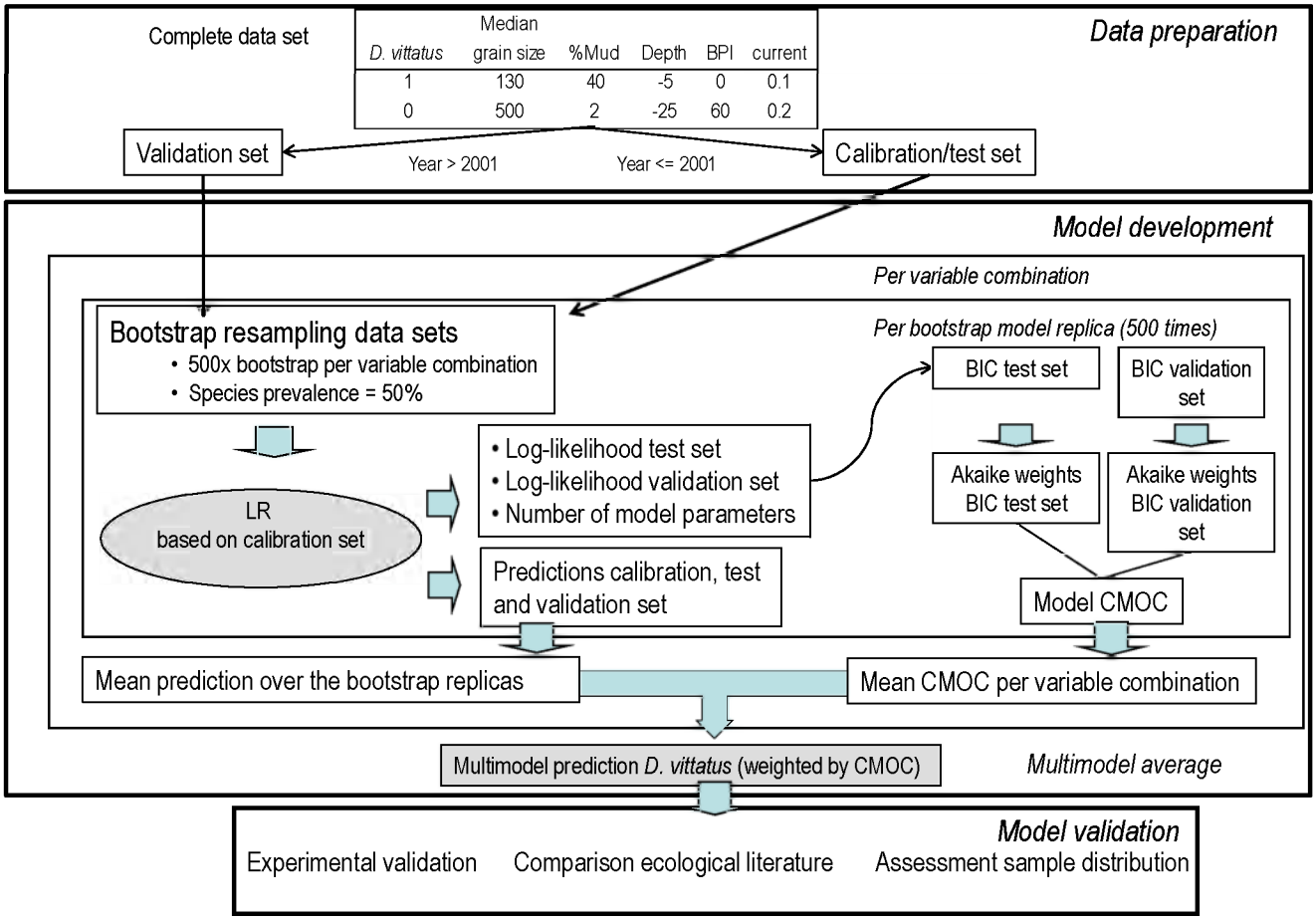

Fig. 4.2. Overview of the steps in the HSM methodology for the species *D. vittatus.* The Combined Model Optimisation Methodology (CMOC) is a weighted average of the Akaike weights of the test and validation set BICs.

In an exhaustive approach, the number of variable combinations rises exponentially with the number of alternative variables considered. For this reason the hierarchical model building approach (Kutner *et al.*, 2005) is used during the generation of alternative models. In the hierarchic model building first all possible combinations of the linear regression terms are tested. From these set of models, the models with a CMOC > 0.05 are kept. To this retained models with linear terms, all combinations of quadratic terms are added. A quadratic term of a variable is only considered in a model if the linear term is already in the model. From these set of models with quadratic term the models with a CMOC > 0.05 are retained. For this set of models all possible interaction terms between the linear terms are considered. The final models are the models that were obtained in this last stage. The hierarchical approach

drastically limits the number of alternative models to be calibrated and compared, while keeping the exhaustive character of the model selection.

To increase the precision of the CMOC estimate, 500 replica models are calibrated per alternative variable combination by bootstrap resampling the pre-2001 data set to obtain a calibration and test set, and the post-2001 data set to obtain a validation set (Fig. 4.2). The mean CMOC per variable combination is calculated as an average of the 500 CMOC values. During the bootstrap resampling, the prevalence of the species is kept at 50%.

For each of the 500 bootstrap replicas per alternative variable combination the log-likelihood and the number of variables *n* are used to calculate the Bayesian Information Criterion (BIC, Shono, 2005). The BIC is a Model Optimisation Criterion (MOC) that allows models to be ranked from most optimal to worst model. The BIC was chosen as a MOC because it is a consistent criterion which will find the true model in case the number of samples *n* goes to infinity (Anderson *et al.*, 1998). Based on artificial species data, BIC proved to find the true model used to generate the species data or a very similar model (see Chapter 3).

An MOC in the general MOC framework consists of three parts: 1) a measure of the model fit to the observations, 2) a measure of the model complexity and 3) a regularisation parameter λ (Reineking and Schroder, 2006; Equation 4.1). The goodness of the model fit is quantified in the MOC framework based on the likelihood model *i* with the parameter vector $\hat{\theta}$ (Equation 4.2). The model complexity term equals the number of model parameters *p*. The regularisation parameter λ determines the relative weight of the model complexity *p* in the MOC formula (Equation 4.1; Reineking and Schroder, 2006). The regularisation parameter λ times the model complexity *p* equals the penalisation term of the MOC (Reineking and Schroder, 2006). A more complex model will thus probably have a higher model fit, but also a higher penalty term because there are more model parameters. The model with the lowest MOC will be a trade-off between maximal model fit (correct predictions) and minimal model complexity (number of model parameters). A model with minimal MOC has maximal parsimony and is assumed to be the most optimal model given the data set.

$$\text{MOC} = \text{goodness of fit} + \lambda \cdot \text{model complexity} \tag{4.1}$$

$$\text{MOC} = -2 \ln\left(L(\hat{\theta})\right) + \lambda p \tag{4.2}$$

$$\text{BIC} = -2 \ln\left(L(\hat{\theta})\right) + \ln(n)\, p \tag{4.3}$$

In the BIC formula (Equation 4.3), the regularisation parameter λ equals the logarithm of the number of samples n (Equation 4.3). The BIC is a consistent MOC, which theoretically selects the true model in case the number of samples n reaches infinity (Hastie *et al.*, 2001). Consistent criteria thus provide an asymptotically unbiased estimate of complexity of the true model (Anderson *et al.*, 1998). When MOCs are applied to large data sets consistency is desirable (Shono, 2005).

The CMOC combines the BICs of the test and validation set. To calculate the CMOC, first the BIC values of the test and validation set are averaged over the 500 model replica created for each variable combination (Equation 4.4). To make these $\overline{BIC_i}$ values relative and independent of the number of samples n, the Akaike Weight (AW, Wagenmakers and Farrell, 2004) of these $\overline{BIC_i}$ values is calculated (Equation 4.5 and 4.6).

$$\overline{BIC_i} = \frac{1}{500}\sum_{j=1}^{500} BIC_{ij} \tag{4.4}$$

$$\Delta\overline{BIC_i} = \overline{BIC}_i - \min(\overline{BIC}) \tag{4.5}$$

$$AW\left(\overline{BIC_i}\right) = \frac{e^{-\Delta\overline{BIC}i/2}}{\sum_1^k e^{-\Delta\overline{BIC}r/2}} \qquad k = nr\ models \tag{4.6}$$

$$CMOC = m \cdot AW\left(\overline{BIC}_{test}\right) + \left(1\text{-}m\right) \cdot AW\left(\overline{BIC}_{validation}\right) \tag{4.7}$$

The CMOC is then calculated as the weighted average of the Akaike weights of the test and validation set $\overline{BIC}$ (Equation 4.7). The relative contribution of the test set and the validation to the CMOC is determined by the weight m, which can be chosen in the interval [0, 1]. In the remainder of this chapter, m = 0.5, which means both data sets have equal contribution. The choice of m will shift the emphasis in the model selection from accurate predictions on the test data (m = 1), to maximal generalisation and thus more accurate prediction for the validation set (m = 0). This will result in a single CMOC value per alternative model that indicates how optimal a model is relative to the other models.

For biological observations the true model can never be known and is of unlimited complexity (Anderson *et al.*, 1998). Therefore the use of one single model is not always realistic, especially not for biological observations (Anderson *et al.*, 2009). Also, the most optimal models selected in the model selection have a similar predictive performance, and choosing only the first one would be assuming that a single true model can be found for biological observations. Therefore a multimodel averaged or ensemble model prediction will be performed in this chapter. Per sample in the data set, the prediction is calculated as the weighted sum of the prediction of each alternative model for that sample. The CMOC is used as weighing factor, to let more optimal models have a higher contribution to the final prediction.

To produce full cover predictions of the distribution of *D. vittatus* in the BPNS, full cover grids of predictive variables are fed into the selected optimal models. Per grid cell a multimodel weighted average is calculated over all the predictions from the different alternative models.

Besides the CMOC, contingency table based model performance indicators are provided as they are often used to express the predictive performance of HSMs. The indicators provided are the Cohen's Kappa (Fielding and Bell, 1997), Normalised Mutual Information criterion (NMI, Fielding and Bell, 1997) and the Area Under the Curve (AUC; Maggini *et al.*, 2006; Table 4.2). Contingency tables are obtained by tabulating the observed species presence against the predicted species presence (Fielding and Bell, 1997). Before tabulation it is necessary to convert the continuous model prediction [0 – 1], to discrete absences and presences [0, 1] by using a cut-off for presence of 0.5 (Liu *et al.*, 2005). The AUC is the surface under the receiver-operator curve, which is constructed by plotting the sensitivity values against 1-specificity for a series of cut-off for presence-values (Swets, 1988). The AUC is 1 for a perfect model and 0.5 for a nonsense model.

4.2.2. Ecology of *Donax vittatus*

Donax vittatus (da Costa, 1778), the banded wedge-shell, belongs to the Mediterranean boreal element of the European fauna, with a distribution extending as far north as the coasts of Norway (Ansell *et al.* 1980b). On the Atlantic coast of France the range of *Donax vittatus* overlaps with the congener *D. trunculus* (Ansell *et al.* 1980b). All the species within this genus *Donax* are rapid burrowers, with an elongated, slender shell that accommodates a large foot (Stanley, 1970). The valves are thick for stability and are smooth for streamlining (Stanley, 1970). *D. vittatus* is a suspension feeder (Yonge, 1949; Pohlo, 1969). In Scottish waters the life span of the species is estimated to be 5 to 7yr (Ansell *et al.*, 1980).

4.2.3. Assessment of the sample distribution

The data set used in this research is a compilation of existing data sets, not specifically collected for the purpose of HSM development (Fig. 4.5). Therefore, it is most relevant to assess the distribution of the samples over the range of each predictive variable. Undersampled, oversampled or unsampled regions in the variable range can be identified. For each of the variables in the data set, the sample distributions over the range of the variable are plotted. As such, the distribution of 1) the complete data set, 2) the samples where the species is present, 3) where the species is absent, and 4) the values of the grid cells in the full cover grids can be compared. The distribution of the observations per variable are converted to sample densities [0-1], to make them relative. The relative sample densities under each curve sum to one.

4.3. Results

4.3.1. Habitat suitability models

4.3.1.1. Model selection

The hierarchical, exhaustive model selection based on the CMOC, resulted in four models that were retained (Table 4.2). These four models are based on only two predictive variables: median grain size and depth. The model with the highest CMOC (0.543) is only using depth and depth². In three out of the four models a unimodal response of the species for depth is modelled. Median grain size is modelled with only a linear response, except for the model with the lowest CMOC (0.015). However, the latter model has a very low contribution to the prediction in a multimodel average, due to the very low CMOC. The contingency based model performance indicators Kappa, NMI and AUC are highly correlated among each other. For the calibration set, these contingency table based indicators match well with the AW BIC values of the calibration set. For the test and validation set, there is no clear relation.

Table 4.3. Models for the prediction of the spatial distribution of *D. vittatus.* The models are the result of a hierarchical, exhaustive model selection based on the Combined Model Optimisation Criterion (Chapter 3). The CMOC is calculated as the weighted average of the Akaike weights of the Bayesian Information Criterion (BIC) of the test and validation set. NMI: Normalised Mutual Information. AUC: Area Under the Curve.

					Contingency table based criteria									
					Calibration set			Test set			Validation set			
	CMOC	AW(BIC calibration)	AW(BIC test)	AW(BIC validation)	Kappa	NMI	AUC	Kappa	NMI	AUC	Kappa	NMI	AUC	Systematic component logistic regression
Model 1	0.54	0.09	0.13	0.95	0.46	0.19	0.73	0.44	0.18	0.72	0.55	0.29	0.77	*Depth + Depth²*
Model 2	0.32	0.67	0.59	0.05	0.54	0.24	0.77	0.50	0.21	0.75	0.59	0.32	0.80	*Median grain size + Depth + Depth²*
Model 3	0.12	0.03	0.24	0.00	0.37	0.12	0.68	0.38	0.12	0.69	0.46	0.19	0.73	*Median grain size*
Model 4	0.01	0.21	0.03	0.00	0.56	0.26	0.78	0.50	0.21	0.75	0.55	0.29	0.77	*Median grain size + Depth + Depth² + Median grain size *Depth*

4.3.1.2. Multimodel prediction species distribution

The multimodel prediction of the species-environment relations of *D. vittatus* per sample is calculated as the weighted sum of the prediction of the four models. The result per sample in the original data of the BPNS (990 samples) is a prediction of the probability of presence of *D. vittatus.* The multimodel response of the four optimal models is visualised in Fig. 4.3. The predicted probability to find *D. vittatus*

is lowest in deeper waters with coarse sediments and is highest in shallow waters with fine sediments. As the interaction is only present in model 4 with the lowest CMOC, interactions are not visible in the multimodel combined prediction, because model 4 has a low contribution to the multimodel prediction. When the presence observations of *D. vittatus* are plotted on the modelled response, they can be found where the probability of presence is highest (Fig. 4.3B). Most of the presence observations are above the cut-off for presence of 0.5, indicating there are few false absences (Fig. 4.3B).

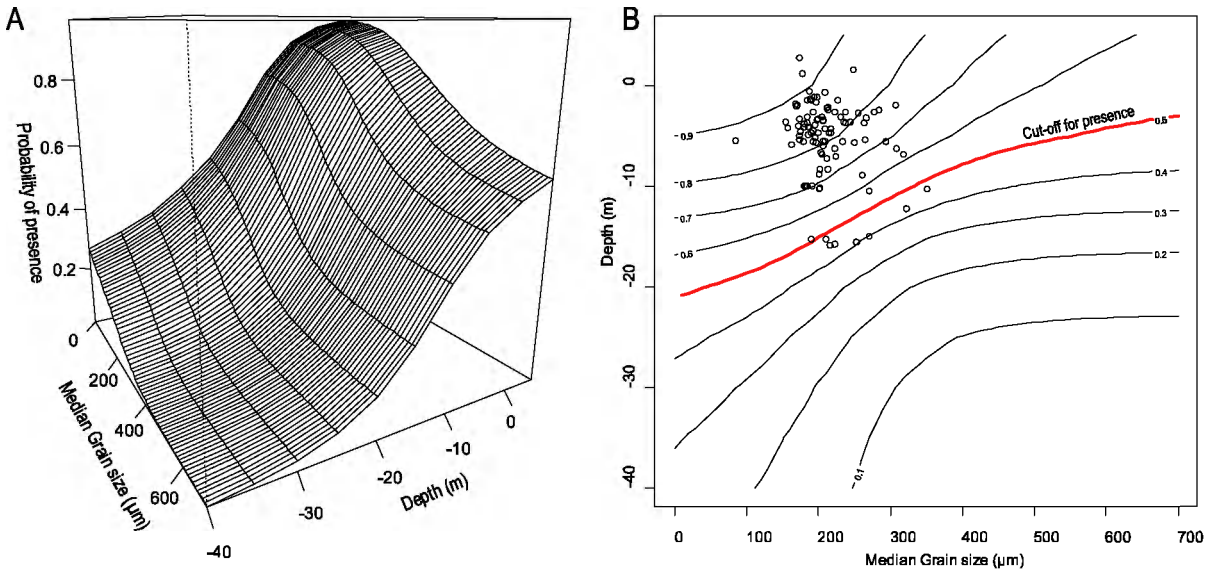

Fig. 4.3. Modelled response of the multimodel prediction of the probability of presence of *D. vittatus*. 4.3A. Visualisation of the modelled response, 4.3B. Contour lines with the probability of presence. Dots indicate the samples where the species was present.

4.3.1.3. Full cover predictions of the species distribution

Full cover grid layers of the variables median grain size and depth are fed into the 500 replica models of each of the four optimal models (Table 4.2, model 1-4). The average predicted probability of presence per sample over each of the 500 replica models is calculated. This results in four predicted probabilities per sample, one for each alternative model (model 1-4). The weighted average over these four predictions per sample is calculated. The result is a full cover prediction of the probability per grid cell (Fig. 4.4). *Donax vittatus* is predicted to be present only close to the coast, because offshore the sediment is generally coarser and the depth increases. The model predictions offshore match well with the observations as the species was never found offshore. In the densely sampled area (Fig. 4.4B), the overall pattern of the predicted distribution (Fig. 4.5), matches quite well with the pattern in the species observations (Fig. 4.1). The predictions could not be validated in the north of the BPNS because no

observations were available there. In the area between Ostend and Zeebrugge only few observations were available for comparison. This area is a relatively muddy area.

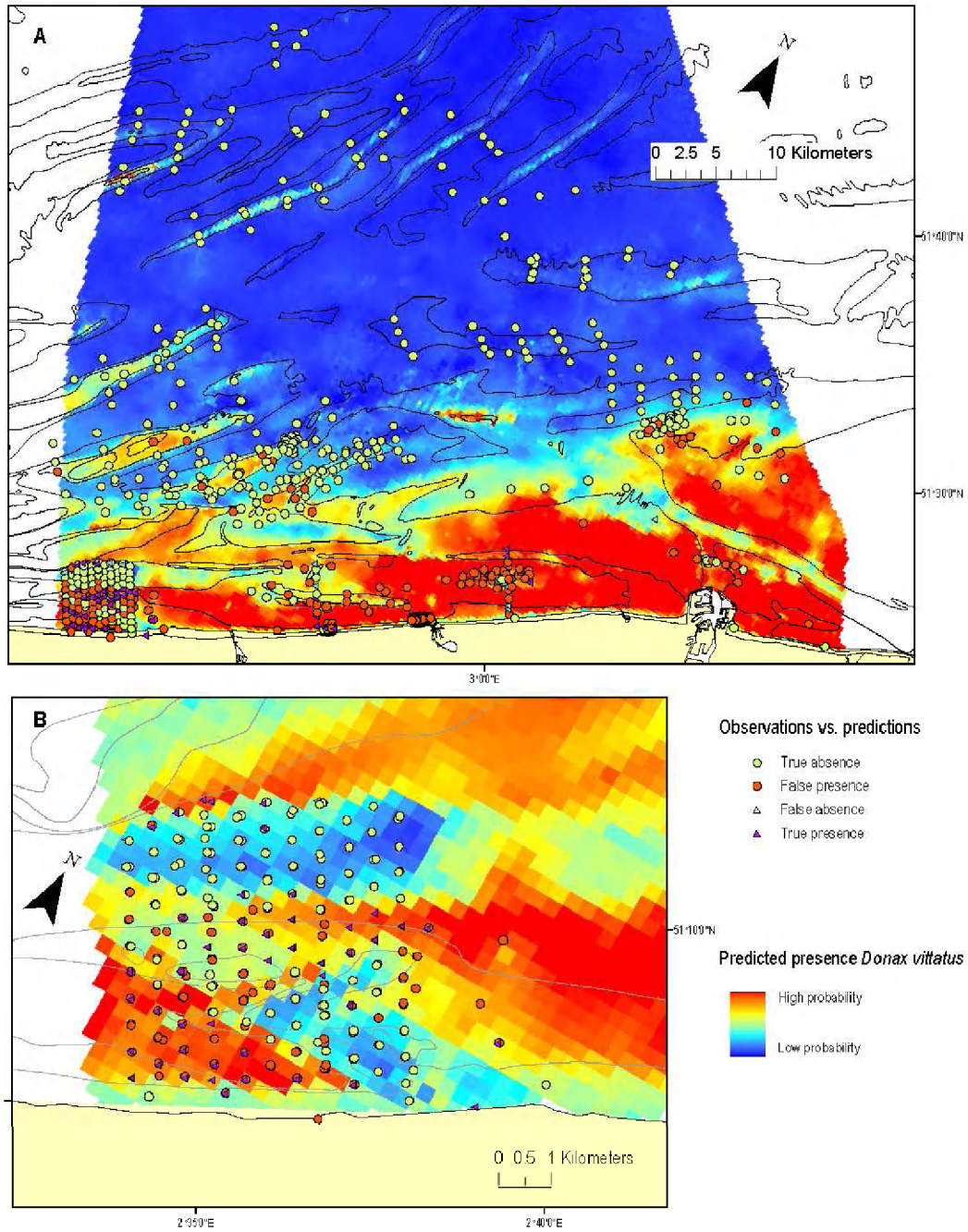

Fig. 4.4. Prediction of the probability of presence of the species *D. vittatus* in the Belgian part of the North Sea based on a multimodel weighted average with the Combined Model Optimisation Criterion (CMOC) as weight. The prediction is based on grid cells of the variables median grain size and depth. Projection UTM 31N. A. Prediction for the Belgian part of the North Sea, B. Prediction for a densely sampled area close to the city of De Panne.

4.3.2. Sampled distribution per variable

Overall the distribution of the complete data set and the absence samples matched well, as the species was absent in 890 of the 990 samples in the complete data set (Fig. 4.5). Per predictive variable the sample distribution is discussed in this section.

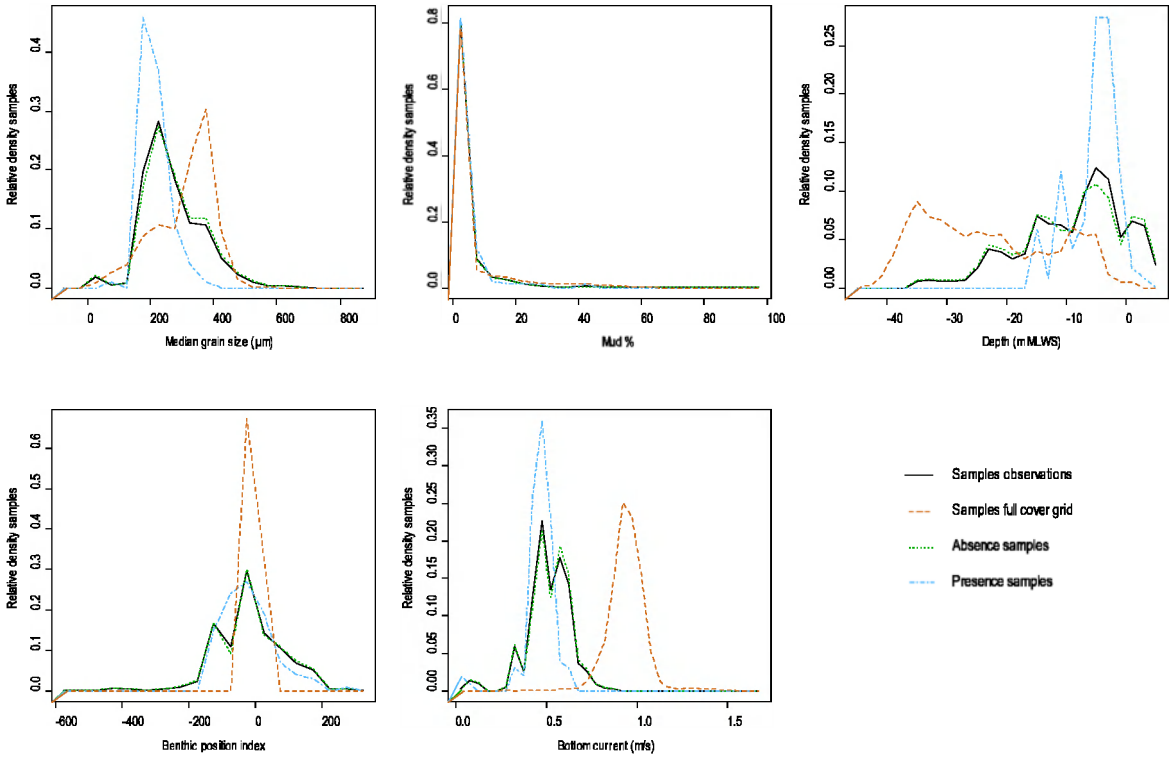

Fig. 4.5. Distribution of the observations over the range of the variables in the data set.

Median grain size

The distributions of presence and absence samples do not fully overlap, the median grain size allows distinguishing them. The peak of the presence sample distribution was found at lower grain sizes. Very few samples were available at very fine or very coarse sediments.

Mud%

The distribution of all samples in the data set, of the presence and absence samples, and of the full cover grid values overlap. The variable mud% does not allow to clearly distinguish between the presence and absence observations in the data set. The full range of the variable mud is sampled, but there are few samples at very high mud percentages.

Depth

The full depth range of the BPNS is sampled, but more samples were collected in the shallower parts. On the beach little or no samples are available. Presence samples have a distinctly different distribution from absence samples, which indicates usefulness of the variable for presence prediction.

BPI

The distribution of all observations, of the presence and absence samples, and of the full cover grid values overlap. The variable is not able to distinguish absence and presence samples for the given data set as these have the same distribution for this variable.

Currents

Samples were collected at locations with low bottom currents, when compared to the full cover grid values. The species was found at all the observed bottom current speeds in the samples, thus this variable is not useful to distinguish presence and absence samples in the available data set.

4.4. Discussion

4.4.1. Habitat suitability models

The four models retained in the hierarchical exhaustive model selection (Table 4.2) are using only two predictive variables, median grain size and depth, although the data set contained five variables. The variables in the final four models, median grain size and depth, will be discussed in detail with an emphasis on the ecological implications of the variables. Further on, the variables in the data set that were not included in the final models, as well as the variables that could be considered for future models for the species, are discussed.

4.4.1.1. Model validation: ecological knowledge

The ecological knowledge on *D. vittatus* obtained from literature was combined in a conceptual scheme (Fig. 4.6). The probability of presence of the species is determined mainly by the local growth and survival of the individuals. The relation between the larval settling and the physical habitat is not well known specifically for this species, but numerous macrobenthic species are known to actively select their habitat during settlement (Snelgrove *et al*., 1999). The local growth of *D. vittatus* is determined by a number of variables: metabolic rate, phytoplankton density, siphon closure time, resurfacing speed, biomass lost by predation, salinity and valve closing energy expenditure. The growth rate is directly related to the metabolic rate, and the latter is influenced by temperature and salinity (Ansell and

Lagardere, 1980). High phytoplankton densities lead to higher growth rates as the species is a suspension feeder (Yonge, 1949; Pohlo, 1969).

The preferred grain size range observed by Stanley (1970) is 150-375 µm. This preferred grain size range is also the range where *D. vittatus* burrows fastest (Ansell and Lagardere, 1980; Alexander *et al.*, 1993; Ansell, 1994). The sediment in this range has a high shear stress and requires the current velocity 0.2 m/s to be eroded (Postma, 1967). The higher shear strength and higher cohesion of these fine sediments in the preferred range are important in helping infaunal bivalves hold their valves closed (Stanley, 1970). Thus less muscle action is needed, which leaves more energy for somatic growth. Coarse sediments with low shear strength and reduced cohesion could increase the energetic costs compared with finer sediments (de la Huz *et al.*, 2002).

If the siphons are closed for longer periods, this decreases the food uptake. Longer closure periods can be due to high mud content that could potentially block the gills. As mud particles stay longer in suspension, the burrowing speed will be lower in a muddy environment, because there will be prolonged siphons closure periods until mud particles settle back to the sediment (Ansell, 1962; Trueman, 1966; Eltringham, 1971). Yonge (1949) argues that *D. vittatus* cannot eliminate the mud particles on its gills fast enough, because the internal anatomy causes the formation of pseudofaeces to be slow in comparison with bivalves that can live in muddy environments (e.g. *Tellina*). The mantle folds and a central channel leading to the waste channel are lacking in *D. vittatus* (Yonge, 1949).

Each species in the *Donax* genus has a specific maximum temperature it can tolerate (Ansell *et al.*, 1980). These maximum temperatures can occur in summer as *D. vittatus* lives in shallow waters (Ansell *et al.*, 1980). The median lethal temperature (LT_{50}) ranges between 24 and 29°C, depending on the prior acclimatisation. Median burial temperature (BT_{50}) is the temperature at which 50% of the individuals retain the ability to burrow is up to 29°C for *D. vittatus* (Ansell *et al.*, 1980). Both the LT_{50} and BT_{50} were significantly lower for individuals acclimatised at lower temperatures (Ansell *et al.*, 1980). Lower salinities decrease the tolerance for high temperatures (Ansell *et al.*, 1980), in agreement with the general principle that fluctuations in the combination of environmental stresses will combine to lower tolerance limits (Kinne, 1970). It is assumed that the lethal temperatures are rarely reached in the field, but that sublethal effects will render populations non-viable (Ansell *et al.*, 1980).

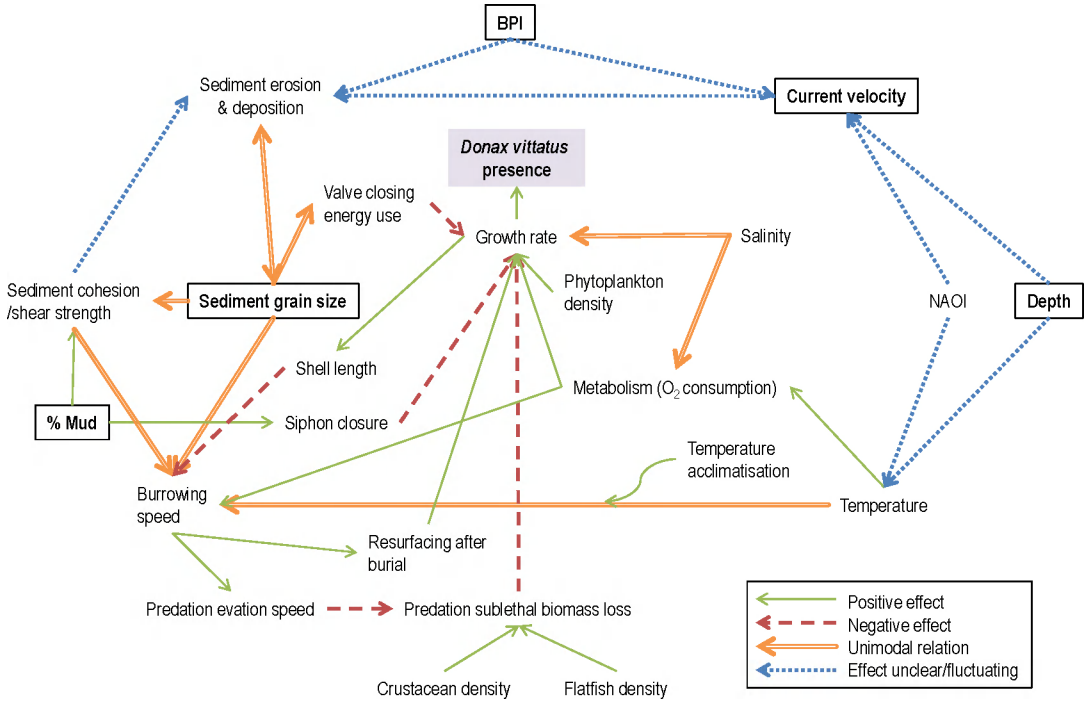

Fig. 4.6. Conceptual scheme of the ecological processes determining the presence of *Donax vittatus*. The boxed variables are in the data set used for prediction. NAOI: North Atlantic Oscillation Index. BPI: Bathymetric Position Index.

The burrowing speed is determined by the bottom shear strength, grain size (Alexander *et al.*, 1993), shell length (de la Huz *et al.*, 2002), temperature (Ansell *et al.*, 1980) and salinity (Ansell *et al.*, 1980). Fast burrowing speeds are important to avoid physical exclusion and burial by waves or current action (Brown and McLachlan, 1990; McLachlan *et al.*, 1995). Bivalves in environments with high current speeds must burrow quickly to resurface after burial. Fast burrowing also implies a higher predation evasion speed which is needed to evade the sublethal foot and siphon predation by crustaceans (Salas *et al.*, 2001) and flatfish (Burrows and Gibson, 1995). Sublethal predation requires regeneration at the expense of somatic growth (Salas *et al.*, 2001). In sediment preference experiments, Gibson and Robb (2000) observed that juvenile plaice had a strong preference for finer sediments, which coincides which the sediment preference of *D. vittatus.*

The North Atlantic Oscillation Index (NAOI) is known to influence temperature, current and wind patterns (Reiss and Kröncke, 2005). Temporal changes of benthic communities have been linked to the NAOI, with an effect probably mediated by rising winter water temperatures (Kröncke *et al.*, 2001; Reiss and Kröncke, 2005; Kröncke *et al.*, *submitted*). Depth has been found to structure benthic communities (Hagberg *et al.*, 2003) and several environmental variables are correlated with depth (e.g. currents, light penetration, water mixing, wave action). As such depth, can act as a proxy for several unmeasured

variables. Pesch *et al.* (2008) used depth to predict the distribution of macrobenthic communities in the North Sea. The relation of the BPI (Lundblad *et al.*, 2006) with macrofauna distribution has not been clearly proven yet. However, the BPI has been used to predict the spatial distribution of cold water corals (Guinan *et al.*, 2008) and rockfish (Iampietro *et al.*, 2008)

4.4.1.2. Experimental model validation: habitat preference experiments

In this research habitat preference experiments from the literature were used to validate the modelled species response for the variable median grain size. The variable sediment grain size was chosen because: it appeared as one of the important variables that determine the distribution of the species (Table 4.2), and experimental results are available for comparison in ecological literature (Alexander *et al.*, 1993; de la Huz *et al.*, 2002). The variable depth was even more important in the predictive models (Table 4.2). However, depth is considered to be a proxy variable that is correlated with several other, often unmeasured, variables. Therefore experiments that test for the isolated effect of the hydrostatic water pressure are not expected to be useful.

The relation between the sediment type and the metabolic and growth rate of *Donax trunculus* was tested in an experimental setup by de la Huz *et al.* (2002). The results for this species were considered for comparison with the model for *D. vittatus* as: 1) both species are very related and have overlap in their distributional range (Ansell and Lagardere, 1980), 2) the sediment preferences based on burrowing experiments (de la Huz *et al.*, 2002) are similar. de la Huz *et al.* (2002) kept individuals in different sediment types at 18°C and fed them with microalgae. The oxygen consumption (µmol O_2 $h^{-1}ind^{-1}$) and the growth (µm day^{-1}) were measured regularly during 88 days. The fluctuations were minor in this period and in this research only the oxygen consumption and growth at day 88 were considered (Fig. 4.7B). The peak in the metabolic rate and growth coincided with the peak in the burrowing rates (Fig. 4.7A, 4.7B), and with the distribution of the presence samples, as well as with the modelled response of the species (Fig. 4.7C). Both the metabolic rate and the growth steeply declined from the optimum of about 200 µm till 750µm and kept a more or less constant level at the coarser grain sizes tested (Fig. 4.7A).

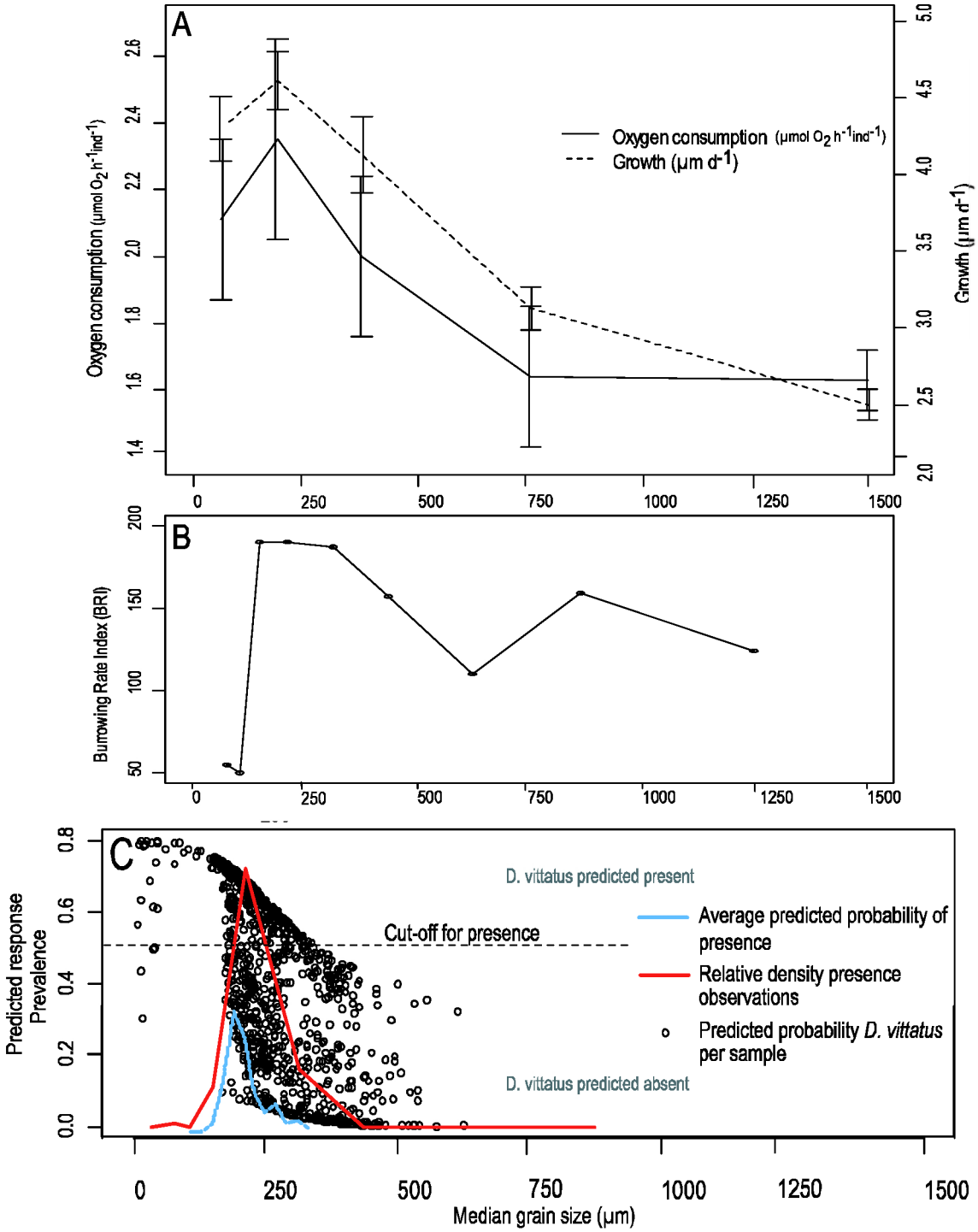

Fig. 4.7. Results of sediment preference experiments for the species *D. vittatus.* A. Metabolism and growth of the species *D. trunculus* for different grain sizes (de la Huz *et al.*, 2002), B. BRI for the species *D. vittatus* (Alexander *et al.*, 1993). BRI = Burrowing Rate Index (Alexander *et al.*, 1993). C. Predicted probability of *D. vittatus* from the multimodel, and the relative density of the presence and absence observations. The average predicted probability of presence for a specific grain size is always below the cut-off for presence because the probability also depends on the depth.

In the burrowing experiments by Alexander *et al.* (1993), the sediment preference of the species is quantified by the burrowing rate, with the highest burrowing rates expected in the most optimal sediment types. High burrow rates are important to quickly resurface after burial (Brown and McLachlan, 1990; McLachlan *et al.*, 1995) and to evade predation of the foot and siphon (Burrows and Gibson, 1995; Salas *et al.*, 2001). The burrowing time is defined as the period between the onset of the burrowing and the moment when the shell completely disappears in the sediment. Because the shell length and mass influences the burrowing speed (de la Huz *et al.*, 2002), the speed per sediment type was converted to the Burrowing Rate Index (BRI; Alexander *et al.*, 1993). The BRI normalises burrowing time for specimen mass (Equation 4.8; Alexander *et al.*, 1993). The burrowing speed of *D. vittatus* was measured by Alexander *et al.* (1993) and the BRI was calculated for several sediment types with half φ intervals (Fig. 4.7B).

$$BRI = \frac{\sqrt[3]{\text{specimen mass (g)}}}{\text{burrowing time (s)}} \cdot 10^4 \qquad (4.8)$$

The optimum of the distribution of the species in field observations (Fig. 4.7C) matches almost exactly with the optimum of the burrowing rate in the experiments by Alexander *et al.* (1993; Fig. 4.7B). The modelled species response and the observed field distribution are zero for grain sizes greater than 400-500 µm (Fig. 4.7C). The experimental BRI results, however, show a decreasing BRI (Fig. 4.7B). Thus the species is still able to burrow in coarser sediments, albeit at much lower speeds. The fundamental niche for the variable burrowing speed is thus wider than the realised niche and the modelled species response based on field observations. In the burrowing experiments a broader optimum was found (125-375 µm; Fig. 4.7B), but the location of the optimum matches quite well with the field observations and with the modelled species response. In the BRI response there appears to be a second local optimum in the BRI at 875 µm, but this is not observed in the growth rate or oxygen consumption, nor in the field observations.

4.4.1.3. Variables in the final models

Median grain size

The importance of median grain size in the determination of the presence of *D. vittatus* is confirmed by the field observations (Fig. 4.5), the burrowing experiment (Fig. 4.7B), the metabolic rate experiments (Fig. 4.7A) and the preferred grain size range in the ecological literature (Stanley, 1970). The species

response to median grain size is modelled in the four final HSMs (Table 4.2) only as a linear response. The probability of presence thus keeps on increasing if the grain size decreases. In field observations, the sediment preference experiments and the preferred range described in the literature (Stanley, 1970), the suitability of the habitat is highest for fine sediments (175-350 µm) and decreases again for very fine, muddy sediments. A unimodal response of the probability of presence for the median grain size is thus observed in the field observations and in the experiments. It can be assumed that the linear response, with no decrease in the probability for low grain sizes, is the result of the skewed distribution of the samples. Few samples were collected in muddy environments, giving less penalty for a bad model fit in muddy environments.

Alexander *et al.* (1993) classified *D. vittatus* as a substrate sensitive species which penetrates a wider range of grain sizes than substrate specialists, but burrowing speeds are lower in coarser sediments compared to generalists. The modelled species response and observed field distribution are zero for grain sizes greater than 400-500 µm (Fig. 4.7C), but few samples are available in these coarser sediments. The experimental burrowing speed, metabolic rate and growth rate results (Fig. 4.7A, 4.7B), however, show low values for coarser grain sizes, but are not zero. Thus the species is still able to burrow and to grow in coarser sediments, albeit at much lower rates. The fundamental niche for the variable median grain size, based on experimental results on the burrowing speed, metabolic rate and growth is thus wider than the observed realised niche in the field observations.

The grain size determines the distribution of *D. vittatus* in different ways:

1) The grain size controls the burrowing rate of the species. In general, organisms tend to occupy sediments where they can burrow faster (Alexander *et al.*, 1993). The ability of an animal to obtain an anchorage to pull the shell downwards plays an important role in the burrowing process and various grain sizes are related to differences in foot anchorage (Trueman *et al.*, 1966). Fast burrowing speeds are important to avoid physical exclusion and burial by waves or current action (Brown and McLachlan, 1990; McLachlan *et al.*, 1995). Fast burrowing also implies a higher predation evasion speed which is needed to evade the sublethal foot and siphon predation by crustacean (Salas *et al.*, 2001) and flatfish (Burrows and Gibson, 1995). Less energy loss through less predation in sediments where the burrowing speed is highest, results in more energy for somatic growth (Salas *et al.*, 2001).

2) The grain size determines the energy needed for holding the valves together. The higher shear strength and higher cohesion of fine sediments in the preferred range are important for infaunal bivalves to reduce the energy cost to hold their valves closed (Stanley, 1970). Because less muscle action is needed in the grain size range 175-350 µm, more energy is left for somatic growth (de la Huz *et al.*, 2002).

3) Very fine sediments clog the gills of the species. In the field observations, the species *D. vittatus* was never found in median grain sizes below 84 µm. The inability to deal with high concentrations of fine particles in the water can be explained by the different internal anatomy of the species compared to other bivalves, which causes less pseudofaeces to be formed. If the siphons remain closed due to high concentrations of fine particles, filtering for food particles is also stopped. Longer siphon closure periods in the finest sediments, cause a lower overall energy uptake (Ansell, 1962; Trueman *et al.*, 1966).

Depth

Depth is the most important variable in the multimodel prediction of the distribution of *D. vittatus,* because the variable was used in each of the selected models and the model which only uses depth has the highest CMOC (model 1; Table 4.2). The importance of depth can clearly be seen in the different distribution of the samples where the species was present or absent along the depth range (Fig. 4.5). Depth is a predictive variable in each of the selected HSMs (Table 4.2). In three out of the four models a unimodal response of the probability for presence for depth is modelled. Although a unimodal relation is modelled, the multimodel averaged probability of presence is not low for locations high up on the beach (>0 m). This can be explained because very few samples were collected high up on the beach, giving less penalty for a bad model fit in these locations. But as little samples were available for model development in the range >0m, the model can only reliably be used in the range [-40m, 0m].

Although depth has been demonstrated earlier to have an effect on the distribution of macrofauna species in the North Sea (Basford *et al.*, 1990; Ysebaert *et al.*, 2002; Willems *et al.*, 2007), experiments are difficult to set up as depth is a proxy for numerous other environmental variables. As no experiments were available, a distinction between the experimental fundamental depth range and the realised depth range from observations was not possible.

Depth is a proxy for the distance to the coast in the BPNS, and several relevant predictive variables also show an onshore-offshore gradient:

1) Bottom currents increase with depth (r = -0.61, Kendall's tau correlation). In the BPNS the bottom currents increase with the distance from the coast (Luyten *et al.*, 2003). The depth increases as well away from the coast. Depth is used as a variable in the hydrodynamic model from which the current speeds are obtained (Luyten *et al.*, 2003). In the Southern North Sea, bottom currents determine via the bed shear stress erosion and deposition (Stanev *et al.*, 2009), both processes which can alter the grain size by erosion or deposition of certain grain sizes. Sediment transport can be a key structuring

factor in infaunal communities, as sediment resuspension and deposition changes the local habitat, and hence the suitability of the habitat (Desroy *et al.*, 2007). Currents also influence the dispersion of postlarval stages (Butman, 1987; Commito *et al.*, 1995), and in that way determine the dispersion limitation of the species, as the small postlarvae cannot swim against the currents.

2) Phytoplankton concentrations are higher in shallow waters. On the BPNS, depth is quite well correlated with phytoplankton densities (Peters *et al.*, 2005), which are highest in shallow water close to the coast. Suspended particulate matter is also highest closest to the coast (Fettweis *et al.*, 2007), where the depth is low. Due to the shallowness of the BPNS, microphytobenthos is also expected to contribute to the primary production close to shore.

4.4.1.4. Variables not included in the models

Mud%, BPI and bottom currents

Three variables were in the data set of the BPNS, but were not used in the selected models: mud%, BPI and bottom current. The mud% was not chosen because this variable could not make a clear distinction between the presence and absence samples (Fig. 4.5), and because the mud% is correlated (r = -0.29) with the median grain size which was retained during the model selection. The BPI showed no clear relation with the distribution of *D. vittatus.* The bottom current is correlated with the median grain size (r = 0.31) and the depth (r = -0.61), and these variables were already in the final models. So redundancy of the variable currents could be the reason this variable was not selected in the model selection.

Other variables

The water temperature would be a useful variable if HSM would be developed at the scale of the North–East Atlantic region. In the BPNS the spatial variance of the temperature was too low. On a large spatial scale the temperature varies significantly and it would be useful to model the temperature-based limitation of the geographical range of *D. vittatus.* The spatial distribution is limited due to the (sub)lethal effect of the temperature on the general metabolic rate and on the burrowing speed, as burrowing decreases with too high and too low temperatures (Ansell *et al.*, 1980).

The NAOI is known to influence temperature, currents and wind patterns (Reiss and Kröncke, 2005). Temporal changes of benthic communities have been linked to the NAOI, with an effect probably mediated by rising winter water temperatures (Kröncke *et al.*, 2001; Reiss and Kröncke, 2005; Kröncke *et al.*, *submitted*). However, to test the NAOI as a predictive variable, repeated samplings of the

locations would be needed. The opportunistic sample compilation in the data set used in this research is not suitable, as it is highly clustered in space and in time.

The inclusion of the variable salinity in the HSMs would be useful if the HMS is to be calibrated and applied to estuarine as well as marine samples (Ansell *et al.*, 1980; Ysebaert *et al.*, 2002). The physiological stress posed by salinity on the growth and metabolic rate limits the distribution of the species in low salinity waters.

Predator densities can be used to model the biomass and energy loss due to sublethal predation. The interaction with burrowing speed needs to be considered, as the burrowing speed determines how quick *D. vittatus* can avoid predators (Burrows and Gibson, 1995; Salas *et al.*, 2001). The major challenge when predator densities are included is that these species are often highly mobile (e.g. juvenile plaice; Burrows and Gibson, 1995), and detailed spatial distribution information is lacking.

Beside the presence and density of *D. vittatus,* the shell length would be very valuable biological information to be collected in future surveys. The shell length affects the burrowing speed (Alexander *et al.*, 1993), and thus each shell length has a slightly different optimal grain size. Ansell and Lagardere (1980) collected shell length information and age rings on the shells of 100 individuals at different sampling locations during different times of the year. This information allowed modelling the growth and size distribution for different periods and locations, by using the Von Bertalanffy growth equation (Ansell and Lagardere, 1980). The combination of age rings and shell length allowed following the growth, recruitment and mortality of a year class. Shell length information would also allow to test the observations by Ansell and Lagardere (1980), who observed that the smallest individuals are found the highest on the beach. de la Huz *et al.* (2002) also made similar observations for the species *D. trunculus*. In this context, the swash exclusion hypothesis states that the individuals which can burrow the quickest, are the best able to withstand the more reflective morphodynamics higher up on the beach (de la Huz *et al.*, 2002). Their burrowing time is the smallest relative to the wave period.

4.4.2. Integrated model validation

4.4.2.1. Ecological scheme

A conceptual scheme allows identifying both the interrelation of variables and how each variable influences the suitability of the habitat in a direct or indirect way. Creating such a scheme will: 1) force researchers to think about the possible causal relation and the direction of the relation between variables and/or the species, 2) help to determine which variables need to be measured (Gibson, 1994) and included in an optimal model (Austin, 2007) and, 3) help to determine which important variables are missing in the data set. For each variable in a data set, the intermediate variables that link a variable

with presence of the species can be determined with the conceptual scheme. As such, the scheme can combine all literature and experimental knowledge on a species, and makes this knowledge available for model validation.

4.4.2.2. Experiments

Several arguments support the use of habitat preference experiments for HSM model validation:

1) Experiments allow proving the causal relation of a variable with the presence of the species. Experiments can thus distinguish proxy variables and causal variables.

2) Experiments can test if each variable in the HSM determines the probability of presence of the species, when tested in isolation of other variables and in the absence of biological interactions. Field observations of the species response are always multivariate observations.

3) Only experiments allow to delimit the fundamental niche of a species for a variable. Some variables are known to have a causal relation with the probability of presence of the species, but in experiments the variable range suitable for the species will often be broader compared to the realised range in field observations. Knowledge of the fundamental niche can be useful in some model applications. For example, it is useful to know the fundamental niche of invasive species to predict areas suitable for colonisation. In these colonised regions abiotic interactions are often less limiting on the fundamental niche, as their predators have not moved along with them (Rodder and Lotters, 2009).

Model selection based on field observations in a data set could be replaced by experimental testing for a set of predictive variables, if they determine the probability of presence of the species. However there are several reasons why a variable selection purely based on experiments would not be beneficial:

1) Each predictive variable needs to be testable experimentally. Some variable that can be tested in experiments are not available in field measurements and vice versa.

2) Experiments are expensive and time consuming while field observations of species and environmental variables are often already available.

3) Experiments provide knowledge on the fundamental niche, while knowledge on the realised niche is more relevant for most model applications.

4.4.2.3. Predictive variables in the selected models

Predictive variables that are used in a model that is calibrated with presence and absence observations are chosen in the variable selection process primarily if they enable the model to distinguish between samples where the species was present or absent. Variables that have a known causal relation with the

distribution of the species can sometimes not be chosen during the variable selection due to several reasons:

1) The variable cannot discriminate presence and absence samples in the range measured. Ideally, sampling for HSMs would be thoroughly planned stratified sampling, with the aim of sampling the whole range of each variable (Araujo and Guisan, 2006). If only a part of the environmental range is measured, it is possible that the species is present along this limited range with an equal probability (e.g. BPI, Fig. 4.5). Under these circumstances, the variable is not useful to distinguish presence and absence samples. For example, the samples where the species was observed to be present and absent had a different distribution for the variables median grain size and depth (Fig. 4.5). These variables were able to discriminate the presence and absence samples well and were also selected in the model selection to be included in the predictive HSMs.

2) The variable is correlated with other predictive variables. If two or more predictive variables in a data set are correlated, the model selection (stepwise or CMOC) will mostly select only one variable for inclusion in the final HSM. For example, the variable mud% has a causal relation with the probability of presence of *D. vittatus* based on the literature, experiments (Fig. 5.6), and field observations (Fig. 4.5), but this variable was correlated with the median grain size. An assessment of the interrelations of the variables in the data set should be performed prior to the model selection. Such an assessment can be based on an analysis of the interrelation of the predictive variables (e.g. principle components analysis and correlation analysis) and on the conceptual scheme, where the interdependency of the variables is assessed based on ecological literature.

3) The variance of the variable is too low. Each predictive variable has a spatial and temporal scale on which it has the most variance. In the BPNS, for example, water temperature will not differ significantly on a meter scale. Salinity can be an important predictor of the presence of macrofauna (Ysebaert *et al.*, 2002), but in the sampled non-estuarine locations salinity is expected to have only minor fluctuations.

4.5. Conclusions

- Integrated validation of HSMs should consider 1) the model validation of observations vs. predictions, 2) ecological knowledge from the literature, 3) habitat preference experiments and, 4) the distribution of the samples over the range of the predictive variables.
- The current knowledge on the species from ecological literature should be combined in a conceptual scheme to allow visualisation of the variable interrelations. Such a scheme can help distinguishing proxy and causal variables.

- Habitat preference experiments allow identifying the fundamental niche for each variable, while field observations can only provide insight in the realised niche. If habitat preference experiments are available, they should be used to assess the causality of variables in the prediction of the species distribution. If no experiments are available, the feasibility of experiments should be considered.
- The distribution of samples over the range of each variable should be compared. In this way it is possible to determine if the sampled range is sufficiently sampled and if a variable is correlated to the presence of the species.
- In a multimodel approach, HSMs were developed for the species *D. vittatus*. In the model selection based on the CMOC four models were retained that use the depth and median grain size as predictive variables. In an integrated discussion the ecological relevance of these variables was discussed. Experimental validation confirmed the modelled sediment grain size response.

Acknowledgements

Special thanks go out to all the contributors to the Macrodat data base. The Management Unit of the North Sea Mathematical Models (MUMM) provided the current speed data. The Renard Centre for Marine Geology (RCMG, UGent) provided the full cover median grain size, depth and bathymetric position index grids. The first author was grant holder of the Institute for the promotion of Innovation through Science and Technology in Flanders (IWT).

Chapter 5. General discussion and conclusion

As the application of HSMs in marine sciences is increasing (Elith and Leathwick, 2009), the need for a sound modelling methodology arises. In the introductory chapter of this thesis (Chapter 1), a state of the art of HSMs was provided. Based on this state of the art, specific objectives were identified for further research in this thesis (Chapter 1). In summary, the methodological research focused on three specific challenges: 1) the choice of modelling techniques, 2) model selection algorithms and 3) integrated model validation. The discussion is also organised around these three challenges. After this, comes a future outlook of habitat modelling. This discussion includes an analysis of the factors determining errors in the predictions of HSM's for marine benthic species. And finally, a discussion on the application of HSMs for macrobenthos species is provided.

5.1. Challenge 1: Which modelling technique to use?

The modelling techniques used in HSMs quantify the relation between the predictive variables and the density of a species or the probability that a species is present. The majority of the HSMs rely on statistical techniques that use the correlation between the predictive variables and the response of the species (Kearney and Porter, 2009). Statistical techniques estimate a number of model parameters based on a calibration data set. In a broad sense neural networks can also be regarded as a statistical technique, as parameters are estimated from a data set. However, a sample distribution is not assumed and no significance testing is done with neural networks. Because of the flexibility of ANNs, they can emulate a logistic regression. When an ANN with one interneuron (identity transfer function) and one output neuron (logistic transfer function) is developed, it produces similar predictions as a LR (see Appendix II. GLM)

In this thesis a comparison was made between Logistic Regression (LR) and Artificial Neural Networks (ANNs) in their ability to predict the spatial distribution of *Lanice conchilega* (Chapter 2). LR and ANNs are in this context used to predict the probability of presence [0-1] and are calibrated with a data set containing presence and absence information (0, 1). Both techniques thus model proportions: the proportion of samples where the species is expected to be present for a given combination of variables. This proportion can also be interpreted as the chance or probability to find a species for a given combination of variables.

ANNs gave a higher predictive performance, based on a number of performance indicators (CCI, AUC, specificity and sensitivity). However, there was a high correlation between the model output of the LR and the ANN. ANNs can be powerful predictors, and the high predictive performance of ANNs

can be explained by the fact that complex functions and variable interaction are inherent to the architecture of the ANN, because of the connections between the neurons. Another advantage of ANNs is that there is a dimension reduction during the modelling process if the number of interneurons is lower than the number of input neurons (Tibshirani, 1994). This is useful if several predictive variables are available. When the number of variables increases, however, the number of model parameters to be estimated increases exponentially. A high number of samples is necessary, otherwise the parameter estimates become unreliable.

ANNs can fit very complex species responses, but this can easily lead to overfitted models, that are not transferable to other regions or periods. In the *Lanice* example in this thesis, there was quite a high dissimilarity between the ANNs produced for each of the three crossvalidation folds. The LRs for the folds were more similar among each other, which points out the higher robustness of LRs. Presently, there is no established theory to determine the number of interneurons or the choice of the transfer functions for ANNs. Also, as no error distributions are assumed, no statistical test are performed, and hence the model output or model parameters are not tested for significance.

From a parsimonious point of view, LRs were superior in the modelling of the response of *Lanice*, as the model was simpler and the ANN only performed slightly better. The advantage is that LRs are a popular technique (Schroder, 2008), with a sound statistical methodology. The significance of the overall model and of the model parameters can be tested easily with LRs. Also the model parameters are directly interpretable, especially if the predictive variables are standardised.

5.2. Challenge 2: What is the most optimal combination of predictive variables? The Combined Model Optimisation Criterion

Model selection is considered a central step in the development of HSMs (Guisan and Zimmermann, 2000; Heikkinen *et al.*, 2006; Franklin and Miller, 2009). Stepwise model selection is most often used in habitat suitability modelling (e.g. Attrill *et al.*, 1999; McBreen *et al.*, 2008), but this approach has a number of disadvantages (see Chapter 1). Therefore, the Combined Model Optimisation Criterion (CMOC) was proposed in this thesis. The CMOC is based on the commonly used framework of the Model Optimisation Criteria. Via the χ^2 distribution the significance level α used in the F-statistic, can be transformed into a λ value. As such, the commonly used F-statistic could also be fitted into the MOC framework. In the CMOC approach, a MOC is calculated for the calibration set and for the test or validation set, depending on the availability of independent data. The properties and advantages of the

CMOC model selection approach, compared to the stepwise model selection approach are discussed below.

5.2.1. The selected models are parsimonious

Because MOCs have a term for the model fit, based on the likelihood, and a penalisation term, based on the number of model parameters, the CMOC approach selects parsimonious models. The selected models are a trade-off between good model fit and low model complexity. The term λ determines in which direction the trade-off will go. A high λ will select simpler models, as the penalisation for model complexity is higher. The concept of model parsimony is linked to the concept of model generalisation. A properly parsimonious model will have higher generalisation ability compared to an over- or underfitted model.

5.2.2. An exhaustive comparison of all alternative models

In the stepwise model selection approach there is only a one-by-one comparison during each step in the model selection. Of all the possible models, the globally most optimal can be chosen in the exhaustive approach, while stepwise approaches can get stuck in local optima (Reineking and Schroder, 2006). The stepwise approach becomes instable when multicollinearity is present, because one of the two correlated variables is chosen for inclusion or removal into the model (Prost *et al.*, 2008). Multicollinearity is not a problem for CMOC, because all variable combinations are calculated separately, and not in sequence as in a stepwise approach. The data generating model (or a very similar model) was found in the virtual species approach (Chapter 3), even if correlated variables were used to generate the virtual species data (e.g. bottom current and depth: r = -0.61).

In the full exhaustive model selection as described in Chapter 3, all possible alternative models are compared by means of their CMOC values. As the number of possible alternative models rises quickly, the hierarchical model building approach (Kutner *et al.*, 2005) was used in Chapter 4. In the hierarchic model the model is built with sequentially more complex terms. A clear advantage is that the number of possible alternative models to be compared remains lower.

5.2.3. The CMOC approach is robust

A single observation is only a snapshot of a dynamical situation and can only give a partial and instantaneous observation of the species-environment relations (Hirzel *et al.,* 2001). Therefore, bootstrap resampling has been used to increase the robustness of the model selection (Araujo and

Guisan, 2006; Prost *et al.*, 2008). In the CMOC approach, bootstrapping increased the robustness because the mean MOC of the numerous replica models was taken, instead of the MOC for one single model.

5.2.4. The CMOC approach uses information theory-based measures

The CMOC uses information theory-based measures (e.g. AIC, BIC) that can be described within the model optimisation framework (Reineking and Schroder, 2006). However, the predictive performance of most HSMs, is assessed with contingency table-based measures (e.g. CCI, Kappa, NMI; Fielding and Bell, 1997; Guisan and Thuiller, 2005).

A first disadvantage of using a contingency table based indicator, in comparison with the MOCs, is the requirement to choose an arbitrary cut-off for presence to convert probabilistic predictions to presence/absence for the contingency table (Barry and Elith, 2006). Barry and Elith (2006) therefore argue that it's better to consider directly the discrepancy between the model and the actual observations. In the MOCs this is accomplished by the calculation of the likelihood of the model given the observations.

A second disadvantage of contingency table-based indicators is the lack of compensation for model complexity. More complex models will always have a better fit, and MOCs use a penalisation term to compensate for this (Reineking and Schroder, 2006). In the virtual species analysis of the CMOC (Chapter 3), the contingency based indicators were only able to find the true model in case this true model was very complex (TM5). In case the model was simpler (TM1 and TM3) there was little match between the rankings based on the CMOCs and the contingency based indicators. AUC suffers from the same disadvantage as there is no compensation for model complexity, but this indicator does not require a cut-off for presence to be chosen (Lobo *et al.*, 2008).

6.2.5. The CMOC model selection is consistent

The model selection is consistent when the true model is found among a set of all possible alternative models, in case the number of samples n goes to infinity (Anderson *et al.*, 1998). Consistency is desirable when MOCs are applied to large data sets (Shono, 2005). The CMOC approach is consistent when it is based on consistent MOCs, such as the BIC and the CAIC. For both these consistent MOCs the regularisation parameter λ is not fixed, but a function of the number of samples n (Table 3.1). The justification to allow more complex models is that greater sample sizes bring on more information on the species-environment relation. More time and area effects may thus be revealed by the data when the number of samples increases (Anderson *et al.*, 1998). Consistent MOCs are thus to be preferred for the

selection of models developed with biological observation. Because the true model has an infinite complexity and therefore the models chosen by the MOC should also get more complex if the number of samples goes to infinity. In the virtual species analysis, the CMOCs based on the consistent MOCs (CAIC and BIC) did a better model selection for all complexities of the true model. A clear relation between the λ and the weighted average of the number of variables over all models was observed for the virtual species (Chapter 3; Fig. 3.7).

5.2.6. Independency of the species prevalence

The prevalence of the species in the calibration set will have a major influence on the fitted model response and the reliability for models calibrated with presence/absence observations. Therefore, the prevalence of the species is compensated in the CMOC approach.

When p/a observations are available the prevalence can be calculated. When a grid-based, exhaustive survey of the areas is available, the Relative Occurrence Area (ROA, Lobo *et al.*, 2008) can also be calculated. The ROA is the ratio between the surface of the grid cells where the species is present to the surface of the whole region (Lobo *et al.*, 2008). The ROA is a measure of the true rareness of the species, while the prevalence is a measure of the rareness of the species in a given data set (Jimenez-Valverde and Lobo, 2006). The prevalence is sometimes mentioned in HSM studies, but the true rareness of the species is seldomly known; still, mentioning of the rareness would greatly increase the comparability between HSMs (Jiménez-Valverde *et al.*, 2009).

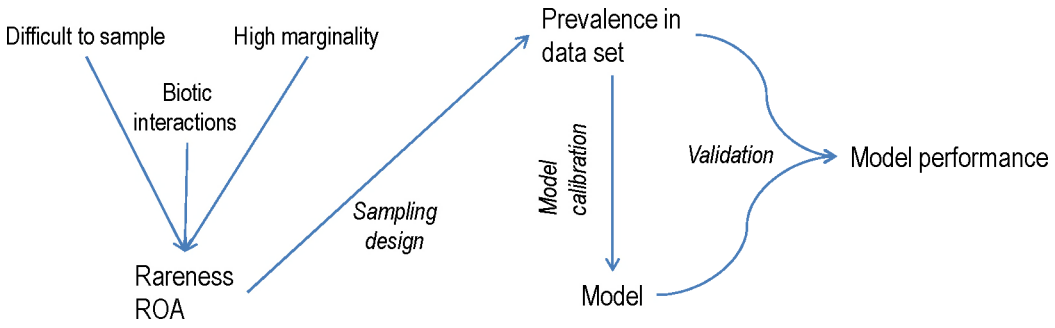

Fig. 5.1. Relation between the rareness and Relative Occurrence Area (ROA, Lobo *et al.*, 2008), the sampling and the model performance for habitat suitability models.

The interaction between the true rareness (i.e. the ROA), the sampling design, the prevalence and the model performance is visualised in a scheme (Fig. 5.1). Several causes can determine the rarity of a species in a region. Some species are difficult to sample and observe, e.g. mobile and cryptic species are difficult to detect, and tend to be underestimated by common field survey techniques.

Another reason of species rarity may be, that the species has a high marginality. This means that species have a narrow preferred range for one or more environmental variables. The marginality is thus the degree of departure of the environmental conditions where the species is present, from the mean environmental conditions in the study region (Hirzel *et al.*, 2002). Thus when the species prefers a very specific habitat and this habitat is rare, this might result in rarity of the species in the area. Marginality can be observed by looking at the distribution of samples where the species is present and absent over the range of each environmental variable. In chapter 4 of this thesis the species *Donax vittatus* showed a moderate marginality for the variables median grain size and depth, while for the other variables there was little marginality. High marginality for a variable is a good thing because it allows variables to distinguish between suitable and unsuitable locations. On the other hand, a species with a high marginality and thus narrow preference range, is more likely to be rare in a region, and thus difficult to observe representatively. Depending on the sampling design, a high rarity/low ROA of a species will result in a low prevalence. In a spatially random design, the species rareness in a region will result in a low prevalence. A stratified survey, where each subhabitat is sampled evenly can lead to a more balanced prevalence, closer to 50%. Ultimately, a sampling design directed to find a certain rare species, based on known habitat preferences (Guisan *et al.*, 2006a), can make rareness and prevalence independent.

After model calibration, a model is validated, this is mostly done by comparing observations and predictions. Modelling techniques used to model p/a data, e.g. logistic regression, model the proportion or probability of occurrence of a species for a given combination of environmental variables. As such the species prevalence will greatly influence the estimated model parameters. The fitted model parameters, rather than reflecting the true rareness of the species, will reflect the overall ratio between the presences and absences in the data set (Barry and Elith, 2006). For example, Barry and Elith (2006) observed that increasing the number of absences reduced the average of the fitted probabilities (Barry and Elith, 2006).

In the CMOC approach, the species prevalence is kept at 50% during bootstrap resampling, because this provides the best trade-off between omission (false absence) and commission (false presence) errors for the LR models (Liu *et al.*, 2005). This is achieved by stratified bootstrap resampling. A prevalence of 50% also allows a more objective comparison of the contingency based model performance criteria kappa and NMI. The advantage of resampling to a prevalence of 50%, is that the modelled response and reliability are less dependent on the factors that determine the prevalence: the sampling design and the rareness of the species.

5.2.7. Maximisation of the model transferability by integration of the model validation

In the proposed CMOC approach the model validation is integrated in the model selection. This is achieved by incorporating the MOC calculated on a test set or on a truly independent validation set. This is a major improvement over the current model selection approach, where only the "best" model is validated with the calibration set (see 5.3.). The incorporation of the model performance on independent data in the model selection process greatly improves the model selection methodology.

By choosing a value for the parameter *m* in the CMOC formula, a higher weight can be given to the calibration set MOC or to the test/validation set MOC. In this way the model selection can be guided to choose a model more suitable for interpolation (higher fit for calibration data), or rather for extrapolation (higher fit for validation data). A high potential for extrapolation, will allow the model to be more reliably transferred to other regions.

5.2.8. The level of automation and expert input can be chosen

Automated model selection should not be seen as a substitution for pre-selecting relevant predictors, based on knowledge of the ecology of species (Araujo and Guisan, 2006). Therefore, it is advisable to perform a first selection of predictive variables based on: 1) the sampled range and spatial coverage, 2) habitat preference experiments if available and, 3) literature and expert knowledge. This limited set of potential predictive variables can be fed into the CMOC approach to determine the most optimal variable combination and modelled response (e.g. linear or unimodal response). A limited set of variable will also limit the number of alternative models to be compared in the exhaustive CMOC approach.

A modeller can thus choose between a fully automated approach where all variables are left in the data set and the CMOC approach selects the most optimal combination, or a semiautomated approach where some decisions can to be taken to include certain variables or not, before the CMOC model selection.

5.2.9. The CMOC approach is generally applicable

The MOCs used in the CMOC approach require the likelihood of a model to be calculated, as well as a measure for the complexity of the model (i.e. the number of parameters). As such the CMOC approach can be applied to each modelling technique for which a likelihood can be calculated. For GLMs, likelihood calculations are possible, both if the model output is a probability of presence [0-1], or a

density (continuous output). Also GAMs and even ANNs can have the likelihood of the model given the data calculated. The CMOC can thus be used for a much wider array of modelling techniques.

The CMOC can be used for modelling techniques that predict the probability of presence, species densities or abundance classes, as long as a likelihood can be calculated with the modelling technique used. GLMs, for example, can predict both densities, probabilities of presence and densities classes. For all these model outputs a likelihood, and thus a CMOC, can be calculated for the GLMs.

5.2.10. Multimodel prediction is possible

The traditional model selection concept of selecting one "best" model is not really realistic for biological observations as it assumes that the variables not included in this model are irrelevant (Burnham and Anderson, 2004). Also, the alternative models compared in a model selection can be so variable as to compromise their usefulness for guiding policy decisions (Araújo and New, 2007). Therefore, a shift in the concept of model selection took place which reduced the reliance on one "best" and true model (Whittingham *et al.*, 2006). As several alternative models can often describe the species-environment relation equally well, the inference and prediction of the species distribution will be based on several models (Olden and Jackson, 2002a; Segurado and Araujo, 2004; Whittingham *et al.*, 2006).

Multimodel approaches consider all possible models and thus require a weighing term that expresses the relative importance of each model. In the CMOC approach this value, is the CMOC value of each model. As such, the CMOC is ideal to calculate the weighted mean prediction for a sampling point over all alternative models.

5.2.11. The CMOC approach has been tested with a virtual species

The use of the simulated data was very useful to test the CMOC approach, as simulated data allowed to control the complexity of the species-environment relation and the relative contribution of each predictive variable. This proved an objective benchmark to test if the CMOC would effectively find the true data generating model among all alternative models. In such a context simulated species observations have been used previously with success (Hirzel *et al.*, 2001; Austin *et al.*, 2006; Meynard and Quinn, 2007).

5.3. Challenge 3: Are the model predictions reliable? Integrated model validation

There is a disproportionally large effort in developing HSM models, compared to the validation of the models (Eastwood *et al.*, 2003). Therefore, the validation methodology of HSMs has been assessed and improved in this thesis. Three types of model validation were applied in this thesis. The HSM predicting the distribution of *Lanice conchilega* (Chapter 2) is validated with a traditional validation by comparing the observations and predictions in a threefold crossvalidation. The CMOC approach (Chapter 3) goes one step further and integrates the validation into the model selection process. Finally, the *Donax vittatus* HSM (Chapter 4) also integrates ecological insights from literature, and habitat preference experiments into the model validation.

5.3.1. Traditional validation of HSMs

The traditional way HSMs are validated is by comparing the predictions of a single "best" model with the observations. Ideally, the type of model validation and the stringency for rejecting a model should be determined by the type of prediction of the HSM (Redfern *et al.*, 2006). In case of interpolation, the predictions are made to new sites within the spatial range of the samples in the calibration set and also within the same time frame (Elith and Leathwick, 2009). This approach can be regarded as model-based interpolation to unsampled sites (Elith and Leathwick, 2009). Interpolation is thus performed in HSM applications where one data set is available, and the HSM is used to obtain information on the species distribution in between the samples or even at a full cover scale. The predictive variables in an interpolation are not necessarily causal, but mostly have a high resolution and full coverage of the region (Hirzel and Le Lay, 2008). It is important that species-environment correlations are stable across the region, in case such non-causal predictive variables are used (Elith and Leathwick, 2009).

Another type of model prediction is model extrapolation, where a HSM is transferred to another region and/or period (Elith and Leathwick, 2009). As the environment where the HSM is used differs, different combinations of environmental predictors may be present compared to the model calibration set. Extrapolation requires more causal environmental variables to be used, as they rely less on local correlations. Hindcasting (e.g. biogeographic research) or forecasting (e.g. climate change and invasive species models) are forms of extrapolation in time and also require special care (Elith and Leathwick, 2009).

For both interpolation and extrapolation there is a matching internal or external model validation procedure (Hirzel and Le Lay, 2008). In case of interpolation there is only one data set and an internal

model validation is performed. Mostly this data set is split and one part is used as calibration set to test the model, while the other part is the test set. But the test set used in the internal validation is only pseudo-independent from the calibration set, as they are both part of the same data set. This pseudo-independence will cause an overestimation of the predictive performance (Beutel *et al.*, 1999), but as the goal is interpolation between the samples in the calibration set, this is less of a problem.

Model extrapolation is mostly validated with an external validation. External validation uses an independent data set for validation that is collected at another point in time (Iampietro *et al.*, 2005), at another location (e.g. Clark *et al.*, 2004), or both (Francis *et al.*, 2005). External validation is linked to the concepts model generalisation and model transferability. A model can reliably be transferred to another period and/or region, if it is validated with data from this period and/or region in an external validation.

In this thesis, the HSM for the species *Lanice conchilega* (Chapter 2) was validated with an internal validation. In a threefold crossvalidation, three replica models were constructed and tested with the test sets. The *Lanice* model is thus validated for use within the same region. As the observations in the calibration set were already from several years, it can be assumed that enough temporal variance was incorporated to apply the model to other periods, but validation with future observations is advised.

5.3.2. CMOC: validation during model selection

The CMOC model selection approach presented in this thesis is an improvement to the traditional model validation of HSMs. Instead of testing only one "best" model after the model selection has taken place, the model validation is incorporated into the model selection. The ranking of the models, based on the CMOC is thus already taking into account the trade-off between good model fit for the calibration data and for the test or validation data, as well as taking into account the model complexity. If independent data in a validation set are used in the model selection process, the models receiving the highest CMOC, will have maximal model generalisation ability and thus can be transferred to other periods or regions.

The HSM for the species *Abra alba* (Chapter 3) was developed with the CMOC model selection. The data set was split based on the sampling year and one part was kept apart to validate the model during the model selection. As the parameter *m* (see Chapter 3) was 0.5, the CMOC per alternative model was calculated as the mean of the calibration and validation set MOC.

5.3.3. Integrated model validation: model selection, samples, ecological literature and experiments

The incorporation of the validation into the model selection as done with the CMOC approach is a major improvement. Further improvement was made in this thesis by striving towards an integrated validation of HSMs. Such an integrated validation considers: 1) model validation with field observations (internal or external validation), 2) ecological knowledge from literature combined in a conceptual scheme, 3) habitat preference experiments and 4) influence of the sampled range of each variable. The traditional model validation was already discussed above (see 5.3.1.).

5.3.3.1. Conceptual scheme to combine ecological knowledge

A conceptual scheme (Fig. 4.6) bringing together the current ecological knowledge of the modelled species will bring more ecological realism to the model validation. The current linkages between habitat suitability modelling practice and ecological theory are often quite weak, hindering progress (Elith and Leathwick, 2009). A conceptual scheme allows determining the directness of the link between variables in the model and the distribution of the species, and as such the causality of the variables. The scheme will also provide guidance on which important variables are missing and which variables should be included in future model building exercises (Gibson, 1994).

In this thesis, it is clearly advocated that the relations between the predictive variables among each other, and with the species, should be explored before the model development. The best approach is to combine the ecological literature on the environmental variables and on the species in a conceptual scheme. Modellers and field ecologists should set up such a scheme in cooperation. Creating such a scheme will: 1) force researchers to think about the possible causal relation and the direction of the relation between variables and/or the species (Austin, 2002), 2) help to determine which variables need to be measured (Gibson, 1994) and included in an optimal model (Austin, 2007) and, 3) help to determine which important variables are missing in the data set.

The conceptual scheme can be used to make a first selection of variables from the available variables. As such the selection is not fully dependent on automatic variable selection and ecological insight can guide the choice of the selection. The limited set of potential predictive variables can then be used in an automated variable selection approach to find the most optimal combination from all possible combinations.

5.3.3.2. Experiments

Habitat preference experiments are rarely performed within the context of HSM validation (e.g. Wright *et al.*, 2000). Experiments allow testing the causal effect of a single variable in isolation of other effects.

This is a major improvement over model validation with field observations. Field observations always have a multivariate nature, mystifying the predictive power of a single variable.

An additional advantage of experiments, is that they allow delimiting the fundamental niche of a species for a variable. Some variables are known to have a causal relation with the probability of presence of the species, but in experiments the variable range suitable for the species will often be broader, compared to the realised range in field observations. Knowledge of the fundamental niche can be useful in some model applications. For example, it is useful to know the fundamental niche of invasive species to predict areas suitable for colonisation. In these colonised regions the biotic interactions are often less limiting on the fundamental niche, as their predators have not moved along with them (Rodder and Lotters, 2009).

In this thesis, habitat preference experiments from the literature allowed determining the fundamental niche of the species *Donax vittatus* for the variable grain size (Chapter 4).This fundamental niche was wider than the niche based on the samples where the species was present. Knowledge of the fundamental niche is useful, but in practice the realised niche, modelled with the field observations, is more useful to predict the spatial distribution of the species.

5.3.3.3. Sampled variable range

The sampled variable range can greatly influence both the estimation of model parameters and determine if a parameter is relevant to predict the species distribution. For the variables in the model, and the variables not chosen by the model selection, the distribution of the samples over the variable range should be assessed. Variables known to be relevant for the species based on literature or experiments can be irrelevant in the given data set, because in the range sampled, they cannot discriminate between samples where the species is present and absent.

Ideally, samples should cover the full range of each environmental variable in the model (Vaughan and Ormerod, 2003). Failing to sample the full range of a variable might lead to bias in the estimation of the importance and modelled response of a variable (Vaughan and Ormerod, 2003), as the observed species response is truncated. If the species is present along the complete measured range of the variable, this variable has no discriminatory value for the species. One solution might be to measure a wider range for this variable, otherwise the variable should not be included in the HSM. The possible bias introduced by sampling only a limited part of the variable range is so large, that any response curve can be obtained depending on the part of the range sampled (Yee and Mitchell, 1991). For example, if the true response of the species for a variable is a Gaussian response, sampling different parts of the complete range may lead to completely different modelled responses (Fig. 5.2).

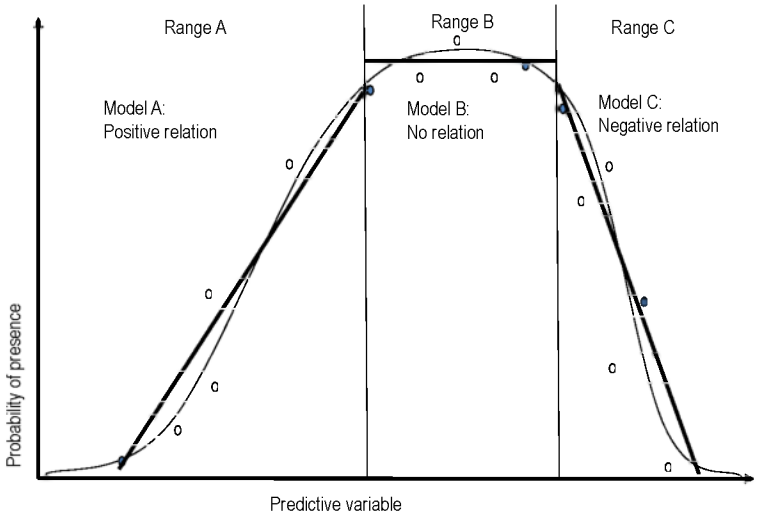

Fig. 5.2. A hypothetical species-environment relation for one variable, and three fitted models based on a different part of the sampled range of the variable.

The chance of missing part of the variable range increases when models are built from a limited geographic region, because the environmental range will probably be smaller in a smaller region (Vaughan and Ormerod, 2003). Environmental stratification is the best approach to attain an optimal sampling of the variable range of each variable (Margules *et al.*, 1994). Stratification has been proven to improve the model performance compared to random sampling (Hirzel and Guisan, 2002).

5.3.3.4. Integrated model validation

Integrated model validation combines a discussion about the different ways the model performance (e.g. based experiments, field observations, etc.) is analysed. This can assess the reliability of the model and can guide future field sampling. For example, it can advice to sample some previously unmeasured variables, or extend the measured range of ecologically relevant variables not chosen in the model selection. For example, on the BPNS more samples could be collected in locations with high bottom currents (Fig. 4.5).

In this thesis an integrated model validation was done for the species *Donax vittatus*. It was possible to couple the physical variables available in the field observations (e.g. median grain size, mud%, currents, etc.) with known ecophysiological processes of the species such as burrowing, somatic growth, holding the valves together. Habitat preference experiments confirmed the causal influence of the median grain size, the variable that was also selected in the model selection.

5.4. Future model improvement

In case a model performs badly in the model validation, an analysis is needed to find the cause of the low model performance due to model error. An understanding of the sources, magnitude and pattern of the errors of HSMs is essential if the models are to be used transparently in decision making (Barry 2006). Despite the wide use of predictive models, many applications give insufficient consideration to model error and uncertainty (Barry and Elith, 2006). Mostly only the compound error is assed during the model validation by comparing species observations and predictions. The compound error is the resulting error caused by the propagation and intermingling of errors during the model development (Guisan *et al.*, 2006b). To improve the future model development and sampling, an assessment was made of the most frequently occurring sources of error and bias in the HSM development process (Table 5.1). This can be used as a start to improve specific steps in the HSM methodology. It can suggest to which steps research time would need to be allocated to get more reliable and ecologically sensible HSMs.

Ultimately a complete analysis of all sources of error and bias can lead to reliable confidence maps for the predictions (Hirzel and Le Lay, 2008). At the moment such maps can be made based on the validation error (observed vs. predicted), but this is only a projection of the compound error, caused by sampling errors, bias in model parameters, etc. Another useful application of an assessment of the sources of error, is the establishment of a theoretical framework to estimate the statistical power of a HSM. The other way round, the number of observations needed to attain a specified power level can also be estimated in such a framework. At the moment, the statistical power of some techniques can be calculated (e.g. logistic regression: Væth and Skovlund, 2004), but much more factors play in the calculation of the power of a HSM. Ecological traits of the modelled species play a major role, but their influence on the model error is often difficult to quantify. One possible way to test the robustness of models to slight changes in the calibration set is the use of bootstrapping, as done in the CMOC model selection approach. Another advantage of bootstrapping is the ability to infer the distribution of a statistic (e.g. predicted probability for a location) and place confidence intervals around it, based on the bootstrap analysis (Carpenter and Bithell, 2000).

Based on the assessment of the sources of error and bias in HSMs, a future outlook on how to further improve HSMs is provided. The emphasis in the future outlook will be on the species ecology, sampling for HSMs and the model development.

Table 5.1. Overview of some of the potential sources of error and bias in the development of HSMs for marine species.

Source of error/bias	Potential problem	Reference
Sampling general		
No time series	Snapshot, miss species presences	Le Pape *et al.* 2007
Data age	Dynamism, environment changed	
Species Observations: influences species ecology		
Dispersion limitation	Suitable habitat unoccupied: false absences	De Marco *et al.* 2008
Patchiness	False absences and spatial autocorrelation	
Broad habitat preference (generalist)	Predictive variables difficult to find	Seoane *et al.* 2005
No biotic interactions modelled	Unaccounted absences	
Source/sink dynamics	Not in equilibrium with environment	
High mobility	Difficult to observe, true absences difficult	Boyce *et al.*, 2002
Migration	Need to model migration triggers	Reyes *et al.* 1994
Rareness	Low prevalence	Jimenez-Valverde *et al.* 2006
Density dependent habitat preference	Not in constant equilibrium with environment	
Ontogenetic shift habitat preference	Separate model larvae/ adults needed	
High trophic level	Shifts between preys	
Change in habitat preference	Niche conservation assumption violated	Wiens *et al.* 2005
Predictive variables		
Narrow variable range sampled	Species response truncated, bias in modelled response	Guisan *et al.* 2005
Spatial resolution insufficient	Spatial mismatch species-variable relation	Austin 2007
Temporal resolution insufficient	Temporal mismatch species-variable relation	
Only proxy variables	Local relation, bad model transferability	
Human disturbance	Disturbance causes false absences, include as a variable	Guisan *et al.* 2005
Model		
Missing variables	Spatial pattern in residuals	
Wrong response modelled	Response too complex or too simple	Austin 2007
Choice modelling technique	Inappropriate technique, difficult interpretation	Guisan *et al.* 2005
Stepwise model selection	Getting stuck in local optimum	
Only correlative relations	More mechanistic model	Kearney *et al.* 2009
Very low/high prevalence	Bias in model parameter estimations	Jimenez-Valverde *et al.* 2006
Model selection sensitive to collinearity	Selection stuck in local optimum, use exhaustive model selection	Reineking *et al.* 2006
Model selection not robust for data variability	Data set once used, use bootstrap resampling	
Contingency table performance indicator	Need to choose cut-off for presence	Liu *et al.*2005
Non-consistent model selection	Too complex models selected when $n \Rightarrow \infty$	Reineking *et al.* 2006
Only best model validated	Model validation not incorporated in model selection	

Model validation	
No spatiotemporal independent observations	Overestimation predictive performance on independent data
No habitat preference experiments	Not able to distinguish realised/fundamental niche
Little ecological literature	Validation against known ecology incomplete
Only post hoc testing one best model	Best model not necessary has high generalisation ability

5.4.1. The influence of species ecology

The species-habitat relations can be mystified or biased due to a number of live history properties of species. Some authors have already pointed out the influence of the migratory status and trophic rank (McPherson *et al.*, 2004). Such ecological properties can influence the model performance and/or generalisation, mostly due to false absence observations that violate the equilibrium or niche conservatism assumption. In general it can be stated that ecological properties of the species should be incorporated in the model and in the sampling methodology.

Interacting species that are known prey, competitors, or predators may be valuable predictors and can increase the variance accounted for by a model in a species distribution pattern (Leathwick and Austin, 2001; Vaughan and Ormerod, 2003). Species interactions may be direct, such as through interference and predation, or indirect, by depleting a common resource or being preyed upon by a common predator. When the biotic interactions are not in a model, the variance explained for by the abiotic variables is overestimated due to the correlation between some variables in the model and the density of some interacting species. Incorporating both dispersion limitation and biotic interactions will also bring the modelled niche closer to the fundamental niche (Guisan and Thuiller, 2005), which will increase the model generalisation on new data and the model transferability to other regions (Randin *et al.*, 2006).

Competitively dominant species might be expected to suffer few biotic constraints – and their modelled realised niche is expected to be closer to the fundamental (Hirzel and Le Lay, 2008). Subordinate species might be expected to undergo strong limitations due to competition, which can cause a high dissimilarity between their fundamental and realised niche (Hirzel and Le Lay, 2008). For example, presence of a superior competitor may prevent a species from occupying some part of its niche, leading to a truncated or even bimodal niche (Austin, 1999).

Very little attention has been given so far to the place of a species in the food chain in relation to the modelling of the species' response to environmental variables. The location of a species in the food chain determines if the species is rather influenced by environmental variables, or rather by the

distribution of its prey in case of a predator. Environmental variables that determine the distribution of a prey species can be causal for the prey species, but are proxy variables for the predator species because they determine the distribution of its prey. Thus, when modelling predator species, often the niche of the prey species is modelled to some extent. But predators can sometimes switch to other prey species, and the predator niche might then become more similar to the niche of the replacing species. One solution is to include the presence of one or more prey species as a predictive variable. Le Pape *et al.* (2007), for example, used macrobenthos densities to predict the distribution of flatfish.

The broad niche of a generalist species poses a challenge in the HSM variable selection. In general the HSMs for generalist species have a lower predictive performance (McPherson *et al.*, 2004), as it is difficult to discriminate between suitable and unsuitable locations.

In contrast to sessile species, it is difficult to obtain valid absences for mobile species (Boyce *et al.*, 2002), e.g. birds or whales (Guisan *et al.*, 2005). So, ideally the sampling resolution should depend on the species' home range (Guisan *et al.*, 2005). The mobility of species can lead to lower performance of HSMs, as the species is not always in its most suitable habitat when it is observed, which is stated in the equilibrium assumption. Boyce *et al.* (2002) further emphasise that mobile animals may not be using the entire suitable habitat at any one time and modelling their habitat requires an appropriate data model and special resource selection functions. Time series of observations would also increase the chance of observing the species in a suitable location.

Through scaling laws the speed and thus mobility of larger animals is larger. As smaller species cannot flee very far if the local environment becomes unsuitable, smaller species, e.g. macrobenthos, are more dependent on the local environmental variables and therefore better to model, also because they are often on a lower level in the food chain. A drawback is that for small species, finer scale predictive variables are needed to match the fine scale distribution of the species (see further: scale mismatch). The species' size thus influences the spatial scale we have to observe species and predictive variables.

Dynamism can be introduced into HSMs by incorporating migration into the modelling framework. In this context, migration can be defined as a movement of individuals towards the region with the highest suitability that is closest by (Reyes *et al.*, 1994; Sundermeyer *et al.*, 2005). Migrating species thus track the environment most suitable for them at a given time. A requirement is that dynamic data of the environmental variables are available. At the moment, marine HSMs that include migration are rare (e.g. Reyes *et al.*, 1994; Rubec *et al.*, 2001). Time series of species and environmental variable would also be needed to estimate the triggers for migration.

Source-sink dynamics can violate the equilibrium assumption as the sink habitats have a low suitability, but nevertheless the species is present (Fig. 5.3). But as the suitability of a habitat is

proportionate to the probability of presence of a species, sink habitat are modelled as suitable habitats. A definition of suitability based on reproductive success would allow to correctly model the suitability of source and sink habitats. However, reproductive success is much more difficult and expensive to observe in marine environments than species presence or density.

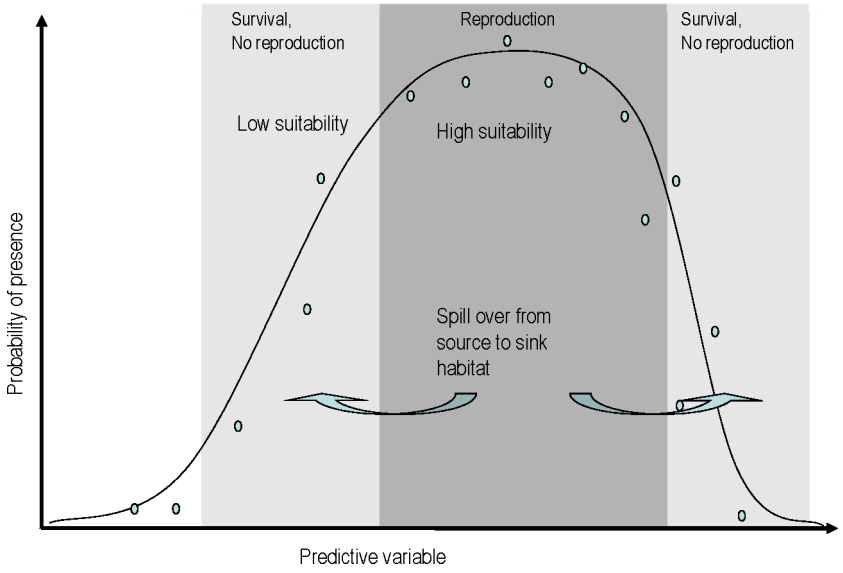

Fig. 5.3. Illustration of source-sink dynamics and the results for habitat suitability models.

Another challenge is to accurately quantify the distribution of species with a patchy distribution or gregariousness, which is the equivalent of patchiness for mobile species. Both phenomena will introduce increased variance in the estimation of the species distribution. The habitat as a whole can be suitable but the species is present only in patches, which leads to much false absence observations. A patchy distribution or gregariousness will thus lead to unexplained variance, as the species is often observed to be absent in a suitable habitat. The spatial scale at which species are observed is crucial to avoid the bias. One option is thus to increase the sampling surface (e.g. larger grab size for macrobenthos) to capture the local variability or to use modelling techniques that allow to make models with presence observations only (Hirzel et al., 2002; Phillips et al., 2006).

Spatial autocorrelation (SA) is an important source of bias in most spatial analyses (Segurado *et al.*, 2006). The definition of SA is that two observations that are close together are more similar (positive SA) or dissimilar (negative SA) than randomly expected (Legendre, 1993; Doniol-Valcroze *et al.*, 2007). Patchiness can be one of the possible causes of SA (Thomson *et al.*, 1996). SA is a problem for statistical modelling because it violates the assumptions of most classical statistical tests used in

ecology: independence of samples (Legendre, 1993). In statistics, each independent observation counts for one degree of freedom, but in the case of autocorrelated predictive variables each observation does not bring on fully independent information and is not worth a full degree of freedom. SA thus reduces the effective number of samples (Thomson *et al.*, 1996). As such, patchiness might be a problem in the model selection, because the consistent MOCs base their penalisation term for model complexity on the number of samples (e.g. BIC: ln(n)). Thus the model selection for a calibration set with SA in the observations would select models that are too simple, in comparison with the models that would have been selected based on the effective sample size.

5.4.2. Sampling for HSM: matching scales of species and predictors

Each variable has a specific resolution at which its variance can contribute most to the prediction of the habitat suitability (Guinet *et al.*, 2001). In order to obtain an optimal HSM the combined variance of the predictive variables should thus cover as much as possible of the spatiotemporal variance of the species (Fig. 5.4). A mismatch can occur between the spatiotemporal resolution at which species data were sampled and the spatiotemporal resolution of environmental variables (Guisan *et al.*, 2005). Doniol-Valcroze *et al.* (2007), for example, stated that the association between whales and thermal fronts was not observed in all studies, due to the spatiotemporal mismatch between the momentaneous whale sightings and temperatures averaged over several days, which missed the dynamics of local fronts. For grid based environmental variables, "released matching" is a problem when several habitat requirements for a species are met in a single grid cell based on the aggregated variable grids, but inside the cell the suitable conditions do not overlap on a microscale (Guisan and Thuiller, 2005).

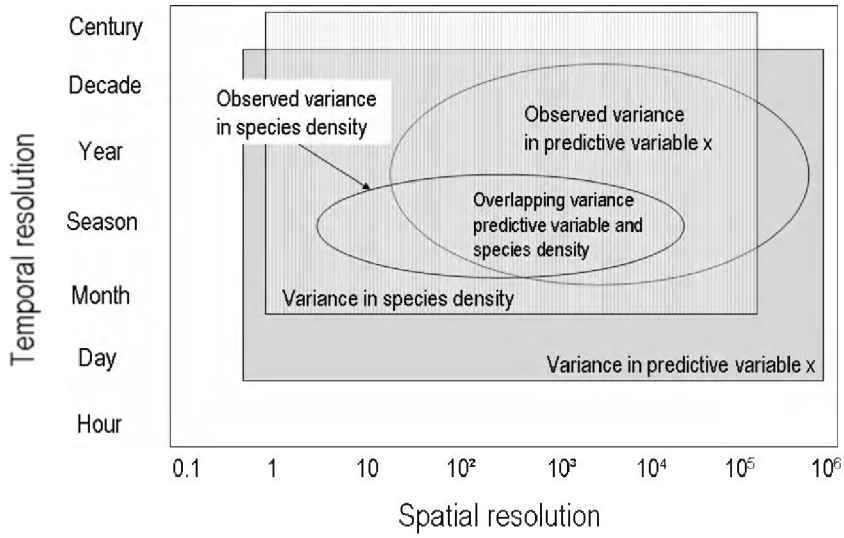

Fig 5.4. Illustration of the mismatch between the spatiotemporal scale of the species observations and the environmental variables in a habitat model.

Temperature, for example, could have a variance on a spatial scale of 5 km (and above), but a species could have a spatial variance on a scale of several meters in case of a small benthic species. This is an example of a mismatch of the large spatial scale of the variance in the variable temperature and the small scale variance in the distribution pattern of a small benthic species. An extra variable, e.g. sediment grain size, would be helpful to predict the finer scale variance of the species, as median grain size could have a finer scale variance. Nevertheless temperature could still contribute to the prediction of the species at a coarse spatial resolution. Together they are able to model the variance of the species over a range of spatial scales. Different variables determining the species' distribution cause different spatial patterns to be observed at different scales, often in a hierarchical manner (Pearson *et al.* 2003). For example, climatic regulators will result in a gradual distribution observed over a large extent and at coarse resolution, whereas patchy distributions observed over a smaller area and at fine resolution are more likely to result from a patchy distribution of resources (Guisan *et al.*, 2005)

The differences in resolution and extent of the predictive variables limit the generalisation, and thus the model transferability to other periods and regions. A small, local survey may fail to capture and predict regional variability (Francis *et al.*, 2005). Extrapolating a model based on such a local survey may be dangerous, as on a large scale different variables will determine the distribution of a species. Therefore, it should be avoided to use a HSM based on local observations on a small spatial scale on a large spatial scale, without proper external validation.

When including biotic interactions in the HSM, scale aspects come into play, as each ecological process and interaction has its proper scale (Guisan and Thuiller, 2005). In general, biotic interactions are mostly expected to have an effect on species distributions on small scales (Hirzel and Le Lay, 2008), while on macro-scale physiological limits are expected to determine the species distributions (Pearson and Dawson, 2003). Indeed, many country- or continent-wide models achieve good accuracy of prediction when based on climatic variables only (Pearson *et al.*, 2002). A mismatch between the spatial distribution of predator and prey should be avoided by using fine resolution observations, otherwise the problem of "released matching" (Guinet *et al.*, 2001) will happen: the prey-predator relation will change when observed at another spatial resolution. At fine resolution the prey avoids the predator, at coarser resolutions the predator seems to attract the prey because they both share the same physiological range limits (Guinet *et al.*, 2001).

As a conclusion, more research is necessary on mismatch of the spatiotemporal resolution of the predictive variables and the observed spatiotemporal variance in the species distribution. This mismatch may be one of the main reasons, together with insufficient sampling of the range per variable, why ecologically relevant variables are not selected during the model selection, as the resolution of the variables the data set was not usable to discriminate between suitable and unsuitable habitats for the

species. Also, research is needed to associate scale domains to predictor variable that are known to determine the distribution of species (Mackey and Lindenmayer, 2001).

5.4.3. Model development and integration

5.4.3.1. Software for model development

During this thesis most of the statistical programming was done in the R software (http://www.r-project.org). This open source software provides a fully customisable environment, which was necessary to program the newly developed HSM modelling methodology. In most currently available statistical programs this level of automation and programming would never have been achieved. Programming the modelling methodology in R forces modellers to consider each step. Readymade modelling programs (e.g. Maxent; Phillips and Dudik, 2008) are faster and easier to implement and use, but come with the risk that nice maps are generated based on unknown assumptions and theories. In future application of HSMs the use of R will also increase the efficiency. In the ideal case, modelling code would access species observations and environmental data from existing data bases directly. Models can then be generated for each species in a more automated way, while still allowing some input (e.g. prior selection of variables before automated selection).

5.4.3.2. Spatially explicit models

The practical integration of habitat suitability, dispersion limitation and biotic interaction modelling can be done in a spatially explicit model. Such integrated models have been used by some authors to model the distribution of fish (e.g. Reyes *et al.*, 1994; Sundermeyer *et al.*, 2005). For each grid cell in the spatially explicit model the suitability of the habitat is modelled based on the physical environment of that cell. Dispersion limitation is explicitly incorporated; leaving suitable grid cells unoccupied if they are not reachable by the species.

Dispersion limitation can be modelled with a least cost migration approach (Adriaensen *et al.*, 2003), where species disperse in the direction of highest suitability (Reyes *et al.*, 1994), but chose the path with the lowest effective distance. In this context the effective distance is the shortest geographic distance, but this distance is multiplied with a weighting term when grid cells are crossed that have a low habitat suitability (Ferrier and Guisan, 2006). When dispersion modelling is to be fully integrated together with habitat suitability modelling, some knowledge on the dispersion abilities of each species is needed. For example, benthic species will only disperse during their pelagic stage, while marine

mammals are very mobile and migrate to follow their most suitable habitat. Seasonal migration can be modelled when dynamic data on the environmental variables is available (e.g. water temperature maps).

Biotic interactions can be included in a spatially explicit modelling approach. The probability of presence of a species thus also depends on the presence of other species in same grid cell or neighbouring cells. The modelling of biotic interactions by including the density of interacting species should be regarded as an approximation when no dynamic population models are available yet. Feedback loops are not possible in this approach. When including biotic interactions in a spatially explicit model, aspects of scale come into play, as each ecological process and interaction has its proper scale (Guisan and Thuiller, 2005). When observed at different scales, prey and predator seem to attract or avoid each other ("released matching"; Guinet *et al.*, 2001). Spatially explicit models thus should receive more attention as they are a method to integrate both the suitability of the environment as well as biotic interactions and dispersion limitations. Incorporating all these effects will allow spatially explicit models to capture and predict more of the observed variation in field observations.

5.4.3.3. Mechanistic models

Most modelling techniques used in HSMs are correlative as they link measured environmental variables to species distributions. An advantage of this approach is that models can be developed straight away from field observations of species and environmental variables. An alternative strategy is to explicitly incorporate the mechanistic links between the functional traits of organisms and their environments into HSMs by using mechanistic or process-based modelling techniques (Kearney and Porter, 2009). A disadvantage of mechanistic models in comparison to correlative models is that they often require more time, effort, resources and data to construct and validate (Kearney and Porter, 2009). This is mainly because mechanistic HSMs require specific data on traits of organisms and extensive field and laboratory validation (Kearney and Porter, 2009). However, the application of mechanistic HSMs has clear advantages over the use of correlative approaches: 1) they provide more understanding of the underlying causal processes and 2) the models can be developed to integrate processes that would otherwise violate the assumptions of correlative models (e.g. modelling of species in non-equilibrium; Kearney and Porter, 2009).

As mechanistic models would make the process of habitat modelling more based on ecological processes rather than correlations, these kind of models should receive more attention in future HSM research. For some well known species it should be possible to bring together all the literature knowledge and experimental results in a mechanistic model that links the distribution of the species to measured environmental variables.

An intermediate approach would be the use of composite variables in a correlative modelling approach. As such, ecological processes (e.g. burrowing, growth for the species *Donax vittatus*, see Chapter 4) can be integrated in the current correlative modelling approach, until mechanistic approaches are fully implemented in habitat modelling. Based on the interrelation between variables, it is possible to create composite variables by combining available variables, which are transformed or combined in a way that they are directly related to the species' morphology, behaviour and physiology (Kearney and Porter, 2009). In that way, composite variables are more causal to the distribution of the species, than the original environmental variable they originate from (Kearney and Porter, 2009). Biophysical processes can thus be integrated into HSMs in a straightforward manner, and the more statistical HSM will begin to resemble more closely mechanistic models.

One example of a composite variable, would be to quantify the relation between the variables that determine the energy needed to hold the valves together for a bivalve species (e.g. grain size and sediment cohesion). As such, a composite variable "energy needed to hold valves together" is created which can be used in a correlative HSM. This is an improvement over modelling the relation between the sediment grain size and the species distribution directly, which is not considering underlying ecological effects of the variable grain size. Ecological interpretation of the model result would benefit greatly from the use of such composite biophysical variables.

5.5. Habitat suitability models for North Sea macrobenthos

5.5.1. Habitat models in this thesis

In this thesis HSMs were developed for three macrobenthos species: the polychaete *Lanice conchilega* and the bivalves *Abra alba* and *Donax vittatus* (Table 5.2). One variable was included in all the models after model selection: the sediment grain size. In the complete data set several other variables were available (e.g. nutrients, chlorophyll and currents), but these variables were not chosen notwithstanding their known ecological role for some macrobenthos species. The advantage of sediment grain size as predictive variable is that this variable is available at a full coverage scale for the BPNS and is relatively easy and cheap to measure.

Although depth has been demonstrated earlier to have an effect on the distribution of macrobenthos species (Basford *et al.*, 1990; Ysebaert *et al.*, 2002) and cold-water corals (Davies *et al.*, 2008) in the North Sea, experiments are difficult to set up as depth is a proxy for numerous other environmental variables. As a consequence the distinction between the fundamental and realised depth range is difficult to make. The danger of using depth as a variable is the lower model transferability of

the models. Depth is inherently a proxy variable in marine systems for numerous other variables (Elith and Leathwick, 2009). In the BPNS, depth is a proxy for the distance to the coast, and several relevant predictive variables also show an onshore-offshore gradient (e.g. chlorophyll; Peters *et al.*, 2005). The bottom currents increase with the distance from the coast (Luyten *et al.*, 2003). These bottom currents determine via the bed shear stress the erosion and deposition (Stanev *et al.*, 2009). Sediment transport can be a key structuring factor in infaunal communities (Desroy *et al.*, 2007). Currents also influence the dispersion of postlarval stages (Butman, 1987; Commito *et al.*, 1995), and in that way depth can be a proxy variable for the dispersion limitation of the species. In the BPNS, the depth is quite well correlated with phytoplankton densities, which are highest in shallow water close to the coast. The suspended particulate matter is also highest closest to the coast (Peters *et al.*, 2005), where the depth is shallow.

Table 5.2. Overview of the habitat suitability models developed in this thesis. BPI: Bathymetric Position Index (Lundblad *et al.*, 2006).

	Lanice conchilega		***Abra alba***	***Donax vittatus***
Modelling technique	GLM	ANN	GLM	GLM
Model selection	Stepwise	based on GLM	CMOC	CMOC
Predictive variable types in data set	Granulometry Currents Nutrients Chlorophyll		Granulometry Currents Depth BPI	Granulometry Currents Depth BPI
Predictive variables in model(s)	Median grain size Mud% Coarse fraction%		Depth Median grain size BPI	Depth Median grain size

Numerous previous studies pointed out the importance of grain size variables in the determination of the spatial distribution of macrobenthos in the North Sea (e.g. Basford *et al.*, 1990; Rees *et al.*, 2002; Degraer *et al.*, 2008; Meißner *et al.*, 2008). In this thesis it could be concluded that median grain size was not only a relevant predictive variable, but also had a causal relation with the distribution of *Donax vittatus.* In experiments in the framework of this thesis and in previous research, the grain size determined the burrowing speed (Alexander *et al.*, 1993), metabolic rate and the growth of *Donax vittatus* ((de la Huz *et al.*, 2002).

For bivalves the grain size determines the distribution of the species in different ways. The grain size controls the burrowing rate, because bivalves tend to occupy sediments where they can burrow faster (Alexander *et al.*, 1993). Faster burrowing speeds are important to avoid predation and burial by waves or current action (Brown and McLachlan, 1990; McLachlan *et al.*, 1995). *Donax vittatus,* for example, will burrow faster in fine sediments. Bivalves must actively hold their valves together. The higher shear

strength and higher cohesion of finer sediments is important for infaunal bivalves to reduce the energy cost to hold their valves closed (Stanley, 1970). But very fine sediments can be negative for some bivalve species. High percentages of mud can clog the gills of *Donax vittatus*, which forces the animals to hold their siphons closed until the fine sediments have settled. If the siphons remain closed, filtering for food particles is also stopped. Longer siphon closure periods in the finest sediments, cause a lower overall energy uptake (Ansell, 1962; Trueman *et al.*, 1966). As a conclusion, sediment grain size determines in many ways the spatial distribution of bivalve species.

Some variables with known ecological relevance for macrobenthos species were not chosen during the model selection. In the model for *Lanice conchilega* currents were not chosen despite the important role currents have on a small scale for the species (Buhr and Winter, 1977). This could be an example of scale mismatch, where the currents play at a small scale, but they were only available at a 250m grid cell resolution. Each variable has a specific resolution at which its variance can contribute most to the prediction of species' habitat suitability (Guinet *et al.*, 2001). Some variables directly influence an individual organisms' neighbourhood (e.g. ability to maintain burrows and ventilate sediment), other factors play both at local and broader scale, while still others are broad-scale factors operating over regional scales (e.g. larval dispersal; Thrush *et al.*, 2005).

5.5.2. Applications of habitat suitability models for macrobenthos

Habitat suitability models have been used to model macrobenthos species in the North Sea (Degraer *et al.*, 2008; Meißner *et al.*, 2008; Pesch *et al.*, 2008; Willems *et al.*, 2008; MeißNer and Darr, 2009), but several promising applications of this modelling technique for macrobenthos species are still waiting to be implemented as tools for marine management. Here a few potential applications of HSMs for North Sea macrobenthos are discussed.

Habitat suitability model-based sampling

At the moment a relatively high number of samples is available of the macrobenthos of the BPNS. This data set can be used to guide the sampling with the aim to produce better HSMs in the future. The range of each predictive variable should be sampled more evenly, and samples should be collected for as much variable combinations as possible. HSM-based sampling (see 1.3.1) allows to increase the sampling efficiency (Guisan *et al.*, 2006a), because fewer samples can be collected, while maintaining the amount of ecological information collected. Based on the model fit, and ideally also on the heterogeneity of the habitat, an optimal number of samples can be collected. Locations with a high

predicted probability of presence but no observations should be sampled to test the model. In locations without samples, but where the presence of the species is predicted, sampling should be planned to validate the predictions. When new samples are collected based on an initial HSM, they can be used to improve this initial model that can then again be used to guide the sampling effort.

Environmental impact assessment and habitat loss

Macrobenthos species are often used in monitoring programs because they are good indicators of the local environment and are relatively easy to sample and to handle (Rees *et al.*, 2002). The current practice to assess the effect of human impacts on macrobenthos species is to compare the species composition in a reference area where the impact is absent, with an impacted area. HSMs can separate the effect of human impacts from other natural causes. The human impact can then be used as predictor in the HSM. The contribution the human impact has in the variance of the species distribution can then be assessed. This is a major improvement over the current practice, where often the habitat difference between reference and impacted regions is neglected. HSM can be used as a simulation tool to simulate the effect of planned works on the distribution of the macrobenthos.

Assessment of the strength of biotic interactions

Macrobenthos species are food sources for fish and seabirds. Including macrobenthos densities in HSMs for these higher predators should be done on a more standardised basis as it will increase the model performance. Le Pape *et al.* (2007), for example, used macrobenthos densities to predict the distribution of flatfish. Some bivalve species are known as important food sources for seabirds on the BPNS, e.g. *Abra alba* and *Spisula subtruncata* for the common scoter *Melanitta nigra* (Degraer *et al.*, 1999b). Mapping the spatial distribution of these species, will thus be very relevant in the management of the common scooter and other seabirds. The HSMs for the species *Donax vittatus* (Chapter 4) could be improved by the addition of the spatial distribution of juvenile flatfish and crustaceans as predictors to the models. Because sublethal predation by these species lowers the somatic growth of *D. vittatus* (Burrows and Gibson, 1995; Salas *et al.*, 2001). The challenge when these predator densities are included is that these species are often highly mobile (e.g. juvenile plaice; Burrows and Gibson, 1995).

Modelling macrobenthos communities

Communities of macrobenthos species have been identified in the BPNS (Van Hoey *et al.*, 2004), and modelled by Degraer *et al.* (2008). Similarly, Pesch *et al.* (2008) modelled macrobenthos communities in the German Bight. The approach in both the papers was “assemble first, predict later” (Guisan and

Thuiller, 2005; Ferrier and Guisan, 2006). The communities were first identified, than modelled. In the future it would be most interesting to test also alternative modelling approaches for macrobenthos communities: "predict first, assemble later" and "assemble and predict simultaneously" (Guisan and Thuiller, 2005; Ferrier and Guisan, 2006). The latter approach was applied by Dunstan *et al.* (*submitted*), and yielded communities of species with similar habitat preferences, named species archetypes. A North Sea-wide HSM for the macrobenthos communities, as identified by (Rachor *et al.*, 2007), would be very relevant for marine management.

5.6. General conclusions

Since the research in this thesis was mainly methodological, the conclusions apply mostly to improvements to the modelling techniques and approaches. The proposed methodological improvements can be applied to develop models for all macrobenthos species in the North Sea.

- When model parsimony is considered important, logistic regression, a type of GLM, is superior, as these models are simpler and the predictive performance was only slightly lower than the ANNs for the species *Lanice conchilega*.
- The Combined Model Optimisation Criterion (CMOC) approach is proposed as a model selection method for HSMs. The advantages of the CMOC over stepwise selection for GLMs was assessed in a model selection with artificial data of a virtual species.
    - The CMOC approach is an exhaustive, robust model selection approach based on information theoretic measures (e.g. BIC), that chooses parsimonious models to maximise the transferability of the models to other regions or periods.
    - The model validation is incorporated in the model selection and there is a resampling to compensate very high/low species prevalence levels in a data set.
    - The CMOC approach can be used for multimodel inference and prediction, and has parameters to shift the emphasis in the model selection between high predictive performance on the calibration data or on the test data.
    - The CMOC approach was successfully tested with artificial species data.
- An integrated validation of the HSMs should compare the modelled species response with 1) field observations (traditional validation), 2) ecological knowledge in a conceptual scheme, 3) habitat preference experiments and 4) distribution of the samples over the range of the variables.

- The current knowledge on the species from the literature should be combined in a conceptual scheme to allow visualisation of the variable interrelations. Such a scheme must be used to put hypotheses forward on the causality of predictive variables.
- Habitat preference experiments allow identifying the fundamental niche for each variable, while field observations can only provide insight in the realised niche. Experiments can test the causality of predictive variables.
- The distribution of the samples where the species is present or absent over the range of each variable should be compared.

- Habitat suitability models were successfully developed for the macrobenthic species *Lanice conchilega*, *Abra alba* and *Donax vittatus*. The sediment grain size was selected in each HSM. The causal effect of this variable could be proven for the species *Donax vittatus*. Other variables that were used in HSMs for macrobenthic species were depth, mud%, coarse fraction% and BPI.
- More ecological insights should be incorporated both in the model selection process and in the model validation. Ecological properties of species often cause assumptions of HSMs to be violated, and can lower the predictive performance of HSMs.
- A theoretical framework should bring together all sources of error and bias in HSMs and this should guide the future improvement effort of the HSM methodology.
- The mismatch between the scale of the spatiotemporal variance in the environmental variables, and the scale of the spatiotemporal of the species distribution should receive more attention when models are developed. This mismatch can cause ecologically relevant variables not to be included in the models.
- Ideally, HSMs should include more mechanistic, process based effects in the predictions of the spatial distribution of species. The models should be spatially explicit and incorporate the habitat, biotic interactions as well as dispersion limitation.
- In the North Sea, HSMs have only been used to produce species distribution maps for macrobenthos. More specific and promising *applicat*ions should be considered such as maps indicating the chance of colonisation for invasive species, calculation of the amount of suitable habitat lost after human alterations, etc.

Appendices

Appendix I. Introduction to Artificial Neural Networks

What are Artificial Neural Networks?

Artificial Neural Networks (ANNs), are increasingly being used in ecological modelling to predict the occurrence of species (Lek and Guegan, 1999). ANNs are nonlinear models based on biological neural networks, which are made up of a set of interconnected neurons. ANNs are a modelling technique that is studied within the field of machine learning, also called artificial intelligence. Neural network algorithms can be divided into supervised and unsupervised training methods. Supervised methods use a training set which has both predictive variables and field observations, while unsupervised methods don't require observations. Unsupervised methods are used mainly as an alternative to commonly known statistical clustering and ordination methods. The supervised ANN models can be used for classification problems (e.g. presence/absence of a species) and regression problems (e.g. density of a species). In this thesis supervised ANNs will be used, as they are suitable for prediction purposes.

Neurons are the building blocks of ANNs that perform the calculations and generate a model output (Lek and Guegan, 1999). The neurons are arranged in layers: an input layer, one or several interlayers (also called hidden layers) and an output layer. In the ANNs used in habitat suitability modelling each neuron is connected to all other neurons in the next and previous layer, but no feedback loops exist. These ANN models are feed forward models: input neurons pass on the information to the interneurons arranged in the interlayer and finally the output layer will generate the model output. The importance or weight of the connection of two neurons in two consecutive layers is expressed in the interconnection weight *W*. A small interconnection weight means a lower relative contribution of a neuron to the output of the neuron in the next layer, and hence to the final model output. The adaptation of the interconnection weight terms will change the model output. The bias term is a constant that is fed into each neuron to increase or decrease the offset, and thus also influences the neurons' output. ANN training thus equals the estimation of interconnection weights and bias terms, as they are the model parameters that determine the model output. Based on a data set with predictive variables and field observations the interconnection weights and bias terms are estimated during the model training process, which is also termed the learning process.

A single neuron (Fig. I.1) receives a number of inputs (p_1, p_2, ..., p_R) and has a bias term, the constant *b*. Each input p_i is multiplied by the respective element W_i of the weights vector *W* and summed, together with the bias term *b.* The result is the weighted sum *n*, which is fed into the transfer function *f*. This function will generate the output *a*, the output of the neuron. One neuron can thus be expressed as a mathematical formula (Formula I.1).

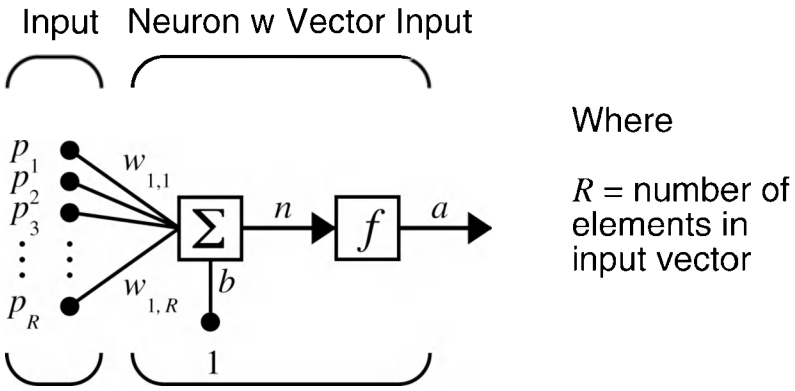

Fig. I.1. Schematic overview of one neuron of an artificial neural network. *p* = input, *b* = bias term, *f* = transfer function, *a* = neuron output. Adapted from Demuth *et al.* (2008).

$$a = f(n) = f\left(b + \sum_{i=1}^{i=R} W_{1i} \cdot p_i \right) \quad \text{(I.1)}$$

The transfer function in a neuron determines the relation between the neurons' inputs and output and thus the model output. Transfer functions can be any function, as long as they are differentiable (Maier and Dandy, 2000), as it is necessary to calculate the gradient during the training process (see further). The linear transfer function (Formula I.2) is the simplest function and the output is just the sum of the inputs, thus similar to the identity link in GLMs (see appendix II). The most frequently used transfer functions are sigmoid ones such as the hyperbolic tangent (Formula I.3) and logistic functions (Formula I.4; Maier and Dandy, 2000). These transfer functions are similar to the linear function around the centre of the function (zero) and approach a minimum and maximum at the extremes of the function range (asymptotic). Limited research on the effect of different transfer functions was carried out by Maier and Dandy (1998). In most ANNs used to predict species, logsig & tansig are used (Maier and Dandy, 2000; Goethals *et al.*, 2007).

$$\text{linear}(x) = x \quad \text{(I.2)}$$

$$\text{tansig}(x) = \frac{2}{1 + e^{-2x}} - 1 \quad \text{(I.3)}$$

$$\text{logsig}(x) = \frac{1}{1 + e^{-x}} \quad \text{(I.4)}$$

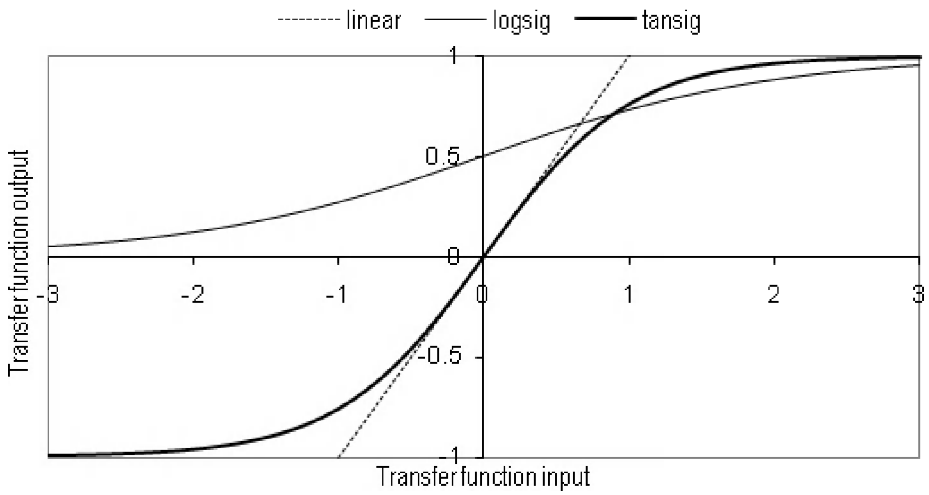

Fig. I.2. The three most commonly used transfer functions in artificial neural networks for species prediction.

The neural network

To produce an ANN, also termed perceptron, several neurons are arranged in layers and connected to one another. As an example an ANN with four inputs and three interneurons one interlayer is discussed (Fig. I.3). The input vector *P* is offered at the input of the ANN. The elements of the vector *P* are multiplied with their respective weights in the weights vector W_1 and summed together with the bias term b_{1j} for each neuron on the interlayer. This sum is then fed into the transfer function of each neuron to generate the output of each neuron. The neuron on the output layer, in this case a single neuron, receives the outputs of the neurons in the interlayer, multiplied with the elements in the weights vector W_2. After summing the weighted inputs from the interlayer and the bias term b_2, the transfer function of the output neuron generates the final model output. Depending on the transfer function of the output neuron, this output can be unconstrained (in case of a linear transfer function, Equation I.2), or bound within the interval [-1, 1] (tansig transfer function, Equation I.3) or the interval [0, 1] (logsig transfer function, Equation I.4). The ANN can be expressed as a formula (Equation I.5) where *f* is the transfer function of the interlayer, and *g* is the transfer function of the output layer.

$$Y = g\left(b_2 + \sum_{j=1}^{j=3} W_{2j} \cdot a_j \right) = g\left(b_2 + \sum_{j=1}^{j=3} W_{2j} \cdot f\left(b_{1j} + \sum_{i=1}^{i=4} W_{1ij} \cdot p_i \right) \right) \qquad \text{(I.5)}$$

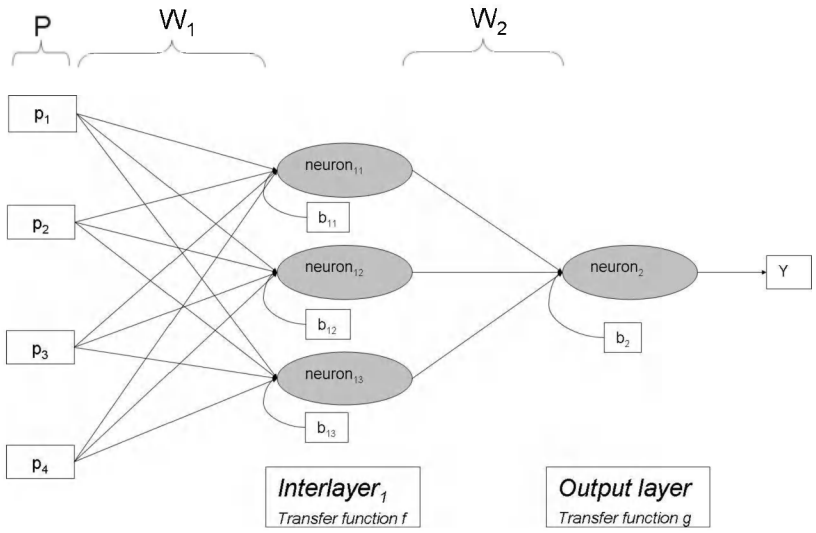

Fig. I.3. Visual presentation of an ANN with four inputs, three neurons on the single interlayer and one output neuron. *P* = input vector, *b* = bias terms, *f* = transfer function interlayer, *g* = transfer function output layer, Y = model output.

The number of input neurons is determined by the number of predictive variables. The number of dependent variables to be predicted determines the number of output neurons, e.g. the presence of a species (one output neuron), or the presence of five species (five output neurons). The number of interlayers and the amount of neurons on each of these layers is to be chosen by the modeller. Although some automated methods exist within the field of machine learning, most species distribution models use a trial and error approach to determine the number of interlayers and interneurons. Goethals *et al.* (2007) provide an overview of the rules of thumb proposed by some authors to determine the number of interlayers and interneurons. A model with more interlayers and interneurons will be able to model more complex relations, but this comes at the risk of overfitting the training data and thus the loss of generalisation of the model. Generalisation means that the model can predict well on new, unseen data that were not used to train the model. Additionally, very complex ANNs have a very high number of parameters, interconnection weights and bias terms to be estimated, which requires a sufficiently large training set. The number of ANN parameters raises quickly and exponentially, as expressed in Equation I.6 which gives the number of parameters for an ANN model with a single interlayer. The rise in the number of parameters is also visualised in Fig. I.4.

ANN parameters = interneurons(input neurons + output neurons + 1) + 1 (I.6)

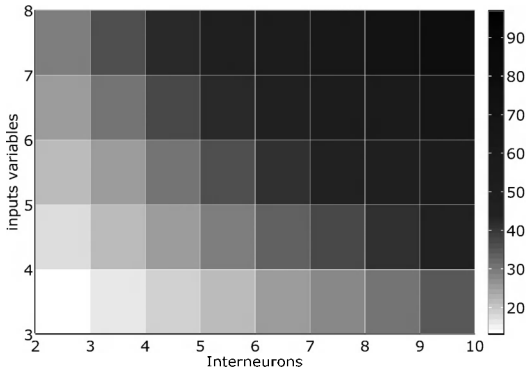

Fig. I.4. Visual presentation of the growing number of ANN parameters, if the number of interneurons and input variables rises. The number ANN parameters is calculated for an ANN with one interlayer.

Neural network training

During the training phase the ANN parameters that determine the relation between the model input and output, the interconnection weights and the bias terms, are estimated based on a training set. This training set contains the predictive variables that are fed into the ANN input, and the field observations that need to be modelled. The observations to be predicted are often called the targets in the machine learning terminology. To assess the predictive performance of the ANN model, the dissimilarity between the targets and the model output, a cost function is used. This can be, for example, the sum of squared errors or the mean absolute error. During the ANN training steps the cost function is to be minimised. Prior to ANN training the input variables need to be transformed to the same range (Özesmi & Özesmi, 1999). This is necessary, as the variable ranges need to match with the limits of the transfer functions used in the neurons (Maier and Dandy, 2000). Maier and Dandy (2000) further mentioned that, if values are scaled to the extreme limits of the transfer function, the size of the weight updates during training is extremely small and flat spots are likely to occur.

At the start of the ANN training, the interconnection weights and bias terms are random small numbers. Through the use of a training (or learning) algorithm, these model parameters will be adapted in the consecutives steps or epochs of the ANN training phase and the cost function will decrease because the prediction will approach the field observations (Lek and Guegan, 1999). The stepwise ANN training procedure is repeated until the cost function becomes small enough, a predefined maximum number of steps is reached or by comparison of the predictive performance on a validation set.

Because of its generality (robustness), and ease of implementation, backpropagation is the best choice for the majority of ANNs. This means that the interconnection weights and bias terms are adapted starting for the back of the ANN, thus from the output. By various techniques, the error in the

cost function is fed back through the network and the training algorithm adjusts the weights of each connection and the bias terms in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training steps, the network will usually converge to some state where the cost function is small. Backpropagation is the superior learning method when a sufficient number of relatively noise-free training examples are available, regardless of the complexity of the specific domain problem (Walczak and Cerpa, 1999). Although backpropagation networks can handle noise in the training data (and may actually generalise better if some noise is present in the training data), too many erroneous training values may prevent the ANN from properly training the desired model. When only a few training examples or very noisy training data are available, other learning methods should be selected instead of backpropagation (Walczak and Cerpa, 1999).

To adjust the weights and bias terms properly, one mostly applies a general method for nonlinear optimisation that is called gradient descent. For this, the derivative of the error function with respect to the network parameters is calculated, and the parameters are then changed such that the cost function decreases (thus declining on the surface of the error function). The transfer functions used should therefore be differentiable. Two settings that need to be determined for the training are the learning rate and the momentum as they help determine if the ANN parameter estimation will converge or not. The learning rate is proportional to the size of the steps taken in weight space during the ANN training phase. Traditionally, learning rates remain fixed during training (Maier and Dandy, 2000) and optimal learning rates are determined by trial and error. However, heuristics have been proposed which adapt the learning rate as training progresses to keep the learning step size as large as possible while keeping learning stable (Hagan *et al.*, 1996). A momentum term is usually included in the training algorithm in order to improve learning speed (Qian, 1999) and convergence (Hagan *et al.*, 1996). The momentum term basically allows a change to the ANN parameters to persist for a number of adjustment steps. The magnitude of the persistence is controlled by the momentum factor. Qian (1999) derived the bounds for convergence on learning rate and momentum parameters, and demonstrated that the momentum term can increase the range of learning rates over which the system converges. The ANN training might converge, but the model might be stuck in a local optimum, which makes the model less useful for prediction of new, unseen data.

What is the optimal ANN?

The main challenge in the use of ANNs for ecological modelling is to choose the most optimal ANN model architecture and model settings. Because ANN architecture is highly problem-dependent, the

determination of the ANN network architecture is one of the most important and difficult tasks in the model building process (Maier and Dandy, 1998; 2000). During ANN development, internal parameter settings are often ignored and users have limited knowledge how to use and optimise ANNs (Maier and Dandy, 1998). The modeller has to choose: 1) the number of interneurons and interlayers, 2) the transfer function for each ANN layer, 3) the training algorithm and algorithm specific settings (i.e. the learning rate and momentum), 4) the cost function. The modelling process is generally poorly described (Maier and Dandy, 2000) and therefore it is difficult to judge the confidence of the model predictions. Maier and Dandy (2000) stated that future research efforts should be directed towards the development of guidelines to assist in the development of ANN-models.

Theoretically, an ANN with one hidden layer can approximate any function as long as sufficient neurons are used in the hidden layer (Lek and Guegan, 1999). ANNs with a high number of interneurons can model more complex functions, but have more parameters to be estimated. In a review of 43 articles on prediction of water resource variables, Maier and Dandy (2000) assessed how the number of interneurons was determined. The majority of the papers (23/43) used trial & error to determine the number of interneurons. The number of hidden neurons was fixed in ten papers, and thus no model optimisation was done at all. Six papers did not mention any method. To obtain an optimal ANN model architecture and transfer function combination a number of approaches are used in the literature. This model optimisation can be carried out a priori or a posteriori. In the literature a number of rules of thumb can be found that suggest a priori the number of interneurons from the number of input variables or samples (Goethals *et al.*, 2007). A posteriori strategies include ANN pruning (Saunders *et al.*, 1994). The ANN architecture is however mostly determined by trial and error (Brosse *et al.*, 2001). In existing methodological reviews on ANNs mostly in ecological modelling one parameter was manipulated (e.g. number of interneurons), while the others were kept constant (Maier and Dandy, 1998; 2000).

The function of the interneurons in the ANN is to reproject the data from a multivariate space of the environmental variables, to a new multidimensional space where each interneuron accounts for one dimension, similar as in ordination techniques. Information reduction is also used in ordination techniques where the data set is projected onto a limited number of ordination axes, which are combinations of the original variables in the data set. If the number of interneurons is higher than the number of variables, the information is projected to a higher dimensional space. In case of a lower number of interneurons, the resulting space has less dimensions and information reduction takes place. In the simple networks (nr. neurons < nr. variables), the low number of interneurons thus reduce the information, which eventually goes into a single output neuron. This output neuron further reduces the information to one number, in this research the habitat suitability.

More complex ANNs imply an increased chance of overfitting on the training data and reducing the model's ability to generalise on new data (Özesmi and Özesmi, 1999). Generalisation is defined (Cheng and Titterington, 1994) as a model's ability to perform well on data that were not used to train it. The ability of ANNs to generalise depends further on: size and representativeness of the training set, data quality, complexity of the problem and the ANN model architecture (Haykin and Network, 1999). As very complex ANNs have numerous model parameters to be estimates, there is a risk that ANN models become underdetermined. This means that there is not enough information in the training data set to let the parameter estimation algorithm converge. There are infinitely many parameter combinations that would generate an equal cost function output. In an overview of ANN applications, Goethals *et al.* (2007) observed several ANN models where the number of model parameters to estimate was higher than the number of samples in the training set.

ANN interpretation

ANN models have been called 'black box' because they provide little explanatory insight into the relative influence of the independent variables in the prediction process (Olden and Jackson, 2002b). The interpretation of the interconnection weights is difficult in comparison with the terms in a regression model that are a more straight forward expression of the variable contribution. The interconnection weights thus lack ecological meaning (Lek and Guegan, 1999). Therefore, several techniques have been developed to assess the relative variable contribution (Olden *et al.*, 2004). Most often a sensitivity analysis is performed: all input variables are kept constant, except one variable that is allowed to change (Olden *et al.*, 2004) to assess its influence on the model output. Gevrey *et al.* (2006) used the 'PaD' method (Dimopoulos *et al.*, 1995) which consists in a calculation of the partial derivatives of the output according to the input variables.

ANN applications

ANN originate from engineering sciences, and recently most applications can still be found in these technical applications of the technique: voice and image recognition, chemical research, robotics,... (Lek and Guegan, 1999). In the biomedical sciences most application are found within molecular biology and medicine. Biological applications are the prediction of algal blooms (e.g. Lee *et al.*, 2003), remote sensing, prediction of biodiversity indices (e.g. Foody and Cutler, 2006) and the prediction of species distributions.

Appendix II. Generalised Linear Models

In this appendix a brief introduction to Generalised Linear Models (GLMs) is provided. A more in-depth introduction can be found in Kutner et al. (2005) and more specifically on logistic regression in Agresti (2002).

At first it is necessary to make a distinction between general linear models and Generalised Linear Models (GLMs). General linear regression models have been commonly used for quite some time now. These models predict the level of a response variable based on one or more predictive variables in a regression formula. It is necessary that both the response and the predictive variables are on a continuous scale. They are called linear models, because the response is predicted by a linear sum of the predictive variables times the regression parameters (β). The general linear models are thus linear in the parameters. But the modelled relations between the response variable and predictive variable are not restricted to straight lines. The general linear model $density = \beta_0 + \beta_1 \cdot depth - \beta_2 \cdot depth^2$, for example, can model a bell shaped response. Another property of general linear models is that they require the errors, the observations minus the predictions, to be normally distributed. This can be a serious drawback as the distribution of biological observations is often skewed, consists of non-negative values, or non-continuous values (e.g. counts: 0, 2, 5, 100, ... individuals). Non-normally distributed data are thus often transformed (e.g. to a *log*-scale) when using general linear models in order to approximate the normality constraint. A drawback of this transformation is that this makes the interpretation difficult, as not the mean expected value but the geometric mean value is modelled (Olivier *et al.,* 2008).

To overcome the limitations of general linear models, GLMs have been introduced. GLMs are a generalisation of general linear models because 1) other distributions can be assumed besides the normal distribution, 2) both the response variable and the predictive variables can be categorical variables and 3) GLMs allow to model also nonlinear functions of the mean. Categorical predictive variables (e.g. male/female; presence/absence), can be introduced in the GLM regression equation by using dummy variables. Categorical response variables, e.g. species present or absent, are dealt with by using the concept of proportions: for a given combination of predictive variables a species is observed to be present in a proportion of the samples. Such information from a data set is used to model the predicted proportion, in this case a probability that a species will be present.

Generalised Linear Models theory

GLMs consist of three elements: 1) a random component, that identifies the response variable Y and assumes a probability distribution for it, 2) a systematic component that specifies in a linear form the explanatory variables used as predictors in the model and 3) a link function that describes the functional

relationship between the systematic component and the expected value (mean) of the random component.

random component = link function(systematic component) (II.1)

1. The random component

The random component identifies the response variable Y and assumes a probability distribution for it. The distribution of Y should be part of the exponential family (Table II.1). In case of a binary outcome (present or absent in a sample), the binomial distribution is used. The *N* observations of the response variable *Y* are expected to be independent: $Y_1, Y_2, \ldots, Y_N$. Examples of random components and assumed distributions are: the proportion of species presence-observations in ten observations in a given habitat (binomial distribution); counted densities of a species (Poisson distribution); size measurements of fish in a population (normal distribution), the observed colour of flowers: blue, red or yellow (multinomial distribution).

2. The systematic component

The systematic component specifies in a linear form the explanatory variables used as predictors in the model. The explanatory variables are added in a linear fashion and for each variable a β_i term is estimated, which quantifies the slope for that predictive variable. The β_0 term is the constant that determines the offset of the slope. Alternatively the systematic component is also called the linear predictor.

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \quad \text{(II.2)}$$

The systematic component η is the quantity that incorporates the information about the independent variables into the model. It is expressed as a linear combination of unknown parameters β_i. It is related to the expected value of the data (thus, "predictor") through the link function (see further). Written as vectors, where *X* is the vector with the explanatory variables and β the vector with the estimated GLM parameters:

$$\eta = X\beta \quad \text{(II.3)}$$

The systematic component is a linear sum of variables, but these variables can be higher order terms, e.g. *depth²* to model curvilinear effects, and can also be interaction terms, e.g. *depth·temperature.*

3. The link function

The link function *g()* describes the functional relationship between the systematic component and the expected value (mean) of the random component (Equation II.4). The expected value of the random component is expressed as *E(Y)* and equals the modelled mean μ (Equation II.5). By taking the inverse of this link function ($g^{-1}()$), the value of the mean can be calculated back to the response scale. It should be mentioned that using a *log* link function is fundamentally different from the transformation of the dependent variable; the error structure is not affected. GLMs with an appropriate link function are to be preferred from general linear models that use transformation to achieve normality.

$$g(\mu) = \eta = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} \quad \text{(II.4)}$$

$$E(Y) = \mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}) \quad \text{(II.5)}$$

The possible choices of the link function are related to the distribution of the random component. In case the random component has a binomial distribution, for example species presence/absence, the predicted model output needs to be bound to the interval [0, 1]. An identity link function (Table II.1) would not be suitable, as the systematic component would have a range of [-∞, +∞] and the modelled response should be bound between zero and one. A logit link (Table II.1) will introduce a sigmoid relation between the systematic component and the mean μ that is limited to the interval [0, 1]. The identity link is the simplest link function and is used to perform a regression with a normal distribution for continuous responses.

The *log* and *logit* link functions allow the mean μ to be nonlinearly related to the predictor variables. The link function $g(\mu) = log(\mu)$ models the log of the mean, while the logit link models the log of the odds. The logit link is used in case the mean μ is between 0 and 1, as when species presence/absence is predicted. For each distribution of the random component *Y* there is one link function that is called the canonical link and that is used by default. In case of the normal distribution, for example, the canonical link is the identity link. The canonical link functions for each potential distribution are provided in Table II.1. Other links are possible, but in practice the canonical link is mostly used.

Table. II.1. The canonical link functions and their inverses for several exponential distributions of the random component. X = matrix with the predictive variables; β = vector with the GLM parameters; μ = the modelled mean of the distribution.

			Canonical Link Functions	
GLM type	**Distribution**	**Name**	**Link Function**	**Inverse Function**
General Linear Model	Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Poisson regression	Poisson	Log	$X\beta = \ln(\mu)$	$\mu = e^{X\beta}$
Logistic regression	Binomial Multinomial	Logit	$X\beta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{e^{X\beta}}{1+e^{X\beta}}$

GLM parameter estimation and performance assessment

The most popular parameter estimation method used for GLMs is the maximum-likelihood method. For a given data set, a set of β parameters will be estimated. Maximum-likelihood means that the likelihood of the estimated β parameters is maximal, given the observations in the data set. When the variance σ *depends* on the values of X, the maximum-likelihood estimation can have smaller standard errors than least squares estimators. The likelihood function that is maximised is mostly the sum of squared errors where a maximum-likelihood coincides with the minimum of Q (Equation II.6, II.7), but other likelihood functions can be used.

$$Q = \sum [Y - (\hat{\beta}X)]^2 \qquad \text{(II.6)}$$

Maximum-likelihood if $\frac{\partial Q}{\partial \hat{\beta}x} = 0$ (II.7)

The peak of the likelihood distribution is numerically approximated in statistical packages, through iterations to determine the β-parameters that best fit the data (Equation II.7). The method of maximum-likelihood provides estimators that have both a reasonable intuitive basis and many desirable statistical properties. The general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference.

Statistical tests

After estimation of the β-parameters the goodness-of-fit of the modelled response needs to be tested. Further, it's necessary to test if the β-parameters differ significantly from zero. The most simple test is

the Wald-test that tests if *H0:* $\beta = 0$ is true by using the large-sample normality approximation of maximum-likelihood estimates:

$$z^2 = \left(\frac{\hat{\beta}}{SE} \right)^2 \qquad \text{(II.8)}$$

SE is the standard error which is not estimated from the observed data, but is based on the properties of the distribution of the random component. This allows the more powerful *z*-test to be used, as opposed to the t-test. The resulting z^2 (Equation II.8) is obtained by dividing a parameter estimate by its standard error and then squaring it. In case H_0 holds, z^2 has an χ^2 distribution with one df. To obtain a *p*-value if the estimated parameters differ from zero, the z^2 value has to be looked up in a table with χ^2-values and the *p*-value is then the right-tail χ^2 probability. The statistic is called the Wald-statistic and is similar to the t-test in general linear models. This test is used for assessing the significance of single predictive variables.

A second test that is commonly used to test if the β-parameters differ significantly from zero is the likelihood-ratio test. For this test the maximised log-likelihood of the model to be tested (ℓ_1) is compared with the log-likelihood (ℓ_0) of a model for which H_0 holds (all parameters, besides the regression constant β_0, are zero) (Equation II.9). The logarithm of the ratio of both likelihoods times -2, has an χ^2-distribution in case H_0 holds, with the number of degrees of freedom equalling the difference in the number of parameters between both the compared models. Based on the χ^2-distribution a *p*-value is provided if the model parameters differ significantly from zero. The model for which H_0 holds is called the null model, reduced model or restricted model. The model with the variable to be tested is called the unconstrained or unrestricted model. The likelihood-ratio test is equivalent to the F-test used in general linear models

$$-2\ln\left(\frac{\ell_0}{\ell_1}\right) = -2[\ln(\ell_0) - \ln(\ell_1)] = -2(L_0 - L_1) \qquad \text{(II.9)}$$

The two statistical tests introduced here work in case the number of samples is sufficiently large.

GLM as a special case of ANNs

In this section, the relation between ANNs, GLM and other statistical techniques will be discussed. Because of the flexibility of their structure, ANNs can be rearranged and simplified until they become functionally equal to a GLM function. The ANNs similar to a GLM will have one interneuron on one hidden layer and one output neuron (Sarle, 1994). To prove that GLMs are a special case of ANNs, the ANN formula is adapted until it becomes similar to the GLM formula.

For GLMs with a link function *f()*:

$$f(Y_{GLM}) = X\beta \Rightarrow Y_{GLM} = f^{-1}(X\beta)$$

Suppose $g() = f^{-1}()$

$$Y_{GLM} = g(X\beta) = g(x_0\beta_0 + x_1\beta_1 + ... + x_n\beta_n) = g(\beta_0 + \sum_{i=1}^{n} x_i\beta_i)$$

The formula for neural networks with a bias vector of the interlayer B_1, a bias vector of the output layer B_2, a weights vector of the interlayer W_1, a weights vector of the output layer W_2 and an input vector X:

$$Y_{ANN} = g(B_2 + W_2\, f(B_1 + W_1 X))$$

Suppose:

- the bias term of the output neuron is zero: $B_2 = 0$;
- the output neuron has the linear transfer function: f(x) = x;
- the elements of weights vector W_2 are one: $W_2 = 1$.

$$Y_{ANN} = g(B_1 + W_1 X)$$

Thus we can conclude that GLMs are a specific case of the more general model family that ANNs are. After rewriting and reducing the ANN formula becomes similar to the GLM formula:

$$Y_{GLM} = g(\beta_0 + \sum_{i=1}^{n} x_i\beta_i) \Leftrightarrow Y_{ANN} = g(B_1 + W_1 X)$$

The function *g()* is the link function in GLM-terminology and the transfer function in ANN-terminology. If the same function can be used, the output will be similar. The GLM linear-link function and the ANN identity transfer function are identical (*f(x) = x*). Fig. II.1 shows the ANN architecture that is equivalent to a GLM.

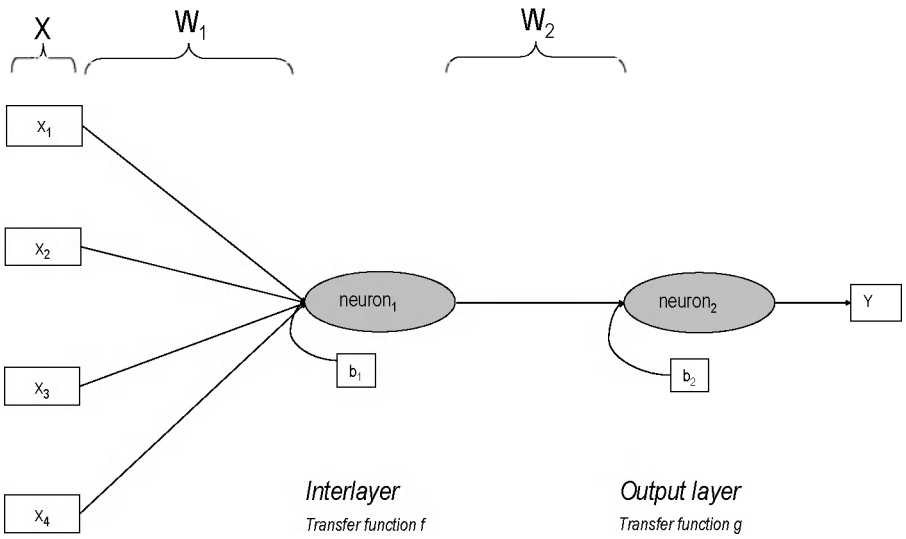

Fig. II.1. An artificial neural network that functionally emulates a general linearised model. *X* = input vector, *b* = bias terms, *f* = transfer function interlayer, *g* = transfer function output layer, Y = model output.

Further similarities between GLMs and ANNs become clear when the terminology is compared. As both techniques emerged in different research fields, ANNs in machine learning and GLMs in statistics, similar concepts have different names. DeVeaux and Ungar (1996) summarised the ANN and GLM terminology (Table II.2). The parameter estimation process is called "learning" or "training" in ANNs. The maximum-likelihood estimation is a numerical approach that works in steps, while the steps in the backpropagating parameter estimation of ANNs are often called epochs.

But the largest dissimilarity is the parameter estimation method used for ANNs and GLMs. For GLM mostly the Fisher Scoring algorithm is used which converges rapidly to the maximum-likelihood estimates. This method is deterministic: with the same data set the same model parameter values will be estimated. Parameter estimation for the ANNs commonly used in HSM is the backpropagation algorithm. This algorithm starts off when all the bias and interconnection terms are random and small, and will search for the best possible parameter estimates. But there is no guarantee that the algorithm will find the best solution as local minima might be present or the method doesn't converge at all. This estimation method is nondeterministic. The algorithm might come up with different parameter estimates even if the same data is used. Other estimation methods could be used for ANNs, but Kutner *et al.* (2004) argues that ANNs are often overparameterised and standard estimation methods will result in fitted models that have poor predictive ability.

Table II.2. Comparison of terms used in statistical and neural network terminology. Adapted from Deveaux *et al*. (1996).

Statistical Term	***Neural Network Term***
coefficient	weight
predictor, explanatory	input
response	output
parameter estimation	training or learning
steepest descent	back propagation
intercept	bias term
step	epoch

Just as GLMs are a special case of ANNs, several linear models are specific cases of GLMs. As mentioned before, GLMs are an extension of general linear models that assume a normal error distribution and an identity link. When the predictive variables are all categorical, GLMs can be used to test if groups differ significantly. Categorical variables are introduced in the GLM formula by using dummy variables. In this way, ANOVA (ANalysis Of VAriance) models are special cases of GLMs. For this modelling purpose the normal distribution is assumed and the identity link is used. The theory of GLMs thus unifies a wide variety of statistical methods: ANOVA, regression and categorical data models. Based on the continuity of their response and predictive variables, different GLMs can be identified (Fig. II.2). For each GLM type the distribution of the random component *Y* is proved and the canonical link function.

Artificial Neural Networks

General Linearised Models

	response	predictive variables	distribution	link function
ANOVA	continuous	categorical	normal	identity
general linear model	continuous	continuous	normal	identity
logistic regression	categorical	both	binomial	logit
multinomial regresion	categorical	both	multinomial	logit
poisson regression	continuous	both	poisson	log

Fig. II.2. Overview of the relation between artificial neural networks, generalised linear models and specific generalised linear models.

Appendix III. Results Combined Model Optimisation Criterion model selection

In this section, overview tables with the results of Combined Model Optimisation Criterion model selection are presented (see Chapter 3). TM1, TM3 and TM5: true models for the virtual species. For each true model, a separate table is provided. AIC: Akaike Information Criterion; AICc: AIC with small sample correction; CAIC: Consistent Akaike Criterion; BIC: Bayesian Information Criterion. NMI: Normalised Mutual Information. AUC: Area Under the Curve. For each virtual species true model the values of the MOCs, CMOCs and the Kappa's, NMIs and AUCs are provided. Additionally the model ranking of the models based on the MOC values and the Kappa, NMI and AUC values is provided.

TM1

Euclean Distance	Underfitting	Overfitting	Deviance based MOCs (expressed as Akaike Weights* 10^3)																								
			Calibration set					Test set					Combined					Calibration set			Test set							
			AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	Kappa	NMI	AUC	Kappa	NMI	AUC	Median grain size	Mud	Depth	BPI	Currents	Median grain size²	Mud²	Depth²	BPI²	Currents²
0	0	0	3.33	3.44	82.93	73.37	24.23	38.81	39.16	95.24	92.28	69.03	21.07	21.30	89.09	82.82	46.63	78.0	51.2	89.0	77.7	50.8	88.8	1	0	0	0	0	0	0	0	0	0
1	0	1	2.43	2.50	3.17	4.63	7.05	14.42	14.43	1.85	2.96	10.22	8.42	8.46	2.51	3.80	8.63	78.1	51.3	89.1	77.5	50.4	88.7	1	0	1	0	0	1	0	0	0	0
1	0	1	2.08	2.14	2.72	3.96	6.03	9.69	9.70	1.25	1.99	6.87	5.88	5.92	1.98	2.98	6.45	78.0	51.1	89.0	77.5	50.4	88.7	1	0	0	1	0	1	0	0	0	0
1	0	1	1.87	1.92	2.44	3.56	5.42	7.79	7.80	1.00	1.60	5.52	4.83	4.86	1.72	2.58	5.47	78.1	51.3	89.0	77.6	50.6	88.8	1	0	0	0	1	1	0	0	0	0
1	0	1	5.38	5.53	7.03	10.25	15.61	4.23	4.24	0.54	0.87	3.00	4.81	4.88	3.79	5.56	9.30	78.3	51.7	89.2	77.6	50.7	88.8	1	1	0	0	0	1	0	0	0	0
1	1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	74.1	45.5	87.1	74.1	45.4	87.1	1	0	0	0	0	0	0	0	0	0
1.41	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	73.2	43.8	86.6	72.9	43.6	86.4	1	1	0	0	0	0	0	0	0	0
1.41	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.1	39.8	85.1	69.8	39.5	84.9	1	0	0	0	1	0	0	0	0	0
1.41	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.6	41.9	85.7	71.2	41.5	85.6	1	0	1	0	0	0	0	0	0	0
1.41	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	74.1	45.4	87.1	73.7	45.0	86.9	1	0	0	1	0	0	0	0	0	0
2	0	2	4.42	4.49	0.30	0.73	5.10	1.60	1.58	0.01	0.03	0.45	3.01	3.04	0.16	0.38	2.78	78.3	51.6	89.2	77.5	50.5	88.8	1	1	0	1	0	1	0	0	0	0
2	0	2	2.68	2.72	0.18	0.44	3.10	2.37	2.34	0.02	0.04	0.67	2.52	2.53	0.10	0.24	1.88	78.6	52.0	89.3	78.0	51.2	89.0	1	1	0	0	0	1	1	0	0	0
2	0	2	2.72	2.77	0.19	0.45	3.14	2.03	2.01	0.01	0.04	0.57	2.37	2.39	0.10	0.24	1.86	78.1	51.2	89.1	77.1	49.7	88.6	1	0	1	0	0	1	0	1	0	0
2	0	2	2.63	2.68	0.18	0.43	3.04	2.11	2.09	0.01	0.04	0.60	2.37	2.39	0.10	0.24	1.82	78.4	51.7	89.2	77.8	50.9	88.9	1	1	1	0	0	1	0	0	0	0
2	0	2	2.52	2.57	0.17	0.42	2.92	0.94	0.93	0.01	0.02	0.26	1.73	1.75	0.09	0.22	1.59	78.1	51.3	89.0	77.2	49.9	88.6	1	0	0	1	1	1	0	0	0	0
2	0	2	1.21	1.23	0.08	0.20	1.40	2.15	2.13	0.01	0.04	0.61	1.68	1.68	0.05	0.12	1.00	78.1	51.3	89.1	77.5	50.4	88.8	1	0	0	0	1	1	0	0	0	1
2	0	2	2.15	2.18	0.15	0.35	2.48	0.96	0.95	0.01	0.02	0.27	1.56	1.57	0.08	0.19	1.38	78.3	51.5	89.1	77.4	50.3	88.7	1	1	0	0	1	1	0	0	0	0
2	0	2	1.62	1.64	0.11	0.27	1.87	1.39	1.38	0.01	0.02	0.39	1.50	1.51	0.06	0.15	1.13	78.1	51.3	89.0	77.3	50.2	88.6	1	0	0	1	0	1	0	0	1	0
2	0	2	1.58	1.61	0.11	0.26	1.83	1.37	1.36	0.01	0.02	0.39	1.48	1.49	0.06	0.14	1.11	78.0	51.2	89.1	77.2	50.0	88.6	1	0	1	1	0	1	0	0	0	0
2	0	2	1.81	1.84	0.12	0.30	2.09	0.86	0.86	0.01	0.02	0.24	1.34	1.35	0.06	0.16	1.17	78.1	51.3	89.1	77.1	49.9	88.6	1	0	1	0	1	1	0	0	0	0

Euclean Distance	Underfitting	Overfitting	Ranking deviance based MOCs															Ranking contingency table based MOCs															
			Calibration set					Test set					Combined					Calibration set			Test set												
			AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	Kappa	NMI	AUC	Kappa	NMI	AUC	Median grain size	Mud	Depth	BPI	Currents	Median grain size²	Mud²	Depth²	BPI²	Currents²
0	0	0	3	3	1	1	1	1	1	1	1	1	1	1	1	1	1	90	75	94	6	5	9	1	0	0	0	0	1	0	0	0	0
1	0	1	10	10	3	3	3	2	2	2	2	2	2	2	3	3	3	83	71	81	15	16	19	1	0	1	0	0	1	0	0	0	0
1	0	1	12	12	4	4	4	3	3	3	3	3	3	3	4	4	4	99	80	93	18	15	25	1	0	0	1	0	1	0	0	0	0
1	0	1	13	13	5	5	5	4	4	4	4	4	4	5	5	5	5	86	72	84	11	10	11	1	0	0	0	1	1	0	0	0	0
1	0	1	1	1	2	2	2	5	5	5	5	5	5	4	2	2	2	51	47	46	9	7	8	1	1	0	0	0	1	0	0	0	0
1	1	0	160	160	144	150	157	157	157	138	140	152	159	159	144	148	155	126	118	129	109	109	109	1	0	0	0	0	0	0	0	0	0
1.41	1	1	121	121	98	107	111	109	109	62	84	109	109	109	94	104	109	138	138	138	132	130	132	1	1	0	0	0	0	0	0	0	0
1.41	1	1	146	146	137	137	141	142	142	137	137	138	145	145	137	137	141	153	153	153	152	153	151	1	0	0	0	1	0	0	0	0	0
1.41	1	1	154	154	146	149	154	154	154	142	143	150	154	154	145	149	154	144	144	144	139	139	140	1	0	1	0	0	0	0	0	0	0
1.41	1	1	161	161	152	157	160	160	160	144	150	157	161	161	151	156	160	130	119	128	110	110	110	1	0	0	1	0	0	0	0	0	0
2	0	2	2	2	6	6	6	10	10	10	10	10	6	6	6	6	6	50	48	45	17	13	14	1	1	0	1	0	1	0	0	0	0
2	0	2	7	7	8	8	8	6	6	6	6	6	7	7	8	8	7	20	20	22	1	1	1	1	1	0	0	0	1	1	0	0	0
2	0	2	6	6	7	7	7	9	9	9	9	9	8	8	7	7	8	80	74	68	66	59	61	1	0	1	0	0	1	0	1	0	0
2	0	2	8	8	9	9	9	8	8	8	8	8	9	9	9	9	9	43	42	40	5	4	5	1	1	1	0	0	1	0	0	0	0
2	0	2	9	9	10	10	10	15	15	14	14	14	11	11	10	10	10	81	66	90	50	38	47	1	0	0	1	1	1	0	0	0	0
2	0	2	31	30	15	15	16	7	7	7	7	7	12	12	15	15	15	82	70	79	19	17	15	1	0	0	0	1	1	0	0	0	1
2	0	2	11	11	11	11	11	14	14	13	13	13	14	14	11	11	11	58	54	62	23	20	23	1	1	0	0	1	1	0	0	0	0
2	0	2	19	18	13	13	13	11	11	11	11	11	15	15	13	13	13	87	67	99	31	24	35	1	0	0	1	0	1	0	0	1	0
2	0	2	20	20	14	14	14	12	12	12	12	12	16	16	14	14	14	92	77	82	43	36	38	1	0	1	1	0	1	0	0	0	0
2	0	2	16	14	12	12	12	16	16	15	15	15	17	17	12	12	12	85	73	83	55	46	52	1	0	1	0	1	1	0	0	0	0

TM3

Euclean Distance	Underfilling	Overfilling	Deviance based MOCs (expressed as Akaike Weights*10²)																			Ranking deviance based MOCs										Ranking contingency lable based MOCs										
			Calibration set					Test set					Combined					Calibration set			Test set												Calibration set			Test set							
			AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	Kappa	NMI	AUC	Kappa	NMI	AUC	Median grain size	Mud	Depth	ER	Currents	Median grain size²	Mud²	Depth²	ER²	Currents²										
0	0	0	3.53	3.65	8.70	11.90	12.40	5.26	5.32	1.45	2.28	6.19	4.39	4.48	5.08	7.09	9.29	67.8	38.0	83.9	67.3	37.4	83.6	1	1	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	½	0	1.57	1.64	73.68	61.08	13.80	17.63	18.05	92.93	88.37	52.12	9.60	9.85	83.30	74.73	32.96	67.3	37.4	83.7	67.0	37.1	83.5	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
1	0	½	3.88	3.95	0.50	1.13	5.43	3.50	3.49	0.05	0.13	1.64	3.69	3.72	0.29	0.63	3.53	67.5	37.5	83.7	66.9	36.9	83.4	1	1	0	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	0	½	0.64	0.66	0.08	0.19	0.90	3.10	3.09	0.04	0.12	1.46	1.87	1.88	0.06	0.15	1.18	67.8	38.0	83.9	67.4	37.6	83.7	1	1	0	1	0	1	1	0	1	0	1	1	0	0	0	0	0	1	0
1	0	½	2.57	2.62	0.33	0.75	3.59	1.11	1.11	0.02	0.04	0.52	1.84	1.87	0.17	0.40	2.06	67.9	38.1	83.9	67.3	37.4	83.6	1	1	1	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.9	37.1	83.4	66.6	36.8	83.3	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.2	36.2	83.6	66.8	35.8	83.4	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	3.45	3.57	8.51	11.63	12.12	6.76	6.84	1.87	2.93	7.96	5.11	5.21	5.19	7.28	10.04	67.2	37.1	83.6	66.6	36.5	83.3	1	1	0	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1.41	1	½	1.04	1.07	2.55	3.49	3.64	8.25	8.34	2.28	3.57	9.71	4.64	4.71	2.42	3.53	6.67	67.4	37.5	83.7	67.2	37.2	83.6	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
1.41	1	½	1.48	1.53	3.65	4.99	5.20	3.49	3.53	0.96	1.51	4.11	2.48	2.53	2.31	3.25	4.65	67.4	37.4	83.7	67.0	37.0	83.6	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0
1.41	1	½	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.2	37.4	83.6	66.7	36.9	83.3	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1.41	1	½	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.7	36.9	83.3	66.5	36.7	83.2	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1.41	1	½	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.2	36.1	83.6	66.7	35.6	83.3	1	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.3	36.2	83.6	66.9	35.9	83.5	1	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.2	36.2	83.6	66.9	35.8	83.5	1	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0
2	0	2	5.27	5.28	0.04	0.13	2.93	2.87	2.81	0.00	0.01	0.54	4.07	4.05	0.02	0.07	1.73	68.1	38.4	84.0	67.6	37.9	83.9	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	0	2	2.33	2.34	0.02	0.06	1.30	2.69	2.64	0.00	0.01	0.50	2.51	2.49	0.01	0.03	0.90	67.6	37.7	83.8	67.1	37.1	83.6	1	1	0	1	1	1	1	0	1	0	1	1	0	0	0	0	0	1	0
2	0	2	3.43	3.44	0.02	0.09	1.91	1.49	1.47	0.00	0.00	0.28	2.46	2.45	0.01	0.05	1.09	67.8	37.9	83.9	67.1	37.1	83.6	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	2	2.83	2.84	0.02	0.07	1.58	0.57	0.56	0.00	0.00	0.11	1.70	1.70	0.01	0.04	0.84	67.7	37.8	83.9	67.0	37.0	83.5	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1
2	0	2	0.90	0.90	0.01	0.02	0.50	0.56	0.55	0.00	0.00	0.10	0.73	0.73	0.00	0.01	0.30	67.9	38.1	83.9	67.4	37.6	83.7	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	1	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.9	37.2	83.4	66.7	36.9	83.3	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.7	35.6	83.4	66.4	35.3	83.2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.4	35.4	83.2	66.1	35.1	83.1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	62.9	34.4	81.5	62.7	34.2	81.3	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Euclean Distance	Underfilling	Overfilling	Ranking deviance based MOCs															Ranking contingency lable based MOCs																										
			Calibration set					Test set					Combined					Calibration set			Test set																						
			AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	AIC	AICc	CAIC	BIC	F-test	Kappa	NMI	AUC	Kappa	NMI	AUC	Median grain size	Mud	Depth	ER	Currents	Median grain size²	Mud²	Depth²	ER²	Currents²										
0	0	0	9	8	2	2	2	8	7	4	4	4	7	6	3	3	3	27	14	25	12	6	16	1	1	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	½	0	26	25	½	½	½	½	½	½	½	½	½	½	½	½	½	51	33	51	32	24	39	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
1	0	½	7	7	7	7	6	10	11	10	10	10	10	10	7	7	7	43	26	43	45	32	46	1	1	0	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	0	½	36	36	13	14	23	12	12	11	11	11	21	20	13	13	15	30	13	26	8	8	9	1	1	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0	1	0
1	0	½	18	18	9	9	10	20	20	12	12	14	23	22	10	10	12	17	10	23	13	11	13	1	1	1	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
1	1	0	50	50	32	38	45	46	46	26	27	40	49	49	31	38	41	87	57	90	75	37	77	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	70	69	53	54	56	56	56	45	50	54	57	57	52	53	55	66	77	64	48	73	56	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	10	8	6	6	6	5	5	6	6	6	4	4	2	2	2	67	52	74	65	52	66	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	32	32	5	5	9	6	6	2	2	2	5	5	4	4	4	47	28	49	19	15	15	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
1.41	1	½	29	27	4	4	6	11	10	5	5	5	16	14	5	5	6	49	30	52	34	25	29	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
1.41	1	½	37	37	26	32	37	37	37	27	30	37	37	37	26	32	37	63	34	67	57	30	61	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1.41	1	½	52	52	45	47	52	44	44	35	38	41	46	46	43	44	47	96	63	96	78	38	79	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1.41	1	½	66	66	59	60	58	55	55	55	56	58	56	56	59	60	58	68	81	57	60	77	60	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1.41	1	½	72	71	61	61	60	63	63	60	62	62	68	68	62	62	62	58	76	60	41	67	40	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1.41	1	½	78	78	63	64	69	64	65	61	63	63	72	72	63	63	65	65	78	65	46	72	43	1	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
2	0	2	4	4	15	15	12	13	13	15	15	13	9	9	15	15	13	8	4	7	2	2	2	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
2	0	2	20	19	20	20	20	14	14	16	16	15	14	15	20	20	18	37	24	37	26	18	20	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
2	0	2	11	10	16	16	15	16	16	18	18	17	17	17	16	16	17	26	18	32	27	19	26	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	2	16	16	18	18	17	26	26	23	22	22	24	24	18	18	20	32	21	29	33	28	41	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
2	0	2	33	33	23	23	26	27	27	24	23	23	34	34	23	23	25	19	11	16	7	5	8	1	1	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
2	2	0	57	57	39	44	54	54	54	25	31	48	54	54	33	40	53	82	49	87	58	29	63	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	88	88	50	52	62	65	64	41	47	55	74	74	48	52	56	93	95	94	83	85	87	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	101	101	89	91	96	100	100	81	87	93	100	100	86	91	93	101	101	103	97	93	96	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	123	123	109	109	111	109	109	108	108	109	109	109	109	109	109	126	124	126	124	116	126	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
2	2	0	234	234	235	234	234	234	234	234	234	234	234	234	234	235	234	237	237	237	237	237	237	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

TM5

Euclidean Distance	Underfitting	Overfitting	Ranking deviance based MOCs: Calibration set: AIC	AICc	CAIC	BIC	F-test	Ranking deviance based MOCs: Test set: AIC	AICc	CAIC	BIC	F-test	Ranking deviance based MOCs: Combined: AIC	AICc	CAIC	BIC	F-test	Ranking contingency table based MOCs: Calibration set: Kappa	NMI	AUC	Ranking contingency table based MOCs: Test set: Kappa	NMI	AUC	Median grain size	Mud	Depth	BPI	Currents	Median grain size²	Mud²	Depth²	BPI²	Currents²
0	0	0	26.82	26.64	1.98	3.82	16.63	16.85	16.69	0.61	1.38	8.75	21.84	21.66	1.30	2.60	12.69	75.16	46.47	87.58	74.39	45.45	87.17	1	1	1	1	1	1	1	1	0	1
1	1	0	37.08	37.59	52.30	61.06	57.71	47.88	48.39	33.16	45.54	62.39	42.48	42.99	42.73	53.30	60.05	75.23	46.62	87.62	74.48	45.61	87.22	1	1	1	1	1	1	0	1	0	1
1	0	1	7.06	6.86	0.03	0.09	1.74	10.76	10.42	0.02	0.08	2.22	8.91	8.64	0.02	0.08	1.98	75.07	46.36	87.57	74.39	45.44	87.21	1	1	1	1	1	1	1	1	1	1
1	1	0	0.17	0.17	0.24	0.28	0.27	1.93	1.95	1.33	1.83	2.51	1.05	1.06	0.79	1.06	1.39	74.34	45.24	87.13	73.79	44.52	86.89	1	1	1	1	1	0	1	1	0	1
1	1	0	0.05	0.05	0.07	0.08	0.08	0.09	0.09	0.06	0.09	0.12	0.07	0.07	0.07	0.09	0.10	74.35	45.32	87.18	73.59	44.31	86.80	1	1	1	1	1	1	1	1	0	0
1	1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	73.42	43.74	86.70	72.62	42.72	86.30	1	1	1	1	1	1	1	0	0	1
1	1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.13	41.00	85.56	70.53	40.26	85.30	1	1	1	0	1	1	1	1	0	1
1.41	1	1	26.09	25.92	1.93	3.71	16.18	15.68	15.53	0.57	1.29	8.14	20.89	20.72	1.25	2.50	12.16	75.13	46.48	87.53	74.37	45.44	87.18	1	1	1	1	1	1	0	1	1	1
1.41	1	1	0.33	0.33	0.02	0.05	0.21	0.31	0.31	0.01	0.03	0.16	0.32	0.32	0.02	0.04	0.18	74.40	45.30	87.18	73.64	44.28	86.79	1	1	1	1	1	0	1	1	1	1
1.41	1	1	0.17	0.17	0.01	0.02	0.10	0.47	0.46	0.02	0.04	0.24	0.32	0.31	0.01	0.03	0.17	74.61	45.70	87.31	73.87	44.73	86.94	1	1	1	1	1	1	1	1	1	0
1.41	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	73.45	43.75	86.70	72.68	42.77	86.32	1	1	1	1	1	1	1	0	1	1
2	2	0	1.46	1.50	39.21	27.77	5.69	3.32	3.42	43.91	36.58	10.87	2.39	2.46	41.56	32.17	8.28	74.57	45.55	87.31	73.81	44.54	86.89	1	1	1	1	1	0	0	1	0	1
2	2	0	0.06	0.06	1.63	1.16	0.24	0.26	0.27	3.48	2.90	0.86	0.16	0.17	2.56	2.03	0.55	74.25	45.19	87.13	73.57	44.30	86.77	1	1	1	1	1	1	0	1	0	0
2	2	0	0.00	0.00	0.06	0.04	0.01	0.01	0.01	0.18	0.15	0.05	0.01	0.01	0.12	0.10	0.03	73.75	44.36	86.89	73.15	43.59	86.61	1	1	1	1	1	0	1	1	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	72.47	42.99	86.23	71.81	42.15	85.95	1	0	1	1	1	1	0	1	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	73.42	43.73	86.71	72.68	42.79	86.36	1	1	1	1	1	1	0	0	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	72.77	42.86	86.37	71.90	41.75	85.96	1	1	1	1	1	0	1	0	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.11	40.76	85.58	70.54	40.06	85.27	1	1	0	1	1	1	1	0	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	72.09	41.86	86.05	71.21	40.77	85.57	1	1	1	1	1	1	1	0	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	69.17	38.63	84.57	68.47	37.79	84.23	0	1	1	1	1	0	1	1	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.97	40.79	85.50	70.38	40.07	85.16	1	1	1	0	1	1	0	1	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.73	40.34	85.33	70.38	39.94	85.17	1	1	1	0	1	0	1	1	0	1
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.42	39.99	85.22	69.77	39.22	84.85	1	1	1	0	1	1	1	1	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.31	36.25	83.66	66.67	35.53	83.31	1	1	1	1	0	1	1	1	0	0
2	2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.70	41.47	85.83	71.23	40.90	85.62	1	1	1	0	1	1	1	0	0	1

Euclidean Distance	Underfitting	Overfitting	Ranking deviance based MOCs: Calibration set: AIC	AICc	CAIC	BIC	F-test	Ranking deviance based MOCs: Test set: AIC	AICc	CAIC	BIC	F-test	Ranking deviance based MOCs: Combined: AIC	AICc	CAIC	BIC	F-test	Ranking contingency table based MOCs: Calibration set: Kappa	NMI	AUC	Ranking contingency table based MOCs: Test set: Kappa	NMI	AUC	Median grain size	Mud	Depth	BPI	Currents	Median grain size²	Mud²	Depth²	BPI²	Currents²
0	0	0	2	2	3	3	2	2	2	9	6	3	2	2	5	4	2	2	3	2	2	2	4	1	1	1	1	1	1	1	0	1
1	1	0	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
1	0	1	4	4	13	11	5	4	4	13	13	6	4	4	13	13	5	4	4	3	3	3	2	1	1	1	1	1	1	1	1	1	1
1	1	0	9	9	10	10	8	6	6	5	5	5	6	6	7	7	6	11	11	11	8	8	8	1	1	1	1	1	0	1	1	0	1
1	1	0	12	12	11	12	12	12	12	12	12	14	12	12	12	12	14	10	8	9	11	10	10	1	1	1	1	1	1	1	1	0	0
1	1	0	26	26	28	28	26	26	26	27	27	27	26	26	28	27	27	19	19	20	20	20	20	1	1	1	1	1	1	1	0	0	1
1	1	0	64	64	81	77	68	73	73	88	87	77	71	71	86	84	73	50	45	52	53	48	52	1	1	1	0	1	1	1	1	0	1
1.41	1	1	3	3	4	4	3	3	3	10	7	4	3	3	6	5	3	3	2	4	4	4	3	1	1	1	1	1	1	0	1	1	1
1.41	1	1	7	7	14	13	10	10	10	16	15	13	9	9	14	14	11	8	9	10	10	12	11	1	1	1	1	1	0	1	1	1	1
1.41	1	1	10	10	16	15	11	9	9	14	14	11	10	10	15	15	12	5	6	6	6	6	5	1	1	1	1	1	1	1	1	1	0
1.41	1	1	20	20	25	22	22	22	22	25	25	23	22	22	25	25	22	18	18	19	19	19	19	1	1	1	1	1	1	1	0	1	1
2	2	0	5	5	2	2	4	5	5	1	2	2	5	5	2	2	4	7	7	5	7	7	7	1	1	1	1	1	0	0	1	0	1
2	2	0	11	11	5	5	9	11	11	4	4	9	11	11	4	6	9	12	12	12	12	11	12	1	1	1	1	1	1	0	1	0	0
2	2	0	15	15	12	14	16	16	16	11	11	15	16	16	11	11	15	15	15	15	15	15	15	1	1	1	1	1	0	1	1	0	0
2	2	0	18	18	17	17	17	17	17	17	17	17	17	17	17	17	17	28	22	28	26	22	26	1	0	1	1	1	1	0	1	0	1
2	2	0	25	25	23	24	24	24	24	24	24	24	24	24	24	24	24	20	20	18	18	18	18	1	1	1	1	1	1	0	0	0	1
2	2	0	27	27	27	27	28	28	28	28	28	28	28	28	27	28	28	22	23	23	25	26	25	1	1	1	1	1	1	0	0	0	1
2	2	0	40	40	44	44	41	40	40	43	43	41	40	40	43	43	41	52	49	50	52	52	53	1	1	0	1	1	1	1	0	0	1
2	2	0	44	44	48	47	45	45	45	51	49	47	45	45	50	49	47	30	31	30	33	36	37	1	1	1	1	1	1	1	0	0	0
2	2	0	48	48	53	53	49	50	50	58	56	52	50	50	56	56	52	88	87	88	88	87	89	0	1	1	1	1	0	1	1	0	1
2	2	0	66	66	73	71	65	67	67	75	75	70	67	67	75	74	70	56	47	57	59	51	60	1	1	1	0	1	1	0	1	0	1
2	2	0	67	67	76	76	72	70	70	79	77	72	70	70	79	77	72	64	58	66	60	56	59	1	1	1	0	1	0	1	1	0	1
2	2	0	74	74	85	84	77	84	84	90	90	87	83	83	90	89	85	70	69	70	73	71	74	1	1	1	0	1	1	1	1	0	0
2	2	0	86	86	92	91	88	90	90	94	93	91	90	90	93	93	91	100	100	99	103	103	103	1	1	1	1	0	1	1	1	0	0
2	2	0	97	97	114	110	101	102	102	119	117	109	101	101	118	117	108	33	34	33	31	32	31	1	1	1	0	1	1	1	0	0	1

Appendix IV. R-code

In this appendix the R-code used in this thesis is provided. All code is an application of the CMOC algorithm proposed in Chapter 3 of this thesis. The code should be considered as an illustration of one approach to model the CMOC methodology. This code is not claimed to be the most efficient or compact, neither to be completely generalistic. Thus other formats of input data would require some code rewriting. The code is mostly not written in separate functions, as this would have complicated the overview during the coding.

```
## ============================================================================
## R-code PhD "Habitat suitability Modelling for the analysis and prediction of
## macrobenthos in the North Sea" Ghent University, 2010
## Wouter Willems
## ============================================================================

## ============================================================================
## Basic functions
## ============================================================================

# Contingency matrix
contmatrix <- function(obs,pred){
  a_TP <- (obs == 1 & pred == 1)
  b_TP <- (obs == 0 & pred == 1)
  c_TP <- (obs == 1 & pred == 0)
  d_TP <- (obs == 0 & pred == 0)

  TP <- sum(a_TP) + 0.0001   # true presence
  FP <- sum(b_TP) + 0.0001
  FN <- sum(c_TP) + 0.0001   # false absence
  TN <- sum(d_TP) + 0.0001
  contmatrix = matrix(data = c(TP, FP, FN, TN), nrow = 2, byrow = TRUE)
}

### Correctly Classified Instances CCI: CCI =(a + d)/n
CCI <- function (cont)
  (cont[1,1] + cont[2,2])/sum(cont)

### Cohens Kappa
kappa.cont <- function(cont){
  a <- cont[1, 1]
  b <- cont[1, 2]
  c <- cont[2, 1]
  d <- cont[2, 2]
  N <- sum(cont)
  ((d + a) - (((a + c)*(a + b) + (b + d)*(c + d))/N))/
     ( N - (((a + c)*(a + b) + (b +d)*(c + d))/N))
}

### Normalised Mutual Information
NMI <- function(cont){
  a <- cont[1, 1]
  b <- cont[1, 2]
  c <- cont[2, 1]
  d <- cont[2, 2]
  N <- sum(cont)
  NMI_temp <- 1 - ((-a*log(a) - b*log(b) - c*log(c) - d*log(d) +
                   (a+b)*log(a+b) + (c + d) * log(c + d))/
                  (N*log(N) - ((a+c)* log(a + c) + (b + d)*log(b + d))))
  if (max(a,d) > max(b, c)) NMI1 <- NMI_temp else NMI1 <- -NMI_temp
  NMI1
}

### Area Under the Curve
AUC <- function(obs,pred){
  Y_sort <- sort(-pred, index.return = TRUE)
  Y   <- Y_sort$x
  idx <- Y_sort$ix
```

```
  obs <- obs[idx]
  tp  <- cumsum(obs)/sum(obs)
  fp  <- cumsum(!obs)/sum(!obs)
  tp  <- c(0, tp, 1)
  fp  <- c(0, fp, 1)
  n   <- length(tp)
  sum((fp[2:n] - fp[1:n-1])*(tp[2:n]+tp[1:n-1]))/2
}

### MOCs
  AIC  <- function (model, nr_sample)   # model= glm
     model$deviance + 2*model$rank

  AICc <- function (model, nr_sample)   # model= glm
     model$deviance + 2*model$rank +
     (2*model$rank)*(model$rank+1)/(nr_sample - model$rank - 1)

  CAIC <- function (model, nr_sample)   # model= glm
     model_train$deviance + model_train$rank*(log(nr_sample)+1)

  BIC <- function (model, nr_sample)    # model= glm
     model_train$deviance + model_train$rank*(log(nr_sample))

  F05 <- function (model)               # model= glm, F-test alpha 0.05
     model_train$deviance + 3.841* model_train$rank

  MOC <- function(model)  {
     nr_sample <- ifelse (is.null(model$nr_sample), length(model$residuals),
     model$nr_sample),  list(AIC  =  AIC(model,  nr_sample),  AICc  =  AICc(model,
nr_sample),
          CAIC  =  CAIC(model,  nr_sample),  BIC  =  BIC(model,  nr_sample),  F05  =
F05(model))
  }

## ============================================================================
## ============================================================================
## CMOC virtual species
## First artificial presence/absence observations are generated for a
## virtual species based on a logistic model.
## Virtual species approach for simulation (Hirzel et al., 2001)
## ============================================================================
## ============================================================================

# ----------------------------------------------------------------------------
# Data treatment
# ----------------------------------------------------------------------------

### load data (environmental variable set from field observations)
data_virt <- read.table("Abraalba_donaxvitt_macomabalthica_cleaned final.txt",
                        header = TRUE)
env1 <- data_virt[,7:ncol(data_virt)]

### transform (standardise) environmental variables (min = -1, max = 1)
env <- env1

for(a in 1:ncol(env1))
  env[,a] <- -1 + 2*((env1[,a] - min(env1[,a]))/
                 (max(env1[,a]) - min(env1[,a])))

# ----------------------------------------------------------------------------
# Calculate "Niche Coefficient"
# ----------------------------------------------------------------------------

# For each of the three alternative models, different weights are used.
# The R-code is similar for the three TMs, only the weights differ.
# The code in the virtual species section below should be repeated for each
# of the three true models (TMs).
```

```
# True model 1
niche_coef_weights <- c(1, 0, 0, 0, 0) # TM1
true_vars <- c(1, 0, 0, 0,0,1,0,0,0,0) # TM1

# True model 3
niche_coef_weights <- c(0.6, 0.3, 0.1, 0, 0) # TM3
true_vars <- c(1, 1, 0, 1,0,1,1,0,0,0) # TM3

# True model 5
niche_coef_weights <- c(0.2, 0.2, 0.2, 0.2, 0.2) # TM5
true_vars <- c(1, 1, 1, 1,1,1,1,1,0,1) # TM5

# number of virtual species
nr_virt_sp_vars <- 5

# environmental response
env_resp <- matrix(data = 0, nrow = nrow(env), ncol = nr_virt_sp_vars)
colnames(env_resp) <- paste("env_resp",1:nr_virt_sp_vars,sep="")

# calculate responses
env_resp[,1] <- -(env[,1]^2 + env[,1])            # median grain size: Gaussian
env_resp[,2] <- -0.2*env[,2] + ((env[,2]-1)^2)/4  # mud: exponential decreasing
env_resp[,3] <- -(env[,3]^2) + env[,3]            # depth: gaussian
env_resp[,4] <- env[,4]                           # BPI: linearly increasing
env_resp[,5] <- -(env[,5]^2 + env[,5])            # currents gaussian relation

# calculate niche coefficient of the virtual species
niche<- colSums( t(env_resp) * niche_coef_weights )/ sum(niche_coef_weights)

# add noise
niche <- niche + rnorm(length(niche), mean = 0, sd = 0.05)

# logistic transform to [0, 1]
niche <- 1/(1 + exp(-niche))

## convert niche coefficient to absence/presence
cutoff_choice <- matrix(data = 0, nrow = 10000, ncol = 2)

for (prev in 1:10000) {
  cutoff_virt_spec <- prev/10000
  cutoff_choice[prev, 1] <- cutoff_virt_spec
  cutoff_choice[prev, 2] <- sum(niche >= cutoff_virt_spec) / length(niche)
}

## cutoff => provides a prevalence closest to 50%
cutoff_virt_def <- cutoff_choice[which.min(abs(cutoff_choice[,2] - 0.5)),1]
Present         <- 1 * (niche >= cutoff_virt_def)
niche_f         <- cbind(env, niche, Present)
print(prevalence_virt_spec <- sum(niche_f[,ncol(niche_f)])/nrow(niche_f))

# ------------------------------------------------------------------------------
# Create all alternative models and save model performance to a file.
# ------------------------------------------------------------------------------
nr_reps <- 1000     # nr replicas
w_TT    <- 0.5      # relative weight training test set [0, 1]
                    # 0.5 = equal weight both data sets

# starttime analysis (for naming of the files)
timenow <- format(Sys.time(), "%d%m%Y_%H%M")

# variable combinations hierarchical model selection (for 5 predictive variables)
a     <- c(0,1)
model <- expand.grid(a, a, a, a, a, a, a, a, a, a)

# hierarchic model selection,
# remove models that lack a lower order term when higher order term is included
```

```
modelstokeep <- model[,1] >= model[,6] &
                model[,2] >= model[,7] &
                model[,3] >= model[,8] &
                model[,4] >= model[,9] &
                model[,5] >= model[,10]

# final models
models_final <- cbind(model[modelstokeep,], 0, 0)

# remove first case were all variables are ommited
models_final <- models_final[-1,]

# nr samples per bootstrap
nr_sample    <- nrow(niche_f)

### cutoff for presence (convert density to a/p)
cutoff_model <- 0.5    # as prevalence is kept at 0.5 in the bootstrap procedure

##  select  predictive  variables  to  test  (=  variables  from  true  model  +  extra
variables)
#### version for virtual species ####

b             <- ncol(niche_f)
envvars_sp_AP <- cbind(niche_f[,1:5], niche_f[,1:5]^2, niche_f[, (b-1):b])
rm(b)

#### start of the bootstrap - model loop
size_model_props   <- 60
nr <- 1
param_combination  <- 1

for(modnr in 1:nrow(models_final)) {
  model_props = matrix(data = NA, nr_reps, size_model_props)
  nr <- 1    # model nr for matrix

  for (reps in 1:nr_reps) {
  ## bootstrap resampling to create training and test set

  # absence data set
    Absence <- envvars_sp_AP[envvars_sp_AP[,ncol(envvars_sp_AP)] == 0,]

    sel_abs_sample <- sample(1:nrow(Absence), nr_sample,
       replace = TRUE, prob = NULL)

    test_abs  <- Absence[sel_abs_sample,]

  # presence data set
    Presence <- envvars_sp_AP[envvars_sp_AP[,ncol(envvars_sp_AP)] == 1,]

    sel_pres_sample <- sample(1:nrow(Presence), nr_sample,
       replace = TRUE, prob = NULL)
#    assign(paste("test_pres"), Presence[sel_pres_sample,])
    test_pres <- Presence[sel_pres_sample,]

  # combine absence and presence samples
    comb_sampled_traintest <- rbind(test_abs, test_pres)

  # construct training and test set
    halfset <- 1:(nr_sample/2)
    traning <- rbind(test_abs[ halfset, ], test_pres[ halfset, ])
    test    <- rbind(test_abs[-halfset, ], test_pres[-halfset, ])

  # step randomise the samples
    z <- nrow(traning)
    train_sel <- sample(1:z, z, replace = FALSE, prob = NULL)
    test_sel  <- sample(1:z, z, replace = FALSE, prob = NULL)
    rm(z)
    train_sample <- traning[train_sel,]
```

```
  test_sample  <- test[test_sel, ]
  train_obs    <- train_sample[,ncol(train_sample)]
  train_var    <- train_sample[,models_final[modnr,]==1]

# construct logistic regression model on training set
  model_train <- glm(formula = train_obs~as.matrix(train_var), family = binomial)

# apply to test set
  test_obs <- test_sample[,ncol(test_sample)]
  test_var <- test_sample[,models_final[modnr,]==1]

# write model properties to matrix
  model_props[nr,1] <- model_train$rank
  model_props[nr,2] <- nr_reps
  model_props[nr,3] <- param_combination
  model_props[nr,4] <- model_train$iter
  model_props[nr,5] <- model_train$df.null
  model_props[nr,6] <- model_train$df.residual
  model_props[nr,7] <- model_train$null.deviance
  model_props[nr,8] <- model_train$deviance

## model optimisation criteria (MOC)
# training set
  model_props[nr,9:13]  <-unlist(MOC(model_train))

# calculate model predictions for the test set
  test_var1 <- cbind(1,test_var)
  pred_test_temp1 <- as.matrix(test_var1) %*% as.matrix(model_train$coef)
  pred_test_temp2 <- exp(pred_test_temp1)/(1 + exp(pred_test_temp1))
  Nan_inf         <- exp(pred_test_temp1) == Inf
  pred_test       <- replace(pred_test_temp2, Nan_inf, 1)

# calculate the loglikelihood for the model given the test data
  model_test <- list(nr_sample=nr_sample)
  model_test$rank <-model_train$rank
  LogLik <- 0
  for (a in 1:length(test_obs)){
    LogLik = LogLik +   test_obs[a] *log(pred_test[a]) +
                  (1-test_obs[a])*log(1 - pred_test[a])
  }
  model_test$deviance <- -2*LogLik

# calculate the model optimisation criteria for the test set
  model_props[nr,14:18] <- unlist(MOC(model_test))

# calculate the combined optimisation criteria
  model_props[nr,18:23] <- w_TT * model_props[nr,8:13] +
                       (1.-w_TT) * model_props[nr,13:18]

### area under the curve measures
  model_props[nr,24] <- AUC(train_obs, (model_train$fitted.values > 0.5))
  model_props[nr,25] <- AUC(test_obs, 1*(pred_test > 0.5))

### contigency based measures
# calculate contingency table
  cont_train <- contmatrix(train_obs,(model_train$fitted.values > 0.5))
  cont_test  <- contmatrix(test_obs,1*(pred_test > 0.5))

# %Correctly Classified Instances, Cohens Kappa and Normalised Mutual Information
  model_props[nr,26] <- CCI(cont_train)
  model_props[nr,27] <- kappa.cont(cont_train)
  model_props[nr,28] <- NMI(cont_train)
  model_props[nr,29] <- CCI(cont_test)
  model_props[nr,30] <- kappa.cont(cont_test)
  model_props[nr,31] <- NMI(cont_test)

  model_props[nr,32] <- cont_train[1,1]
```

```
    model_props[nr,33] <- cont_train[1,2]
    model_props[nr,34] <- cont_train[2,1]
    model_props[nr,35] <- cont_train[2,2]
    model_props[nr,36] <- cont_test[1,1]
    model_props[nr,37] <- cont_test[1,2]
    model_props[nr,38] <- cont_test[2,1]
    model_props[nr,39] <- cont_test[2,2]

  ### attach parameter selection to model_props
    model_props[nr,40:49] <- as.matrix(models_final[modnr,1:10])

  ### attach parameters estimates
  # the selected variables in the model
    sel_varinmodel <- which(models_final[modnr,]==1)

  # add a 1 to select the intercept as well
    sel_varinmodel2 <- cbind(1, 1 + t(as.matrix(sel_varinmodel)))
  # a matrix with zeros
    model_param_temp <- matrix(data = 0, nrow = 1, ncol = ncol(models_final)-1)
  for(a in 1:model_train$rank) {
    model_param_temp[,sel_varinmodel2[a]] <- model_train$coefficients[a]
  }

  rm(a)
  # attach the parameters estimates
  model_props[nr,50:60] <- model_param_temp
  nr = nr + 1
 }   # END REPLICATES
 param_combination <- param_combination + 1
 write(t(model_props), paste("modelprops_",timenow, ".txt", sep=""),
       ncol = ncol(model_props), append = TRUE)
}   # END MODEL LOOP



# ------------------------------------------------------------------------------
# The mean over all the bootstrap replicas per alternative variable combination
# is calculated.
# ------------------------------------------------------------------------------
props_all <- read.table (paste("modelprops_",timenow, ".txt", sep=""),
                         header = FALSE)
props_mean <- matrix(data = NA, nrow(models_final), ncol(props_all))

sel_props  <- as.matrix(rowMeans(props_all)== Inf)
all_props  <- na.omit(props_all[!sel_props,])

for (var_comb1 in 1:nrow(models_final)) {
  props_mean[var_comb1,] <- colMeans(all_props[all_props[,3] == var_comb1,],
                                     na.rm = TRUE)
  props_mean[var_comb1,4] <- max(all_props[all_props[,3] == var_comb1,4],
                                     na.rm = TRUE)
}

### Akaike weights
props_mean_Akweights1 <- props_mean_Akweights <- props_mean

for (a in 9:33) {
  props_mean_Akweights1[,a] = ((props_mean[,a] - min(props_mean[,a])))
  props_mean_Akweights[,a] = exp(-0.5*props_mean_Akweights1[,a]) /
     sum(exp(-0.5*props_mean_Akweights1[,a]))
}

for (i in 19:23) {
  props_mean_Akweights[,i] = w_TT    * props_mean_Akweights[,i-10] +
                             (1-w_TT)* props_mean_Akweights[,i-5]
}

write.table(props_mean, "modelprops_mean_virtspec.txt")
```

```
write.table(props_mean_Akweights, "modelprops_mean_Akweights_virtspec.txt")

# Assessment of the matching between the true model ranking and the ranking
# based on the CMOC model selection.
# The rest of the code can only be used for the virtual species modelling.

var_match2 <- matrix(data = NA, nrow = nrow(props_mean_Akweights), ncol = 1)

var_match_over <- var_match_under <- var_match <- var_match2
vars_true_match <- props_mean_Akweights[,40:49]

for (rowvar in 1:nrow(props_mean_Akweights)) {
  Select <- props_mean_Akweights[rowvar,40:49]

  # best model = 0, overfitting > 0, underfitting < 0
  var_match[rowvar]        <- sum(Select - true_vars)
  # % correct, not distinguishing between over- and underfitting
  var_match2[rowvar]       <- sum(Select == true_vars)
  # nr. variables overfitting
  var_match_over[rowvar]   <- sum(Select == 1 & true_vars == 0)
  # nr. variables underfitting
  var_match_under[rowvar]  <- sum(Select == 0 & true_vars == 1)
  vars_true_match[rowvar,] <- Select == true_vars
}

# Euclidean distance form the best model (0, 0) in the overfit/underfit scatterplot
var_match_eucl <- sqrt(var_match_over^2 + var_match_under^2)
sort_var_match_eucl <- sort(var_match_eucl, decreasing = FALSE,
                            index.return = TRUE)
sort_var_match_eucl <- sort_var_match_eucl$ix
best_model_row_eucl <- sort_var_match_eucl[1]

### ranking models according to the different model optimisation criteria
ranking_opt_crit2 <- matrix(data = NA, nrow = nrow(props_mean_Akweights), ncol = 23)

ranking_opt_crit <- ranking_opt_crit2

# Loglikelihoodbased criteria (should be minimal)
for (col_sort in 1:10) {
  ranking_opt_crit[,col_sort] <- sort(props_mean_Akweights[,col_sort+8],
                                      decreasing = FALSE, index.return = TRUE)$ix
}

# Loglikelihoodbased criteria Combined MOC (should be minimal)
for (col_sort in 11:15) {
  ranking_opt_crit[,col_sort] <- sort(props_mean_Akweights[,col_sort+8],
                                      decreasing = FALSE, index.return = TRUE)$ix
}

# contingency based criteria (should be maximal)
# CCI, Kappa, NMI training
for (col_sort in 16:18) {
  ranking_opt_crit[,col_sort] = sort(props_mean_Akweights[,col_sort+10],
                                      decreasing = TRUE, index.return = TRUE)$ix
}

# AUC train
ranking_opt_crit[,19] <- sort(props_mean_Akweights[,24],
                              decreasing = TRUE, index.return = TRUE)$ix

# CCI, Kappa, NMI test
for (col_sort in 20:22) {
  ranking_opt_crit[,col_sort] <- sort(props_mean_Akweights[,col_sort + 9],
                                      decreasing = TRUE, index.return = TRUE)$ix
}

# AUC test
```

```
ranking_opt_crit[,23] <- sort(props_mean_Akweights[,25],
                              decreasing = TRUE, index.return = TRUE)$ix


### select the models that have an Euclidean distance <= 2 from the true model
# ------------------------------------------------------------------------------
props_mean_Akweights_rank <- cbind(props_mean_Akweights[,9:23],
                                   props_mean_Akweights[,26:33])
props_mean_Akweights_rank_rankings <- props_mean_Akweights_rank
for (a in 1:15){
  props_mean_Akweights_rank_rankings[,a] <- rank(-props_mean_Akweights_rank[,a])
}
for (a in 16:23){
  props_mean_Akweights_rank_rankings[,a] <- rank(props_mean_Akweights_rank[,a])
}
## combine the distances to the true model,
# the Akaike weights and the rankings per MOC in one table

MOCs_weights_rankings <- cbind(var_match_eucl, var_match_under, var_match_over,
                               props_mean_Akweights[,9:23], props_mean[,26:33],
                               props_mean_Akweights_rank_rankings[,1:23],
                               props_mean_Akweights[,40:49])

# select the models with an Euclidean distance <= 2 to the true model
select_eucldistMods <- MOCs_weights_rankings[,1]<= 2

# calculate the relative variable contribution a = sum(AW(CMOCi) |variable a in
model_i)
var_contr_AW <- matrix(data = NA, nrow = 15, ncol = 10)
for (i in 1:15) {
  for (j in 1:10) {
    var_contr_AW[i, j] <- sum(MOCs_weights_rankings[, i+3] *
                              MOCs_weights_rankings[,j+49])
  }
}

# weighted mean of number of variables
aver_nr_vars <- matrix(
  data           =           colSums(rowSums(MOCs_weights_rankings[,           50:59])*
MOCs_weights_rankings[,4:18]),
  ncol = 5, nrow = 3, byrow = T)

colnames(aver_nr_vars) <- c("AIC", "AICc", "BIC", "CAIC", "F-test")
rownames(aver_nr_vars) <- c("Calibration set", "Test set", "Combined set")

# write
write.table(MOCs_weights_rankings,
            "MOCs_weights_rankings_TM5_Akaike_correctALL.txt")
write.table(MOCs_weights_rankings[select_eucldistMods,],
            "MOCs_weights_rankings_TM5_Akaike_correct.txt")
write.table(aver_nr_vars,
            "Av_nr_vars_TM5.txt")

# compare the n best models
nbest   <- 25
model_rankings_all <- cbind(sort_var_match_eucl, ranking_opt_crit[,19:23])

# percentage of time a variable is chosen in the n best models
nbest_temp <- 1:nrow(model_rankings_all)
var_prec_chosen <- matrix(data = NA, ncol = 10, nrow = length(nbest_temp))
for (nbest_mod in 1:length(nbest_temp)) {
   var_prec_chosen[nbest_mod,] <- colSums(as.matrix(
   props_mean[model_rankings_all[1:nbest_temp[nbest_mod],2], 40:49]))/nbest_mod
}
var_prec_chosen[1,] = props_mean[model_rankings_all[1,2], 40:49]
```

References

Adriaensen, F., Chardon, J.P., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H. and Matthysen, E., 2003. The application of least-cost modelling as a functional landscape model. Landscape and Urban Planning, 64:233-247.

Agresti, A., 2002. Categorical Data Analysis. Wiley, Hoboken, 734 p.

Akaike, H., 1973. Information theory and an extension of the likelihood ratio principle. pp. 267–281.

Akaike, H., 1974. New Look at Statistical-Model Identification. Ieee Transactions on Automatic Control, 19:716-723.

Alexander, R.R., Stanton, R.J. and Dodd, J.R., 1993. Influence of sediment grain size on the burrowing of bivalves - correlation with distribution and stratigraphic persistence of selected neogene clams. Palaios, 8:289-303.

Armstrong, M.J., 1982. The predator-prey relationships of Irish Sea poor-cod (Trisopterus minutus L.), pouting (Trisopterus luscus L.) and cod (Gadus morhua L.). ICES Journal of Marine Science, 40:135.

Anderson, B.J., Akcakaya, H.R., Araujo, M.B., Fordham, D.A., Martinez-Meyer, E., Thuiller, W. and Brook, B.W., 2009. Dynamics of range margins for metapopulations under climate change. Proceedings of the Royal Society B-Biological Sciences, 276:1415-1420.

Anderson, D.R., Burnham, K.P. and White, G.C., 1998. Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. Journal of Applied Statistics, 25:263-282.

Ansell, A.D., 1962. Observations on burrowing in the Veneridae (Eulamellibranchia). Biological Bulletin, 123:521–530.

Ansell, A.D., 1994. In-situ activity of the sandy beach bivalve *Donax vittatus* (Bivalvia Donacidae) in relation to potential predation risks. Ethology Ecology & Evolution, 6:43-53.

Ansell, A.D., Barnett, P.R.O., Bodoy, A. and Masse, H., 1980. Upper temperature tolerances of some European molluscs .2. *Donax vittatus*, *Donax semistriatus* and *Donax trunculus*. Marine Biology, 58:41-46.

Ansell, A.D. and Lagardere, F., 1980. Observations on the biology of *Donax trunculus* and *Donax vittatus* at Ile-Doleron (French atlantic coast). Marine Biology, 57:287-300.

Araujo, M.B. and Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography, 33:1677-1688.

Araújo, M.B. and New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution, 22:42-47.

Araujo, M.B. and Rahbek, C., 2006. How does climate change affect biodiversity? Science, 313:1396-1397.

Attrill, M.J., Power, M. and Thomas, R.M., 1999. Modelling estuarine Crustacea population fluctuations in response to physico-chemical trends. Marine Ecology-Progress Series, 178:89-99.

Austin, M., 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling, 200:1-19.

Austin, M.P., 1999. A silent clash of paradigms: some inconsistencies in community ecology. Oikos, 86:170-178.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling, 157:101-118.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D. and Luoto, M., 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. Ecological Modelling, 199:197-216.

Avissar, N.G., 2006. Modeling potential impacts of beach replenishment on horseshoe crab nesting habitat suitability. Coastal Management, 34:427-441.

Ballance, L.T., Pitman, R.L. and Fiedler, P.C., 2006. Oceanographic influences on seabirds and cetaceans of the eastern tropical Pacific: A review. Progress in Oceanography, 69:360-390.

Barry, S. and Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology, 43:413-423.

Basford, D., Eleftheriou, A. and Raffaelli, D., 1990. The infauna and epifauna of the northern North Sea. Netherlands Journal of Sea Research, 25:165-173.

Bax, N., Williamson, A., Aguero, M., Gonzalez, E. and Geeves, W., 2003. Marine invasive alien species: a threat to global biodiversity. Marine Policy, 27:313-323.

Beaumont, L.J., Hughes, L. and Pitman, A.J., 2008. Why is the choice of future climate scenarios for species distribution modelling important? Ecology Letters, 11:1135-1146.

Beaumont, N.J., Austen, M.C., Atkins, J.P., Burdon, D., Degraer, S., Dentinho, T.P., Derous, S., Holm, P., Horton, T., van Ierland, E., Marboe, A.H., Starkey, D.J., Townsend, M. and Zarzycki, T., 2007. Identification, definition and quantification of goods and services provided by marine biodiversity: Implications for the ecosystem approach. Marine Pollution Bulletin, 54:253-265.

Bekkby, T., Rinde, E., Erikstad, L., Bakkestuen, V., Longva, O., Christensen, O., Isaeus, M. and Isachsen, P.E., 2008. Spatial probability modelling of eelgrass (Zostera marina)

distribution on the west coast of Norway. ICES Journal of Marine Science. 65(7):1093-1101.

Bello, P.J., Rios, L.V., Liceaga, C.M.A., Zetina, M.C., Cervera, C.K., Arceo, B.P. and Hernandez, N.H., 2005. Incorporating spatial analysis of habitat into spiny lobster (*Panulirus argus*) stock assessment at Alacranes reef, Yucatan, México. Fisheries Research, 73:37-47.

Beutel, T.S., Beeton, R.J.S. and Baxter, G.S., 1999. Building better wildlife-habitat models. Ecography, 22:219-223.

Borja, A., Franco, J. and Pérez, V., 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin, 40:1100-1114.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. and Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecological Modelling, 157:281-300.

Bradshaw, C.J.A., Davis, L.S., Purvis, M., Zhou, Q.Q. and Benwell, G.L., 2002. Using artificial neural networks to model the suitability of coastline for breeding by New Zealand fur seals (*Arctocephalus forsteri*). Ecological Modelling, 148:111-131.

Brosse, S., Giraudel, J.L. and Lek, S., 2001. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. Ecological Modelling, 146:159-166.

Brotons, L., Thuiller, W., Araujo, M.B. and Hirzel, A.H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography, 27:437-448.

Brown, A.C. and McLachlan, A., 1990. Ecology of sandy shores. Amsterdam, 373 p.

Buhr, K.J. and Winter, J.E., 1977. Distribution and maintenance of a *Lanice conchilega* association in the Weser estuary (FRG), with special reference to the suspension-feeding behaviour of *Lanice conchilega*.

Burnham, K.P. and Anderson, D.R., 2004. Multimodel inference - understanding AIC and BIC in model selection. Sociological Methods & Research., 33:261-304.

Burrows, M.T. and Gibson, R.N., 1995. The effects of food, predation risk and endogenous rhythmicity on the behavior of juvenile plaice, *Pleuronectes platessa*. Animal Behaviour, 50:41-52.

Butman, C.A., 1987. Larval settlement of soft-sediment invertebrates—the spatial scales of pattern explained by active habitat selection and the emerging role of hydrodynamical processes. Oceanography and Marine Biology, 25:113-165.

Cabeza, M., Araujo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. and Moilanen, A., 2004. Combining probabilities of occurrence with spatial reserve design. Journal of Applied Ecology, 41:252-262.

Cadrin, S.X. and Secor, D.H., 2009. Accounting for spatial population structure in stock assessment: past, present, and future. The future of fisheries science in North America. . Springer, New York, pp. 405–426.

Canadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E. and Hammond, P.S., 2005. Habitat preference modelling as a conservation took proposals for marine protected areas for cetaceans in southern Spanish waters. Aquatic Conservation: Marine and Freshwater Ecosystems, 15:495-521.

Capinha, C., Leung, B. and Anastacio, P., *accepted*. Invasiveness models: an evaluation of using different calibration sets. Ecography.

Carpenter, J. and Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine, 19:1141-1164.

Carr, M.H., Neigel, J.E., Estes, J.A., Andelman, S., Warner, R.R. and Largier, J.L., 2003. Comparing marine and terrestrial ecosystems: Implications for the design of coastal marine reserves. Ecological Applications, 13:90-107.

Chang, Y.C., Hong, F.W. and Lee, M.T., 2008. A system dynamic based DSS for sustainable coral reef management in Kenting coastal zone, Taiwan. Ecological Modelling, 211:153-168.

Chefaoui, R.M. and Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling, 210:478-486.

Cheng, B. and Titterington, D.M., 1994. Neural Networks - a Review from a Statistical Perspective. Statistical Science, 9:2-30.

Cheung, W.W.L., Lam, V.W.Y., Sarmiento, J.L., Kearney, K., Watson, R. and Pauly, D., 2009. Projecting global marine biodiversity impacts under climate change scenarios. Fish and Fisheries, 10:235-251.

Clark, R.D., Christensen, J.D., Monaco, M.E., Caldwell, P.A., Matthews, G.A. and Minello, T.J., 2004. A habitat-use model to determine essential fish habitat for juvenile brown shrimp (*Farfantepenaeus aztecus*) in Galveston Bay, Texas. Fisheries Bulletin, 102:264-277.

Commito, J.A., Thrush, S.F., Pridmore, R.D., Hewitt, J.E. and Cummings, V.J., 1995. Dispersal dynamics in a wind-driven benthic system. Limnology and Oceanography:1513-1518.

Costanza, R., d'Arge, R., De Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V. and Paruelo, J., 1998. The value of the world's ecosystem services and natural capital. Ecological Economics, 25:3-15.

Davies, A.J., Wisshak, M., Orr, J.C. and Roberts, J.M., 2008. Predicting suitable habitat for the cold-water coral *Lophelia pertusa* (Scleractinia). Deep-Sea Research Part I-Oceanographic Research Papers, 55:1048-1062.

de la Huz, R., Lastra, M. and Lopez, J., 2002. The influence of sediment grain size on burrowing, growth and metabolism of *Donax trunculus* L. (Bivalvia : Donacidae). Journal of Sea Research, 47:85-95.

de Segura, A.G., Hammond, P.S., Canadas, A. and Raga, J.A., 2007. Comparing cetacean abundance estimates derived from spatial models and design-based line transect methods. Marine Ecology-Progress Series, 329:289-299.

Dedecker, A.P., Goethals, P.L.M., Gabriels, W. and De Pauw, N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecological Modelling, 174:161-173.

Degraer, S., Mouton, I., De Neve, L. and Vincx, M., 1999a. Community structure and intertidal zonation of the macrobenthos on a macrotidal, ultra-dissipative sandy beach: Summer-winter comparison. Estuaries, 22:742-752.

Degraer, S., Van Lancker, V., Moerkerke, G., Van Hoey, G., Vanstaen, K., Vincx, M. and Henriet, J.-P., 2003. Evaluation of the ecological value of the foreshore: habitatmodel and macrobenthic side-scan sonar interpretation: extension along the Belgian Coastal Zone. Final report. Ministry of the Flemish Community, Environment and Infrastructure Department Waterways and Marine Affairs Administration, Coastal Waterways.63 p.

Degraer, S., Van Lancker, V., Moerkerke, G., Van Hoey, G., Vincx, M., Jacobs, P. and Henriet, J.-P., 2002. Intensive evaluation of the evolution of a protected benthic habitat: HABITAT. Final report. Federal Office for Scientific, Technical and Cultural Affairs (OSTC) – Ministry of the Flemish Community, Environment and Infrastructure Department Waterways and Marine Affairs Administration, Coastal Waterways.124 p.

Degraer, S., Verfaillie, E., Willems, W., Adriaens, E., Vincx, M. and Van Lancker, V., 2008. Habitat suitability modelling as a mapping tool for macrobenthic communities: An example from the Belgian part of the North Sea. Continental Shelf Research, 28:369-379.

Degraer, S., Vincx, M., Meire, P. and Offringa, H., 1999b. The macrozoobenthos of an important wintering area of the common scoter (*Melanitta nigra*). Journal of the Marine Biological Association of the United Kingdom, 79:243-251.

Degraer, S., Wittoeck, J., Appeltans, W., Cooreman, K., Deprez, T., Hillewaert, H., Hostens, K., Mees, J., Vanden Berghe, E. and Vincx, M., 2006. The macrobenthos atlas of the Belgian part of the North Sea D:2005/1191/5. Belgian Science Policy Brussels, Belgium, 164 p.

Demuth, H., Beale, M. and Hagan, M., 2008. Neural Network Toolbox 6 User's Guide. Matlab:1-907.

Derous, S., Agardy, T., Hillewaert, H., Hostens, K., Jamieson, G., Lieberknecht, L., Mees, J., Moulaert, I., Olenin, S., Paelinckx, D., Rabaut, M., Rachor, E., Roff, J., Stienen, E.W.M., van der Wal, J.T., van Lancker, V., Verfaillie, E., Vincx, M., Weslawski, J.M. and Degraer, S., 2007. A concept for biological valuation in the marine environment. Oceanologia, 49:99-128.

Desroy, N., Janson, A.L., Denis, L., Charrier, G., Lesourd, S. and Dauvin, J.C., 2007. The intra-annual variability of soft-bottom macrobenthos abundance patterns in the North Channel of the Seine estuary. Hydrobiologia, 588:173-188.

DeVeaux, R.D. and Ungar, L.H., 1996. Neural networks in applied statistics - Discussion. Technometrics, 38:215-218.

Diaz, R.J. and Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science, 321:926.

Dimopoulos, Y., Bourret, P. and Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters, 2:1-4.

Doniol-Valcroze, T., Berteaux, D., Larouche, P. and Sears, R., 2007. Influence of thermal fronts on habitat selection by four rorqual whale species in the Gulf of St. Lawrence. Marine Ecology-Progress Series, 335:207-216.

Douvere, F., Maes, F., Vanhulle, A. and Schrijvers, J., 2007. The role of marine spatial planning in sea use management: the Belgian case. Marine Policy, 31:182-191.

Dreyfus-Leon, M. and Kleiber, P., 2001. A spatial individual behaviour-based model approach of the yellowfin tuna fishery in the eastern Pacific Ocean. Ecological Modelling, 146:47-56.

Ducrotoy, J.P., Elliott, M. and De Jonge, V.N., 2000. The North Sea. Marine Pollution Bulletin, 41:5-23.

Dunstan, P.K., Foster, S.D. and Darnell, R., *submitted*. Model Based Grouping of Species across Environmental Gradients.

Dzeroski, S. and Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Islands. Ecological Modelling, 170:219-226.

D'heygere, T., Goethals, P.L.M. and De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecological Modelling, 160:291-300.

Eastwood, P.D., Meaden, G.J., Carpentier, A. and Rogers, S.I., 2003. Estimating limits to the spatial extent and suitability of sole (*Solea solea*) nursery grounds in the Dover Strait. Journal of Sea Research, 50:151-165.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. and Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29:129-151.

Elith, J. and Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology Evolution and Systematics, 40:677-697.

Eltringham, S.K., 1971. Life in mud and sand. English Universities Press, London, 218 p.

Engler, R., Guisan, A. and Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Ecology, 41:263-274.

Ferrier, S. and Guisan, A., 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology, 43:393-404.

Fettweis, M., Nechad, B. and Van den Eynde, D., 2007. An estimate of the suspended particulate matter (SPM) transport in the southern North Sea using SeaWiFS images, in situ measurements and numerical model results. Continental Shelf Research, 27:1568-1583.

Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24:38-49.

Foody, G.M. and Cutler, M.E.J., 2006. Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. Ecological Modelling, 195:37-42.

Forbes, A.D., 1995. Classification-Algorithm Evaluation - 5 Performance-Measures Based On Confusion Matrices. Journal of Clinical Monitoring, 11:189-206.

Forster, S. and Graf, G. 1995. Impact of irrigation on oxygen flux into the sediment: Intermittent pumping by *Callianassa subterranea* and ‚piston-pumping' by *Lanice conchilega*. Marine Biology, 123: 335-346.

Francis, M.P., Morrison, M.A., Leathwick, J., Walsh, C. and Middleton, C., 2005. Predictive models of small fish presence and abundance in northern New Zealand harbours. Estuarine Coastal and Shelf Science, 64:419-435.

Franklin, J. and Miller, A.J., 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge, 320 p.

Frid, C.L.J., Harwood, K.G., Hall, S.J. and Hall, J.A., 2000. Long-term changes in the benthic communities on North Sea fishing grounds. ICES Journal of Marine Science, 57:1303-1309.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I. and Thuiller, W., *in press*. Predicting potential distributions of invasive species: where to go from here? Diversity and distributions.

Gevrey, M., Dimopoulos, I. and Lek, S., 2006. Two-way interaction of input variables in the sensitivity analysis of neural network models. Ecological modelling, 195:43-50.

Gibson, R.N., 1994. Impact of habitat quality and quantity on the recruitment of juvenile flatfishes. Netherlands Journal of Sea Research, 32:191-206.

Gibson, R.N. and Robb, L., 2000. Sediment selection in juvenile plaice and its behavioural basis. Journal of Fish Biology, 56:1258-1275.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S. and De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology, 41:491-508.

Gogina, M., Glockzin, M. and Zettler, M.L., 2010. Distribution of benthic macrofaunal communities in the western Baltic Sea with regard to near-bottom environmental parameters. 2. Modelling and prediction. Journal of Marine Systems, 80:57-70.

Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J. and Moritz, C., 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution, 58:1781-1793.

Graham, M.H., Kinlan, B.P., Druehl, L.D., Garske, L.E. and Banks, S., 2007. Deep-water kelp refugia as potential hotspots of tropical marine diversity and productivity. Proceedings of the National Academy of Sciences of the United States of America, 104:16576-16580.

Gray, J.S., 1981. The Ecology of Marine Sediments. Cambridge Studies in Modern Biology, Cambridge University Press, Cambridge. 185 p.

Gremare, A., Labrune, C., Berghe, E.V., Amouroux, J.M., Bachelet, G., Zettler, M.L., Vanaverbeke, J., Fleischer, D., Bigot, L., Maire, O., Deflandre, B., Craeymeersch, J., Degraer, S., Dounas, C., Duineveld, G., Heip, C., Herrmann, M., Hummel, H., Karakassis, I., Kedra, M., Kendall, M., Kingston, P., Laudien, J., Occhipinti-Ambrogi, A., Rachor, E., Sarda, R., Speybroeck, J., Van Hoey, G., Vincx, M., Whomersley, P., Willems, W., Wlodarska-Kowaiczuk, M. and Zenetos, A., 2009. Comparison of the performances of two biotic indices based on the MacroBen database. Marine Ecology-Progress Series, 382:297-311.

Guinan, J., Grehan, A.J., Dolan, M.F.J. and Brown, C., 2008. Quantifying relationships between video observations of cold-water coral cover and seafloor features in Rockall Trough, west of Ireland. Marine Ecology Progress Series, 375:125-138.

Guinet, C., Dubroca, L., Lea, M.A., Goldsworthy, S., Cherel, Y., Duhamel, G., Bonadonna, F. and Donnay, J.P., 2001. Spatial distribution of foraging in female Antarctic fur seals *Arctocephalus gazella* in relation to oceanographic variables: a scale-dependent approach using geographic information systems. Marine Ecology-Progress Series, 219:251-264.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A. and Zimmermann, N.E., 2006a. Using niche-based models to improve the sampling of rare species. Conservation Biology, 20:501-511.

Guisan, A., Graham, C.H., Elith, J. and Huettmann, F., 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions, 13:332-340.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M.C., Aspinall, R. and Hastie, T., 2006b. Making better biogeographical predictions of species' distributions. Journal of Applied Ecology, 43:386-392.

Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8:993-1009.

Guisan, A. and Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135:147-186.

Haelters, J., Vigin, L., Stienen, E.W.M., Scory, S., Kuijken, E. and Jacques, T.G., 2004. Ornithologisch belang van de Belgische zeegebieden. Identificatie van mariene gebieden die in aanmerking komen als speciale beschermingszone in uitvoering van de Europese vogelrichtlijn, MUMM and INBO, Brussels, Belgium.

Hagan, M.T., Demuth, H.B. and Beale, M.H., 1996. Neural network design. PWS Boston, MA. 421 p.

Hagberg, J., Jonzen, N., Lundberg, P. and Ripa, J., 2003. Uncertain biotic and abiotic interactions in benthic communities. Oikos, 100:353-361.

Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., Fujita, R., Heinemann, D., Lenihan, H.S., Madin, E.M.P., Perry, M.T., Selig, E.R., Spalding, M., Steneck, R. and Watson, R., 2008. A global map of human impact on marine ecosystems. Science, 319:948-952.

Haputhantri, S.S.K. and Jayawardane, P.A.A.T., 2006. Predictive models for penaeid shrimp abundance in the seas off Negombo and Hendala, Sri Lanka. Fisheries Research, 77:34-44.

Hartmann-Schröder, G. (1996): Annelida, Borstenwürmer, Polychaeta. – In: Dahl, F. (Ed.): Die Tierwelt Deutschlands und der angrenzenden Meeresteile, 58, 2. Aufl., 648 pp.; Jena (Fischer).

Hastie, T., Tibshirani, R. and Friedman, J., 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag. 764 p.

Haykin, S. and Network, N., 1999. Neural Networks: a comprehensive foundation. Prentice Hall. 842 p.

Heglund, P.J., 2002. Foundations of species-environment relations. In: J.M. Scott (Editor), Predicting species occurences. Island Press, Washington, pp. 35-41.

Heikkinen, R.K., Luoto, M., Araujo, M.B., Virkkala, R., Thuiller, W. and Sykes, M.T., 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography, 30:751-777.

Heuers, J., Zühlke, R., Dittmann, S., Günther, C.P., Hildenbrandt, H., Grimm, V. and Jaklin, S., 1998. A model on the distribution and abundance of the tube building polychaete *Lanice conchilega* (Pallas, 1766) in the intertidal of the Wadden Sea. Verhandlungen Gesellschaft Ökologie:28,207-215.

Hirzel, A. and Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling, 157:331-341.

Hirzel, A.H. and Arlettaz, R., 2003. Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. Environmental Management, 32:614-623.

Hirzel, A.H., Hausser, J., Chessel, D. and Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology, 83:2027-2036.

Hirzel, A.H., Helfer, V. and Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecological Modelling, 145:111-121.

Hirzel, A.H. and Le Lay, G., 2008. Habitat suitability modelling and niche theory. Journal of Applied Ecology, 45:1372-1381.

Hosmer, D.W., Jovanovic, B. and Lemeshow, S., 1989. Best subsets logistic regression. Biometrics, 45:1265-1270.

Houziaux, J.-S., Kerckhof, F., Degrendele, K., Roche, M. and Norro, A., 2007. The Hinder banks: yet an important area for the Belgian marine biodiversity?, Belgian Science Policy, Brussels, Belgium.

Huettmann, F. and Diamond, A.W., 2001. Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic. Ecological modelling, 141:261-298.

Hurvich, C.M. and Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika, 76:297.

Huxham, M. and Richards, M., 2003. Can postlarval bivalves select sediment type during settlement? A field test with *Macoma balthica* (L.) and *Cerastoderma edule* (L.). Journal of Experimental Marine Biology and Ecology, 288:279-293.

Iampietro, P., Young, M.A. and Kvitek, R.G., 2008. Multivariate Prediction of Rockfish Habitat Suitability in Cordell Bank National Marine Sanctuary and Del Monte Shalebeds, California, USA. Marine Geodesy, 31:359-371.

Iampietro, P.J., Kvitek, R.G. and Morris, E., 2005. Recent advances in automated genus-specific marine habitat mapping enabled by high-resolution multibeam bathymetry. Marine Technology Society Journal, 39:83-93.

Inglis, G.J., Hurren, H., Oldman, J. and Haskew, R., 2006. Using habitat suitability index and particle dispersion models for early detection of marine invaders. Ecological Applications, 16:1377-1390.

IPCC, 2007. Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate

Change. Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.), IPCC, Geneva, Switzerland.

Jensen, O.P., Seppelt, R., Miller, T.J. and Bauer, L.J., 2005. Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. Marine Ecology-Progress Series, 299:239-255.

Jimenez-Valverde, A. and Lobo, J.M., 2006. The ghost of unbalanced species distribution data in geographical model predictions. Diversity and Distributions, 12:521-524.

Jiménez-Valverde, A., Lobo, J.M. and Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecology, 10:196-205.

Jones, S.E. and Jago, C.F. 1993. In situ assessment of modification of sediment properties by burrowing invertebrates. Marine Biology, 115: 133-142.

Kaiser, M.J., 2005. Are marine protected areas a red herring or fisheries panacea? Canadian Journal of Fisheries and Aquatic Sciences, 62:1194-1199.

Kearney, M. and Porter, W., 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters, 12:334-350.

Kinne, O., 1970. Temperature: animals-invertebrates. Marine ecology, 1:407-514.

Koubbi, P., Loots, C., Cotonnec, G., Harlay, X., Grioche, A., Vaz, S., Martin, C., Walkey, M. and Carpentier, A., 2006. Spatial patterns and GIS habitat modelling of *Solea solea*, *Pleuronectes flesus* and *Limanda limanda* fish larvae in the eastern English Channel during the spring. Scientia Marina, 70:147-157.

Kröncke, I., Reiss, H., Eggleton, J.D., Bergman, M., Cochrane, S., Craeymeersch, J., Degraer, S., Desroy, N., Dewarumez, J.M., Duineveld, G., Essink, K., Hillewaert, H., Lavaleye, M., Moll, A., Nehring, S., Newell, R.C., Oug, E., Pohlmann, T., Rachor, E., Robertson, J.C., Rumohr, H., Schratzberger, R.S., Smith, R., Vanden Berghe, E., Van Dalfsen, J., Van Hoey, G., Vincx, M., Willems, W. and Rees, H., *submitted*. Changes in North Sea macrofauna communities and species distribution between 1986 and 2000. ICES Journal of Marine Science.

Kröncke, I., Zeiss, B. and Rensing, C., 2001. Long-term variability in macrofauna species compositon off the island of Norderney (east Frisia, Germany) relation to changes in climatic and environmental conditions. Senckenbergiana maritima, 31:65-82.

Kupschus, S., 2003. Development and evaluation of statistical habitat suitability models: an example based on juvenile spotted seatrout Cynoscion nebulosus. Marine Ecology-Progress Series, 265:197-212.

Kutner, M.H., Nachtsheim, C.J., Li, W. and Neter, J., 2005. Applied linear statistical models. ASA, 1396 p.

Lacroix, G., Ruddick, K., Ozer, J. and Lancelot, C., 2004. Modelling the impact of the Scheldt and Rhine/Meuse plumes on the salinity distribution in Belgian waters (southern North Sea). Journal of Sea Research, 52:149-163.

Le Pape, O., Baulier, L., Cloarec, A., Martin, J., Le Loc'h, F. and Desaunay, Y., 2007. Habitat suitability for juvenile common sole (*Solea solea*, L.) in the Bay of Biscay (France): A quantitative description using indicators based on epibenthic fauna. Journal of Sea Research, 57:126-136.

Le Pape, O., Chauvet, F., Mahevas, S., Lazure, P., Guerault, D. and Desaunay, Y., 2003. Quantitative description of habitat suitability for the juvenile common sole (*Solea solea*, L.) in the Bay of Biscay (France) and the contribution of different habitats to the adult population. Journal of Sea Research, 50:139-149.

Le Pape, O., Guerault, D. and Desaunay, Y., 2004. Effect of an invasive mollusc, American slipper limpet Crepidula fornicata, on habitat suitability for juvenile common sole *Solea solea* in the Bay of Biscay. Marine Ecology-Progress Series, 277:107-115.

Leathwick, J.R. and Austin, M.P., 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology, 82:2560-2573.

Lee, J.H.W., Huang, Y., Dickman, M. and Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecological Modelling, 159:179-201.

Legendre, P., 1993. Spatial autocorrelation - trouble or new paradigm. Ecology, 74:1659-1673.

Lek, S. and Guegan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling, 120:65-73.

Liu, C.R., Berry, P.M., Dawson, T.P. and Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28:385-393.

Lobo, J.M., Jimenez-Valverde, A. and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17:145-151.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. and Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology, 17:1591-1600.

Louzao, M., Hyrenbach, K.D., Arcos, J.M., Abello, P., De Sola, L.G. and Oro, D., 2006. Oceanographic habitat of an endangered Mediterranean procellariiform: Implications for marine protected areas. Ecological Applications, 16:1683-1695.

Lundblad, E., D.J.Wright, Miller, J., Larkin, E.M., R. Rinehart, Battista, T., Anderson, S.M., Naar, D.F. and Donahue, B.T., 2006. A benthic terrain classification scheme for American Samoa. Marine Geodesy, 29:89-111.

Luoto, M., Kuussaari, M. and Toivonen, T., 2002. Modelling butterfly distribution based on remote sensing data. Journal of Biogeography, 29:1027-1037.

Lutolf, M., Kienast, F. and Guisan, A., 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. Journal of Applied Ecology, 43:802-815.

Luyten, P.J., Jones, J.E. and Proctor, R., 2003. A numerical study of the long- and short-term temperature variability and thermal circulation in the North Sea. Journal of Physical Oceanography, 33:37-56.

Mackey, B.G. and Lindenmayer, D.B., 2001. Towards a hierarchical framework for modelling the spatial distribution of animals. Journal of Biogeography, 28:1147-1166.

MacLeod, C.D., Weir, C.R., Pierpoint, C. and Harland, E.J., 2007. The habitat preferences of marine mammals west of Scotland (UK). Journal of the Marine Biological Association of the United Kingdom, 87:157-164.

Maes, F., Schrijvers, J., Van Lancker, V., Verfaillie, E., Degraer, S., Derous, S., De Wachter, B., Volckaert, A., Vanhulle, A. and Vandenabeele, P., 2005. Towards a spatial structure plan for sustainable management of the sea. MA/02/006, Ghent University, Ghent, Belgium.

Maes, J., Van Damme, S., Meire, P. and Ollevier, F., 2004. Statistical modeling of seasonal and environmental influences on the population dynamics of an estuarine fish community. Marine Biology, 145:1033-1042.

Maggini, R., Lehmann, A., Zimmermann, N.E. and Guisan, A., 2006. Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography, 33:1729-1749.

Maier, H.R. and Dandy, G.C., 1998. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. Environmental Modelling & Software, 13:193-209.

Maier, H.R. and Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software, 15:101-124.

Manel, S., Williams, H.C. and Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology, 38:921-931.

Maravelias, C. and Papaconstantinou, C., 2003. Size-related habitat use, aggregation patterns and abundance of anglerfish (*Lophius budegassa*) in the Mediterranean Sea determined by generalized additive modelling. Journal of the Marine Biological Association of the United Kingdom, 83:1171-1178.

Maravelias, C.D., Haralabous, J. and Papaconstantinou, C., 2003. Predicting demersal fish species distributions in the Mediterranean Sea using artificial neural networks. Marine Ecology-Progress Series, 255:249-258.

Margules, C.R., Austin, M.P., Mollison, D. and Smith, F., 1994. Biological Models for Monitoring Species Decline: The Construction and Use of Data Bases [and Discussion]. Philosophical Transactions: Biological Sciences, 344:69-75.

McBreen, F., Wilson, J.G., Mackie, A.S.Y. and Aonghusa, C.N., 2008. Seabed mapping in the southern Irish Sea: predicting benthic biological communities based on sediment characteristics. Hydrobiologia, 606:93-103.

McLachlan, A., Jaramillo, E., Defeo, O., Dugan, J., de Ruyck, A. and Coetzee, P., 1995. Adaptations of bivalves to different beach types. Journal of Experimental Marine Biology and Ecology, 187:147-160.

McPherson, J.M., Walter, J. and David J, R., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? Journal of Applied Ecology, 41:811-823.

MeißNer, K. and Darr, A., 2009. Distribution of *Magelona* species (Polychaeta: Magelonidae) in the German Bight (North Sea): a modeling approach. Zoosymposia, 2:567-586.

Meißner, K., Darr, A. and Rachor, E., 2008. Development of habitat models for *Nephtys* species (Polychaeta: Nephtyidae) in the German Bight (North Sea). Journal of Sea Research, 60:271-286.

Meynard, C.N. and Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography, 34:1455-1469.

Miller, A.J., 2002. Subset selection in regression. Chapman & Hall, 256 p.

Myers, R.A. and Worm, B., 2003. Rapid worldwide depletion of predatory fish communities. Nature, 423:280-283.

Norcross, B.L., Blanchard, A. and Holladay, B.A., 1999. Comparison of models for defining nearshore flatfish nursery areas in Alaskan waters. Fisheries Oceanography, 8:50-67.

Olden, J.D. and Jackson, D.A., 2002a. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology, 47:1976-1995.

Olden, J.D. and Jackson, D.A., 2002b. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling, 154:135-150.

Olden, J.D., Joy, M.K. and Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling, 178:389-397.

Olivier, F. and Wotherspoon, S.J., 2006. Modelling habitat selection using presence-only data: Case study of a colonial hollow nesting bird, the snow petrel. Ecological Modelling, 195:187-204.

Olivier, J., Johnson, W.D. and Marshall, G.D., 2008. The logarithmic transformation and the geometric mean in reporting experimental IgE results: what are they and when and why to use them? Annals of Allergy, Asthma and Immunology, 100:333-337.

Olsson, P., Folke, C. and Hughes, T.P., 2008. Navigating the transition to ecosystem-based management of the Great Barrier Reef, Australia. Proceedings of the National Academy of Sciences, 105:9489.

OSPAR, 2000. Quality Status report 2000, London, United Kingdom.

OSPAR, 2008. Assessment of the environmental impact of dredging for navigational purposes, OSPAR, London, United Kingdom.

Özesmi, S.L. and Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecological modelling, 116:15-31.

Paruelo, J.M. and Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: A comparison of artificial neural networks and regression models. Ecological Modelling, 98:173-186.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. and Torres, F., Jr., 1998. Fishing Down Marine Food Webs. Science, 279:860-863.

Pearce, J. and Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling, 133:225-245.

Pearson, R.G., 2007. Species' Distribution Modeling for Conservation Educators and Practitioners Synthesis. American Museum of Natural History Available at http://ncep.amnh.org.

Pearson, R.G. and Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography, 12:361-371.

Pearson, R.G., Dawson, T.P., Berry, P.M. and Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecological modelling, 154:289-300.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. and Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography, 34:102-117.

Pesch, R., Pehlke, H., Jerosch, K., Schroeder, W. and Schlueter, M., 2008. Using decision trees to predict benthic communities within and near the German Exclusive Economic Zone (EEZ) of the North Sea. Environmental Monitoring and Assessment, 136:313-325.

Peters, S.V.M., Eleveld, M., Pasterkamp, H., Van der Woerd, H., DeVolder, M., Jans, S., Park, Y., Ruddick, K., Block, T., Brockmann, C., Doerffer, R., Krassemann, H., Schoenfeld, W., Jørgenson, P.V., Tislstone, T., Moore, G., Sørensen, K., Hokedal, J. and Aas, E., 2005. Atlas of the chloropyll-a concentration in the North Sea based on MERIS imagery of 2003, Edition 1.0, Vrije Universiteit, Amsterdam. 117 pp.

Peterson, A.T. and Vieglais, D.A., 2001. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. Bioscience, 51.

Phillips, S.J., Anderson, R.P. and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190:231-259.

Phillips, S.J. and Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31:161-175.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. and Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications, 19:181-197.

Piet, G.J., Rijnsdorp, A.D., Bergman, M.J.N., van Santbrink, J.W., Craeymeersch, J. and Buijs, J., 2000. A quantitative evaluation of the impact of beam trawling on benthic fauna in the southern North Sea. ICES Journal of Marine Science, 57:1332-1339.

Pittman, S.J., Christensen, J.D., Caldow, C., Menza, C. and Monaco, M.E., 2007. Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. Ecological Modelling, 204:9-21.

Pohlo, R., 1969. Confusion concerning deposit feeding in the Tellinacea. Journal of Molluscan Studies, 38:361.

Postma, H., 1967. Sediment transport and sedimentation in the estuarine environment. In: G.H. Lauff (Editor), Estuaries. American Association for the Advancement of Science, Washington, D.C., pp. 158-179.

Prost, L., Makowski, D. and Jeuffroy, M.-H., 2008. Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. Ecological modelling, 219:66-76.

Pulliam, H.R., 2000. On the relationship between niche and distribution. Ecology Letters, 3:349-361.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural Networks, 12:145-151.

Rabaut, M., Guilini, K., Van Hoey, G., Magda, V. and Degraer, S., 2007. A bio-engineered soft-bottom environment: The impact of *Lanice conchilega* on the benthic species-specific densities and community structure. Estuarine Coastal and Shelf Science, 75:525-536.

Rachor, E., Reiss, H., Degraer, S., Duineveld, G.C.A., Van Hoey, G., Lavaleye, M., Willems, W. and Rees, H.L., 2007. Structure, distribution, and characterizing species of North Sea macro-zoobenthos communities in 2000.

Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M. and Guisan, A., 2006. Are niche-based species distribution models transferable in space? Journal of Biogeography, 33:1689-1703.

Raxworthy, C.J., Ingram, C.M., Rabibisoa, N. and Pearson, R.G., 2007. Applications of ecological niche modeling for species delimitation: A review and empirical evaluation using day geckos (Phelsuma) from Madagascar. Systematic Biology, 56:907-923.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A. and Peterson, A.T., 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature, 426:837-841.

Redfern, J.V., Ferguson, M.C., Becker, E.A., Hyrenbach, K.D., Good, C., Barlow, J., Kaschner, K., Baumgartner, M.F., Forney, K.A., Ballance, L.T., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A.J., Qian, S.S., Read, A., Reilly, S.B., Torres, L. and Werner, F., 2006. Techniques for cetacean-habitat modeling. Marine Ecology-Progress Series, 310:271-295.

Rees, H., Cochrane, S., Craeymeersch, J., De Kluijver, M., Degraer, S., Desroy, N., Dewarumez, J.M., Duineveld, G., Essink, K. and Hillewaert, H., 2002. The North Sea benthos project: planning, management and objectives.

Rees, H.L., Eggleton, J.D., Rachor, E. and Vanden Berghe, E., 2007. Structure and dynamics of the North Sea benthos. ICES Cooperative research report, Kopenhagen.

Reineking, B. and Schroder, B., 2006. Constrain to perform: Regularization of habitat models. Ecological Modelling, 193:675-690.

Reiss, H. and Kröncke, I., 2005. Seasonal variability of infaunal community structures in three areas of the North Sea under different environmental conditions. Estuarine Coastal and Shelf Science, 65:253-274.

Reiss, H., Meybohm, K. and Kröncke, I., 2006. Cold winter effects on benthic macrofauna communities in near-and offshore regions of the North Sea. Helgoland Marine Research, 60:224-238.

Reyes, E., Sklar, F.H. and Day, J.W., 1994. A regional organism exchange model for simulating fish migration. Ecological Modelling, 74:255-276.

Rice, J.C., 2005. Understanding fish habitat ecology to achieve conservation. Journal of Fish Biology, 67:1.

Rijnsdorp, A.D. and Vingerhoed, B., 2001. Feeding of plaice *Pleuronectes platessa* L. and sole *Solea solea* (L.) in relation to the effects of bottom trawling. Journal of Sea Research, 45:219-229.

Rissler, L.J. and Apodaca, J.J., 2007. Adding more ecology into species delimitation: Ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). Symposium on Species Delimitation - New Approaches for Discovering Diversity. Taylor & Francis Inc, Stony Brook, NY, pp. 924-942.

Rodder, D. and Lotters, S., 2009. Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (*Hemidactylus turcicus*). Global Ecology and Biogeography, 18:674-687.

Rooper, C.N., Zimmermann, M. and Spencer, P.D., 2005. Using ecologically based relationships to predict distribution of flathead sole *Hippoglossoides elassodon* in the eastern Bering Sea. Marine Ecology-Progress Series, 290:251-262.

Rubec, P.J., Coyne, M.S., McMichael, R.H. and Monaco, M.E., 1998. Spatial methods being developed in Florida to determine essential fish habitat. Fisheries, 23:21-25.

Rubec, P.J., Smith, S.G., Coyne, M.S., White, M., Sullivan, A., MacDonals, T.C., McMichael Jr., R.H. and Wilder, D.T., 2001. Spatial modeling of fish habitat suitability in Florida estuaries. Spatial Processes and Management of Marine Populations:1-18.

Rykiel, E.J., 1996. Testing ecological models: The meaning of validation. Ecological Modelling, 90:229-244.

Salas, C., Tirado, C. and Manjon-Cabeza, M.E., 2001. Sublethal foot-predation on Donacidae (Mollusca : Bivalvia). Journal of Sea Research, 46:43-56.

Sandman, A., Isaeus, M., Bergström, U. and Kautsky, H., 2007. Spatial predictions of Baltic phytobenthic communities: Measuring robustness of Generalized Additive Models based on transect data.

Sarle, W.S., 1994. Neural networks and statistical models.

Saunders, G.M., Angeline, P.J. and Pollack, J.B., 1994. Structural and behavioral evolution of recurrent networks. Advances in Neural Information Processing Systems:88-88.

Schroder, B., 2008. Challenges of species distribution modeling belowground. Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde, 171:325-337.

Segurado, P. and Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography, 31:1555-1568.

Segurado, P., Araujo, M.B. and Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology, 43:433-444.

Shono, H., 2005. Is model selection using Akaike's information criterion appropriate for catch per unit effort standardization in large samples? Fisheries Science, 71:978-986.

Smith, L.I., 2002. A tutorial on principal components analysis. Cornell University, USA, 51:52.

Snelgrove, P.V.R., 1999. Getting to the bottom of marine biodiversity: Sedimentary habitats - Ocean bottoms are the most widespread habitat on Earth and support high biodiversity and key ecosystem services. Bioscience, 49:129-138.

Snelgrove, P.V.R., Grassle, J.P., Grassle, J.F., Petrecca, R.F. and Ma, H., 1999. In situ habitat selection by settling larvae of marine soft-sediment invertebrates. Limnology and Oceanography:1341-1347.

Snelgrove, P.V.R. and Butman, C.A., 1994. Animal-sediment relationships revisited: cause versus effect. Oceanography and Marine Biology, 32:111-177.

Stanev, E.V., Dobrynin, M., Pleskachevsky, A., Grayek, S. and Günther, H., 2009. Bed shear stress in the southern North Sea as an important driver for suspended sediment dynamics. Ocean Dynamics, 59:183-194.

Stanley, S.M., 1970. Relation of shell form to life habits in the Bivalvia (Mollusca). Geological Society of America memoirs, 125:296.

Stauffer, 2002. Linking populations to habitats: where have we been? Where are we going? In: J.M. Scott (Editor), Predicting species occurences. island Press, Washington, pp. 53-61.

Stergiou, K.I. and Browman, H.I., 2005. Bridging the gap between aquatic and terrestrial ecology - Introduction. Marine Ecology-Progress Series, 304:271-272.

Stevens, M.A. and Boness, D.J., 2003. Influences of habitat features and human disturbance on use of breeding sites by a declining population of southern fur seals (*Arctocephalus australis*). Journal of Zoology, 260:145-152.

Stevens, T. and Connolly, R.M., 2004. Testing the utility of abiotic surrogates for marine habitat mapping at scales relevant to management. Biological Conservation, 119:351-362.

Stockwell, D. and Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science, 13:143-158.

Sundermeyer, M.A., Rothschild, B.J. and Robinson, A.R., 2005. Assessment of environmental correlates with the distribution of fish stocks using a spatially explicit model. Ecological Modelling, 197:116-132.

Swets, J.A., 1988. Measuring the Accuracy of Diagnostic Systems. Science, 240:1285-1293.

R core developers team, 2009. R: A Language and Environment for Statistical Computing. Vienna. http://www.r-project.org

Thomson, J.D., Weiblen, G., Thomson, B.A., Alfaro, S. and Legendre, P., 1996. Untangling multiple factors in spatial distributions: Lilies, gophers, and rocks. Ecology, 77:1698-1715.

Thrush, S.F., Hewitt, J.E., Herman, P.M.J. and Ysebaert, T., 2005. Multi-scale analysis of species-environment relationships. Marine Ecology-Progress Series, 302:13-26.

Thrush, S.F., Hewitt, J.E., Norkko, A., Nicholls, P.E., Funnell, G.A. and Ellis, J.I., 2003. Habitat change in estuaries: predicting broad-scale responses of intertidal macrofauna to sediment mud content. Marine Ecology-Progress Series, 263:101-112.

Tibshirani, R., 1994. Comment on 'Neural networks: A review from statistical. perspective' by B. Cheng and DM Titterington. Statistical Science, 9:48–49.

Trexler, J.C. and Travis, J., 1993. Nontraditional Regression-Analyses. Ecology, 74:1629-1637.

Trueman, E.R., 1966. Bivalve Mollusks: Fluid Dynamics of Burrowing. Science 152:523.

Trueman, E.R., Brand, A.R. and Davis, P., 1966. The effect of substrate and shell shape on the burrowing of some common bivalves. Journal of Molluscan Studies, 37:97.

Turkheimer, F.E., Hinz, R. and Cunningham, V.J., 2003. On the Undecidability Among Kinetic Models; From Model Selection to Model Averaging. Journal of Cerebral Blood Flow & Metabolism, 23:490-498.

UNEP, 2004. Fifty key facts about seas and oceans. World Enviroment Day. United Nations Environment Programme, p. 4.

Václavík, T. and Meentemeyer, R.K., 2009. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecological modelling, 220:3248-3258.

Væth, M. and Skovlund, E., 2004. A simple approach to power and sample size calculations in logistic regression and Cox regression models. Statistics in Medicine, 23:1781-1792.

Van Dalfsen, J.A., Essink, K., Madsen, H.T., Birklund, J., Romero, J. and Manzanera, M., 2000. Differential response of macrozoobenthos to marine sand extraction in the North Sea and the Western Mediterranean. ICES Journal of Marine Science, 57:1439.

Van Hoey, G., Degraer, S. and Vincx, M., 2004. Macrobenthic community structure of soft-bottom sediments at the Belgian Continental Shelf. Estuarine Coastal and Shelf Science, 59:599-613.

Van Hoey, G., Vincx, M. and Degraer, S., 2005. Small- to large-scale geographical patterns within the macrobenthic *Abra alba* community. Estuarine Coastal and Shelf Science, 64:751-763.

van Katwijk, M.M., Hermus, D.C.R., de Jong, D.J., Asmus, R.M. and de Jonge, V.N., 2000. Habitat suitability of the Wadden Sea for restoration of *Zostera marina* beds. Helgoland Marine Research, 54:117-128.

Van Lancker, V.R.M., Du Four, I., Verfaillie, E., Deleu, S., Schelfaut, K., Fettweis, M., Van den Eynde, D., Francken, F., Monbaliu, J. and Giardino, A., 2007. Management, research and budgetting of aggregates in shelf seas related to end-users (Marebasse), Belgian Science Policy, Brussels, Belgium.

Van Tomme, J., Willems, W., Vincx, M. and Degraer, S., *submitted*. Modelling the relative importance of biotic interactions of macrobenthic species on West-European sandy beaches.

Vaughan, I.P. and Ormerod, S.J., 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conservation Biology, 17:1601-1611.

Verbruggen, H., Tyberghein, L., Pauly, K., Vlaeminck, C., Nieuwenhuyze, K.V., Kooistra, W., Leliaert, F. and Clerck, O.D., 2009. Macroecology meets macroevolution: evolutionary niche dynamics in the seaweed Halimeda. Global Ecology and Biogeography, 18:393-405.

Verfaillie, E., Degraer, S., Schelfaut, K., Willems, W. and Van Lancker, V., 2009a. A protocol for classifying ecologically relevant marine zones, a statistical approach. Estuarine Coastal and Shelf Science, 83:175-185.

Verfaillie, E., Du Four, I., Van Meirvenne, M. and Van Lancker, V., 2009b. Geostatistical modeling of sedimentological parameters using multi-scale terrain variables: application along the Belgian Part of the North Sea. International Journal of Geographical Information Science, 23:135-150.

Verfaillie, E., Van Lancker, V. and Van Meirvenne, M., 2006. Multivariate geostatistics for the predictive modelling of the surficial sand distribution in shelf seas. Continental Shelf Research, 26:2454-2468.

Verween, A., Hendrickx, F., Vincx, M. and Degraer, S., 2007. Larval presence prediction through logistic regression: an early warning system against *Mytilopsis leucophaeata* biofouling. Biofouling, 23:25-35.

Wagenmakers, E.J. and Farrell, S., 2004. AIC model selection using Akaike weights. Psychonomic Bulletin & Review, 11:192.

Walczak, S. and Cerpa, N., 1999. Heuristic principles for the design of artificial neural networks. Information and Software Technology, 41:107-117.

Walmsley, J., Coffen-Smout, S., Hall, T. and Herbert, G., 2007. Development of a human use objectives framework for integrated management of the Eastern Scotian Shelf. Coastal Management, 35:23-50.

White, D., 1998. Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. Journal of Applied Statistics, 25:263-282.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. and Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology, 75:1182-1189.

Wiens, J.J. and Graham, C.H., 2005. Niche conservatism: Integrating evolution, ecology, and conservation biology. Annu. Rev. Ecol. Evol. Syst., 36:519-539.

Willems, W., Goethals, P., Van den Eynde, D., Van Hoey, G., Van Lancker, V., Verfaillie, E., Vincx, M. and Degraer, S., 2008. Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*. Ecological Modelling, 212:74-79.

Willems, W., Rees, H.L., Vincx, M., Goethals, P. and Degraer, S., 2007. Relations and interactions between environmental factors and biotic properties. In: H.L. Rees and J.D. Eggleton (Editor), Structure and dynamics of the North Sea benthos. ICES, Kopenhagen, pp. 69-90.

Wilson, K.A., Westphal, M.I., Possingham, H.P. and Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biological Conservation, 122:99-112.

Wilson, M.F.J., O'Connell, B., Brown, C., Guinan, J.C. and Grehan, A.J., 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Marine Geodesy, 30:3-35.

Woodin, S.A., 1978. Refuges, Disturbance, and Community Structure - Marine Soft-Bottom Example. Ecology, 59:274-284.

Worm, B., Hilborn, R., Baum, J.K., Branch, T.A., Collie, J.S., Costello, C., Fogarty, M.J., Fulton, E.A., Hutchings, J.A. and Jennings, S., 2009. Rebuilding global fisheries. Science, 325:578.

Wright, P.J., Jensen, H. and Tuck, I., 2000. The influence of sediment type on the distribution of the lesser sandeel, *Ammodytes marinus*. Journal of Sea Research, 44:243-256.

Yee, T.W. and Mitchell, N.D., 1991. Generalized additive models in plant ecology. Journal of Vegetation Science:587-602.

Yonge, C.M., 1949. On the structure and adaptations of the Tellinacea, deposit-feeding Eulamellibranchia. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 234:29-76.

Ysebaert, T., Meire, P., Herman, P.M.J. and Verbeek, H., 2002. Macrobenthic species response surfaces along estuarine gradients: prediction by logistic regression. Marine Ecology-Progress Series, 225:79-95.

Zühlke, R., 2001. Polychaete tubes create ephemeral community patterns: *Lanice conchilega* (Pallas, 1766) associations studied over six years. Journal of Sea Research, 46:261-272.

Zühlke, R., Blome, D., van Bernem, K.H. and Dittmann, S., 1998. Effects of the Tube-Building Polychaete *Lanice conchilega* (Pallas) on Benthic Macrofauna and Nematodes in an Intertidal Sandflat. Senckenbergiana maritima, 29:131-138.

Curriculum Vitae

Personal Details

Wouter Martha Martinus Willems
Frederick Burvenichstraat 23
9050 Ledeberg
Mobile: +32494876198
E-mail: wouter.willems@ugent.be; wouterwillems100@hotmail.com
Date of Birth: 25 august 1981, Hasselt

Education

Hasselt University
Bsc. Biology (1999-2001), Distinction

Ghent University
Msc. Zoology (1999-2003), Distinction

Thesis: Cladistic analysis of the taxon Trigonostomidae (Typhloplanoida, Platyhelminthes) based on morphological data

Promotor: Prof. Dr. Magda Vincx, copromotor: Prof. Dr. Ernest Schockaert, supervisor: Dr. Wim Willems

Msc. Marine and Lacustrine Sciences (2003-2004), Great Distinction

Thesis: A GIS-approach to assess the impact of two pulp mills on intertidal biodiversity in the Howe Sound region (British Columbia, Canada)

Promotor: Prof. Dr. Ann Vanreusel, copromotor & supervisor: Prof. Dr. Shannon Bard

Work experience

2004-2010: PhD research, Ghent University, Marine Biology Section

Habitat Suitability Models for the analysis and prediction of macrobenthos in the North Sea

Promotor: Prof. Dr. Magda Vincx, copromotors: Prof. Dr. Steven Degraer, Prof. Dr. Ing. Peter Goethals, Prof. Dr. Vera Van Lancker

April 2009-October 2009: Research scientist at ILVO, Ostend

- Data management of fisheries dependent data collection and fisheries-independent scientific surveys
- Involved in ICES working groups on beam trawling and shrimp fisheries

Courses

- Summer school Neural Networks (2005, Porto, Portugal).
- Modelling human induced change in ecosystems (2007, Lund University, Lund, Sweden)
- ICES Stock assessment course (2009, ICES, Copenhagen, Denmark)

Publications

Peer reviewed publications (A1)

Published

Arvanitidis, C., P. J. Somerfield, H. Rumohr, S. Faulwetter, V. Valavanis, A. Vasileiadou, G. Chatzigeorgiou, E. V. Berghe, J. Vanaverbeke, C. Labrune, A. Gremare, M. L. Zettler, M. Kedra, M. Wlodarska-Kowalczuk, I. F. Aleffi, J. M. Amouroux, N. Anisimova, G. Bachelet, M. Buntzow, S. J. Cochrane, M. J. Costello, J. Craeymeersch, S. Dahle, S. Degraer, S. Denisenko, C. Dounas, G. Duineveld, C. Emblow, V. Escavarage, M. C. Fabri, D. Fleischer, J. S. Gray, C. H. R. Heip, M. Herrmann, H. Hummel, U. Janas, I. Karakassis, M. A. Kendall, P. Kingston, L. Kotwicki, J. Laudien, A. S. Y. Mackie, E. L. Nevrova, A. Occhipinti-Ambrogi, P. G. Oliver, F. Olsgard, R. Palerud, A. Petrov, E. Rachor, N. K. Revkov, A. Rose, R. Sarda, W. C. H. Sistermans, J. Speybroeck, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, and A. Zenetos. 2009. Biological geography of the European seas: results from the MacroBen database. Marine Ecology-Progress Series 382:265-278.

Berghe, E. V., S. Claus, W. Appeltans, S. Faulwetter, C. Arvanitidis, P. J. Somerfield, I. F. Aleffi, J. M. Amouroux, N. Anisimova, G. Bachelet, S. J. Cochrane, M. J. Costello, J. Craeymeersch, S. Dahle, S. Degraer, S. Denisenko, C. Dounas, G. Duineveld, C. Emblow, V. Escaravage, M. C. Fabri, D. Fleischer, A. Gremare, M. Herrmann, H. Hummel, I. Karakassis, M. Kedra, M. A. Kendall, P. Kingston, L. Kotwicki, C. Labrune, J. Laudien, E. L. Nevrova, A. Occhipinti-Ambrogi, F. Olsgard, R. Palerud, A. Petrov, E. Rachor, N. Revkov, H. Rumohr, R. Sarda, W. C. H. Sistermans, J. Speybroeck, U. Janas, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, M. Wlodarska-Kowalczuk, A. Zenetos, M. L. Zettler, and C. H. R. Heip. 2009. MacroBen integrated database on benthic invertebrates of European continental shelves: a tool for large-scale analysis across Europe. Marine Ecology-Progress Series 382:225-238.

Degraer, S., E. Verfaillie, W. Willems, E. Adriaens, M. Vincx, and V. Van Lancker. 2008. Habitat suitability modelling as a mapping tool for macrobenthic communities: An example from the Belgian part of the North Sea. Continental Shelf Research 28:369-379.

Escaravage, V., P. M. J. Herman, B. Merckx, M. Wlodarska-Kowalczuk, J. M. Amouroux, S. Degraer, A. Gremare, C. H. R. Heip, H. Hummel, I. Karakassis, C. Labrune, and W. Willems. 2009. Distribution patterns of macrofaunal species diversity in subtidal soft sediments: biodiversity-productivity relationships from the MacroBen database. Marine Ecology-Progress Series 382:253-264.

Gremare, A., C. Labrune, E. V. Berghe, J. M. Amouroux, G. Bachelet, M. L. Zettler, J. Vanaverbeke, D. Fleischer, L. Bigot, O. Maire, B. Deflandre, J. Craeymeersch, S. Degraer, C. Dounas, G. Duineveld, C. Heip, M. Herrmann, H. Hummel, I. Karakassis, M. Kedra, M. Kendall, P. Kingston, J. Laudien, A. Occhipinti-Ambrogi, E. Rachor, R. Sarda, J. Speybroeck, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, M. Wlodarska-Kowaiczuk, and A. Zenetos. 2009. Comparison of the performances of two biotic indices based on the MacroBen database. Marine Ecology-Progress Series 382:297-311.

Verfaillie, E., S. Degraer, K. Schelfaut, W. Willems, and V. Van Lancker. 2009. A protocol for classifying ecologically relevant marine zones, a statistical approach. Estuarine Coastal And Shelf Science 83:175-185.

Webb, T. J., I. F. Aleffi, J. M. Amouroux, G. Bachelet, S. Degraer, C. Dounas, D. Fleischer, A. Gremare, M. Herrmann, H. Hummel, I. Karakassis, M. Kedra, M. A. Kendall, L. Kotwicki, C. Labrune, E. L. Nevrova, A. Occhipinti-Ambrogi, A. Petrov, N. K. Revkov, R. Sarda, N. Simboura, J. Speybroeck, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, and M. Wlodarska-Kowalczuk. 2009. Macroecology of the European soft sediment benthos: insights from the MacroBen database. Marine Ecology-Progress Series 382:287-296.

Willems, W., P. Goethals, D. Van den Eynde, G. Van Hoey, V. Van Lancker, E. Verfaillie, M. Vincx, and S. Degraer. 2008. Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*. Ecological Modelling 212:74-79.

Submitted for publication

Kröncke, I., H. Reiss, J. D. Eggleton, M. Bergman, S. Cochrane, J. Craeymeersch, S. Degraer, N. Desroy, J. M. Dewarumez, G. Duineveld, K. Essink, H. Hillewaert, M. Lavaleye, A. Moll, S. Nehring, R. C. Newell, E. Oug, T. Pohlmann, E. Rachor, J. C. Robertson, H. Rumohr, R. S. Schratzberger, R. Smith, E. Vanden Berghe, J. Van Dalfsen, G. Van Hoey, M. Vincx, W. Willems, and H. Rees. *Submitted*. Changes in North Sea macrofauna communities and species distribution between 1986 and 2000. ICES Journal of Marine Science.

Van Tomme, J., W. Willems, M. Vincx, and S. Degraer. *submitted*. Modelling the relative importance of biotic interactions of macrobenthic species on West-European sandy beaches. Journal of Sea Research

Verfaillie, E., Degraer, S., Du Four, I., Rabaut, M., Willems, W., Van Lancker, V. The relevance of ecogeographical variables for marine habitat Suitability modelling of *Owenia fusiformis*. Estuarine, Coastal and Shelf Science.

In preparation

Willems, W., Degraer, S., Van Lancker, V., Vincx, M., Goethals, P. Applications of habitat suitability models in marine management and research

Willems, W., Degraer, S., Van Lancker, V., Vanaelst, S., Vincx, M., Goethals, P. Improved model selection in habitat suitability modelling

Willems, W., Degraer, S., Salembier, Y., Van Lancker, V., Vincx, M., Goethals, P. Integrated validation of marine habitat suitability models

Van Helmond, T. M., Aarts, G., Vandemaele, S., Willems, W. and J. J. Poos. Estimating spatial and temporal variability in juvenile plaice abundances from vessels-of-opportunity observations.

National, non-peer reviewed publications (A4)

Willems, W., Degraer, S., Vincx, M., Goethals, P. 2005. Predicting the occurrence of benthic species in the North Sea. Communications in Agricultural and Applied Biological Science. 70(2): 293-296.

Reports

Degraer, S., Willems, W., Adriaens, E., Vincx, M. 2005. Ecological zonation, in: Maes, F. *et al.* (Ed.) (2005). Towards a Spatial Structure Plan for Sustainable Management of the Sea: Mixed actions - Final report: SPSD II (MA/02/006). pp. 14-22

Willems, W., Rees, H., Vincx, M., Goethals, P., Degraer, S. 2007. Relations and interactions between environmental factors and North Sea macrofauna. *in* "Structure and dynamics of North Sea benthos" ICES cooperative research report. no. 288. pp. 69-90

Willems, W., Rees, H., Vincx, M., Goethals, P., Degraer, S. 2007. Habitat Suitability Modelling with the ICES North Sea Benthos Data Set. *in* "Structure and dynamics of North Sea benthos" ICES cooperative research report. no. 288. pp. 179-187

Scientific activities

Scientific exchanges

- University of Thessaloniki (Thessaloniki , Greece)
  - July 2003
  - Sampling and identification of meiofauna (interstitial Turbellaria)
- Dalhousie University (Halifax, Canada)
  - February-April 2004
  - Master thesis in laboratory of Ecotoxicology (Prof. Dr. Shannon Bard)

International congresses and workshops presentations

10th Benelux Congress of Zoology (Leiden, The Netherlands, 7 - 8 November 2003)
Presentation: Wouter, W., Willems, W., Artois, T. and Schockaert E. Cladistic analysis of the taxon Trigonostomidae Graff, 1905 (Turbellaria, Platyhelminthes) using morphological data

Mapping European Sea Habitats Workshop (Nottingham, United Kingdom, 24- 26 November 2004)
Presentation: Predictive Modelling: Overview of Techniques

ICES North Sea Benthos Survey, Workshop (Copenhagen, Denmark, 12 - 15 April 2005)
Presentation: Habitat suitability models for the analysis and prediction of macrobenthos in the North Sea.

European Congress on Ecological Modelling (Puschino, Russia, 19 - 23 September 2005)
Presentation: Willems, W., Goethals, P., Van den Eynde, D., Van Hoey, G., Van Lancker, V., Verfaillie, E., Vincx, M. & Degraer, S. Where is the worm? Predictive modelling of the habitat preferences of the tube-building polychaete *Lanice conchilega*

ICES North Sea Benthos Survey, Workshop (Oostende, 16 - 18 November 2005)
Presentation: Habitat suitability modelling of North Sea Macrobenthos

International Congress on Ecological Modelling (Yamaguchi, Japan, Augustus 28 – September 1 2006)
Presentation: Willems, W. Degraer, S., Vincx, M., Goethals, P. The use of standard neural network architecture and training in ecological modelling: Lessons learned from a predictive modelling exercise of North Sea macrobenthos.

Mapping European Sea Habitats (MESH). Final conference. (Dublin, 14-15 March 2007)
Co-author presentation: Verfaillie, E., Degraer, S., Long, D., Maljers, D., Schelfaut, K., Willems, W., Van Heteren, S., Van Lancker, V. Towards high resolution habitat maps of the Southern North Sea.

ICES Annual Science Congress (Helsinki, 17-21 September 2007)
Presentation: Willems, W., Rees H., Vincx, M., Goethals P. Degraer S. Relations and interactions between environmental factors and biotic properties

ICES Working Group on Marine Habitat Mapping (Horta, Azores, Portugal, 31 March - 4 April 2008)

Presentation 1: Habitat suitability models for application in marine management: a review

Presentation 2: Confidence and sources of error in spatial models

ICES workshop - Climate related Benthic Processes in the North Sea (Wilhelmshaven, Germany, 8-11 December 2008)

Presentation: Habitat suitability models for application in marine management: a review

International Congress on Ecological Modelling (Quebec, Canada, 6-9 October 2009)

Presentation: Willems, W. Degraer, S., Verfaillie, E., Vincx, M., Goethals, P. Applications of habitat suitability models in marine management and research

National congresses and symposia presentations

Vliz Young Scientists Day, Bruges, 25 February 2005.

Willems. W., Goethals, P., Van Lancker, V., Verfaillie, E., Vincx, M. & Degraer, S. Poster: Habitat Suitability Modelling of the North Sea Macrobenthos: Data Exploration

11th PhD Symposium on Applied Biological Sciences 2005, Leuven, 6 October 2005.

Willems, W., Goethals, P., Van Lancker, V., Verfaillie, E., Vincx, M. & Degraer, S. Habitat Suitability Modelling of the North Sea Macrobenthos

Fourth Marine Biology Section internal Symposium (FOMBSS), Ghent, 10 March 2006

Willems, W., Goethals, P., Van Lancker, V., Verfaillie, E., Vincx, M. & Degraer, S. Habitat suitability models for the prediction and analysis of macrobenthos in the North Sea

SIxth Marine Biology Section internal Symposium (SIMBSS), Ghent, 2 February 2008

Willems, W., Goethals, P., Van Lancker, V., Verfaillie, E., Vincx, M. & Degraer, S. Habitat suitability models for application in marine management: a review

Thesis supervision and reading committees

Supervision students

Cnudde, Clio & Heynssens, Els. Oligochaeten als voedselbron voor vogels in het natuurreservaat het Zwin. Academic year 2005-2006.

Promotor: Dr. Steven Degraer, CoPromotor: Prof. Dr. Magda Vincx, supervisor: Carl Van Colen, Wouter Willems. Student paper. *Msc. Zoology*. UGent, Ghent.

Salembier, Yves. Het experimenteel testen van de habitatpreferentie van macrobenthossoorten uit de Noordzee Academic year 2006-2007

Promotor: Prof. Dr. Ir. Peter Goethals, Dr. Steven Degraer, supervisor: Wouter Willems. *Msc. in Environmental sanitation.* UGent, Ghent.

Chinh Tran, Khuong. Modelling the distribution of selected benthic species (*Echinocardium cordatum, Ophelia borealis and Paramphinome jeffreysii)* of the North Sea. Academic year 2006-2007.

Promotor: Dr. Steven Degraer, supervisor: Wouter Willems. *Msc. Ecological Marine Management*, VUB, Brussels.

Reading committee

Tyberghien, Lennert. Combining phylogenetic, geographic and macro-ecological data to determine niche evoloution and biogeography of the marine green algal genus *Halimeda.* Academic year 2007-2008.

Promotor: Dr. Heroen Verbruggen, supervisor: Klaas Pauly. *Msc. in Marine and Lacustrine Sciences*, UGent, Ghent.

Van Nieuwenhuyze, Katrien. Evolution of the macro-ecological niche in the marine green algal genus *Codium*. Academiejaar 2007-2008.

Promotor: Dr. Heroen Verbruggen, supervisor: Klaas Pauly. *Msc. in Marine and Lacustrine Sciences*, UGent, Ghent.

Scientific memberships

- ICES North Sea Benthos Project (project finished in 2008)
- ICES Working Group on Marine Habitat Mapping
- Flanders marine institute (VLIZ)

Education

- Guest lecture 1st Master Bio-Engineering
  "Habitat Suitability modelling as tools in marine management"

- 2004-2008: Teaching assistant practical exercises Biostatistics 1st Master Zoology/Botany (Prof. Dr. Ann Vanreusel)