Universität Rostock Traditio et Innovatio

Modeling the biogeography of pelagic diatoms of the Southern Ocean

Dissertation

zur
Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Rostock

vorgelegt von
Stefan Pinkernell

Bremerhaven, 16.11.2017

https://doi.org/10.18453/rosdok_id00002404

Gutachter

1. Prof. Dr. Ulf Karsten
   Universität Rostock, Institut für Biowissenschaften
   Albert-Einstein-Str. 3, 18059 Rostock

2. Prof. Dr. Anya Waite
   Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung
   Am Handelshafen 12, 27570 Bremerhaven

Datum der Einreichung: 03.08.2017
Datum der Verteidigung: 11.12.2017

Abstract

Species distribution models (SDM) are a widely used and well-established method for biogeographical research on terrestrial organisms. Though already used for decades, experience with marine species is scarce, especially for protists. More and more observation data, sometimes even aggregated over centuries, become available also for the marine world, which together with high-quality environmental data form a promising base for marine SDMs. In contrast to these SDMs, typical biogeographical studies of diatoms only considered observation data from a few transects.

Species distribution methods were evaluated for marine pelagic diatoms in the Southern Ocean at the example of *F. kerguelensis*. Based on the experience with these models, SDMs for further species were built to study biogeographical patterns. The anthropogenic impact of climate change on these species is assessed by model projections on future scenarios for the end of this century.

Besides observation data from public data repositories such as GBIF, own observations from the Hustedt diatom collection were used. The models presented here rely on so-called presence only observation data. For this simple data type, Maxent has been proven to be a good modeling method.

SDM seems a suitable modeling method to study biogeography of marine pelagic diatoms in the Southern Ocean. Models of decent quality could be built, despite partly poor data. Future projections indicate a moderate decrease of the suitable areas towards the end of the century for most of the investigated species.

Zusammenfassung

Spezies Verbreitungsmodelle (Species Distribution Models/ SDM) sind eine vor allem für terrestrische Organismen etablierte und weit verbreitete Form von Habitat Modellen. Obwohl diese Modelle schon seit Jahrzehnten verwendet werden gibt es damit bisher nur wenig Erfahrung in der Modellierung mariner Organismen, insbesondere von Protisten. Immer mehr Observationsdaten werden veröffentlicht, die teilweise über lange Zeiträume gesammelt wurden. Zusammen mit hochqualitativen Umweltdatensätzen bilden sie eine vielversprechende Datenbasis für marine Verbreitungsmodelle. Studien zur biogeografischen Verbreitung basieren im Gegensatz zu diesen Verbreitungsmodellen typischerweise nur auf Daten aus wenigen Transekten.

Diese Art von Verbreitungsmodellen für marine pelagische Diatomeen im Süd Ozean wurde zunächst am Beispiel von *Fragilariopsis kerguelensis* evaluiert. Basierend auf diesen Erfahrungen wurden Modelle für weitere Spezies für vergleichende Studien erstellt. Mit Hilfe von Modellprojektionen auf Zukunftsszenarien für das Ende dieses Jahrhunderts wurden Auswirkungen des Klimawandels auf die potentielle Verbreitung dieser Arten abgeschätzt.

Neben öffentlichen Datenbeständen, wie z.B. GBIF, wurde auch die Hustedt Diatomeen Sammlung für Observationsdaten genutzt. Die Modelle basieren auf sogenannten 'presence only' Daten, bei denen nur die Anwesenheit einer Spezies dokumentiert wird, nicht Abundanz oder gar Abwesenheit. Für diesen simplen Datentyp hat sich Maxent als geeignete Modellierungsmethode etabliert.

Verbreitungsmodelle haben sich für biogeographische Studien an pelagischen Diatomeen im Südozean als geeignet erwiesen. Trotz der teilweise nachwievor dürftigen Datenlage konnten qualitativ hochwertige Verbreitungsmodellen erstellt werden. Modellprojektionen auf Zukunftsszenarien deuten für die meisten untersuchten Arten einen moderaten Rückgang der Verbreitungsgebiete bis zum Ende des Jahrhunderts an.

Eigenständigkeitserklärung

Doktorandinnen/Doktoranden-Erklärung gemäß §4 Absatz 1 Buchstaben g und h der Promotionsordnung der Mathematisch- Naturwissenschaftlichen Fakultät der Universität Rostock

Name: *Stefan Pinkernell*
Anschrift: *Pestalozzistrasse 14, 27568 Bremerhaven*

Ich habe eine Dissertation zum Thema:

Modeling the biogeography of pelagic diatoms of the Southern Ocean

am Alfred Wegener Institut Helmholtz Zentrum für Polar und Meeresforschung angefertigt. Dabei wurde ich von Herrn *Professor Dr. Ulf Karsten* (Universität Rostock) und Herrn *Dr. Bánk Beszteri* (AWI) betreut.

Ich gebe folgende Erklärung ab:

1. Die Gelegenheit zum vorliegenden Promotionsvorhaben ist nicht kommerziell vermittelt worden. Insbesondere habe ich keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen/ Betreuer für die Anfertigung von Dissertationen sucht oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.

2. Ich versichere hiermit an Eides statt, dass ich die vorliegende Arbeit selbstständig angefertigt und ohne fremde Hilfe verfasst habe. Dazu habe ich keine außer den von mir angegebenen Hilfsmitteln verwendet und die den benutzten Werken inhaltlich und wörtlich entnommenen Stellen habe ich als solche kenntlich gemacht.

Bremerhaven, den 03.08.2017

Publications and contribution to conferences

Publications

- Pinkernell, S. and B. Beszteri (2014). "Potential effects of climate change on the distribution range of the main silicate sinker of the Southern Ocean." Ecology and Evolution 4(16): 3147-3161.

Conferences

- Conference-Talk
  Pinkernell, S and Beszteri, B. (2015), "Re-using diatom observation data: species distribution modeling of pelagic marine diatoms", 9th Central European Diatom Meeting, Bremerhaven, Germany, 10 March 2015 - 13 March 2015.

- Conference-Talk
  Pinkernell, S. and Beszteri, B. (2014) , "Modeling the biogeography of Antarctic phytoplankton", ECEM 2014 - 8th European Conference on Ecological Modeling, Marrakesh, Morocco, 27 October 2014 - 30 October 2014.

- Conference-Poster
  Pinkernell, S. and Beszteri, B. (2013), "A species distribution model of the marine diatom Fragilariopsis kerguelensis", EurOceans Hot Topics Conference - A Changing Ocean, Las Palmas, Gran Canaria, 6 November 2013 - 8 November 2013.

- Conference-Poster
  Beszteri, B. , Pinkernell, S. and Hinz, F. (2012), "Exploring the possibilities of collection-based ecosystem research at the Hustedt Centre", International Diatom Symposium, Ghent, Belgium, 26 August 2012 - 31 August 2012.

- Conference-Talk
  Pinkernell, S. and Beszteri, B. (2012), "Species distribution modeling of marine pelagic diatoms", 22nd International Diatom Symposium, Ghent, Belgium, 27 August 2012 - 31 August 2012.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the priority program "Antarctic Research with comparative investigations in Arctic ice areas" by grant number BE 4316/2-1.

Danksagung

Zunächst einmal möchte ich Professor Ulf Karsten von der Universität Rostock danken, der mich in seiner Arbeitsgruppe aufgenommen und damit diese Arbeit überhaupt erst möglich gemacht hat.

Mein ganz besonderer Dank gilt Bánk Besteri, meinem Betreuer am Alfred Wegener Insitut, für die ständige Diskussions- und Gesprächsbereitschaft, den fachlichen Rat, viele aufmunternde Worte und seinen Einsatz für das Gelingen meiner Doktorarbeit. Ich habe mich in dieser Arbeitsgruppe immer sehr wohl und gut betreut gefühlt.

Einen großen Dank auch an die Mitglieder des Hustedt Diatomeen Zentrums: Friedel Hinz, Michael Kloster, Sarah Olischläger, Fenina Buttler, Ute Postel und Lena Eggers. Neben fachlichem Rat und technischer Unterstützung konnte ich auch immer auf Eure Aufmunterung bei unseren gemeinsamen Kaffeepausen zählen. Dies gilt auch für die Mitarbeiter der Sektion Polare Biologische Ozeanographie am Alfred Wegener Institut, besonders aber für Erika Allhusen, die mich bei meinen Experimenten unterstützt hat. Ein großer Dank gilt auch Fabian Altvater, meinem HiWi, der mich bei der Mikroskopie-Arbeit unterstützt hat. Für die vielen Diskussionen und den fachlichen Rat möchte ich mich ebenfall bei meinem PhD-Kommitee bedanken: Christian Schäfer-Neth, Kristine Klaas, Kerstin Jerosch und Christoph Völker. Vielen Dank auch den Mitarbeitern des Rechenzentrums, besonders aber Stephan Frickenhaus und Antonie Haas. Herzlichen Dank auch an das gesamte Polmar-Team, sowie an alle Dozenten für die sehr guten Kurse.

Ich möchte mich auch bei allen Freunde bedanken, die mir die Zeit in Bremerhaven sehr angenehm gemacht haben. Ein besonderer Dank gilt natürlich meiner Freundin Ingrid, die mich immer wieder motiviert hat diese Arbeit zu beenden. Ein großer Dank geht auch an meine Eltern für Unterstützung, aufmunternde Worte, Ermutigungen und Rückhalt in allen Lebenslagen. Schließlich möchte ich auch allen hier nicht genannten Kollegen, Freunden und Bekannten danken, die mich unterstützt und zum Erfolg dieser Arbeit beigetragen haben.

Contents

List of Figures

List of Tables

1 Introduction

During the last decades, species distribution models (SDM) became more and more popular to study the biogeography of animals and plants, and to forecast potential range shifts due to climate change. Besides more sophisticated modeling approaches, rapidly growing collections of observation data and high-quality environmental data helped to improve model quality continuously. The majority of the SDMs were built to model terrestrial species. For diatoms, this method was used first to model the potential distribution of the freshwater diatom *Didymosphenia geminata* (*Kumar et al.*, 2009). Lately, distribution models were also applied to the marine world. For phytoplankton, only a few studies exist till now, and experience with protists is still scarce, especially for marine pelagic diatoms.

Biodiversity networks, such as the Global Biodiversity Information Facility (GBIF) and the Ocean Biogeographic Information System (OBIS), have become well established and are frequently used for biogeographical studies. They provide bundled information from various collections, e.g., from natural history museums, sampling campaigns, expeditions, etc. Made easily accessible over common web frameworks, they provide a tremendous data pool. The Hustedt diatom collection with over 100.000 slides for light microscopy is a potential data provider for these kinds of networks.

The many data sources lead to several different types and quality levels of data. Presence-only data is the simplest data type that can be derived from any other data type and records just an observation at a certain site and time.

Species distribution models were developed to deal with presence-only data and perform relatively well if abundance and absence data are not available. Especially the maximum entropy algorithm, implemented in the Maxent software, seems very powerful. Of course, models based on data of higher quality, e.g., absence information or abundance data, are better and desirable, but they would exclude the majority of existing observation records, sometimes collected over centuries.

This thesis discusses the use of species distribution models for pelagic marine diatoms, with a focus on the Southern Ocean. Data from public repositories are complemented by the systematic use of samples from the Hustedt Diatom Collection. The aim is to evaluate data quality and availability as well as distribution modeling techniques to model the biogeography of pelagic diatoms on the example of selected species from the Southern Ocean. Modeled and previous knowledge of diatom biogeography shall be compiled to a better understanding of current distribution patterns. As the distribution models can also be projected on modeled environmental conditions for future scenarios, potential range shifts due to global climate change and ocean warming shall be assessed.

1.1 The Southern Ocean

The Southern Ocean (SO) surrounds the Antarctic continent, and, with an area of approximately 20 mio. km^2 is the second smallest of the world's oceans. It ranges from the Subtropical front in the north to the Antarctic continent as its southern boundary. The most important feature of the SO and the adjacent ocean basins is the Antarctic Circumpolar Current (ACC), forced by strong winds. This eastward current around the Antarctic continent is not affected by any barriers, except for the Drake Passage, a bottleneck between South America and the Antarctic Peninsula. This way, the ACC connects the Atlantic, Indian and Pacific Oceans in the SO, and plays an important role, as it acts as a hub for nutrients in the ocean. The ACC has a total transport volume of 130 - 140 Sverdrup (*Pollard et al.*, 2002), which makes it the strongest of all of the ocean currents. In the strict sense, it is not a single ocean current, but rather made up of smaller, but intense ocean currents and jets. The boundaries that separate these water masses of different temperature and salinity are called fronts and build a strong frontal system in the SO (*Whitworth and Nowlin*, 1987; *Nowlin and Klinck*, 1986; *Orsi et al.*, 1995; *Belkin and Gordon*, 1996). *Deacon* (1982) described the system based on wind-driven convergences and divergences in the surface layer. *Pollard et al.* (2002) later argued, not to stick to the latter approach anymore, as the fronts of the Southern Ocean are much sharper than the wide bands of wind-driven convergences and divergences. The two descriptions go along with different terminology, with both found in literature, sometimes even mixed.

The average positions of these fronts according to *Orsi et al.* (1995) are plotted in figure 1.1; from north to south these are the Subtropical Front (STF), the Subantarctic Front (SAF), the Polar Front (PF), the Southern Antarctic Circumpolar Current Front (sACCF), and the Southern Boundary of the Antarctic Circumpolar Current. The westward flowing Antarctic Coastal Current is located between the Southern Boundary and the Antarctic continent.

The Southern Ocean is affected by extreme environmental conditions such as seasonal sea ice and partly winter darkness. Typically, the maximum sea ice extent is reached in September at the end of the austral winter, the minimum extent in February. A significant fraction of the sea ice is located at the lower latitudes and melts during the summer. Figure 1.1 shows a typical summer and winter sea ice extent. South of the polar circle (66°33'46.1"S), the sun can stay above or below the horizon the whole day. Depending on the latitude, e.g., at 60°S day length lasts only about six hours in the winter, whereas in the summer up to 19 hours can be reached. In contrast, at 40°S variation in day length is much less, ranging from 10 hours in the winter up to 15 hours in the summer.

Mixed layer depth is an important feature of the ocean's biology, especially for passively drifting organisms like diatoms. Wind stress and heat exchange at the surface are responsible for turbulent mixing of the upper water masses. *Dong et al.* (2008) determined the MLD from Argo float profiles based on density ($\Delta\rho$= 0.03 kg * m $^{-3}$) and temperature (ΔT= 0.2 °C) difference criteria. A strong seasonality for the MLD

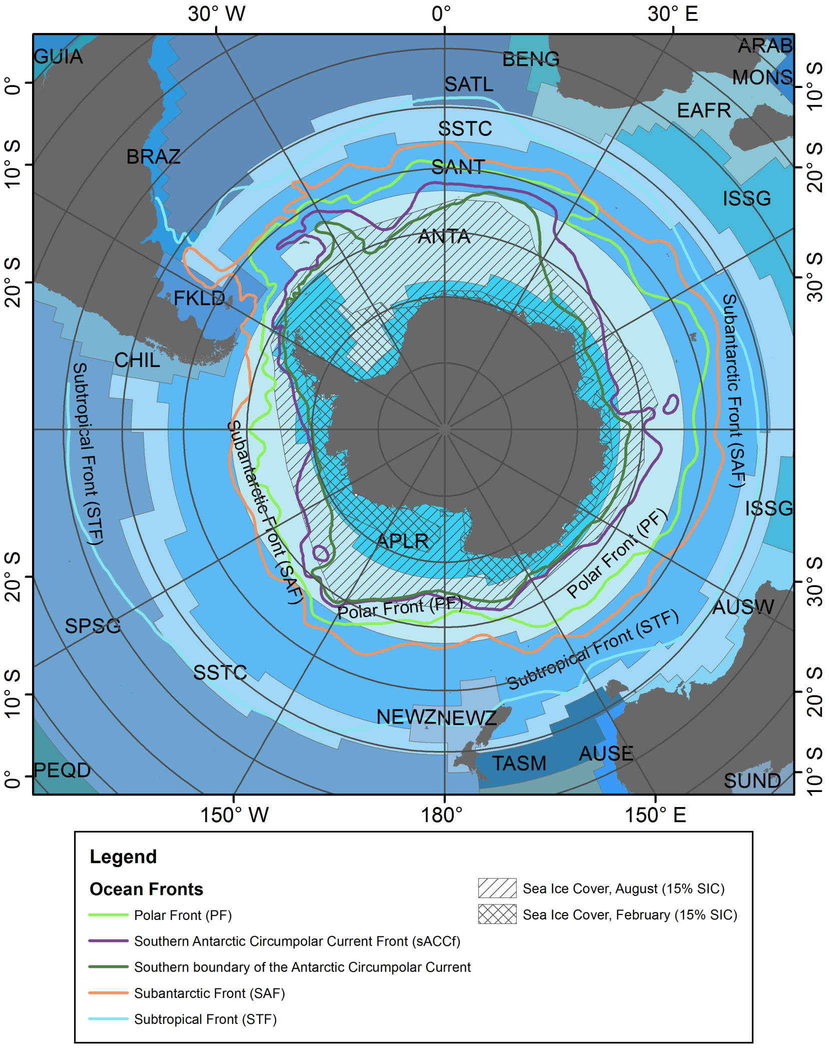

Figure 1.1: Longhurst provinces, frontal system and seasonal variability of sea ice cover in the Southern Ocean. The fronts, according to *Orsi et al.* (1995) are (from north to south): Subtropical Front (STF), Subantarctic Front (SAF), Polar Front (PF), Southern Antarctic Circumpolar Current Front (sACCF), and Southern Boundary of the Antarctic Circumpolar Current. Shaded areas indicate the extent of sea ice (*Rayner*, 2003) for summer (February) and winter (August) conditions with sea ice concentrations ≥ 15% based on an aggregated monthly dataset from 1990 to 2014. The color code indicates the ecological provinces of the World Oceans as defined by *Longhurst* (2010); in the Southern Ocean, these are the South Subtropical Convergence Province (SSTC), the Subantarctic Water Ring Province (SANT), the Antarctic Province (ANTA), and the Austral Polar Province (APLR).

was observed in the Southern Ocean, with its maximum in August/ September and the minimum in February/ March.

Due to upwelling, the Southern Ocean is well supplied with nutrients. Isopycnals in the ACC slope upwards to the south and raise nutrient-rich water to the surface (*Pollard et al.*, 2002). Along the isopycnals, diapycnal mixing with silicate-rich bottom water leads to an increase of silicate towards the south. Nitrate concentrations decrease northwards as well but to a much lesser degree.

Further, *Pollard et al.* (2002) describe difficulties to associate observed frontal jets to certain fronts. They use stratification patterns to define circumpolar features of the Southern Ocean, which are a result of changes in the relative contribution of temperature (dominating towards the equator) and salinity (dominating towards the pole). The transport of 130 - 140 Sv is mainly wind-driven, but the partitioning between the fronts is determined by the bathymetry. For this reason, the latitudinal location, the number and the strength of fronts/jets vary with longitude. They defined four zones based on the contribution of temperature and salinity to stratification and independent of the frontal jets. The Subantarctic Zone (SAZ) ranges from the Subtropical to the Subantarctic Front, with stratification mainly dominated by temperature. South of the SAZ in the Polar Frontal Zone (PFZ), temperature and salinity contribute equally to stratification. This zone reaches to the Polar Front in the south. In the Antarctic Zone (AAZ), south of the Polar Front, salinity dominates stratification. The AAZ extends to the Southern Boundary of the ACC. The Antarctic Circum Polar Current (ACC) stretches over the SAZ and the PFZ. The southernmost zone in this scheme is the zone south of the ACC (SACCZ). Its northern boundary is the so-called southern terminus of the Upper Circumpolar Deep Water (UCDW) which is equal to the Southern Boundary in the definition of *Orsi et al.* (1995) and is characterized by the lack of the subsurfaces nitrate maximum of the UCDW. *Pollard et al.* (2002) state that the exact positions of the fronts do not need to be known for this zonation, but the zones can be identified by them, still.

Longhurst (2010) separated the global ocean surface into 56 ecological partitions (see fig. 1.1 for the partitions of the SO). Interesting for this thesis are two biomes: the Antarctic Polar Biome and the Antarctic Westerly Winds Biome. The latter is divided into the South Subtropical Convergence Province (SSTC) and the Subantarctic Water Ring Province (SANT). The SSTC is the zone north of the Subtropical Front up to the subtropical gyres. South of the SSTC, the SANT ranges from the subtropical front to the polar frontal zone. Longhurst suggests to further divide the SANT into the Subantarctic Zone (SAZ) and the Polar frontal zone itself. The Antarctic Polar Biome is partitioned into two provinces, too: The Antarctic Province (ANTA) and the Austral Polar Province (APLR). ANTA reaches from the Polar front to the southern boundary of the ACC (Antarctic divergence). APLR region, south of ANTA, is the region affected by seasonal sea ice cover, mainly south of the Antarctic convergence. This includes the region around the Antarctic Peninsula and the islands (though north of the Antarctic convergence) that are affected by sea ice. As most of this province is located south of the Antarctic Circle, it is also affected by total winter darkness.

Reygondeau et al. (2013) refined Longhurst's static boundaries using a statistical method and the four environmental parameters chlorophyll *a* concentration, bathymetry, salinity, and surface temperature over ten years (1997-2007). This new model distinguishes the same 56 biogeochemical provinces but takes seasonal and interannual changes (e.g., El Niño/La Niña events) into account.

Two big gyres exist in the Southern Ocean, the Weddell Gyre in the Weddell Sea and the Ross Gyre in the Ross Sea. Both clockwise rotating gyres are connected to the ACC in the north and the Antarctic continental shelf in the south. Due to upwelling, these waters are rich in nutrients, whereas production is relatively low.

The organisms and models analyzed in this study are not limited to the Southern Ocean as defined above. The area of interest includes the adjoined southern parts of the Atlantic-, Indian- and Pacific Oceans up to approximately 30°S. However, all analyses were conducted on a world-map to account for potential global distribution patterns.

1.2 Diatoms

1.2.1 Introduction to diatoms

Diatoms, a major group within the protists, are unicellular heterokont algae, sometimes forming colonies or chains. They can be found in most aquatic environments, from freshwater to marine environments, in sea ice, in pelagic and benthic sites, but also in wetlands and soil. The main identifying feature of diatoms is the complex ornamented frustule, a cell wall of opaline silica. Frustules consist of two overlapping and usually equally shaped thecae, differing just in size: the bigger epitheca and the smaller hypotheca. Diatoms are classified based on morphological features of the diatom frustule. Species identification by light microscopy is still widely used, though many species cannot be distinguished this way taxonomically. Scanning electron microscopes (SEM) further allow the visualization of important details, which are not visible by light microscopy. Lately, also molecular techniques are used to identify cryptic species, morphologically indistinguishable but genetically different species (*Evans et al.*, 2008; *Mann and Vanormelingen*, 2013; *Rovira et al.*, 2015; *D'Alelio et al.*, 2009).

Though essential, a clear definition of a species concept in diatoms is still missing. *Mann* (1999) discussed various concepts in detail, with special regard to historical contexts and traditional approaches in diatom taxonomy. Traditionally, a morphological concept is used, though crossing experiments and the use of molecular markers can lead to a different view.

The group of diatoms is often classified as class *Bacillariophyceae* Haeckel in the phylum *Heterokontophyta* and is divided into two main groups: the order *Centrales* covers centric diatoms with radial symmetry, whereas the order *Pennales* covers the pennate diatoms with bilateral symmetry. In 1990, the diatoms were reclassified by Round, Crawford and Mann into three classes: class *Coscinodiscophyceae* Round and Crawford that covers the centric diatoms, class *Fragilariophyceae* Round, covering the araphid pennate diatoms, and class *Bacillariophyceae* Mann (emend. Haeckel 1878) covering the raphid pennate diatoms. In the currently valid classification (*Adl et al.*, 2012), diatoms are found under *Diatomea* Dumortier 1821 (equal to *Bacillariophyta* Haeckel, 1878) and are separated into two groups: *Coscinodiscophytina* Medlin & Kaczmarska 2004 with six subgroups, and *Bacillariophytina* Medlin & Kaczmarska 2004 with two subgroups.

Morphological features of the frustule are still the main criterion for diatom taxonomy. This is insofar problematic, as the underlying causes of the variations are not fully understood. *Vanormelingen et al.* (2008) give various examples, where this results in misleading perceptions, e.g., when a new species description is based on morphological differences between allopatric populations. Such differences might be purely phenotypic (see also *Cox* (1995)), and lead to an impression of restricted geographic distribution patterns.

They further mention that reproductive, molecular-genetic, physiological and ecological variations often are correlated with subtle morphological differences, which had previously been assumed to have no taxonomic significance (see also *Mann* (1999)). Algaebase[1] lists 11.199 species for the class *Bacillariophyceae* Haeckel, but up to 200.000 are estimated to exist (*Mann and Droop*, 1996). *Norton et al.* (1996) mention a range of 10.000 to 12.000 recognized species and even 100.000 to 10.000.000 estimated species for the class *Bacillariophyceae*.

Diatoms play a key role in the global biogeochemical cycles and ecology in the ocean. They account for 20–25% of the globally fixed carbon and atmospheric oxygen (Mann 1999). About 46% of the global carbon production of 105 Pg per year is attributed to the marine realm, of which diatoms are accounted for 40-45% (up to 20 Pg) (*Mann* (1999) citing *Field et al.* (1998) and *Nelson et al.* (1995)). Besides the carbon export, they also play a substantial role in the export of the other macronutrients as nitrate, phosphate, and silicate.

1.2.2 The role of diatoms in the Southern Ocean

In the Southern Ocean, diatoms are the most important group of primary producers. Approximately 150 pelagic diatom species and many more benthic species are known to occur in this region. The SO hosts several remarkably strongly silicified species, of which *Fragilariopsis kerguelensis* is the most prominent. Other highly abundant species in this area are *Eucampia antarctica*, several species of the genera *Fragilariopsis*, *Thalassiosira*, *Rhizosolenia*, *Proboscia*, *Corethron*, and *Chaetoceros*.

While diatoms in iron-replete regions such as the North Atlantic and the continental margins contribute strongly to carbon export, diatoms in iron limited regions such as the Southern Ocean mainly export silicate and much less carbon (*Assmy et al.*, 2013). Playing an important role in carbon export from surface waters, their role in the transport of carbon to the deep sea is not as high as thought before (*Ragueneau et al.*, 2006), though. According to *Klaas and Archer* (2002), calcifiers may play a more important role than diatoms in exporting carbon to the deep ocean.

[1] http://www.algaebase.org on 16. March 2016

The publication of the iron hypothesis (*Martin*, 1990) about the high nutrient, low chlorophyll paradox led to several studies on the effect of iron on phytoplankton growth and biogeochemical processes in High Nutrient Low Chlorophyll (HNLC) areas, such as the Southern Ocean. A synthesis of iron fertilization experiments is given in *De Baar et al.* (2005) and *Boyd et al.* (2007). The thick shells that are typical for several diatom species in the Southern Ocean are a result of the higher Si:N ratios reached during iron limitation, as silica deposition rates within the internal vesicles are not slowed by low iron availability (*Smetacek* (1999) citing *Hutchins and Bruland* (1998) and *Takeda* (1998)). *Timmermans and Van Der Wagt* (2010) studied the effect of iron limitation on *F. kerguelensis* regarding morphological changes and changes in nutrient depletion ratios: Iron limitation led to a decrease in growth rate, smaller cells and shorter chains, and a change in nutrient depletion rates towards an increased Si:N depletion ratio and a decreased N:P depletion ratio. In the European Iron Fertilization Experiment (EIFEX) in the Southern Ocean in 2004, a stock of thick-shelled diatoms stayed in the surface layer. Many empty shells sank continuously and contributed to a massive silicate export. In contrast, thin-shelled diatoms sank by forming aggregates and led to strong carbon export (*Assmy et al.*, 2013).

Diatoms are the main source of biogenic silica in the Southern Ocean. They transform dissolved orthosilicic acid into the hydrated amorphous silica of their frustules and contribute more to the silica cycle than silicoflagellates and radiolaria (*Treguer and Jacques*, 1992). Biogeochemical cycles of silicate and carbon are strongly decoupled, and the Southern Ocean is characterized by the highest silicate to carbon flux ratio of all ocean basins (*Ragueneau et al.*, 2002; *Dunne et al.*, 2007). Since the Southern Ocean plays an important role in the distribution of nutrients into the large ocean basins, this massive silicate export also affects the adjacent ocean basins. Several studies suggest *F. kerguelensis* to be the main player in the Southern Oceans silicon cycle (*Cortese and Gersonde*, 2007), and to dominate diatom assemblages in the water column as well as in the sediments in the Southwestern Atlantic Ocean (*Romero and Hensen*, 2002; *Olguín et al.*, 2006). Lately, *Thalassiosira lentiginosa* (Janisch) Fryxell was found to play an even more important role in silicate export in the Southern Ocean than *F. kerguelensis* (*Shukla et al.*, 2016). Regardless of which species contributes most to silicate export, the impact of diatoms on the decoupling of silicate and other macronutrients is strong. Thick-shelled diatoms, favored due to the low iron concentrations, are responsible for a massive consumption of silicate and this way decrease the silicate to nitrate ratios in surface waters. Currents in the Southern Ocean act as a hub for macronutrients, but silicate is mostly consumed here. A remarkable amount of nitrate and phosphate is transported northwards by Ekman transport (*Assmy et al.*, 2013).

Currently, the Southern Ocean is affected less by climate change than other parts of the ocean, e.g., the Arctic, but for the future, drastic changes are expected. With a global coupled climate and ocean biogeochemical model, *Bopp et al.* (2005) could explore the impact of climate change upon diatom distributions, and found it to lead to a decrease in diatom productivity. Diatoms are considered as an important functional group in the models, but a differentiated view on the species level was not part of their models.

1.2.3 Diatom Biogeography

Typically, studies on marine pelagic diatom biogeography are based on data collected on transects. Abundance, composition, and distribution of the phytoplankton of the Pacific Southern Ocean was analyzed by *Hasle* (1969) based on the material of the Brategg expedition (1947-1948). During the 1970s, she compiled observation data from expeditions to generate distribution maps, e.g., in *Hasle* (1976) distribution patterns of 26 species of morphologically and taxonomically closely related groups. Newer studies exist, too. *Cefarelli et al.* (2010) studied *Fragilariopsis* species on transects in the Argentine Sea and Antarctic waters, and *Mohan et al.* (2011) used data on a transect from 25-56°S along 45°E.

Spring phytoplankton of the Brazilian Current, the Malvinas Current, and the Drake Passage was analyzed by *Olguín et al.* (2006) and summer phytoplankton by *Olguín and A. Alder* (2011). The Brazilian and Malvinas currents adjoin in the so-called Brazil-Malvinas Confluence Zone (BMCZ) at approximately 34-35°S. While warm and temperate water species were found in the northern stations, affected by the Brazilian Current, cold water species, especially *F. kerguelensis* became more frequent south of the BMCZ (*Olguín Salinas et al.*, 2015). They list several species to become dominant south of the Subantarctic front, most notably *Asteromphalus hookeri*, *A. parvulus*, *Dactylosolen antarcticus*, *Eucampia antarctica*, *Rhizosolenia simplex*, *Thalassiosira gravida*, *T. lentiginosa*, and *T. tumida*.

Semina (2003) analyzed 75 diatom samples from globally distributed stations using a scanning electron microscope (SEM). She distinguishes three main phytogeographical regions: Arcto-Boreal in the north, Notal-Antarctic in the south, and the tropical region in between. In between these regions, mixing zones exist, where tropical species mix with the northern or southern ones. The Notal-Antarctic region is divided into two parts: the High-Antarctic (HA) and Low-Antarctic (LA), which partly includes the mixing zone, too. *Semina* (2003) states that endemism in phytoplankton usually occurs at the species level, not for higher taxa such as families and orders. For the Notal-Antarctic region, she mentions 30 endemic species citing *Hasle* (1968) and *Hasle* (1969). Further, she suggests that for the geographical distribution it should also be considered whether the species are neritic (inhabiting waters of the shelf regions) or oceanic (living in the open sea). She classified most species as neritic, with several of them also being panthalassic (open sea and shelf zones), and most of those species that are living south of the Antarctic Convergence as ice-neritic.

Together with data from museums, data along transects form a great data basis for further biogeographical studies. Nowadays, observation data from collections of, e.g., natural history museums find their way into public data repositories such as the GBIF network or OBIS. Updated distribution maps, as created by Hasle, can be generated based on these observation data, but distribution models, as used in this thesis, promise an improved mapping of distribution patterns. By model projections on future scenarios, these models also allow forecasting of potential range shifts caused by climate change.

Biogeography also plays an important role in a paleological context. Surface sediments of the Southern Ocean are widely used for paleoenvironmental reconstructions (*Zielinski and Gersonde*, 1997). *Crosta et al.* (2005) studied Southern Ocean sediments to estimate biogeography of open-ocean-related diatom species, and *Armand et al.* (2005) did the same for sea ice related diatoms. Based on diatom distributions in the sediment, Antarctic Pleistocene sea ice could be reconstructed (*Esper and Gersonde*, 2014a) and a new transfer function was developed to estimate quaternary surface water temperature for the Southern Ocean (*Esper and Gersonde*, 2014b). Diatom distributions in the southeastern Pacific surface sediments were related to current environmental variables in *Esper et al.* (2010) and *Ren et al.* (2014).

1.3 Species distribution models

1.3.1 Niche theory

The aim of species distribution models (SDM) is to predict the likelihood of a species occurrence based on environmental variables (*Guisan and Zimmermann*, 2000; *Hirzel and Le Lay*, 2008). The idea that a species' distribution is related to its environment is old and was already mentioned by Grinnell in the early 20th century (*Grinnell*, 1914, 1917b) as discussed in *Soberon and Nakamura* (2009). SDMs go along with niche concepts, where two classical ones, that of *Grinnell* (1914, 1917a,b) and that of *Elton* (1927) have to be mentioned.

Grinnell's concept is based on the requirements of the species, linking the fitness of individuals to their environment. Elton's theory in contrast also covers the relationships to other species and the impact a species has on its environment. Also, the variables that form the axes of the multidimensional niche space are different: While the Grinellian niche is typically based on scenopoetic variables, ecological variables that do not interact with others and change very slowly, the Eltonian niche is based on spatially fine-grained variables showing temporal dynamics related to ecological interactions and resource consumption (*Soberon and Nakamura*, 2009).

Hutchinson (1957) later introduced the concept of the niche as a multidimensional hyperspace, permitting positive growth, and differentiated the fundamental and the realized niche. The fundamental niche describes the range of environmental conditions in which a species could survive, whereas the realized niche describes the range of environmental conditions in which a species is really found. The realized niche is a subset of the fundamental niche, limited by dispersal, environmental conditions or biotic interactions. Grinnell and Elton both attributed the niche to the environment, whereas Hutchinson attributed the niche to the species instead and developed the abstract concept of a multidimensional hyperspace. He defined the concept of a niche duality, in which the multidimensional niche space and the physical space a species lives in are mutually connected (see also *Colwell and Rangel* (2009)). His duality now provides a powerful way to link to study environmental patterns and their relation to biogeographical distributions.

The requirement based niche theory based on *Grinnell* (1917a) and *Hutchinson* (1957) links the fitness of a species to the properties of the environment it's living in. A comprehensive overview of the numerous niche concepts can be found in *Chase and Leibold* (2003). However, the development of modeling techniques such as species distribution models resulted in an ongoing discussion (*McInerny and Etienne*, 2012a,b,c; *Soberón and Higgins*, 2014) about these niche concepts again, especially in relation to distribution models. Species distribution models are used to characterize the realized ecological niche.

A species distribution is determined by the distribution of environmental conditions this species can persist at, its niche. These conditions are investigated with species distribution models. It's clear that species distribution models go along with all the assumptions about niches, a species distribution is related to its niche, and that several assumptions go along with this. Of course, avoiding the term 'niche' and favoring the term 'species distribution model' instead does not really provide relief from these assumptions (*Warren*, 2012).

1.3.2 Overview of species distribution models

Species distribution models have a history of several decades now. In aggregate, they represent a well-established method in ecological research to study population dynamics, conservation biology, biogeography, and evolutionary ecology. A comprehensive review of SDM studies and modeling algorithms used up to the year 2000 can be found in *Guisan and Zimmermann* (2000) and later in *Guisan and Thuiller* (2005) and *Elith and Graham* (2009). Further, they are discussed in detail in various books, e.g., in *Raven et al.* (2002), *Franklin* (2010) and *Peterson et al.* (2011). Different names for these kinds of models can be found in the literature: species distribution models, ecological niche models, habitat suitability models, envelope models, etc., each of them with a slightly different focus. Following *Franklin* (2010) and *Elith and Graham* (2009), I use the term species distribution model (SDM) throughout this thesis. The number of papers about SDMs has massively increased in the recent years, e.g., 1886 citations where listed in December 2013 in Web of science for "SDM" (*Fourcade et al.*, 2014), and an article search for "species distribution model maxent" in Google Scholar in 2017 listed more than 8000 new entries during the last 5 years.

The main idea of a species distribution model is to estimate the suitability of environmental conditions for a species, based on species occurrence sites and presumed ecologically relevant environmental conditions at these sites.

The geographic and the environmental space have to be clearly distinguished (*Elith and Graham*, 2009). The geographic space usually is characterized by two dimensions, i.e., latitude and longitude. This is the space in which observation records are usually located. Depending on the study, a third or more dimensions can be necessary. In this study, the sampling month is taken into account, forming the third dimension. For marine species, sampling depth or bathymetry are a potential fourth dimension.

The environmental space is usually of a higher dimensionality, set up by the predictor variables. This is the spaces in which the models are calculated. Predictors can be direct data, e.g., the (mean) temperature over a certain time span, or further processed values, e.g., the minimum and maximum values of the temperature over a certain time span.

Conversion between the geographic and environmental space is easy. The predictor values for each observation record can be extracted from gridded data products, to transform the observation data to the environmental space. The model, once calculated in the environmental space, consists of density functions of the predictors. A model projection on a stack of spatial predictor maps results in a spatial map showing the suitability of the environment to the ecological needs of a species. This way the model - calculated in the environmental space - is mapped in geographical space, which is possible even across space and time.

A wide range of different algorithms and modeling approaches are available. Bioclim was one of the first methods used successfully in this field and according to *Booth et al.* (2014) was cited first in *Nix* (1986). It describes the ranges of predictor variables where a species typically occurs. This is very close to the n-dimensional hyperspace of Hutchinson.

In recent years, Maxent (*Phillips et al.*, 2004, 2006; *Phillips and Dudik*, 2008) has become one of the most important methods for distribution modeling. In contrast to the much older Bioclim, Maxent calculates response curves for predictor variables and indicates, in which range environmental variables are most suitable. Maxent is very strong in modeling distributions based on so-called "Presence Only" data, which is typical for e.g., collections in natural history museums. Several studies confirm a very good performance, e.g., in comparison with Maxlike (*Merow et al.*, 2013; *Fitzpatrick et al.*, 2013), with GARP (*Townsend Peterson et al.*, 2007), with GAM and MD (*Villarino et al.*, 2015).

The majority of distribution models are used to study terrestrial organisms, as the citations so far confirmed. But also studies about marine organisms exist: *Bombosch et al.* (2014) modeled the distribution of humpback and minke whales in the Southern Ocean, *Villarino et al.* (2015) compared different approaches (Maxent, GM, MD) to study the future biogeography of zooplankton in the North Atlantic and *Verbruggen et al.* (2013) and *Neiva et al.* (2014) modeled the distribution of seaweed.

Kumar et al. (2009) published one of the first studies where a species distribution model (Maxent) is used for a diatom species. They modeled the potential habitat distribution of the highly invasive freshwater diatom *Didymosphenia geminata*. Recently, a few studies on marine protists were also published: *Weinmann et al.* (2013) analyzed the distribution of Foraminifera in the Mediterranean sea with Maxent and predicted their distribution for the years 2050 and 2100, *Brun et al.* (2015) used Maxent models for open ocean phytoplankton taxa to classify ecological strategies of microalgal groups (of diatoms, coccolithophores, diazotrophs, phaeocystis, and picophytoplankton) according to Reynolds' C-S-R model, and *Irwin et al.* (2012) used Maxent to estimate phytoplankton niches from field data.

1.3.3 How Maxent works

Various modeling approaches and algorithms have been developed for species distribution models. During the last decade, Maxent, and especially the implementation written by S. Phillips, M. Dudik and R. Schapire (*Phillips et al.*, 2004), became well-established in this field. The aim and big advantage of this method is its performance with so-called presence-only data, the main case for records of herbaria and museums. Often collected over a long time span, these collections are an important source for occurrence data. Systematically collected data with presence and absence data, or even abundance data, are preferable and would enable more sophisticated models, but are usually not available from these repositories.

There are two views on Maxent: In the machine learning view, Maxent is described to estimate a distribution across the geographic space (*Phillips et al.*, 2006; *Phillips and Dudik*, 2008), whereas in the statistical view, the probability densities in the covariate (i.e., the environmental) space are compared (*Elith et al.*, 2011). Both views are mathematically equivalent, because niche spaces can be mapped into geographic space and vice versa. In the following, Maxent is explained by the latter approach, following the description in the paper by *Elith et al.* (2011).

It is assumed that presence-only data from locations within L, the landscape of interest, are used. $y = 1$ indicates presence records, $y = 0$ absence records, and z a vector of environmental covariates. These independent variables (also called covariates, predictors, environmental space/ -conditions) include marine environmental properties, such as water temperature and salinity, mixed layer depth, or nutrient concentrations. The probability density of covariates across L is noted by $f(z)$, over locations where a species is present by $f_1(z)$, and where a species is absent by $f_0(z)$.

$P(y = 1|z)$, the probability of a species' presence conditioned on the environment, shall be estimated. A restriction of Maxent is that with presence-only data only $f_1(z)$ can be modeled which on its own cannot model the probability of presence. In contrast, a method that uses presence and background data would allow to model $f_1(z)$ and $f(z)$. Bayes' rule says: $P(y = 1|z) = f_1(z)P(y = 1)/f(z)$. In this equation, only the term $P(y = 1)$, the so-called prevalence, which describes the proportion of occupied sites, is missing. It cannot be derived from presence-only data. Presence-absence data, as opposed to presence-only, do contain information on prevalence, but *Elith et al.* (2011) argue that this should be taken with caution since detection probability of a species is mostly not 1 and can even vary across the distribution area, leading to biased estimates of prevalence.

A species response to the covariates shall be modeled, which practically can be rather complex and is fit by nonlinear functions. Transformations of the covariates are used, instead of the covariates directly. In machine learning these transformations and basis functions are called features. Maxent combines these features to complex models. In the end, a fitted function is based on many of such features and typically has more features than covariates itself.

Six feature classes are available in Maxent: linear, quadratic, hinge, product, threshold, and categorical. The user can decide, which of them are allowed to find a good fitting function. In automatic mode, Maxent decides which of them are allowed depending on the number of observations.

To characterize $f(z)$, Maxent relies on random background data. In practice, this is a subsample over L and is independent of the locations with observations. In the model, the subsamples are used for comparison with the occupied sites, $f_1(z)$. First, Maxent's so-called "raw output", the ratio $f_1(z)/f(z)$, is calculated based on the covariate data from the occurrence records and the background sample. This gives an overview about the important covariates and indicates the relative suitability of one place over the other. Constraints are used to assure that information from presence records is reflected in the chosen solution. This means that the distribution, where the mean of a covariate like salinity, temperature, nutrient concentrations, etc. for $f_1(z)$ is close to the mean of the sites with observations. As Maxent uses features (i.e., transformations of covariates) instead of the covariates, these constraints are applied to the means of the features instead of the means of the covariates. The vector of features is denoted as $h(z)$, and the coefficients vector is β. First, all features are rescaled to a range of 0-1 , before an error bound λ_j is calculated for each feature. From the many different possible distributions for $f_1(z)$, Maxent chooses the one closest to $f(z)$. In Maxent, this distance is called the relative entropy (Kullback-Leibler divergence), and $f(z)$ can be considered as a null model of $f_1(z)$. Finding the closest model is critical, as, e.g., in a model without any occurrence records, there would be no reason to prefer any environmental conditions over others, which would lead to a prediction proportionally to the environmental conditions over L.

As a workaround for the unknown prevalence, Maxent uses the so-called "logistic output" $\eta(z) = \log(f_1(z)/f(z))$ as a logit score. The intercept is calibrated so that the implied presence probability at typical sites is equal to the parameter τ. The true value of τ is unknown, and per default set to 0.5 in Maxent.

Minimizing the relative entropy results in a Gibbs' distribution with $\eta(z) = \alpha + \beta * h(z)$ and α a normalizing constant to ensure $f_1(z)$ integrates to 1 (*Elith et al.* (2011) citing *Phillips et al.* (2006)). The ratio $f_1(z)/f(z)$ is a log-linear model, estimated by $e^{\eta(z)}$.

A good fit means, to find a model with a good tradeoff between having all constraints satisfied and at the same time avoiding an overfitting. The model would not be able to generalize anymore if matched too closely.

A sampling bias has a stronger effect on models based on presence-only data than on a model based on presence-absence data. In the latter case, as presences and absences are affected by the bias the same way, the effect cancels out. For presence-only data, the sampling bias $s(z)$, which usually occurs in geographical space, might also be transferred to the environmental space, leading to a biased model of $f_1(z)s(z)$. This can be interpreted as a combination of the species distribution and the sampling effort.

Several studies confirmed a better performance of Maxent compared to GLMs (*Gibson et al.*, 2007; *Roura-Pascual et al.*, 2009). *Renner and Warton* (2013) proved Maxent and point process models (also called log-linear models) to be mathematically equal. Reasons for Maxent's better performance are several further techniques, e.g., feature boosting and use of regularization.

Maxent is controversial, as the software is a "black box" which is even admitted by its authors (*Phillips et al.*, 2006). On the other hand, it is still is considered a good method due to its good performance and its "minimum assumption" approach.

1.3.4 Tuning and testing a model

Several steps are necessary to create a reliable species distribution model. Here these main steps and aspects are reviewed and the limits of these models are pointed out. Sampling bias can be a serious issue in distribution models. Tuning of model parameters and settings, such as a proper selection of background data, predictor set, feature classes, and regularization parameters are necessary to find a good tradeoff between a good fit to the data on the one hand and good generalizing capabilities on the other. Further, appropriate model evaluation is an important issue. A broad overview of the various steps and choices to be made can be found in the practical modeling guide by *Merow et al.* (2013). SDMs for predictions across space and time are discussed in *Elith and Graham* (2009), and for modeling range shifts in *Elith et al.* (2010).

Sample size and sampling bias *Wisz et al.* (2008) compared the effect of sample size for 12 algorithms. Models for 46 species based on sets of 10, 30, and 100 observations were compared using the area under the receiver operating characteristic curve (AUC). Maxent had the best predictive power across all sample sizes. It achieved good results, even for small sample sizes (n<30), but the authors warn that models based on small sample sizes are not consistently good and should be used carefully.

Sampling bias seems to be a largely underestimated problem that frequently occurs in SDM studies. *Yackulic et al.* (2013) systematically reviewed 108 SDM papers on studies that used Maxent and found indications for a sampling bias in 87 % of them. Several strategies are available to treat bias in observation data. *Kramer-Schadt et al.* (2013) analyzed how to correct the sampling bias in Maxent models by spatial filtering and by background manipulation. They found that spatial filtering could minimize omission errors (false negatives) as well as commission errors (false positives) if the sample size was high enough. In that case, they recommend adjusting the background dataset by introducing a bias file, which in their opinion is still better than not correcting the bias. *Syfert et al.* (2013) also found that using a sampling bias grid to correct for the sampling bias has a positive effect on model performance, but cannot correct the bias completely. In that study, the choice of feature types was also analyzed, but only negligible effects on the model's predictive power were found.

Fourcade et al. (2014) systematically tested five different strategies to deal with sampling bias in Maxent models. Artificial datasets with four bias types were derived from three original datasets. The strategies are a systematic sampling, a bias file, a restricted background, a cluster, and splitting. All models were evaluated by AUC, the overlap in the geographical and environmental space and the overlap between binary maps. They found a surprisingly low decline in AUC values for the biased datasets. Their study shows that the different kinds of sampling bias are a serious problem. Though correction often may have a positive effect on the model, none of the tested strategies to deal with the bias can be recommended in general. While in some cases a correction method could help to correct the bias, in other cases it led to the worst model.

Merckx et al. (2011) used a different approach to check their SDMs for a spatial bias in observation data: In null-models, an 'imaginary' species is created by randomly selecting spots as occurrence points, as a subset of the real observation records as well as across the entire study area (see also *Raes and ter Steege* (2007)). By comparing these model variants, they could identify a sampling bias in their data.

Background data Maxent is not a presence-absence method but uses so-called presence-only data in combination with background data. Background data are drawn randomly and are a subset of the complete study area. Selection of background data has an effect on the shape of the response curve, depending on how tight the area, the background data is drawn from, is selected. The region that should be covered by the background data depends on the question to be answered by the model. It can be restricted to the region accessible via dispersal or, disregarding dispersal limits, up to a global scale. The latter is common when model projections across time and space are of interest.

Prevalence Prevalence describes the proportion of sampled sites where a species is present and has a strong impact on the predictive power of an SDM (*Santika*, 2011). Prevalence can be set by a factor (per default 0.5) in a Maxent model. It is dependent on the species detectability, as a species, though present, might not be detected well by a survey method, especially in marine phytoplankton (*Cermeno et al.*, 2014). The spatial scale and the time over which observation records are aggregated have to be considered, too. The number of observations in a presence only distribution model does not matter: a grid cell is occupied, or not. Thus poor detectability might be compensated by a coarse grid.

Predictor set and features Model complexity is dependent on the set of predictors and the selection of features (transformations of covariates) that are allowed to be used in the model. Two contrary ideas appear about selection strategies: The first approach is to preselect predictors, e.g., by eliminating the correlation between predictors, and to reduce complexity, e.g., by dimension reduction techniques (PCA, clustering, etc.). This approach is common to the more statistics oriented view on Maxent. Alternatively, the more machine learning view on Maxent suggests to keep in all reasonable predictors and to let the algorithm decide which ones to use.

Predictors should be chosen by their ecological relevance, which in practice is often limited by data availability. For model projections, all predictors also need to be available for future scenarios, etc. *Verbruggen et al.* (2013) identified the selection of a reduced predictor set as the most important factor in their study about modeling the distribution of an introduced species, a highly invasive seaweed in this case. Other techniques, such as occurrence thinning, model complexity and background choice were found to have a much lower impact.

Regularization Besides the selection of a predictor set and allowed feature types, regularization also has a strong influence on the model's performance. The problem of over-fitting is not Maxent specific, same as the technique of regularization to overcome this issue. A model that was fitted too tight to the data ends up to be far too complex to be useful. Response curves of over-fitted models are hard to interpret, and their projections often show very patchy distribution maps. Regularization is a way of smoothing the model's response. Maxent's parameter for that, the beta-multiplier, acts as a penalty to shrink the coefficients (in Maxent called the betas). It helps to avoid over-fitting and to make the model more general. This way a good balance between model fit and complexity can be achieved.

Output type Maxent offers three output types: raw, cumulative and logistic. They vary in their scaling and are monotonically related, so rank based metrics such as AUC are not affected by the choice of the output type. Raw output is the most basic one, as it was not treated by any post-processing, and can be interpreted as the relative occurrence rate. The probabilities, values between 0 and 1, sum up to 1 over all cells used for training, and typically are rather small. The cumulative output is rescaled and can be interpreted as an omission rate. The value of a grid cell is the sum of the probabilities of all grid cells with lower probabilities than that grid cell, multiplied by 100. As a result, the grid cell with the best conditions reaches a value of 100, cells with unsuitable conditions reach a value close to zero. The logistic output results of a transformation that includes a value for prevalence. Assuming this value (per default set to 0.5 in Maxent) was selected correctly, the logistic output can be interpreted as the predicted probability of presence. The true prevalence is usually unknown in practice. In the literature, this output type often is interpreted as the relative habitat suitability.

Model evaluation Several ways of testing a model are possible: First, the fit of the model gives a good hint and how well the model can explain the data that are used to build (train/ construct) the model. Second, a prediction of the model on independent data is used. Data used to build the model are called training data, data for testing are called test data. If no independent test data are available, the available data can be separated into a test and a training dataset. Maxent offers three built-in resampling methods: cross-validation, bootstrapping and sub-sampling. In case of cross-validation, the samples are divided into replicate folds, of which each fold, in turn, is used as test data. In the bootstrapping method, the replicates are chosen by sampling with replacement. This is useful in case of a small number of observations but loses the independence of training and test data. In the sub-sampling method, the replicate sample sets are chosen by randomly selecting a certain percentage of the observations as test data without replacement. Third, jackknife tests, a special case of bootstrapping, can be used to estimate the importance of variables. For each variable, two extra models are trained, one with the single variable on its own and a second with that variable omitted. The models for the isolated variables indicate how much information a variable contains by itself. The models omitting a single variable indicate if a variable contains information that is not present in any of the other variables.

Metrics for model fit are needed to compare different models. Maxent calculates the receiver operating characteristic (ROC) curve, a plot of the model sensitivity (1-omission rate) versus the fractional predicted area (1-specificity). The area under this curve (AUC) is a measurement of model quality that is commonly used for Maxent in the literature. This value can reach a maximal value of one. A random prediction would lead to an AUC value of 0.5, which is the worst value a model could reach, as it would be no better than random. The AUC-ROC value is dependent on the species and background records. Despite several drawbacks, it is commonly used to compare model versions.

Several tools are available to support model evaluation. The R package ENMeval helps by partitioning the data for a k-fold cross analysis (*Muscarella et al.*, 2014). It offers six different partitioning algorithms and provides further evaluation metrics, such as Akaike information criterion corrected for small sample sizes (AICc). *Radosavljevic et al.* (2014) reported good results using k-fold cross-validation with masked geographically structured partitioning.

1.4 Research questions

1. Species distribution models are widely used in ecological research for several decades now with a focus on terrestrial organisms. For the marine realm, just a few studies exist. Species distribution models (SDM) using presence-only data shall be evaluated in this study. Are these models suitable to study the distribution area of marine pelagic diatoms in the Southern Ocean?

2. At first glance, public biodiversity networks give a good overview of the spatial distribution of taxon observation records. Especially for species with a small number of observation records, distribution models might give a better understanding of (macro) distribution patterns. Further, these models might be useful to reveal problematic data, e.g., in the case of cryptic species. The goal of the models is a quantitative description of distribution patterns, which are still poor for many species. What can we learn about the modeled biogeography of pelagic Antarctic diatoms? Are species distinguishable by their distribution patterns? Can oceanographic properties, used as predictors here, explain these distribution patterns? Are these modeled distribution patterns in line with Longhurst's ecological provinces and previous work (e.g., the studies by Grete Hasle)?

3. Climate change is made responsible for massive changes that became visible in Arctic regions. So far, changes of this dimension cannot be observed in the Antarctic. Projections of the distribution models shall help to estimate if, and

to what extent, global change affects the distribution range of pelagic diatoms of the Southern Ocean. What do these models forecast about range shifts regarding climate change? How are these species affected?

1.5 Trajectory

A series of species distribution models was used to answer these questions. The detailed modeling workflow is presented for *F. kerguelensis* first. This includes several aspects, such as the comparison of different modeling algorithms, an assessment of the effect of a massively improved observation dataset (among others, many new own observations from the Hustedt diatom collection), a comparison of the effects of further predictors, the influence of each of the seven predictors in single predictor models, and an investigation of the influence of problematic predictors on future scenarios. Next, a full model for *F. kerguelensis*, based on all available data, was created, which was also used for the projections on future scenarios. Based on the experience with the *F. kerguelensis* models, current and future predictions for further 20 species are presented and analyzed. Further, a perturbation experiment regarding the upper temperature limit of *F. kerguelensis* was conducted with clonal cultures to estimate how realistic the models are. Model results were compared to previous knowledge from literature and findings from field expeditions.

Chapter 2 describes all models, including the used environmental and observation data, the perturbation experiment, and the metadata scheme used to document the modeling effort. Chapter 3 lists all results of the distribution models and the perturbation experiment are presented. This includes projections of the various *F. kerguelensis* models, future projections, a series of projections of the remaining 20 species, and a comparison of modeled distributions with their raw data. Chapter 4 first discusses the methodical aspects of the distribution models, mainly the data situation and the model evaluation. Next, the findings about the distributions of all investigated species are summarized. The modeled distribution patterns of *F. kerguelensis* are discussed in the light of findings from recent cruises, and the results of the experiment on temperature tolerance. Finally, the pattern analysis and the future projections are discussed. This thesis closes with a synthesis by answering the research questions and a summary and outlook.

2 Material and Methods

This chapter first introduces the observation- and environmental data used for the models (chapter 2.1). Chapter 2.2 gives an overview of the models created for this thesis. Chapter 2.3 describes a metadata model, developed to describe and store steps in the process of model generation in a standardized manner. Chapter 2.4 describes the perturbation experiment about temperature tolerance of *F. kerguelensis*.

Preparatory work about distribution models for *F. kerguelensis* was partly published in *Pinkernell and Beszteri* (2014), where especially algorithm comparison and sampling bias in public data sources were discussed. These two aspects of the paper were complemented here by updated Maxent models, based on further environmental predictors and an enhanced observation data set. Based on the experience with the *F. kerguelensis* models, similar distribution models were generated for various other taxa.

2.1 Data

2.1.1 Observation data

For the models described in this thesis, so-called presence-only data were used. Observation data were obtained from three public databases: the Global Biodiversity Information Facility[2] (GBIF), the Ocean Biogeographic Information System[3] (OBIS) and the Global Diatom Database (GDD) (*Leblanc et al.*, 2012), and complemented by observation data from literature and samples from the Hustedt Diatom Collection (herbarium code BRM) at the Alfred Wegener Institute. The species and the number of observations in the various sources are listed in table 2.1.

The Hustedt Collection was searched for samples from the Southern Ocean and adjacent ocean basins up to a northern limit of 20°S. In total, 256 slides were screened for taxa of interest by light microscopy at 200x magnification. Photos were taken with a Zeiss Plan Apochromat objective with 63x magnification, NA=1.4, and a digital camera to document observation records. In the photos, 10 µm were represented by 98 pixels.

For *Fragilariopsis kerguelensis*, observations from additional three transects at 90°W, 120°W, and 150°W have been used, based on a station list published in (*Hasle*, 1969) and a related map of occurrences of *F. kerguelensis* at these stations (*Hasle*, 1968). Further, observation data from three transects across the Weddell Sea, the Drake Passage, and the Argentine Sea were used for genus *Fragilariopsis* (*Cefarelli et al.*, 2010).

[2] http://www.gbif.org

[3] http://www.iobis.org

A local database was used to aggregate and manage observation data for further processing. GBIF was harvested last on 27th January 2016 for various genera (see table 2.2), and OBIS last on 20th January 2016 for the same taxa. GDD entries were included in the local database completely. The local database contains references to the original entries in the data repositories, literature, and Hustedt Collection, respectively, to keep information on data provenance and citeability. For the observation data harvested from GBIF, DOIs exist for each download (see table 2.2). OBIS did not provide DOIs at time of download (and currently still does not). The GDD dataset[4] was taken as a whole from Pangaea. A complete list of the used light microscopy slides from the Hustedt Collection can be found in table 3.1, together with a list of identified taxa. Voucher images were archived at the collection database and will be available online[5].

Many samples appeared in the local database as replicates, e.g., when they were found in more than one of the data repositories. Just a reduced extract was used for the distribution model, containing only the species name, latitude, longitude, as well as the sampling month. As the environmental data (see chapter 2.1.2) had a spatial resolution of 1x1°, the latitude and longitude of the observation data were rounded to a full degree, too. Presence-only data covers just the information that a species had been detected at a site. Thus, only one observation per grid cell and month was used and repeated entries of a sample, e.g., from different data repositories, so close to each other that they were located in the same grid cell, were eliminated. In contrast, entries from the same position (i.e., the same grid cell), but from different sampling month were treated as two distinct observations. Samples were integrated over several decades, so entries from the same grid cell and sampling month, but from different years were also treated as one sample. As depth was not included as a model predictor, only observations from the surface were used.

2.1.2 Environmental data

The models were based on up to seven predictor variables, selected based on ecological relevance and availability for current conditions and future scenarios: sea surface temperature, salinity, mixed layer depth, sea ice-, silicate-, nitrate- and iron-concentrations (see also table 2.3). The monthly averaged predictor datasets were each harmonized for a global extent, with monthly resolution, WGS1984 coordinate reference system and a grid cell size of 1°x1°. Units and data sources are listed in table 2.3.

Additionally, a second version of the environmental predictor set was prepared, consisting of the minimum, the mean, and the maximum values of each of the seven predictor values for each grid-cell over the 12 monthly layers. This second dataset of 21 layers is called the *yearly dataset* in this thesis.

For the projections on future scenarios, modeled monthly environmental data for the year 2100 from five different models of the CMIP5 comparison project (*Taylor et al.*, 2012) was used. Those atmosphere-ocean global climate models (AOGCMs) that also include biogeochemical components for carbon fluxes between the atmosphere, the

[4] DOI: doi:10.1594/PANGAEA.777384

[5] http://hustedt.awi.de and https://doi.org/10.1594/PANGAEA.878263

Table 2.1: Number of entries in observation data for selected taxa

Species	GBIF	GDD	OBIS	Literature	Hustedt	Total	Rounded
Asteromphalus							
A. heptactis	976	51	2704	0	0	3731	1094
A. hookeri	49	113	402	0	43	604	301
A. hyalinus	15	344	121	0	20	499	83
A. parvulus	2	22	96	0	0	120	76
A. roperianus	0	0	31	0	16	47	38
Azpeitia							
A. tabularis	40	281	52	0	0	373	36
Corethron							
C. pennatum	315	322	13109	0	17	13759	3950
Dactyliosolen							
D. antarcticus	3924	341	3656	0	87	7961	1590
Eucampia							
E. antarctica	223	99	391	0	51	750	255
Fragilariopsis							
F. curta	95	85	836	28	185	1106	235
F. cylindrus	429	48	1032	21	25	1542	304
F. kerguelensis	299	652	1777	58	651	2954	712
F. linearis	15	0	66	0	0	81	14
F. nana	0	0	0	0	266	55	49
F. obliquecostata	0	191	327	16	18	547	108
F. pseudonana	69	35	133	24	39	275	81
F. rhombica	27	171	228	24	77	489	160
F. ritscheri	5	0	7	15	18	39	33
F. separanda	0	0	4	6	16	23	19
F. sublinearis	45	0	327	15	12	396	102
F. vanheurkii	0	0	0	5	0	5	5

Table 2.2: GBIF DOIs, harvested on 27th January 2016

Taxon	DOI
Azpeitia Peragallo	http://doi.org/10.15468/dl.x46vch
Corethron Castracane	http://doi.org/10.15468/dl.ejt1vo
Dactyliosolen Castracane	http://doi.org/10.15468/dl.xtjpyh
Eucampia Ehrenberg	http://doi.org/10.15468/dl.lhns5r
Fragilariopsis Hustedt	http://doi.org/10.15468/dl.9ef3gs
Asteromphalus Ehrenberg	http://doi.org/10.15468/dl.03rnyo

Table 2.3: Overview of predictor variables for current environmental conditions

parameter	unit	data source
Sea surface temperature	°C	World Ocean Atlas 2009 (*Locarnini et al.*, 2010)
Salinity	PSU	World Ocean Atlas 2009 (*Antonov et al.*, 2010)
Mixed layer depth	m	Naval Research Laboratory (NRL)[8] (*Kara*, 2003)
Sea ice concentration	percent	Met Office, Hadley Centre[9] (*Rayner*, 2003)
Silicate	µmol * l^{-1}	World Ocean Atlas 2009 (*Garcia et al.*, 2009)
Iron	µmol * l^{-1}	modeled data (IPSL-CM5A-LR) (*Dufresne et al.*, 2013)
Nitrate	µmol * l^{-1}	World Ocean Atlas 2009 (*Garcia et al.*, 2009)

ocean, and the terrestrial biosphere, are also called Earth system models (ESMs). For simplicity, all models are referred to as GCMs (general circulation models) in the following. These are CESM1-BGC (*Long et al.*, 2013), IPSL-CM5A-LR (*Dufresne et al.*, 2013), MPI-ESM-ME (*Giorgetta et al.*, 2013), HadGEM2-ES (*Jones et al.*, 2011; *Martin et al.*, 2011), and Nor-ESM1-LR (*Tjiputra et al.*, 2013). In climate models so-called representative concentration pathways (RCPs) are used, of which four RCP-scenarios are defined: RCP2.6, 4.5, 6 and 8.5. The number stands for the radiative forcing in W/m² for the year 2100. Two of these future scenarios were used for future projections: RCP4.5 and RCP8.5, with radiative forcings of 4.5 W/m² (~650 ppm CO_2 equivalent) and 8.5 W/m² (~1370 ppm CO_2 equivalent), respectively (*van Vuuren et al.*, 2011).

The datasets were downloaded from CERA data repository at the DKRZ[6] last in October 2015. All datasets were regridded with the software UV-CDAT[7] (version 2.2.0) to a grid size of 1°x1°, covering the whole world. Further, the units were harmonized with the datasets of current environmental conditions used for model training (and same as listed in table 2.3).

[6] http://cera-www.dkrz.de/CERA/index.html

[7] http://uvcdat.llnl.gov/index.html

[8] http://www7320.nrlssc.navy.mil/nmld/nmld.html.030214

[9] Hadley Centre for Climate Prediction and Research (2006): Met Office HadISST 1.1 - Global sea-Ice coverage and Sea Surface Temperature (1870-2015). NCAS British Atmospheric Data Centre, April 2015. http://catalogue.ceda.ac.uk/uuid/facafa2ae494597166217a9121a62d3c

2.2 Models

The model aims to find correlations between the presence of a species at a certain site and month with the predominant environmental conditions at that site and month. This correlation, the model, in the next step is being projected on a set of environmental conditions. This can be the same set used to train the model or different ones, e.g., a future scenario. This way the modeled distribution of that species is mapped.

The process of modeling a distribution range for a species is exemplarily described for *Fragilariopsis kerguelensis*. Selection of a modeling algorithm, studies about the data quality (e.g., effects of a potential sampling bias) and influence of various (environmental) predictor variables, and tests of different evaluation metrics are covered. For *F. kerguelensis* more than 100 individual models with (slightly) different settings were calculated and compared. In the following, a few of them will be presented in detail to point out relevant aspects. For the other species (see table 2.1), projections of final models are presented. These models are based on the experiences gained in the *F. kerguelensis* models.

2.2.1 Algorithm comparison

As described in *Pinkernell and Beszteri* (2014), OpenModeller (v.1.3.0), an open framework with a broad variety of implemented algorithms (*Morin and Thuiller*, 2009; *de Souza Muñoz et al.*, 2009), was used to compare different methods for distribution modeling. This resulted in 16 models, using the following algorithms: Artificial Neural Network, Bioclim, Climate Space Model, Ecological-Niche Factor Analysis, Environmental Distance (using four different distance metrics), Envelope Score, GARP (using two different implementations and subsampling strategies), Niche Mosaic, Random Forests and Support Vector Machines. For each grid cell, the number of models indicating the presence of the species with a threshold of 0.2 was counted and mapped in figure 3.1.

2.2.2 Detailed models for *Fragilariopsis kerguelensis*

Previous model versions were published in *Pinkernell and Beszteri* (2014), with a focus on the effects of observation record coverage and sampling bias. Three model versions based on different observation datasets were compared, with (a) observation records from public repositories that showed a bias due to missing observations in the Pacific sector of the Southern Ocean, (b) a dataset with three added transects in the Pacific sector, and (c) a dataset from the Hustedt collection with additional observation records from the boundary regions predicted by previous models. These models were already able to predict the distribution area of the species *F. kerguelensis* reasonably and were also used for prediction on future scenarios for the end of this century.

Insufficient mapping of the species southern distribution boundary was a weak point of these models. In the following, several model variants are compared based on a further improved observation dataset and three additional environmental predictors.

This includes further comparative model projections for the influence of the iron predictor on future projections and measurements of the predicted distribution areas for current and future scenarios.

Improved observation dataset

To improve the models, the observation dataset was extended by samples from the Hustedt collection by 256 slides from regions all over the Southern Ocean and adjacent ocean basins. Further, observation records from, in the meantime improved, public repositories were added, including records from OBIS and further entries from GBIF.

As a first step, the impact of the improved observation dataset was analyzed. The ‘best’ model from the paper, in this thesis called model 1, was compared with model 2a that was based on the new dataset. Both models were based on the same environmental data and model settings; so except for the observation records they were identical. Four environmental predictors were used: nitrate and silicate concentration, sea surface temperature, and salinity. In contrast to the other Maxent models, only hinge features were used and a beta multiplier of 1. The model projections of both models are plotted in figure 3.2 for winter and summer conditions.

Improved environmental predictor set

Next, three additional predictors were included. Predictors were added separately to the previous four predictors first. Mixed layer depth in model 2b, iron concentrations in model 2c, and sea ice concentration in model 2d. Figure 3.5 plots the predictions of model 2a and 2d in comparison. Both models used the improved observation dataset and the same setting as model 2a: a beta multiplier of 1 and only hinge features.

Individual environmental predictors

The impact of the individual environmental predictors was analyzed by single predictor models. A model based on the isolated predictor was built for each of the seven predictors to characterize its autecological relevance. Compared with the full model (described below), the individual models also give insights into the predictor’s contribution to the model’s response. Summer and winter projections, as well as the resulting response curves, are plotted in figure 3.4. For these models, all feature sets were allowed, and a beta multiplier of 2 was used.

Full model

After investigating isolated changes due to improved observation and additional environmental data, model 3 was built including all available data. In contrast to the model in *Pinkernell and Beszteri* (2014), now all feature classes are allowed with an auto-selection by the Maxent algorithm, and the beta multiplier was set to 2. The global projections for February and August are plotted in figure 3.6, and a set of monthly projections are plotted in figure 3.7 with a focus on the Southern Ocean. Figure 3.8 shows the results of the Jackknife test for the predictor influence, and figure 3.9 shows detailed response curves for all predictors. This model is also used for the future projections.

Maxent can be seen in a statistical and a machine learning view. In the first case, isolated model runs are common, with further predictors added step by step. In contrast, the 'machine learning way' is to feed the model with all available information and to let the model decide.

Yearly averaged models

So far, all Maxent models were calculated based on datasets with a monthly resolution. In contrast, most SDM studies in the literature (as described in the introduction) use yearly predictors. To compare both approaches, a derived dataset with minimum, mean, and maximum values of the environmental predictors was prepared. Model 6, which was built and projected on these derived yearly environmental predictors, was compared to a projection of a monthly model (model 3) on a yearly averaged dataset. The projections are plotted in figure 3.10.

Future projections

Future model projections for the year 2100 were mapped based on the RCP4.5 and RCP8.5 scenarios. Five GCMs were chosen for the modeled environmental data, but the mixed layer depth data was not available for the HadGEM2-ES model. Thus, some models were projected on four GCMs only, depending on the use of the mixed layer depth predictor in the Maxent model.

A strong variation in the model outputs among the five GCMs could be observed in the models with iron. An extreme example is plotted for model 4a in figure 3.12. For comparison, the same plot was generated for model 4c - a similar model, but without iron - in figure 3.13. Both models were built without the mixed layer depth predictor so that all five GCMs could be used. Both of them used all feature classes and a beta multiplier of 2.

Future projections of three models are plotted in comparison in figure 3.14 for summer and winter conditions. For the plot in figure 3.14 and the area measurements in table 3.2, the median of these projections was calculated, and a threshold of 0.2 was applied. In the figures, the future distribution is mapped as hatched areas. For comparison, the current distribution was plotted using the same color code as in the other Maxent plots and the corresponding 0.2 iso-line in red. The full model (model 3) and the model without iron (model 4b) were projected on 4 GCMs. Model 4c, a model without the iron and mixed layer depth predictors, was projected on all 5 GCMs.

Overview of Maxent models for *F. kerguelensis*

During the modeling process, a multitude of (slightly) different distribution models are generated and compared. Only a few models are chosen and presented here for discussion. The models described and used within this thesis are listed in the following:

- *F. kerguelensis* model 1
  A model with a reduced observation dataset and four environmental predictors. This model is equal to the 'best' model version in *Pinkernell and Beszteri* (2014).
- *F. kerguelensis* model 2
  All models are based on the improved, full observation dataset, including further data from the Hustedt Collection and the OBIS network.
  - 2a: Four environmental predictors are used (as in model 1): silicate, nitrate, sea surface temperature and salinity.
  - 2b: Similar to model 2a, but including a mixed layer depth predictor.
  - 2c: Similar to model 2a, but including an iron predictor.
  - 2d: Similar to model 2a, but including a sea ice concentration predictor.
- *F. kerguelensis* model 3
  The model includes the full observation dataset and all environmental predictors. It was used for the projections on monthly and on yearly datasets, as well as for the projection on future scenarios. In the following, it is also called the 'full model'.
- *F. kerguelensis* model 4
  Both models include the full observation dataset, but a different set of predictors. They are used to compare the influence of predictors in projections on future scenarios. MLD was excluded for comparative models, as it was not available in all GCMs, and iron, because it showed a strong variation among the GCMs.
  - 4a: This model uses all predictors except mixed layer depth.
  - 4b: This model uses all predictors except iron.
  - 4c: This model uses all predictors except iron and mixed layer depth. In contrast to model 2d, it uses all feature layers and a beta multiplier of 2.
- *F. kerguelensis* model 5
  The models use the full observation dataset and are each based on single predictor layers.
  - Model 5a: silicate
  - Model 5b: nitrate
  - Model 5c: iron
  - Model 5d: salinity
  - Model 5e: mixed layer depth (MLD)

  - Model 5f: sea ice concentration (SIC)
  - Model 5g: sea surface temperature (SST)

- *F. kerguelensis* model 6
  This model uses the full observation dataset and yearly predictors (with minimum, mean, and maximum values).

2.2.3 Models for other species

Models for the other species listed in table 2.1 were calculated. Settings for these models have been chosen based on the experience gained with the *F. kerguelensis* models. Though problematic for future projections (see chapters Results and Discussion), iron was included in these models. All models were calculated with the same settings as used for *F. kerguelensis* model 3: i.e., all feature types were allowed, including automatic feature type selection. The beta multiplier was set to 2, and a set of 12.000 background points was used (the same set for all of the models, including the *F. kerguelensis* one). For validation, the second run with 20-fold cross-validation was used. The spatial projections shown in the results section are based on the models with the full set of observation records (from a run without cross-validation to have all observation records included).

The models were also projected on future environmental conditions for the end of the century. In contrast to the future projections of *F.kerguelensis* model 3, these models are only projected on the February data of the RCP4.5 and RCP8.5 scenarios. Sea ice concentration is negligible in February, and the projections are close to the maximum distribution range. After applying a threshold of 0.2, the modeled species distribution area was measured for current and future (year 2100) February projections (see table 3.3). Only the areas covering the Southern Ocean and the adjacent ocean basins were considered for these measurements, not the northern regions.

2.2.4 Model comparisons

The resulting spatial patterns of all Maxent models were compared systematically based on the raw data used for model training and the relative contribution of model predictors.

Different global distribution patterns can be identified in the spatial model predictions (figs. 3.16 to 3.20), which were clustered in a dendrogram. In a first step, a new map is created for each species with the mean values of the February and August distribution. This map shows the integrated maximum distribution area over the year. In a second step, a threshold of 0.2 is applied, to remove noise. Third, a distance matrix is calculated using Manhattan distance method. Finally, this distance matrix is used for a hierarchical clustering (fig. 3.22) using the complete clustering method.

Additional to the spatial predictions the relative predictor contributions gives further insights on the model's interna. Similar to the analysis of global distribution patterns, the relative importance of variable contribution for each of the models is analyzed by hierarchical clustering (fig. 3.22) and a log ratio analysis (LRA). In an LRA, the dataset is being log-transformed first. A weighted double centering is applied next, followed by a principal components analysis. The result is shown in a biplot in fig. 3.23.

Finally, the raw data that was used for model training is analyzed in the same way as the model's resulting distribution patterns and the model's predictor importance. The dataset contains the median values of environmental conditions of the occurrence sites of the respective month. First, a distance matrix is calculated using Manhattan distance method. The distance method, again, is used for the hierarchical clustering using the complete clustering method (fig. 3.22).

2.3 Metadata

Many different models were produced during the modeling process. As there are different questions to answer, this multitude of sometimes just slightly different models is needed. It became apparent during this process that keeping track of these large numbers of models and their slight differences in how they were constructed is necessary. To address this, a metadata framework was developed, which also supports a transparent archival of model results.

This metadata model describes three categories of entities. The first category belongs to the model's input data: the observation data as well as the environmental predictors. As SDMs rely on a very basic observation dataset, data and metadata are sometimes the same for observation data, i.e., the identified species, the coordinates, and sampling time. Further, information about the data source (e.g., a GBIF or a Pangaea reference), the collector of the data, the responsible person for species identification, the data collection itself (e.g., for plankton the mesh size and type of the net), and the tools used for species identification (e.g., light or electron microscopes) are needed. Environmental predictors usually consist of several layers of raster files, for which metadata about the spatial extent and resolution, the geographical coordinate system, etc. are necessary. Further, references to the data source can be stored. The second category deals with the model itself and contains information regarding the modeling algorithm and model settings. The third category contains information about the model's output, including a list of all spatial projections. These maps have similar metadata like the environmental predictors describing the raster. Further, the environmental dataset used for projection can be referenced.

A simple software for the metadata management was developed in R to print the information in a tree-like structure. The software provides a simple syntax to build and visualize the tree and can export and import the data in XML format.

It's recommended to generate the metadata file together with the models. This file should be kept up to date when further projections are added.

2.4 Perturbation experiment

Four diatom cultures from different positions in the South Atlantic Ocean were used for this experiment. They were collected by a hand net from the surface water (up to 10 meters depth) during Polarstern cruise PS81 (ANT XXIX/5 in April and May 2013) at the stations listed in table 2.4. The hand net had a mesh size of 20 µm.

Table 2.4: Station list of geographic origin of the samples used for the perturbation experiment

Station	Date	Latitude	Longitude
301	21.04.2013	51° 21.23' S	54° 37.44' W
364	30.04.2013	51° 4.48' S	49° 24.29' W
374	02.05.2013	50° 57.8' S	46° 59.79' W
404	05.05.2013	50° 37.36' S	40° 22.95' W

The cultures were isolated on-board R/V Polarstern and stored at 4° C in culture flasks with F/2 medium (*Guillard and Ryther*, 1962), which was also used during this experiment. The medium was prepared with Antarctic water and added silicate, nutrients, and vitamins.

During the experiment, a day-night cycle with a daytime of 16 hours and an illumination intensity of 67.5 µM/s m2 was used. Temperatures were raised every 14 to 20 days by one degree Celsius, depending on the inoculation date of the culture flasks. For inoculation, a few ml of cell suspension was transferred to a new flask with fresh medium. Cultures were controlled twice a week by light microscopy using an inverted microscope (Zeiss Axiovert) for survival. Both, the old and new inoculated culture flasks were examined in this control.

3 Results

The model projections and metadata, as well as the voucher images, are available in Pangaea[10] under https://doi.org/10.1594/PANGAEA.878263. The voucher images will also be stored in the Hustedt database. The images can be identified by the slide number (see table 3.1 in the following section).

3.1 Diatom slides

The following table (table 3.1) lists all slides together with the identified taxa. Further information about the slides, samples, and voucher images can be found in the Hustedt Collection's database under http://hustedt.awi.de.

On the following slides, no valves of interest where found.
January: Hasle17-87, Hasle22-88, Hasle22-96, Hasle22-97, Hasle27-100, Sim55-61, Sim55-63, Sim55-72, Sim55-73, Sim55-74, Sim55-75, Sim55-79, Sim55-80, ZU5-73.
February: 275-01, 275-83, Hasle22-41, Hasle22-45, Hasle22-77, Hasle22-82, Hasle22-99, Hasle29-27.
March: Hasle17-82, Hasle20-01, Hasle20-03, Hasle20-76, Hasle22-01, Hasle26-86.
April: Hasle20-05, Hasle20-06, Hasle20-07, Hasle20-08, Hasle20-11, Hasle20-12, Hasle20-13.
May: Hasle20-14, Hasle20-15, Hasle20-72, Hasle31-04.
June: 275-33, 275-34, 277-28, Hasle18-06.
August: 275-90, 284-47, 284-52.
September: Hasle17-75, Hasle17-76, Hasle17-79, Hasle18-07, Hasle20-26, ZU7-53.
November: Hasle17-53, Sim58-83, Sim58-85, ZU6-16.
December: 275-52, 283-38, 283-40, Hasle22-05, Sim49-82, Sim55-17, Sim55-32, Sim55-40, Sim55-42, Sim55-43, Sim55-45, Sim55-46, Sim55-48, Sim55-53.
Month unknown: 283-05, 283-10, 400-2b, 90-36, Hasle17-63, Hasle17-98, Hasle18-01, Hasle21-51, LA-52.

[10] http://www.pangaea.de

Table 3.1: Diatom observations in samples from the Hustedt collection

Slide	Month	*Fragilariopsis curta*	*Fragilariopsis clindrus*	*Fragilariopsis kerguelensis*	*Fragilariopsis linearis*	*Fragilariopsis nana*	*Fragilariopsis obliquecostata*	*Fragilariopsis pseudonana*	*Fragilariopsis rhombica*	*Fragilariopsis ritscheri*	*Fragilariopsis separanda*	*Fragilariopsis sublinearis*	*Fragilariopsis vanheurkii*	*Asteromphalus heptactis*	*Asteromphalus hookeri*	*Asteromphalus hyalinus*	*Asteromphalus parvulus*	*Asteromphalus roperianus*	*Azpeitia tabularis*	*Corethron pennatum*	*Dactyliosolen antarcticus*	*Eucampia antarctica*
275-08	2			x											x						x	
275-14	10			x											x						x	
275-22	10			x											x						x	
275-30	6			x																		
275-35	6																				x	
275-36	6																				x	
275-37	6																				x	
275-41	11														x							
275-49	12														x							
275-60	8														x							
275-72	7																	x				
276-12b	2	x		x											x					x		
276-17a	2	x																				
276-17b	2																	x		x		
276-19a	2	x		x		x		x	x							x						x
276-19b	2	x		x		x		x														x
276-39a	2			x																	x	
276-39b	2			x																	x	x
276-43a	2			x																		
276-49a				x																		
276-49b				x																		
276-50				x																		
282-87				x																		x
282-88				x											x							x
282-89				x																		x
283-03				x																		x
283-04															x							x
283-09				x																		
283-11				x																		x

283-32	12			x																		x
283-33	12			x																		x
283-34	12			x																		
283-35	12			x																	x	
283-36	12			x																	x	
283-37	12			x											x							
283-39	12			x											x							
283-55	2			x																	x	
283-58	2			x																	x	
283-61	2			x																	x	
283-66	2	x		x																	x	x
283-70	2			x																		x
284-27	3	x				x																x
284-29	3	x		x																		x
284-30	3																				x	
284-31	4			x											x							x
284-49	8														x							
400-02b	12																				x	
400-04a	2	x		x											x							
400-13	2	x		x						x						x						
400-17	1			x																		
400-26a	1	x		x		x		x	x	x										x		
400-32a	1			x		x																
450-67	2			x								x			x							
457-02	5			x											x							
457-09	5			x														x				
457-43	5																				x	
457-75	6																				x	
458-67	12			x																	x	
Hasle17-64															x							
Hasle17-74	10			x																		
Hasle17-77	11			x																		
Hasle17-78	12			x																	x	
Hasle17-81	3			x					x													x
Hasle17-86	1										x											
Hasle17-97				x																		
Hasle18-05	9																				x	
Hasle20-02	3			x																		
Hasle20-04	3			x																		
Hasle20-09	4			x																		
Hasle20-10	4											x										
Hasle20-25	9														x							
Hasle20-27	9														x							

Hasle20-73	5		x																			
Hasle20-74	3			x																		
Hasle20-75	3			x																		
Hasle22-02	12			x																		
Hasle22-03	12			x																		x
Hasle22-04	12			x																		
Hasle22-07	12			x																		x
Hasle22-08	12			x																		
Hasle22-09	12	x		x																		
Hasle22-10	12			x											x	x					x	
Hasle22-100	2	x		x		x		x	x	x											x	
Hasle22-11	12			x		x															x	x
Hasle22-13	1	x	x	x		x		x	x	x										x		
Hasle22-16	1	x		x		x										x				x		
Hasle22-18	1	x		x					x							x						x
Hasle22-19	1	x		x		x			x		x				x	x		x		x	x	x
Hasle22-20	1	x		x		x			x		x					x						
Hasle22-21	1	x	x	x		x									x	x						
Hasle22-23	1			x																		x
Hasle22-25	1			x																	x	
Hasle22-28	1			x																		
Hasle22-29	1			x																		
Hasle22-30	1			x					x		x				x							
Hasle22-31	1	x		x		x			x													x
Hasle22-36	1	x		x		x	x		x	x					x							
Hasle22-37	1	x	x	x		x			x						x			x				x
Hasle22-38	2	x		x		x		x	x	x						x						
Hasle22-39	2	x		x		x									x							
Hasle22-40	2	x		x		x														x		
Hasle22-42	2	x		x		x			x	x											x	x
Hasle22-43	2	x		x		x			x													
Hasle22-44	2	x	x	x		x			x						x							
Hasle22-47	2	x		x		x	x		x	x	x					x						x
Hasle22-48	2	x	x	x		x	x	x	x	x												
Hasle22-51	2	x		x		x										x				x		
Hasle22-52	2	x																				
Hasle22-53	2																			x		
Hasle22-70	1			x																		
Hasle22-71	1			x																		
Hasle22-74	1			x																	x	
Hasle22-76	2																				x	
Hasle22-84		x		x		x			x													
Hasle22-85	12			x			x															

Hasle22-86	1	x		x		x	x		x													
Hasle22-87	1	x		x		x	x		x	x												
Hasle22-89	1	x		x		x			x	x		x										
Hasle22-90	1	x		x		x			x		x											x
Hasle22-91	1	x		x		x			x		x				x	x					x	x
Hasle22-92	1	x		x		x		x	x		x					x				x		x
Hasle22-93	1	x		x		x		x	x		x										x	
Hasle22-94	1	x		x		x			x													
Hasle22-95	1	x	x	x		x		x	x		x	x				x						
Hasle22-98	2	x		x		x		x	x													
Hasle26-03	12			x		x	x															
Hasle26-04	11	x	x																			
Hasle26-100	3	x		x		x									x						x	
Hasle26-20	2	x		x		x		x	x													
Hasle26-25	3			x							x											
Hasle26-26	1	x		x					x													x
Hasle26-35	1					x																
Hasle26-36	1	x		x		x		x							x							
Hasle26-37	1	x		x																		
Hasle26-38	1			x																		
Hasle26-39	2			x		x										x						
Hasle26-66	2			x																		
Hasle26-72	2						x		x			x										
Hasle26-73	3					x																x
Hasle26-76	3	x	x			x						x			x							
Hasle26-78	3	x		x		x						x			x			x				
Hasle26-82	3	x				x			x									x				
Hasle26-93	12			x		x												x				
Hasle26-96	12			x				x														
Hasle27-02	10	x	x			x						x										
Hasle27-04	11			x		x			x						x							
Hasle27-05	3	x	x									x						x				
Hasle27-98	12																				x	
Hasle28-02	2			x		x										x						x
Hasle28-03	2	x	x	x		x			x			x						x				
Hasle28-09	1			x								x										
Hasle28-11	12			x																		
Hasle28-12	3	x		x												x						
Hasle28-13	3			x																		
Hasle28-14	3			x					x			x										
Hasle28-20	1	x		x		x				x		x						x				x
Hasle28-26	3			x																		
Hasle28-27	3	x		x		x						x										

Hasle28-33	1			x																		
Hasle28-42	2	x		x					x													
Hasle30-51	12	x				x	x											x				
Hasle30-52	12	x				x	x															
Hasle30-56	3			x																		
Hasle30-58	3			x											x							x
Hasle30-59	3			x							x											
Hasle30-60	4			x							x										x	
Hasle30-63	4			x																		
Sim55-20	12														x							
Sim55-24	12														x							
Sim55-35	12														x							
Sim55-44	12			x																		
Sim55-52	12																	x				
Sim55-58	12																	x				
Sim55-65	1			x																		
Sim55-66	1			x														x			x	
Sim55-67	1			x																	x	
Sim55-68	1			x																		
Sim55-69	1			x																		
Sim55-70	1			x																		
Sim55-71	1			x																		
Sim55-77	1			x																		
Sim55-78	1			x																		
Sim55-81	2			x																		
Sim55-82	2			x																		
Sim55-83	2			x																		
Sim55-84	2			x																		
Sim55-85	2			x																		
Sim55-86	2			x																		
Sim55-87	2			x																		
Sim55-88	2			x																		
Sim55-89	2			x																x	x	
Sim55-90	2			x					x													
Sim55-91	2			x																		
Sim55-92	2			x																		
Sim55-93	2			x												x						
Sim55-94	2			x																x		
Sim55-95	2			x																		
Sim55-96	2			x												x						
Sim55-97	2			x																		
Sim55-98	2			x					x												x	
Sim56-01	2			x																		

Sim56-07	2			x											x							
Sim56-09	2			x											x							
Sim56-14	2			x		x																
Sim58-87	11																	x				
Sim63-28	2			x																		
Sim63-34	2	x		x																	x	
Sim63-38	3			x																		
Sim63-53	12			x																		x
W7-11a		x																		x		
ZU4-89	11			x											x						x	x

3.2 Models for *Fragilariopsis kerguelensis*

The diatom *Fragilariopsis kerguelensis* is a dominant diatom species in the Southern Ocean and coined to be one of the main silicate sinkers in this region (*Zielinski and Gersonde*, 1997; *Smetacek*, 1999). Because of its importance in Southern Ocean ecology and its contribution to silicate and carbon export it was selected for detailed modeling studies presented in this section. First, a comparison of different algorithms was conducted using OpenModeller. Sampling biases and gaps in the observation coverage found in public databases and repositories were studied. The influence of different predictors was compared in single predictor models. Finally, full models, using all available observation and environmental data were projected on two future scenarios (RCP4.5 and RCP8.5), to assess potential range shifts for this species towards the end of this century.

3.2.1 Algorithm selection

Distribution models for *Fragilariopsis kerguelensis* were calculated with OpenModeller based on nine different algorithms and annually averaged environmental data. The map in figure 3.1 shows the combined results of 16 different models and model variants. Most models predicted *F. kerguelensis* to occur in a "belt" around the Antarctic continent. The Weddell Sea and the Ross Sea were not covered by all models. Except for these two regions, most models agree well on the Antarctic continent to be the southern distribution boundary. The agreement on the northern boundary is much lower. While most models agree that *F. kerguelensis* occurs in the region south of the Subantarctic Front in the belt up to 45-50°S, a few models predict a wider range up to 25-30°S.

These distribution models, though built on an early observation dataset, already map the distribution conforming with previous knowledge about the distribution of *F. kerguelensis.* The sensitive regions, such as the "belt" north of the Subantarctic Front, as well as the Ross and the Weddell Sea, where the models differed from each other, need to be further analyzed.

3.2.2 Improved observation dataset

In *Pinkernell and Beszteri* (2014) it was shown that observation records for *F. kerguelensis* from public databases were strongly biased towards the Atlantic sector of the Southern Ocean. Three models were compared based on different observation datasets. The first dataset, from public repositories, had no observations from the Pacific sector of the Southern Ocean. Three transects from the Pacific were added for the second dataset, and further observations from the Hustedt collection from the northern regions of the predicted areas of previous models for the third dataset. The three added transects in the second model had hardly any effect on the predicted distribution. In contrast, the third model showed a northward shifted distribution boundary for the current projections. None of the three models could explain a reasonable southern dis-

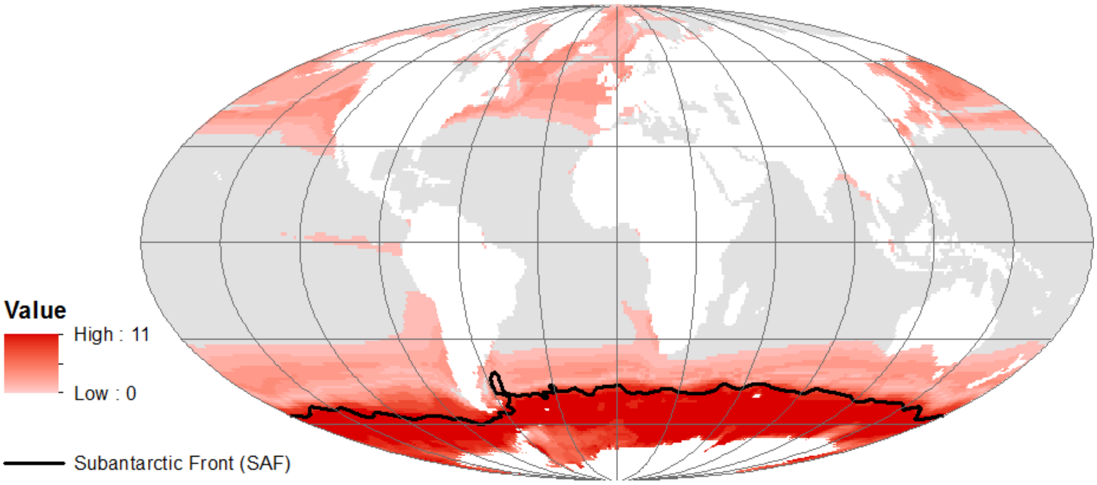

Figure 3.1: Modeled distribution of *Fragilariopsis kerguelensis* as projected on current yearly averaged environmental conditions. This consensus plot of different yearly models (calculated by different algorithms using yearly averages, see chapter 2 for details) shows the number of models where a threshold of 0.2 was reached. The average position of the Subantarctic Front is shown by a black line.

tribution boundary. In the future projections on scenarios for the end of this century, the differences in the predicted northern distribution boundaries were even stronger.

The models presented in this thesis rely on a further improved observation dataset, as further samples from the Hustedt collection were analyzed and more records in the public repositories became available. First, the impact of the improved observation dataset was analyzed. Model 1 was based on the best dataset published in *Pinkernell and Beszteri* (2014) (dataset D in the paper), model 2 based on a further improved dataset with additional samples from the Hustedt Collection and the OBIS network. Both models rely on the same set of environmental predictors and model parameters. Figure 3.2 shows winter (August) and summer (February) projections of two models in comparison:

Both models predicted very similar distribution areas, with almost identical northern distribution boundaries. The main difference in the summer projections was the more pronounced gap in the Ross Sea for model 2, the one with the improved observation data set. None of the models could predict a plausible southern distribution boundary for the winter scenarios. Distribution areas stayed similar for both models, except for the slightly more connected and bigger gaps around Antarctica in the model with improved observation data.

The improved observation dataset neither changed the relative predictor contributions, nor the shape of the response curves. Nitrate, in both models the most important predictor, became slightly more important at the expense of salinity and silicate. Sea surface temperature stayed at second place unaltered.

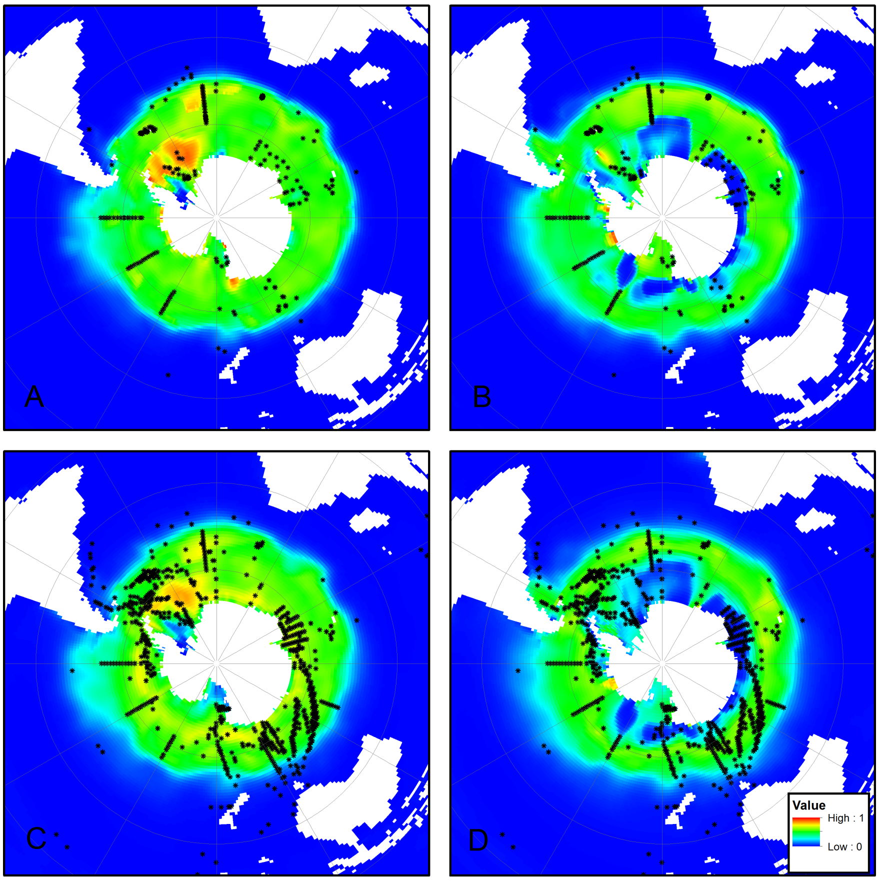

Figure 3.2: Projection of modeled *F. kerguelensis* distribution for summer conditions (February, left images) and winter conditions (August, right images). The models are based on the full dataset (model 2a, bottom row) as well as on the older dataset (model 1, upper row), as used in *Pinkernell and Beszteri* (2014).

Despite the added observations, the sector from 90° to 180° West and the sector from 0° to 60° East are still underrepresented in the observation records. Beside this spatial bias, a temporal bias can be observed. Figure 3.3 shows the monthly distribution of all *F. kerguelensis* observation records, with a strong bias towards an under-representation of winter samples.

The enhanced observation dataset included many positions not covered by the modeled distribution area (see fig. 3.2 C+D). E.g., east of South America, between Australia and New Zealand, in the Pacific, and in the Indian Ocean.

In conclusion, though several biases in the observation data are obvious, just moderate differences in the projections on current environmental data could be observed. The updated dataset used for this study could slightly improve the model. Earlier models showed that a good coverage of the northern and southern distribution areas is

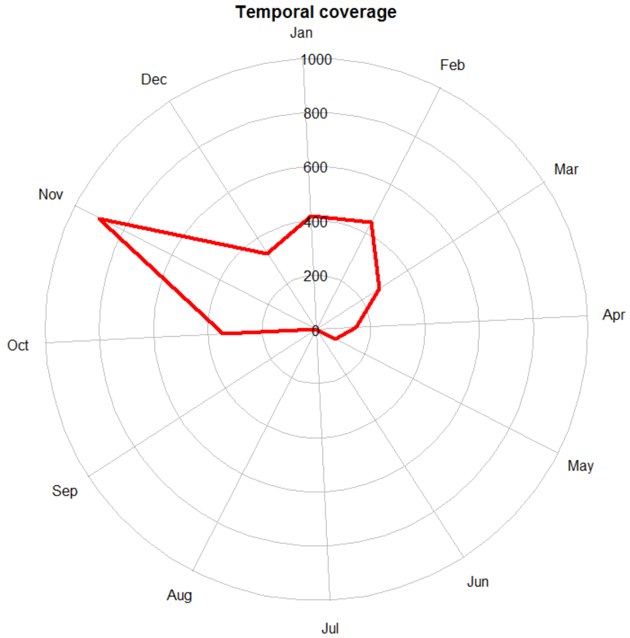

Figure 3.3: Monthly distribution of *F. kerguelensis* observations. Months with only small numbers of observation records are difficult to see in this plot: June has four observation records, and July has one. For August and September, no observation records are counted.

more important than a complete circumpolar coverage, but this bias is more difficult to detect (see also *Pinkernell and Beszteri* (2014)).

3.2.3 Effects of individual predictor variables

Though the predictors used throughout this study affect the physiology of diatoms, this does not mean that all of them are good predictors to model the species distribution. To characterize the autecology of the species and to better understand isolated predictor influence on the distribution, single predictor models (models 5a-g) were built and global scale summer and winter projections compared in figure 3.4.

F. kerguelensis is known to be endemic to the Southern Ocean, but all models predicted a broader distribution range. As the models respond on correlations, instead of ecophysiological requirements, they might give unexpected results. The nitrate model (model 5b) was close to the real distribution, except for the prediction in the North Pacific during (northern) winter. Silicate instead, though essential for diatoms to build their frustules, performed worse than nitrate. It led to a poleward shifted northern distribution boundary in the Southern Ocean and occurrence signals in several regions where *F. kerguelensis* is not known to occur. Sea surface temperature (model 5g) predicted occurrence in the Arctic Ocean throughout the year. Other predictors

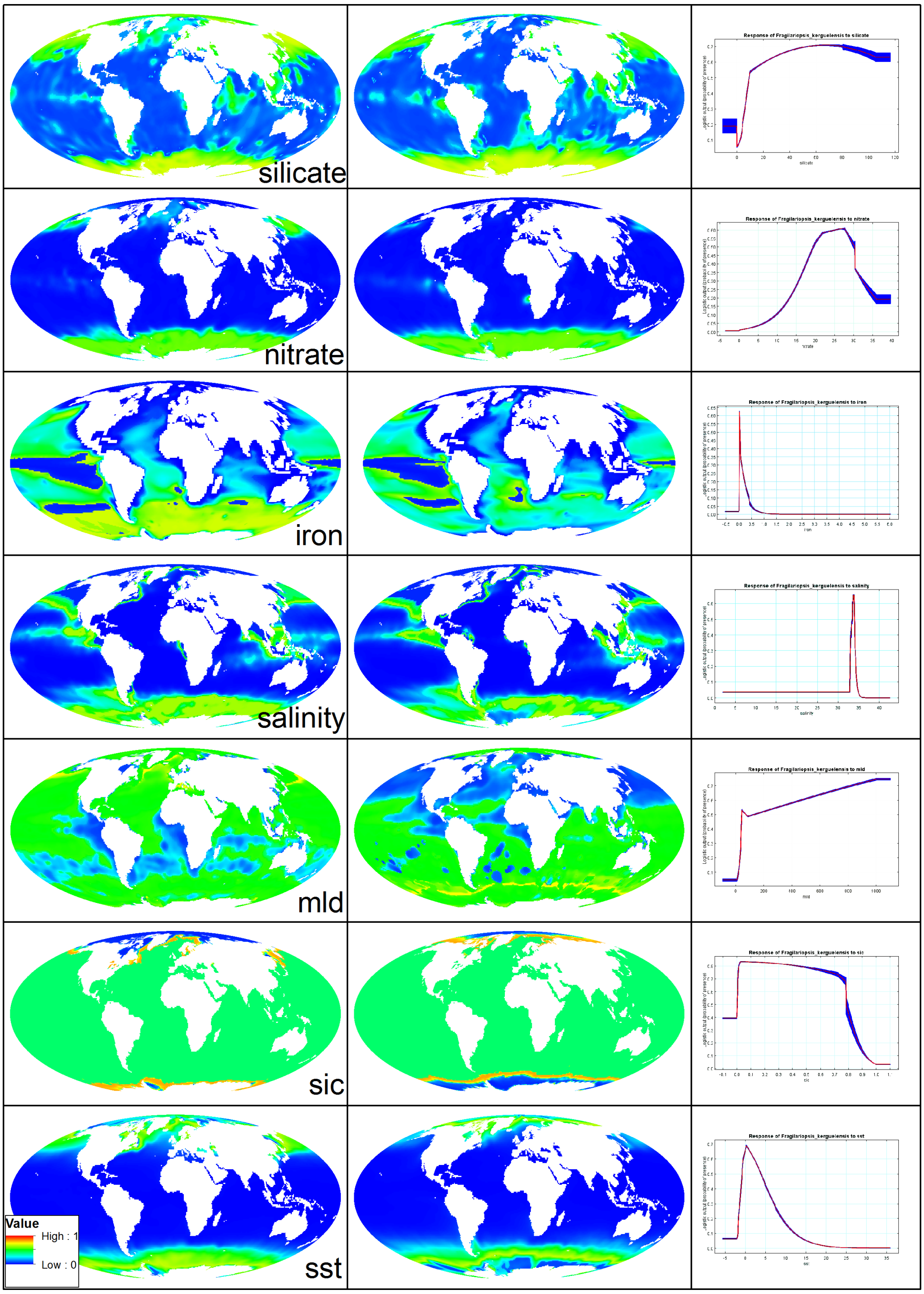
Response of Fragilariopsis_kerguelensis to silicate
silicate
Response of Fragilariopsis_kerguelensis to nitrate
nitrate
Response of Fragilariopsis_kerguelensis to iron
iron
Response of Fragilariopsis_kerguelensis to salinity
salinity
Response of Fragilariopsis_kerguelensis to mld
mld
Response of Fragilariopsis_kerguelensis to sic
sic
Response of Fragilariopsis_kerguelensis to sst
sst
Value
High : 1
Low : 0

Figure 3.4: Projections of model 5, *F. kerguelensis* distributions for February (left) and August (right) based on single predictors. For better visibility, the response curves are plotted again in the right column in figure 3.9.

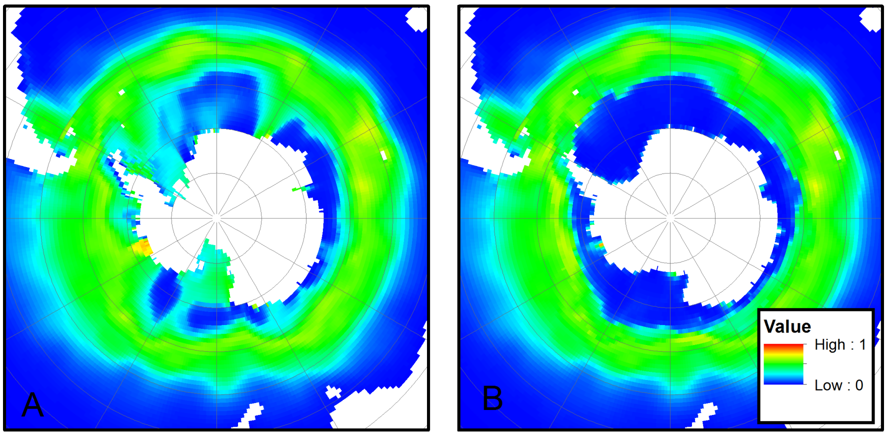

Figure 3.5: Projection of *F. kerguelensis* on August conditions. A: Model 2a with four environmental predictors (sea surface temperature, salinity, silicate, and nitrate). B: Model 2d, which also includes sea ice concentration.

give information for special regions: Sea ice concentration delimited the predicted southern distribution boundary, resulting in a gap in the distribution area where sea ice concentrations are above 79 % (model 5f). A common threshold to consider an area as sea ice covered is a sea ice concentration of 15 %. 26 of the scanned slides fell in this area, and only four of them were from areas with more than 79 %: Hasle26/73 (March 1, 1968), Hasle26/66 (February 19, 1968), Hasle26/72 (February 28, 1968), and Hasle26/76 (March 4, 1968), all from the Weddell Sea. *F. kerguelensis* valves were found on 15 of the 26 slides, and only in one case at more than 79 % ice coverage (Hasle26/66). During Austral summer, iron concentrations in the Southern Ocean are lower than during the winter. As most of the observations were from the summer month, the model responded to low iron concentrations, which leads to contrary model outputs than expected by the knowledge about diatom's physiology. For the summer, when iron concentrations are lower, the model predicts F. kerguelensis to be present in the Southern Ocean. For the winter, though the iron concentrations are higher, the models predict the Southern Ocean to be less suitable for F. kerguelensis. For mixed layer depth, the model responds to a threshold of approximately 50 meters. This covers almost the entire ocean, resulting in a strongly overrated distribution pattern. For the Southern Ocean, this model predicts suitable conditions throughout the year. Many models predict suitable conditions for *F. kerguelensis* in the North Pacific. In the models on mixed layer depth, nitrate, and sea surface temperature, these regions are excluded for the (northern) summer conditions (August projection).

3.2.4 Adding further predictors

Models 1 and 2a both contained four environmental predictors: sea surface temperature, salinity, nitrate, and silicate. Model 1 could already predict a reasonable distribution area (*Pinkernell and Beszteri*, 2014), but did not give a proper explanation of the

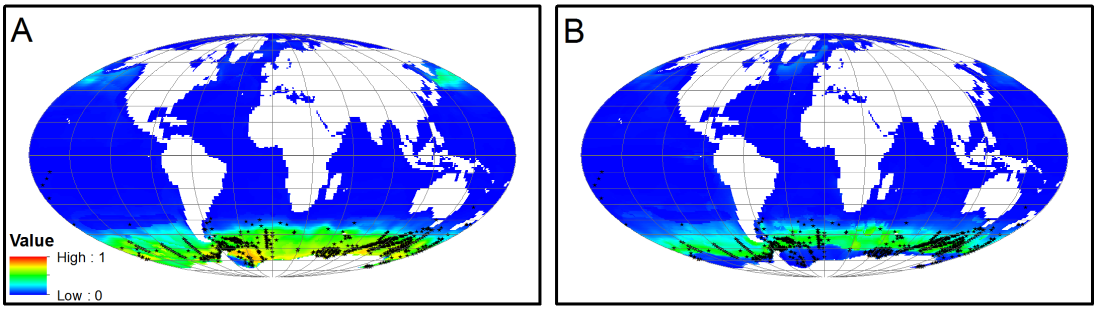

Figure 3.6: Global projection on February (A) and August (B) conditions for *Fragilariopsis kerguelensis*. Positions of the observation records are indicated by the black dots.

southern boundary, especially in regions around 30°E, where the suitable area reaches up to the continent (fig. 3.5A). Adding further observation records in model 2a did not improve the distribution patterns. Next, three predictors were added: mixed layer depth (model 2b), iron (model 2c), and sea ice concentration (model 2d). In model 2b, mixed layer depth reached a variable importance of 5.7 %, and in model 2c iron reached 7.9 %. Both did not change the southern distribution boundary. Adding sea ice concentration as a predictor in model 2d, which reached a relative importance of 5.7 %, significantly changed the predicted southern distribution boundary (fig. 3.5).

This pattern agrees well with reports from literature (*Hasle*, 1976; *Cefarelli et al.*, 2010), but potentially is the result of a bias, as samples from ice-covered regions are rare. In some of those samples, however, *F. kerguelensis* valves are present but occur just sporadically.

3.2.5 Full model

For model 3, all available observation data and all predictors were used; it is also referred to as the "full model" in the following. Global projections of this model for February and August are plotted in figure 3.6, more detailed monthly projections of the Southern Ocean regions in figure 3.7. This model showed a strong seasonality in the predicted habitat suitability. Nitrate had the strongest relative contribution to this model (68.8 %), followed by sea surface temperature (15 %), sea ice concentration (4.8 %), iron (4.4 %), mixed layer depth (3.7 %), salinity (2.5 %) and silicate (0.8 %). Results of the Jackknife test of variable importance are plotted in figure 3.8. The bars in that graph indicate how much the model fit is better than random. A high gain for a particular predictor indicates a greater predictive power. The nitrate predictor, when used isolated, reached the highest gain, which therefore appeared to have the most useful information by itself, followed by sea surface temperature. Excluding a single predictor did not strongly affect the training gain in any case. Mixed layer depth decreases the gain the most when was omitted, and therefore appeared to have the most information that is not present in the other variables.

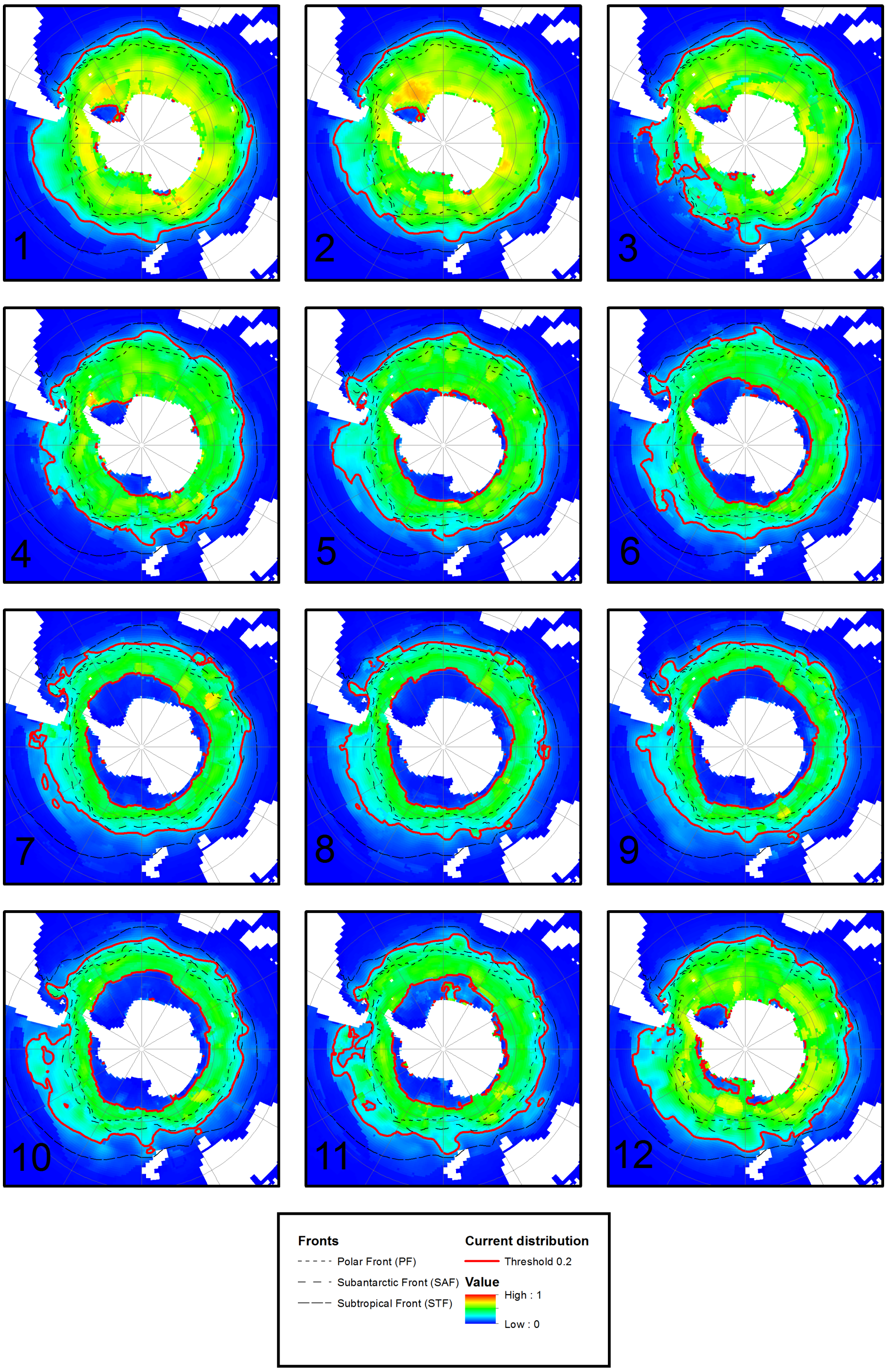

Figure 3.7: Monthly projections of model 3 for the Southern Ocean. *F. kerguelensis* distribution projected on monthly, current environmental data.

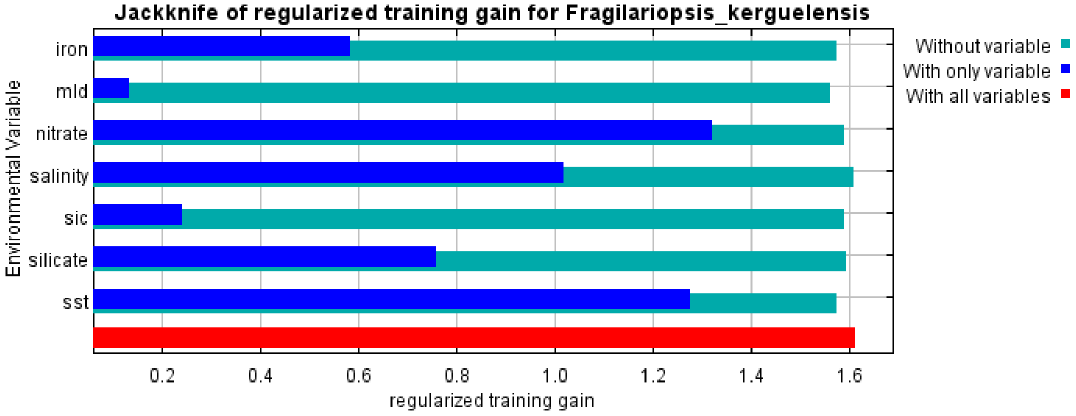

Figure 3.8: The figure shows the results of the Jackknife test for model 3.

Model 3 showed a characteristic southern distribution boundary, forming a belt-shaped gap around the Antarctic continent during the winter season. The contribution of sea ice concentration to the model was less than 5 %. Still, this variable was responsible for this gap. The model's response curve on sea ice concentration showed a significant drop at 79 %, like in model 5f (see fig. 3.4), based just on sea ice concentration. As mentioned before, sampling was strongly biased towards the summer. This can also be interpreted as a sampling bias towards regions with an only small amount of sea ice. The northern boundary could not be explained that clearly by a single predictor. Nitrate concentrations and sea surface temperature play the most important role. Measured distribution areas of models 3, 5 and 6 are listed in table 3.2.

Response curves of all predictors are plotted in figure 3.9 for the full model (model 3), the second version of that model with 20x cross-validation, and the single predictor models in comparison (model 5). The three curves for the predictors showed a very similar shape, except for silicate and mixed layer depth. Response curves for nitrate showed typical saturation curves as expected for nutrients. These curves described the correlation between a species occurrence and the conditions at the occurrence site, which explained the drop at higher nitrate concentrations, in this case slightly above 30 µmol l^{-1}. The silicate curve on the right (single variable model) also showed such a curve, with an increasing logistic output signal towards increasing silicate concentrations and a maximum between 60 and 70 µmol l^{-1}. In contrast, the first and second silicate response curves showed a U-shaped curve with a minimum between 50 and 60 µmol l^{-1} for model 3, and between 60 and 70 µmol l^{-1} for the cross-validation run. The latter response curve showed an increasing standard deviation towards higher silicate concentrations. This pattern can be interpreted as a sign for auto-correlation with another predictor. The peaks for iron and salinity were narrow and in both cases were even sharper for the single predictor model. Sea surface temperature had a maximum at 0 °C. Compared to the single predictor model, the curve was narrower in the full model. The logistic output signal fell under of 0.2 at 6 °C in the first model and at 9 °C in the latter one. A significant drop in the sea ice concentration curves could be found in all of the three curves at 79 %. Mixed layer depth showed different patterns for the single predictor model (model 5) and the full model (model 3).

3.2.6 Yearly averaged projections

Model 6 was built on a dataset of yearly averaged data and minimum, mean, and maximum values of the monthly dataset. The projection is shown in figure 3.10 B. Figure 3.10 A shows model 3, built on a predictor set of monthly resolution and projected on a dataset of yearly means. Further models with just a subset of the yearly predictors, e.g., only mean, and only minimum and maximum were generated and compared (spatial predictions not shown).

Nitrate had most influence in all of the models: 83.4 % for nitrate mean, 78.8 % in total in the min-/max- model (60 % nitrate min, 18.8 % nitrate max), and 70.8 % in the min-/ mean-/ max- model (39 % nitrate min, 27.4 % nitrate mean and 3.8 % nitrate max). Salinity had the second most contribution in all models with 7 % salinity mean, 8.5 % salinity min+max, and 9.8 % salinity min+mean+max. The minimum value was most important in case of nitrate (39 %), salinity (6.8 %), iron (4 %), and mixed layer depth (1.4 %). For sea surface temperature the maximum value (4.4 %) was most important, and the mean value for sea ice concentration (3.5 %). For silicate min and max both contributed with 0.4 %, and the mean silicate concentration dataset with 0.3 %. All models reached high AUC-ROC values (ROC: receiver operating characteristic, a plot of models sensitivity (omission rate) vs. model specificity (fractional predicted area); AUC: area under the curve). The model with minimum, mean and maximum values reached the highest AUC-ROC value of 0.926, followed by the model with minimum and maximum values (AUC-ROC = 0.918) and the model with only mean values (AUC-ROC = 0.91).

The yearly projections showed a similar pattern like the consensus plot of the OpenModeller models in figure 3.1. The northern boundary was shifted north, mainly due to an improved observation dataset with several new observations in the northern regions of the predicted spatial distribution. The maps showed gaps in the Weddell- and Ross-Sea, similar to the December projection of model 3 (the full model based on monthly environmental data, figure 3.7). The northern distribution boundary was more pronounced in the monthly model (model 3) projected on the annual dataset (figure 3.10A than in the projection of the yearly model (fig. 3.10B).

In conclusion, the projection of the monthly model (model 3) on a yearly averaged dataset predicted a reasonable spatial distribution. The model based on yearly minimum, mean, and maximum values showed a similar, but a patchier pattern. Further, direct predictors appeared more useful to analyze, e.g., a response curve on sea surface temperature is more informative than three curves describing the response to maximum-, mean-, and minimum values.

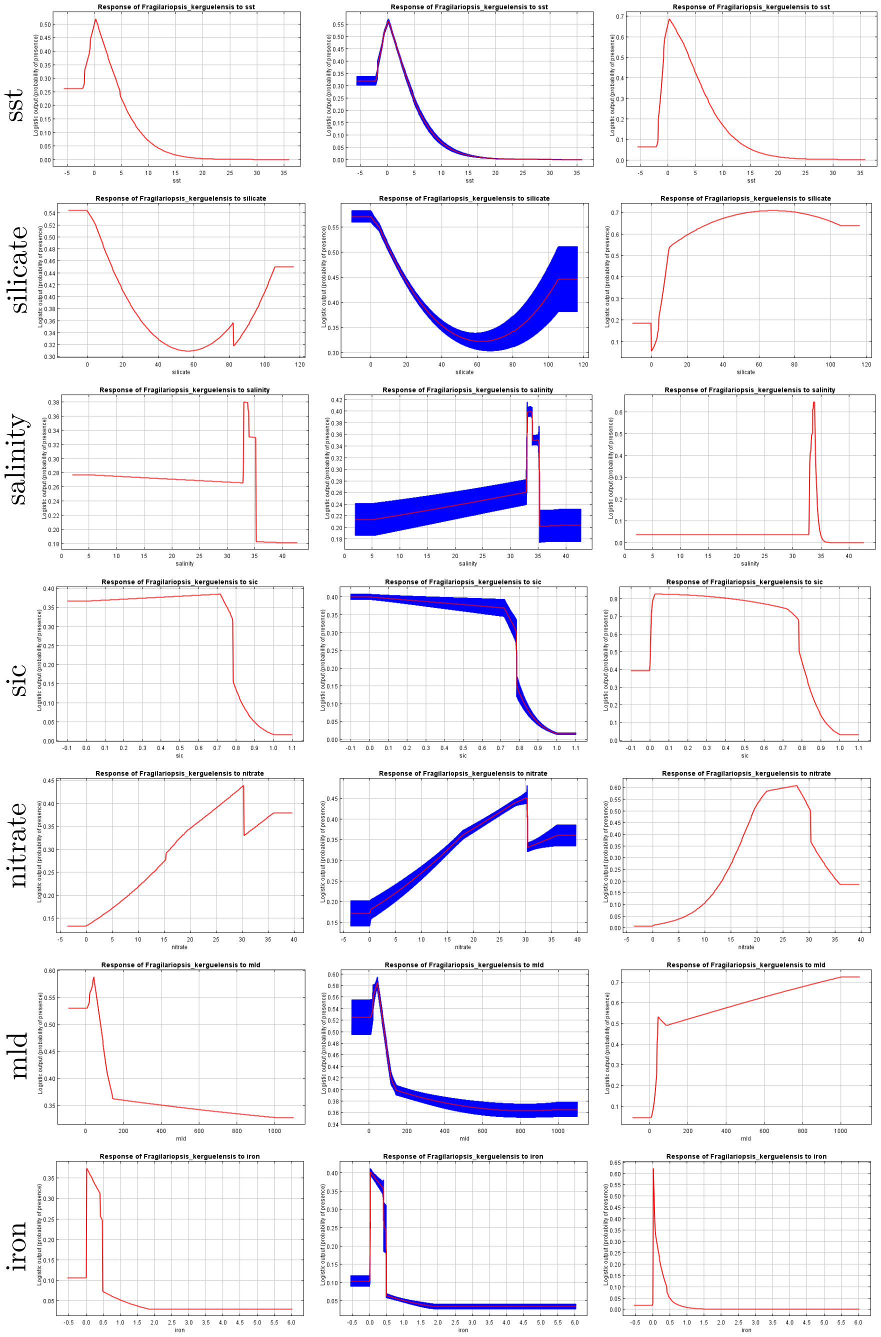

Figure 3.9: The left graphs show the response curves for model 3. Graphs in the middle column show the mean response of 20 replicate Maxent runs (red) and the mean standard deviation (blue). The graphs in the right column belong to the models based on a single predictor.

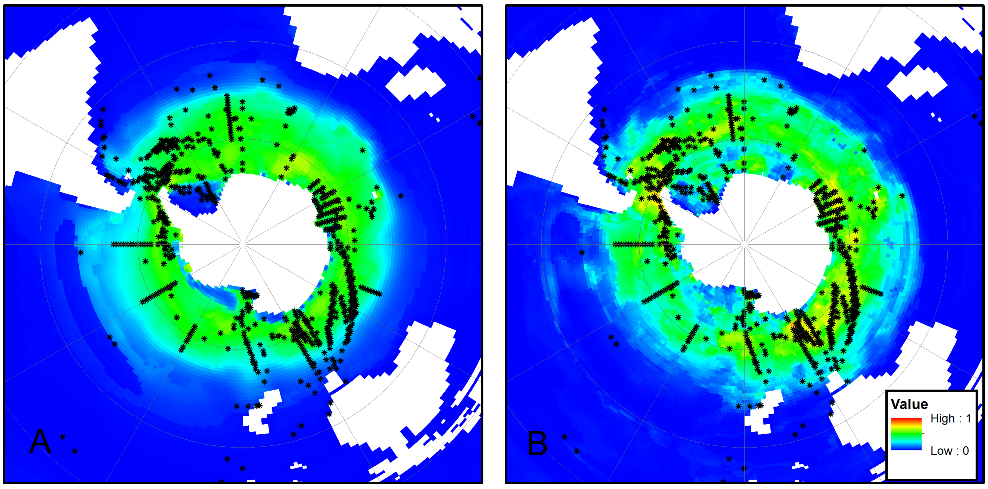

Figure 3.10: Projection of *F. kerguelensis* model 3 (A) and model 6 (B) on a yearly averaged predictor set. Both models used the same set of observation records (indicated by black dots). Model 3 was trained using a monthly dataset, whereas model 6 was directly trained on yearly minimum, maximum and mean values for each predictor.

3.2.7 Future projections

Model 3, the full model for *F. kerguelensis*, was also used for projections on future scenarios for the end of the century. The RCP4.5 and RCP8.5 scenarios were chosen for the year 2100. For the Hadley GCM model, the mixed layer depth dataset was not available. Thus, model 3 was projected on the remaining four GCMs. A reduced model without MLD was created for projection on all five GCMs (model 4a). Iron had only a small effect on the spatial predictions, e.g., in the area between South Argentina and the Falkland Islands. In contrast, the difference in the future projections of models with and without iron was huge. Figure 3.12 shows the projections of model 4a - a model with iron - for the RCP8.5 scenario for August 2100 for each of the five GCMs together with the projection on current August conditions. Figure 3.13 shows the same plots for model 4c - a model without iron. The plots in the latter figure show much less variation among the different GCMs. For better understanding, the value ranges for iron in the "belt" between 40° and 70° South for current and future conditions are plotted in figure 3.11. The current data for model training came from the IPSL-CM5A-LR model and had a similar value range as in its future scenarios. Iron was a good predictor for current and future projections if the latter ones where from the same GCM (see fig. 3.12D). Due to the high variation in iron concentrations between the other GCMs, iron is less useful as a predictor than its ecological relevance suggests.

Figure 3.14 shows combined model projections of the full model (model 3), the model without iron (model 4b) and the model without iron and MLD (model 4c) for February and August. Each plot shows current and future projections for the year 2100

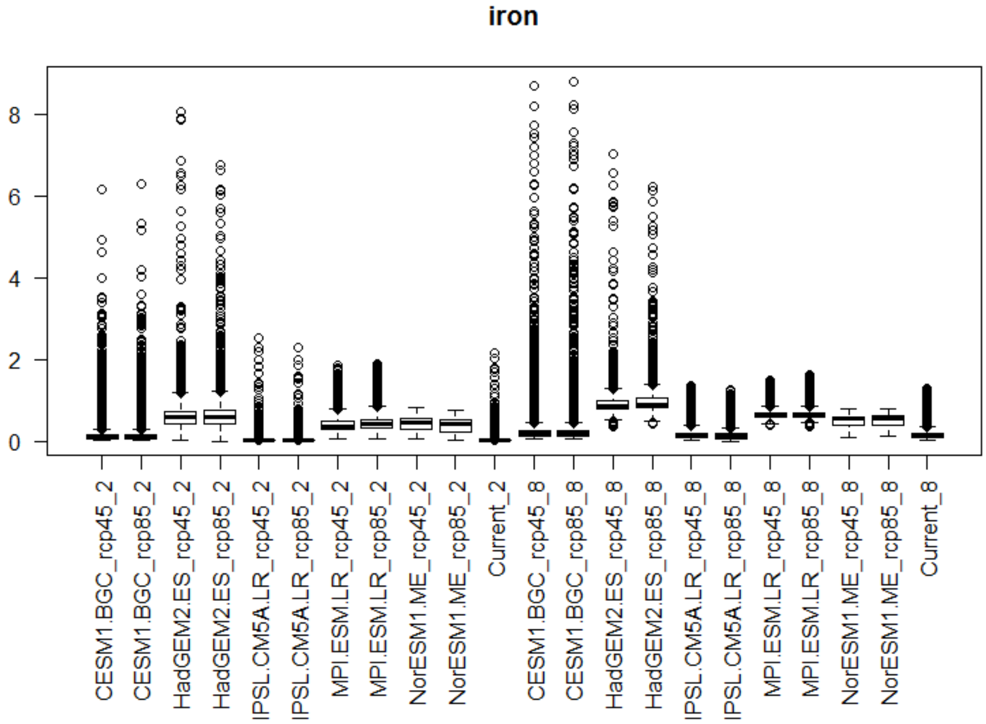

Figure 3.11: February and August iron concentrations in µmol l^{-1} in the "belt" of 40 - 70°S of current environmental data and GCMs for 2100 in the RCP 4.5 and 8.5 scenarios. The name under each boxplot indicates the used GCM model (e.g., HadGEM2-ES), followed by the RCP scenario (e.g., RCP4.5) and the month (e.g., 2 for February). The current dataset are named accordingly.

for the RCP4.5 and RCP8.5 scenario. The measured distribution areas are listed in table 3.2. In all three models, the northern distribution boundary shifted polewards, especially in the summer conditions. During the winter month, a belt-shaped gap around Antarctica remained in all models, again with the boundary shifted polewards. The measured areas, given a threshold of 0.2, are listed in table 3.2.

Thresholds are necessary to calculate the distribution area from the model outputs. The distribution boundaries for five thresholds ranging from 0.1 to 0.5 are plotted in figure 3.15 for comparison. Main differences were found in the Pacific sector of the Southern Ocean, and in between the 0.1 and 0.2 thresholds also in the Indian Ocean sector, where the iso-lines were more distant than in the other ocean basins. A threshold of 0.2, as used throughout the thesis, resulted in an area of 51.61 million square kilometers. Reducing the threshold to a value of 0.1 increased the resulting distribution area to 61.35 million square kilometers. This increase of 9.74 million

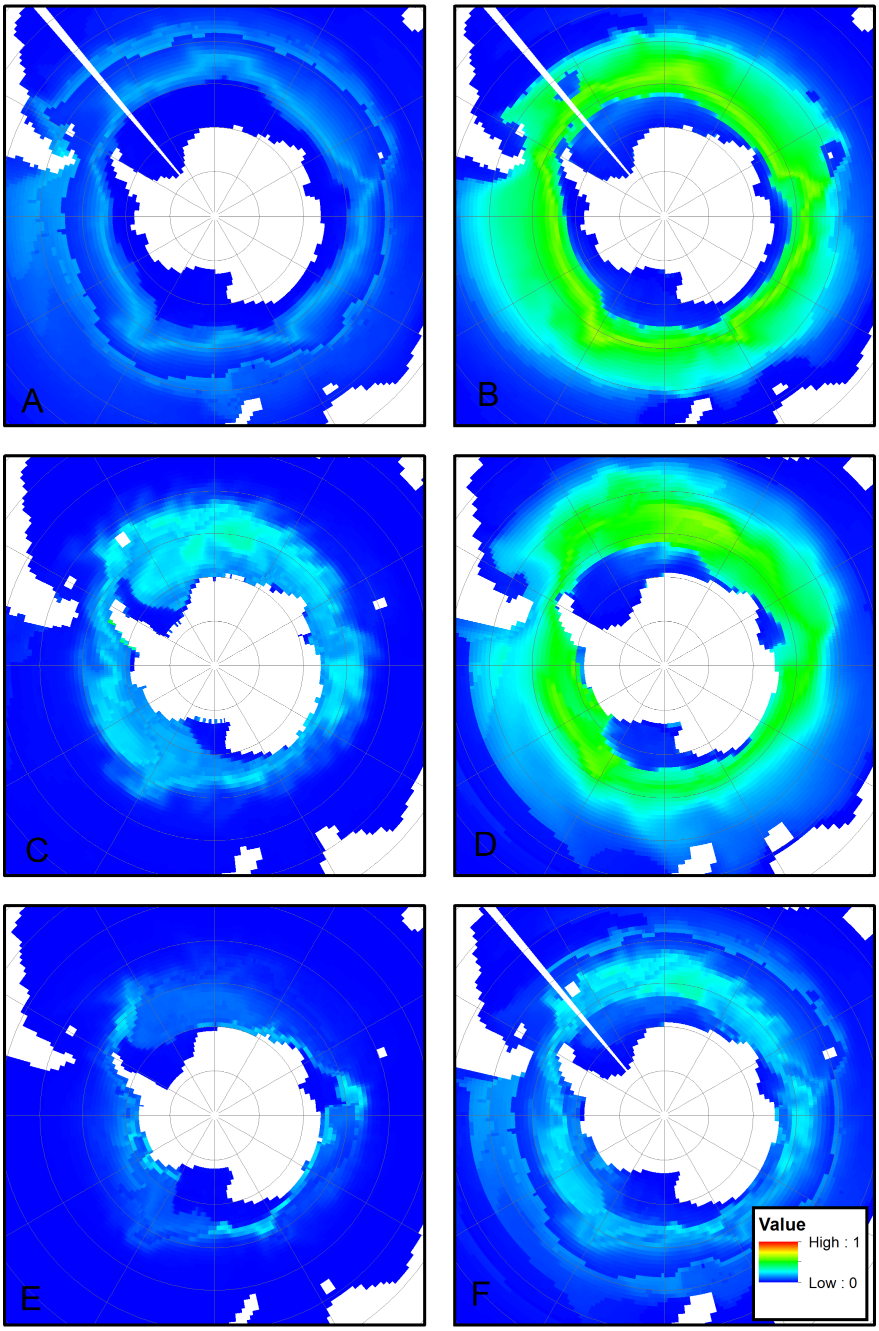

Figure 3.12: Projections of model 4a (with iron, no MLD) on five different GCMs for RCP8.5 scenario for August 2100.
A) NorESM1-ME B) CESM1-BGC C) MPI-ESM-LR D) IPSL-CM5A E) HadGEM2-ES F) Median.

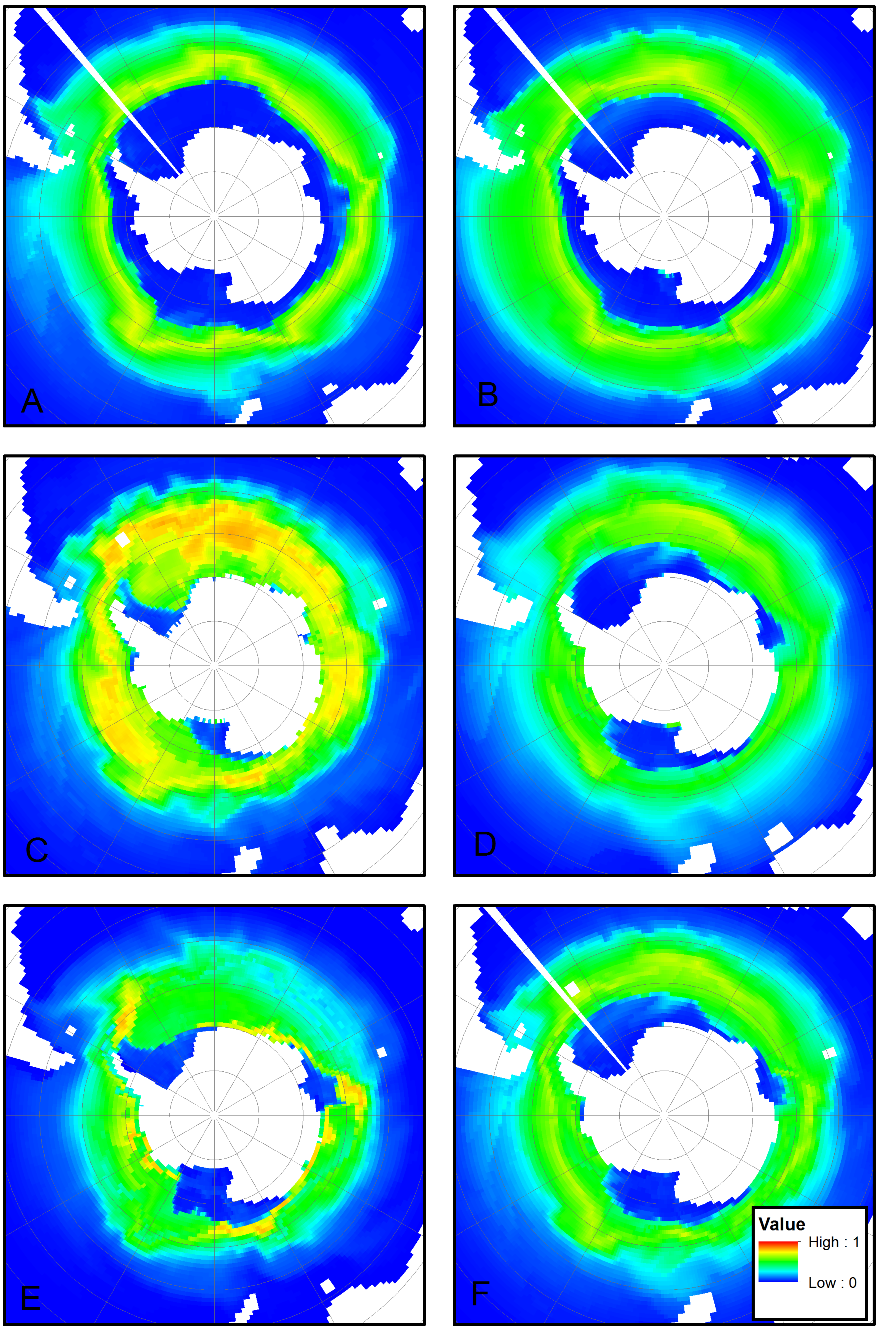

Figure 3.13: Projections of model 4c (no iron, no MLD) on five different GCMs for RCP8.5 scenario for August 2100.
A) NorESM1-ME B) CESM1-BGC C) MPI-ESM-LR D) IPSL-CM5A E) HadGEM2-ES F) Median.

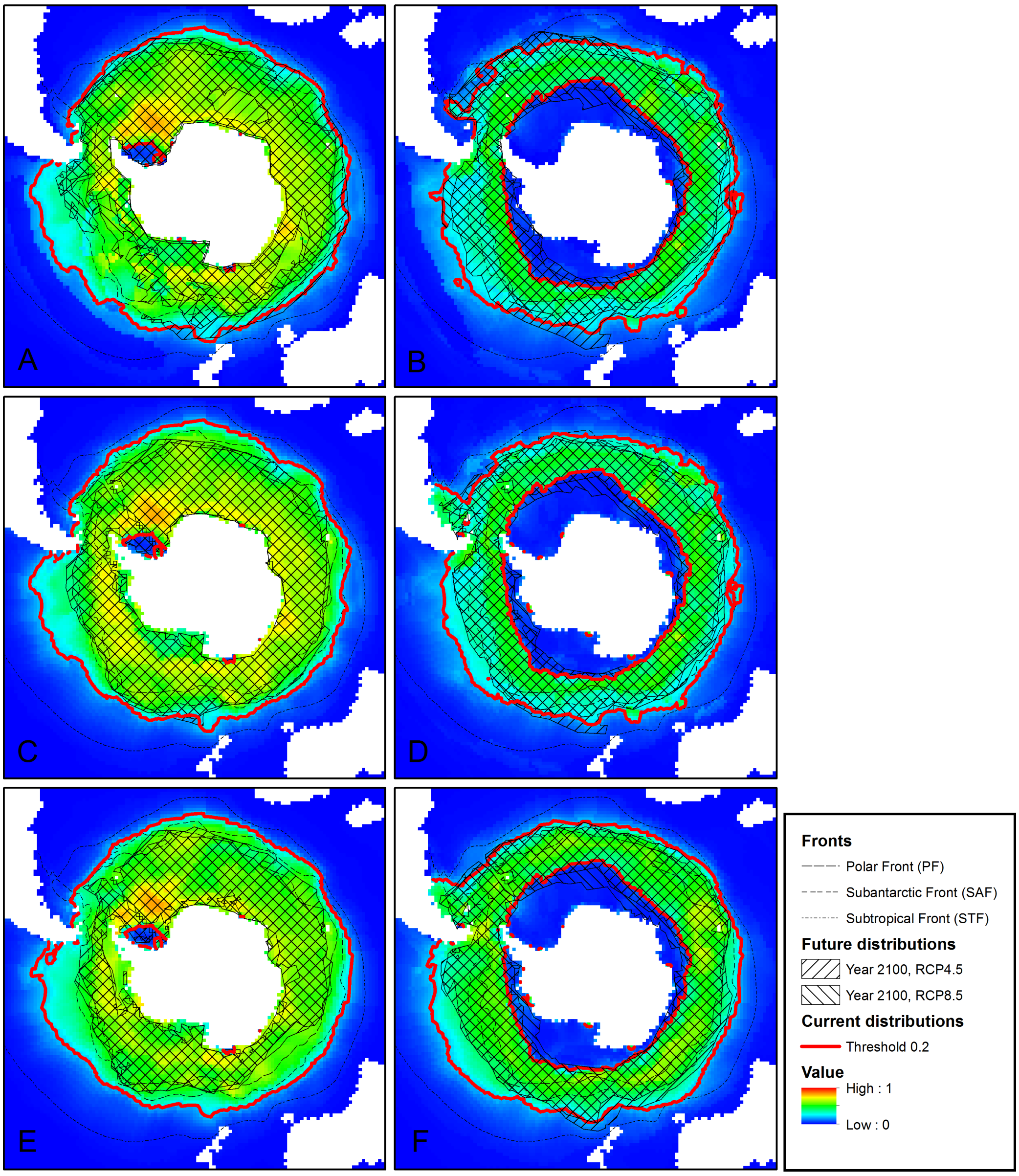

Figure 3.14: Projection of modeled *F. kerguelensis* distribution for February (left column) and August (right column). The red lines indicate the distribution boundaries regarding a threshold of 0.2. The shaded areas indicate the projections on future scenarios for the year 2100 based on the RCP 4.5 and 8.5 scenarios. The measured areas are listed in table 3.2.
A+B) Model 3, future projections on 4 GCMs. C+D) Model 4b, without iron, future projections on 4 GCMs. E+F) Model 4c, no iron and no MLD, future projections on 5 GCMs.

Table 3.2: Measured areas of *F. kerguelensis* predictions (in million km^2). The values belong to the maps in fig. 3.14.

Projection	Model 3	Model 4b	Model 4c
	(full model)	(no iron)	(no iron, no MLD)
Current, February	51.61	50.65	50.03
RCP4.5, February	39.35	37.21	30.00
RCP8.5, February	32.11	33.63	26.21
Current, August	35.62	36.52	40.87
RCP4.5, August	37.94	33.66	36.41
RCP8.5, August	34.23	31.78	35.86

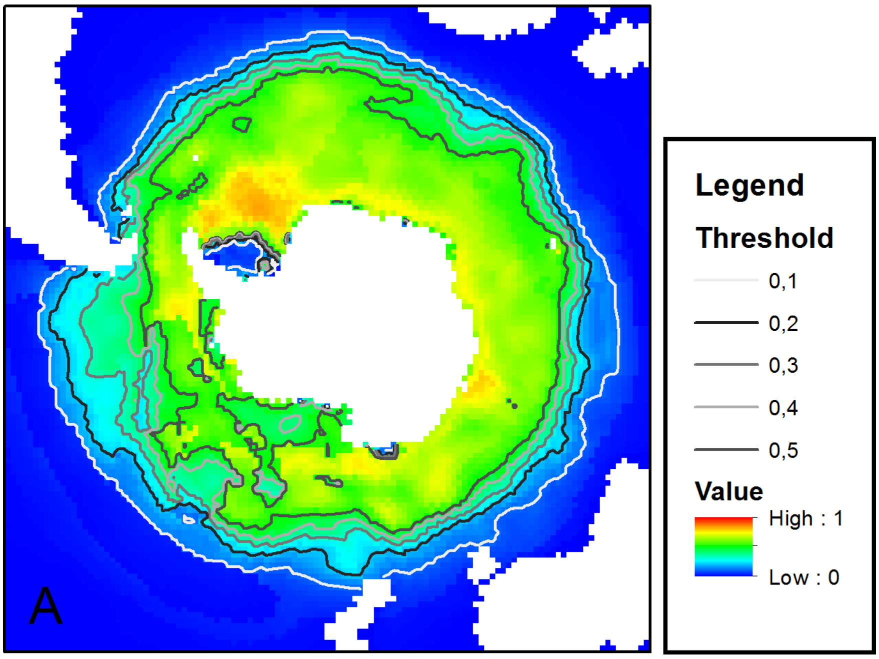

Figure 3.15: Iso-lines according to threshold values from 0.1 to 0.5 for the projection of model 3 on February conditions.

square kilometers (+18.9%) matches roughly the size of the USA. On the other hand, a threshold of 0.3 resulted in an area of 44.61 million square kilometers. Compared to a threshold of 0.2 this is an area loss of 7 million square kilometers (-13%), almost the area of Australia (7.7 million square kilometers). A threshold of 0.4 results in 38.25 million square kilometers, 0.5 in 30.78 million square kilometers. Despite these huge areas, the three relevant iso-lines (threshold of 0.1 - 0.3) are varying by just a few degrees in latitude which - one degree latitude (=60 nautical miles) equals 111.12 km - end up in just a few hundred kilometers. Threshold selection is further discussed in chapter 4.1.2 on page 75.

In conclusion, all future predictions showed a decreased distribution area compared to the current distribution. As expected, for the RCP8.5 scenario, the decrease is stronger than for the RCP4.5 scenarios. Though not all predictors were available, and iron appeared to be problematic for some GCMs, spatial projection can give some hints about potential range shifts and future species distribution.

3.3 Models for other species

Distribution models were calculated for 20 further species listed in table 2.1. Figures 3.16 to 3.20 plot current February and August model projections and projections on the RCP8.5 scenario for February 2100. Current and future (according to the RCP8.5 scenario) distribution areas were measured for the February projections (see table 3.3). Several models indicate suitable habitat conditions in the northern hemisphere for arctic cold water-masses or other high nutrient low chlorophyll (HNLC) regions, e.g., in the northern Pacific, an HNLC region with similar characteristics like the Southern Ocean. For this study, the measurements are limited to the region of the Southern Ocean and adjacent ocean basins. For three species (*Asteromphalus heptactis*, *Corethron pennatum*, and *Dactyliosolen antarcticus*) the areas could not be measured, as no meaningful northern distribution boundary could be selected. Only in two cases, the models predict an increased distribution area: *Fragilariopsis vanheurckii* by 15 % and *Asteromphalus hookeri* by 30.2 %. This is insofar surprising as these two models are based on a very different set of relative predictor contributions (see figure 3.21). In the future distribution of *F. vanheurckii*, the gaps in the circum-continental distribution are closed in the future projections (see fig. 3.18 E), but the northern distribution boundary does not change much. In contrast, the northern boundary of *A. hookeri* is shifted northwards.

According to this model projection, *F. linearis* would completely fade away from the Southern Ocean for the end of the century. Here, again a threshold of 0.2 is used. The model, however, predicts a signal lower than this, but with an even smaller area than the 4.53 million km^2 of the current February prediction (again threshold of 0.2). Further, the explanatory power of this model is probably strongly limited, as it is based on only 13 usable observation records. The range of area loss among the species is high, ranging from only 0.9 % for *F. pseudonana* to 67.7 % for *F. sublinearis*. In comparison, model 3 for *F. kerguelensis* predicted a medium area loss of 37.8 %.

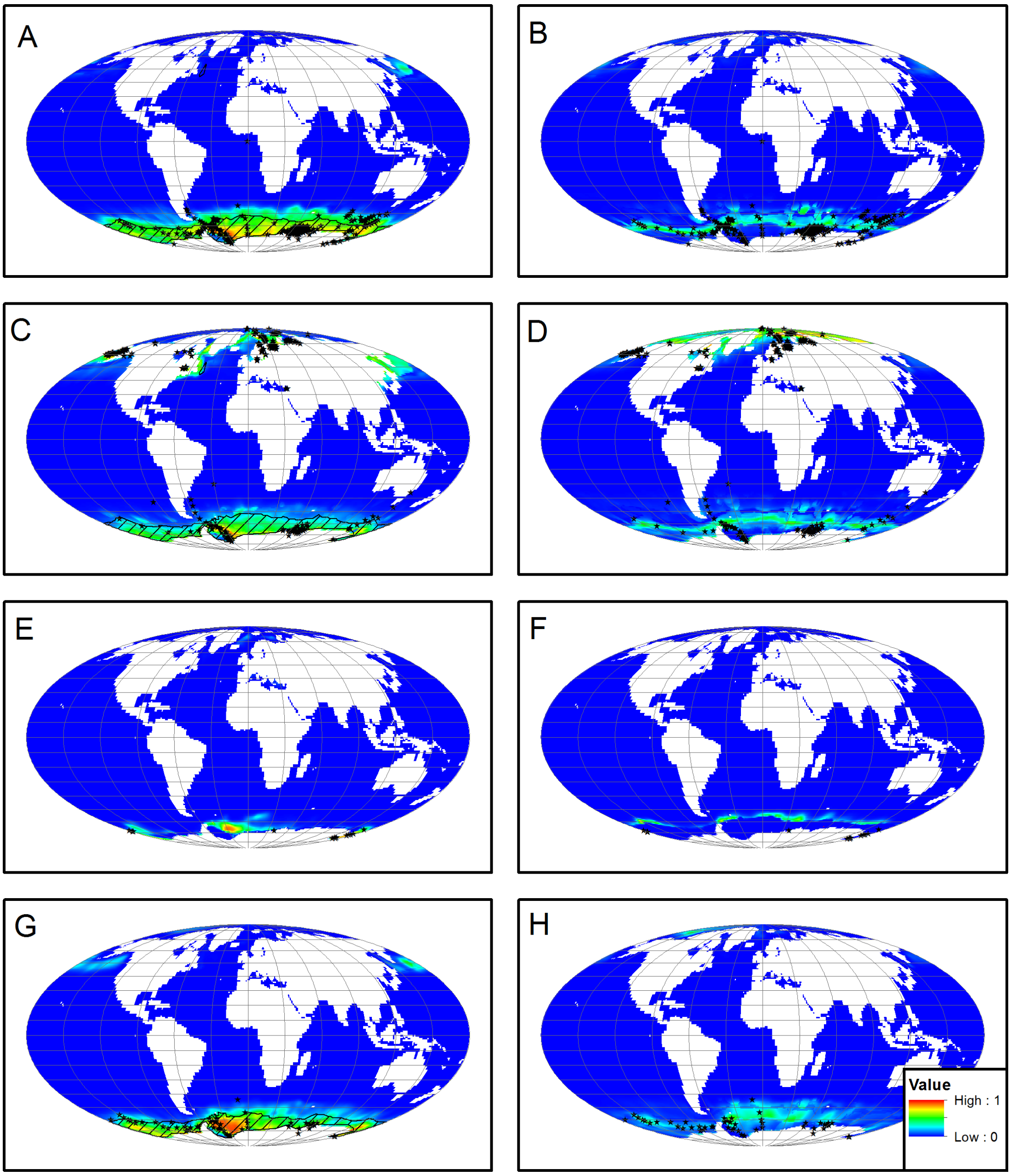

Figure 3.16: Projection on February (left column) and August (right column) conditions for A+B) *Fragilariopsis curta*, C+D) *Fragilariopsis cylindrus*, E+F) *Fragilariopsis linearis*, G+H) *Fragilariopsis nana*. The hatched areas indicate projected future distributions for February 2100 according to RCP8.5 scenario.

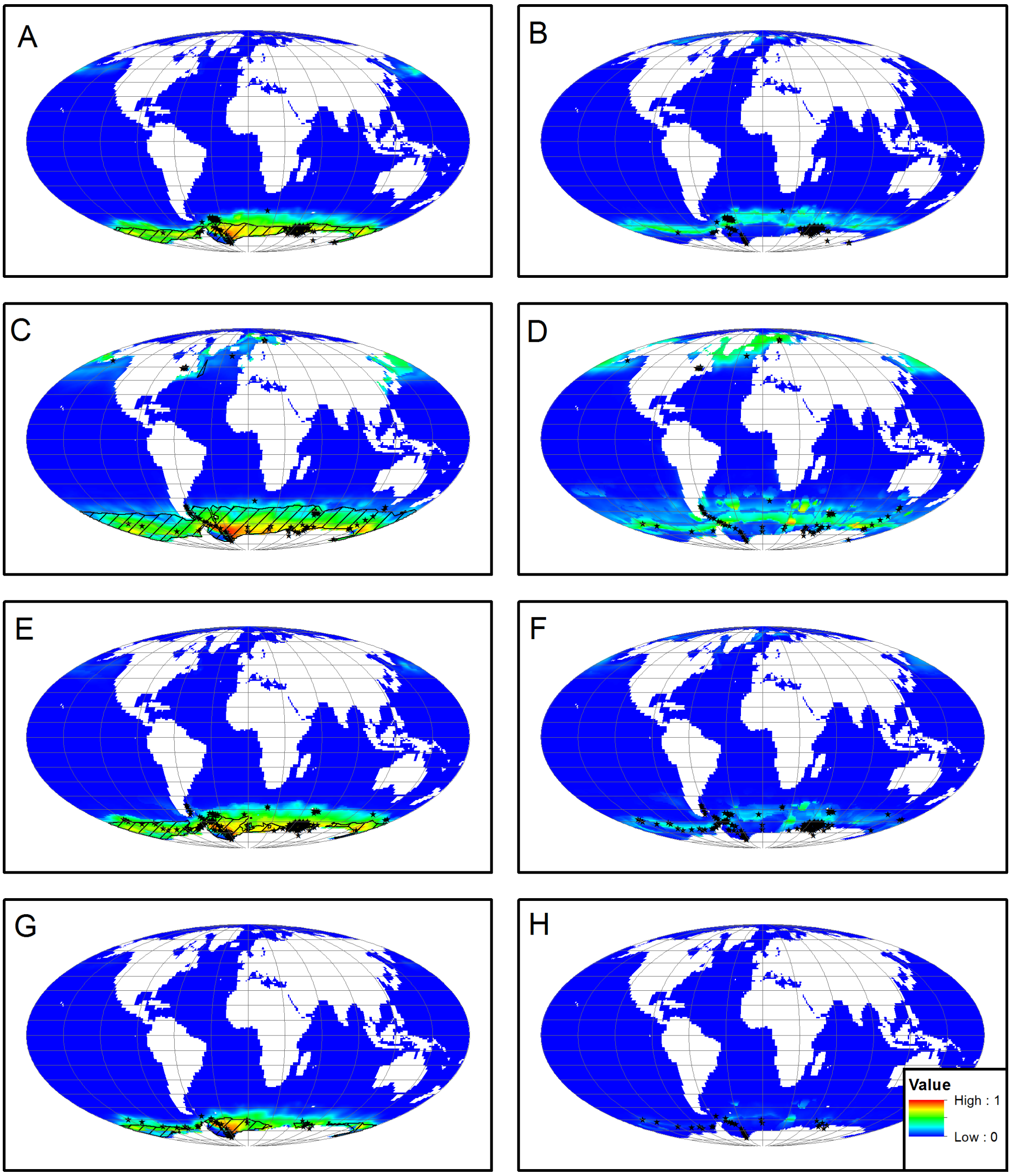

Figure 3.17: Projection on February (left column) and August (right column) conditions for A+B) *Fragilariopsis obliquecostata*, C+D) *Fragilariopsis pseudonana*, E+F) *Fragilariopsis rhombica*, G+H) *Fragilariopsis ritscheri*. The hatched areas indicate projected future distributions for February 2100 according to RCP8.5 scenario.

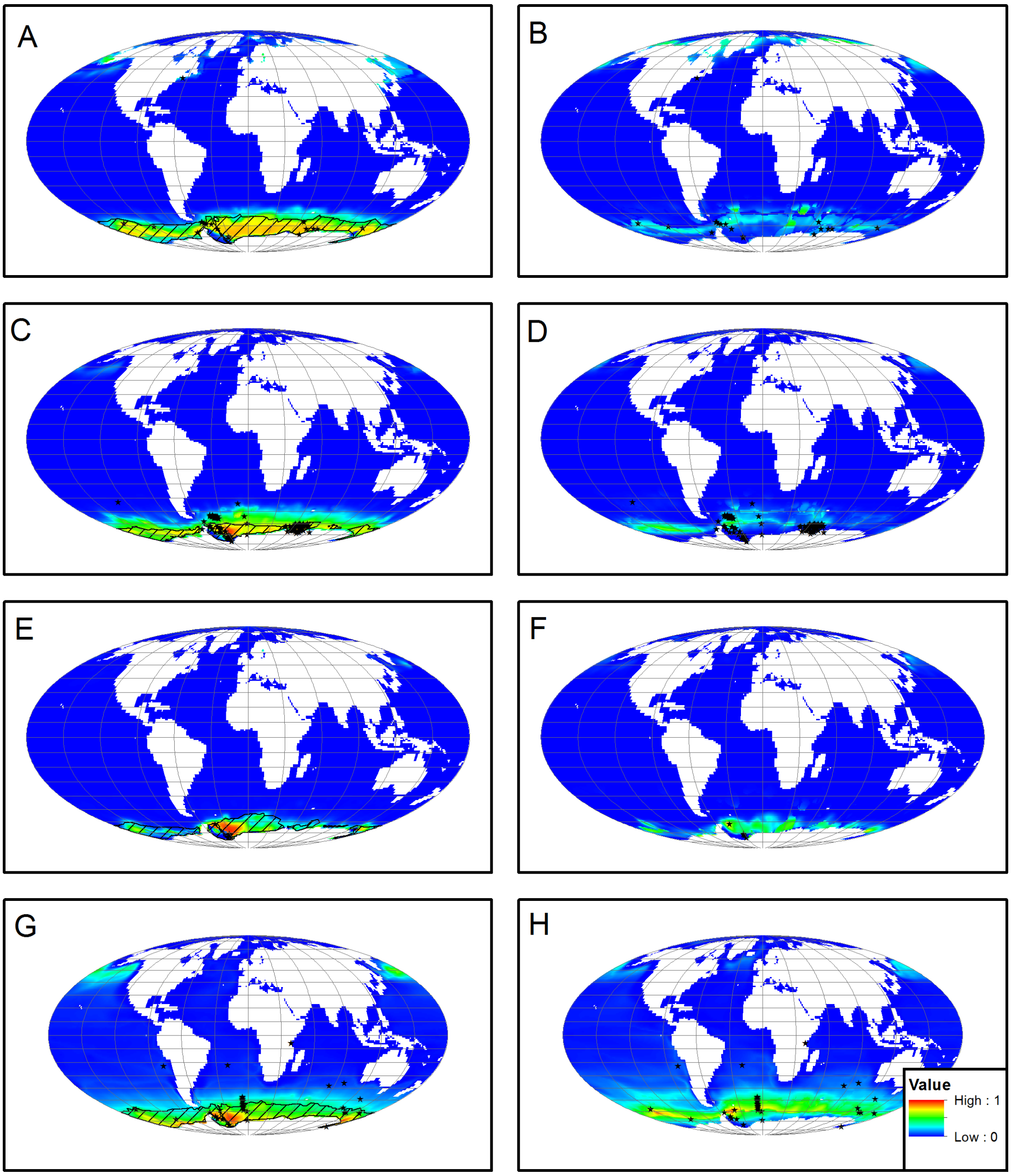

Figure 3.18: Projection on February (left column) and August (right column) conditions for A+B) *Fragilariopsis separanda*, C+D) *Fragilariopsis sublinearis*, E+F) *Fragilariopsis vanheurkii*, G+H) *Asteromphalus roperianus*. The hatched areas indicate projected future distributions for February 2100 according to RCP8.5 scenario.

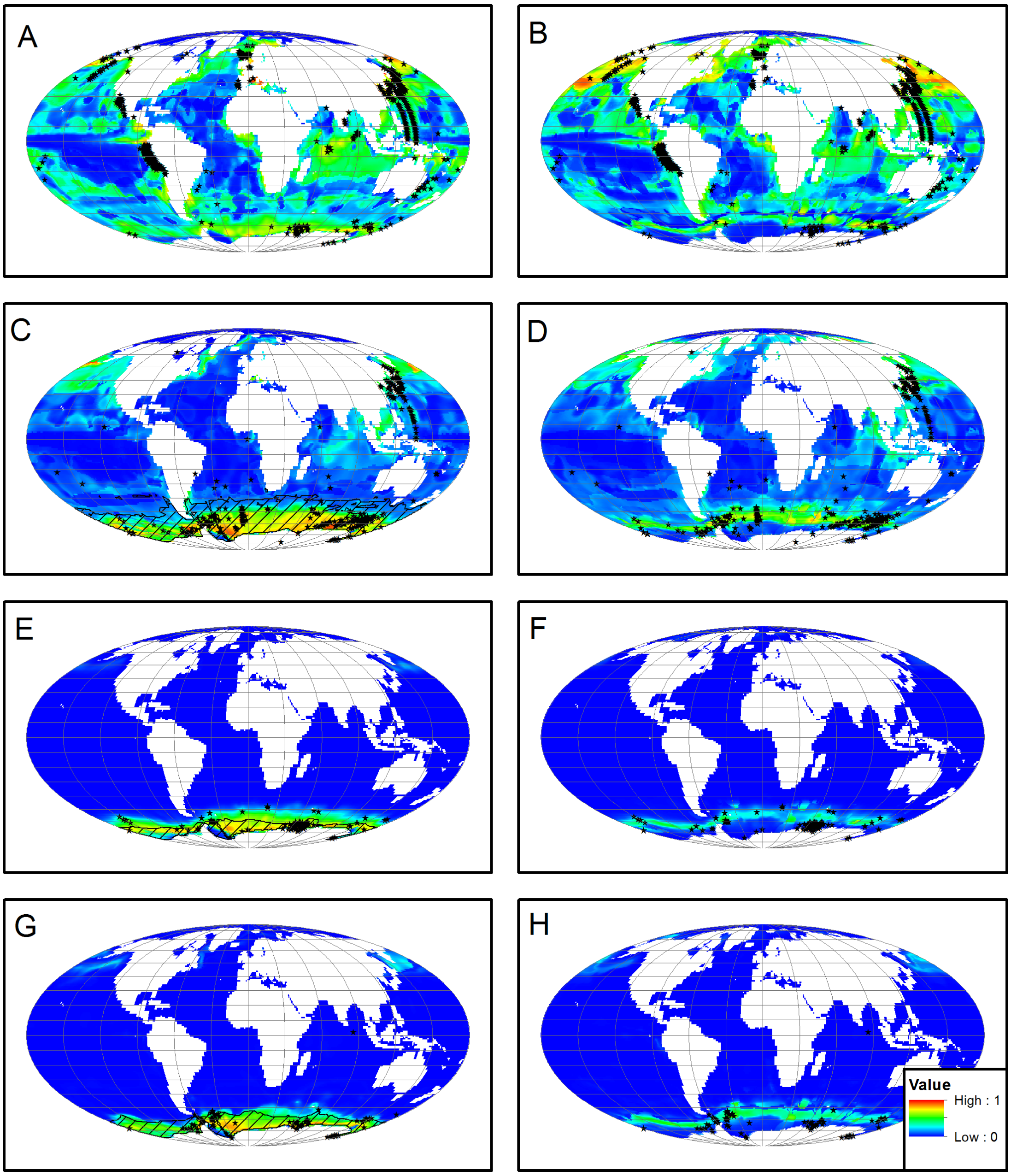

Figure 3.19: Projection on February (left column) and August (right column) conditions for A+B) *Asteromphalus heptactis*, C+D) *Asteromphalus hookeri*, E+F) *Asteromphalus hyalinus*, G+H) *Asteromphalus parvulus*. The hatched areas indicate projected future distributions for February 2100 according to RCP8.5 scenario.

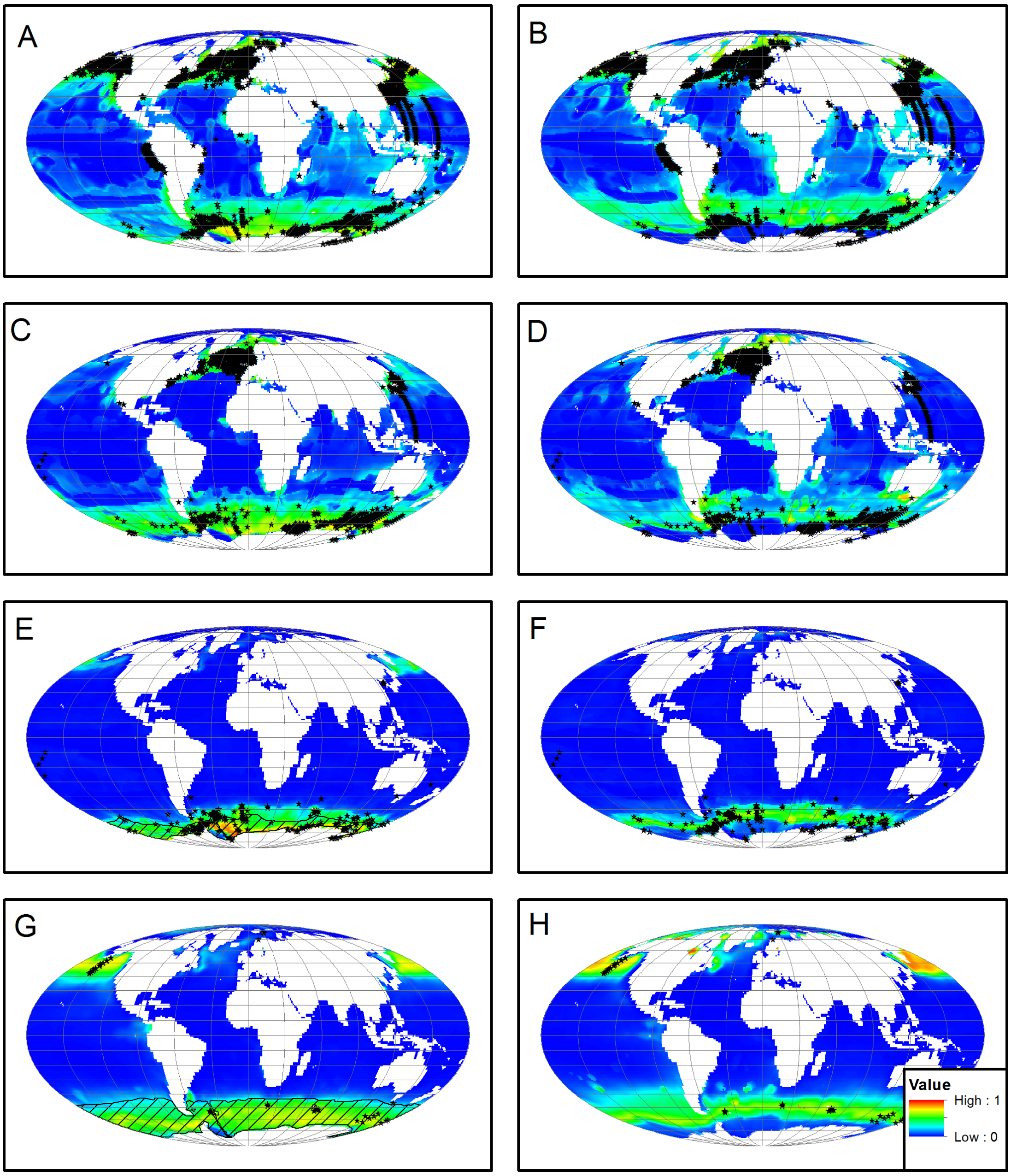

Figure 3.20: Projection on February (left column) and August (right column) conditions for A+B) *Corethron pennatum*, C+D) *Dactyliosolen antarcticus*, E+F) *Eucampia antarctica*, G+H) *Azpeitia tabularis*. The hatched areas indicate projected future distributions for February 2100 according to RCP8.5 scenario.

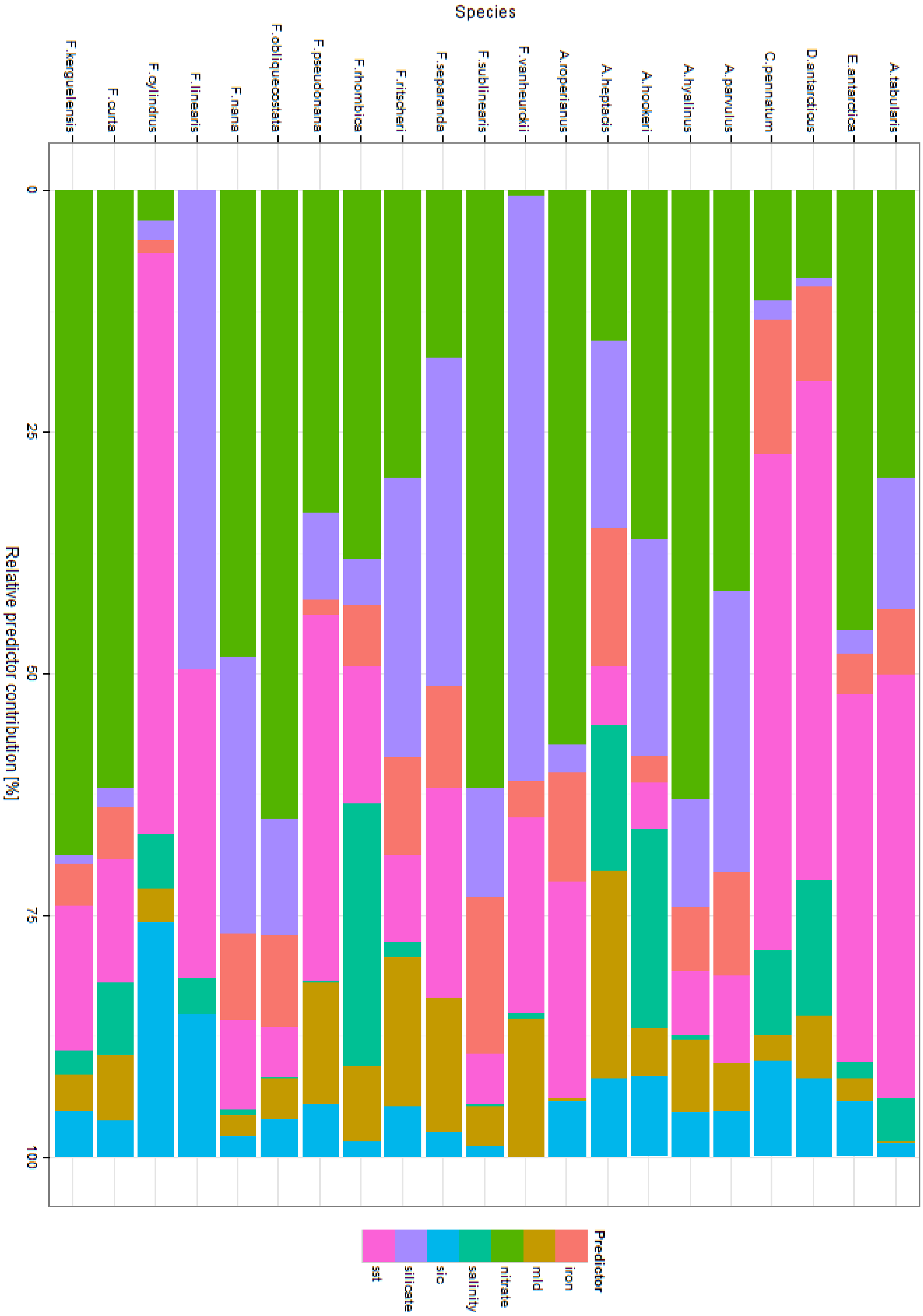

Figure 3.21: Overview of relative predictor contributions in the SDMs for all 21 species.

Table 3.3: Measured areas (in million km^2) of the models February predictions for current and modeled future environmental conditions, as well as the percentage loss of area. The values belong to the maps in fig. 3.16 to 3.20. The measurements were based on a threshold of 0.2 and are limited to the regions in the Southern Ocean. For three species (*Asteromphalus heptactis*, *Corethron pennatum*, and *Dactyliosolen antarcticus*) areas where not measured, as no meaningful northern boundary could be set.

Species	February (current)	February 2100 RCP4.5	Loss of area [%]	February 2100 RCP8.5	Loss of area [%]
F. curta	40.00	32.43	18.9	33.11	17.2
F. cylindrus	24.66	27.36	-10.9	23.85	3.3
F. linearis	4.53	0.00	100.0	0.00	100.0
F. nana	26.42	24.34	7.9	19.88	24.8
F. obliquecostata	29.48	19.39	34.2	15.37	47.9
F. pseudonana	45.56	50.93	-11.8	45.17	0.9
F. rhombica	36.39	16.45	54.8	15.14	58.4
F. ritscheri	19.55	12.54	35.9	9.01	53.9
F. separanda	35.47	30.85	13.0	27.85	21.5
F. sublinearis	33.26	14.10	57.6	10.73	67.7
F. vanheurckii	12.19	16.56	-35.9	14.02	-15.0
A. hookeri	44.14	63.55	-44.0	57.50	-30.2
A. hyalinus	29.73	16.95	43.0	12.88	56.7
A. parvulus	26.45	25.73	2.7	22.01	16.8
A. roperianus	44.59	29.89	33.0	26.45	40.7
A. tabularis	60.51	60.21	0.5	56.30	6.9
E. antarctica	35.62	23.15	35.0	19.46	45.4

The species can be grouped into endemic to the Southern Ocean, bipolar, and cosmopolitan by their distribution pattern. Three models were based on bipolar observation data and predicted a bipolar distribution: *Fragilariopsis cylindrus*, *Fragilariopsis pseudonana*, and *Azpeitia tabularis.* For the majority of the remaining species, observation data were available only for the southern hemisphere, but only five models predicted a distribution limited to the Southern Ocean: *F. linearis*, *F. rhombica*, *F. ritscheri*, *F. vanheurckii*, and *Asteromphalus hyalinus.* The remaining nine models, based on species with occurrences only in the South, predicted a weak occurrence signal for the North Pacific like *Fragilariopsis curta*, *F. nana*, *F. obliquecostata*, and *F. sublinearis*, some even a strong occurrence signal like *Fragilariopsis separanda*, *Asteromphalus hookeri*, *A. roperianus*, *A. parvulus*, and *Eucampia antarctica. Fragilariopsis kerguelensis* belongs to the latter category, too. Three models predicted a wider distribution: *Asteromphalus heptactis*, *Corethron pennatum*, and *Dactyliosolen antarcticus.*

To summarize and compare, all distribution patterns were used for a hierarchical clustering, applied to the integrated maximum distribution areas of the February and August projections (see figure 3.22). The clustering reflected just the projected distribution patterns, not if observation records in the north were existing. Thus, truly bipolar species were not clearly separated from species with falsely predicted occurrences in the north.

A cluster with *F. linearis*, *F. ritscheri*, and *F. vanheurckii*, representing a distribution pattern limited to the Southern is formed. Its sister cluster of twelve species consists mainly of bipolar distribution patterns. In just one case, this reflects a truly bipolar distribution (*F. pseudonana*). Three species, *A. hyalinus*, *F. sublinearis*, and *F. rhombica*, are falsely included, as they show a distribution pattern endemic to the Southern Ocean. The remaining eight species show a bipolar distribution pattern, though the observation records indicate them to be endemic to the Southern Ocean. The globally distributed species clustered well, but with *A. tabularis* a bipolar species was included in this cluster. Also, *A. hookeri* clustered here, probably due to the strong signal in the north-eastern part of the Indian Ocean. *F. cylindrus*, another truly bipolar species, stayed outside any bigger cluster.

The environmental conditions at the observation sites were clustered similarly, using a Manhattan distance matrix and complete hierarchical clustering method (figure 3.22). In contrast to the previously described dendrogram, the models resulting spatial patterns are not accounted, but only the models training data. The spatial patterns, identified before were only partially identified. The three global species, *A. heptactis*, *D. antarcticus*, and *C. pennatum*, clustered like before. *F. cylindrus* again was not included in any of the big groups. A big cluster with the remaining 17 species contained the bipolar and Southern Ocean only distribution patterns, but could not distinguish them.

The number of observations taken in to account for the models varied from just five for *Fragilariopsis vanheurkii* up to 1606 for *Corethron pennatum.* All models reached high AUC values, ranging from 0.849 to 0.998. While the composition of variable importance was similar across the various *Fragilariopsis kerguelensis* models, it's more diverse for the different species. Figure 3.21 shows the percentage of variable con-

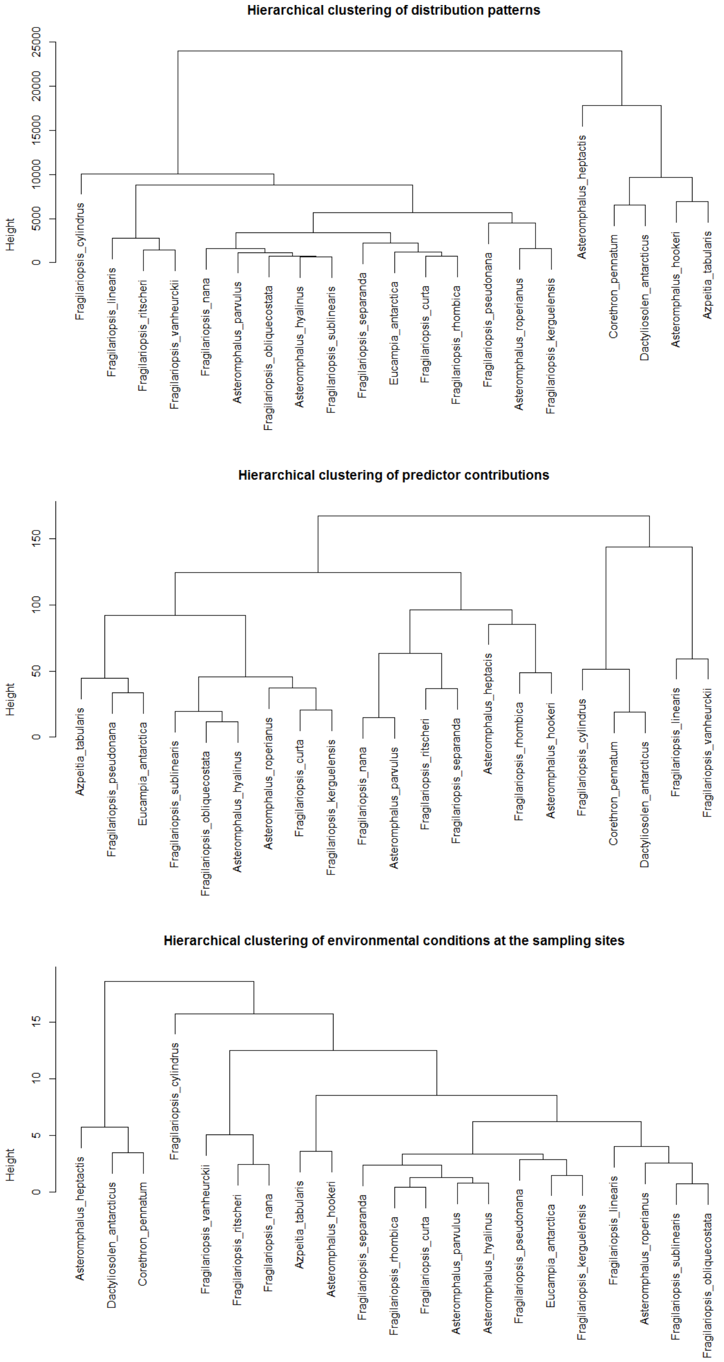

Figure 3.22: Hierarchical clustering of distribution patterns, relative predictor contribution and environmental conditions at the observation sites using a Manhattan distance and complete clustering method.

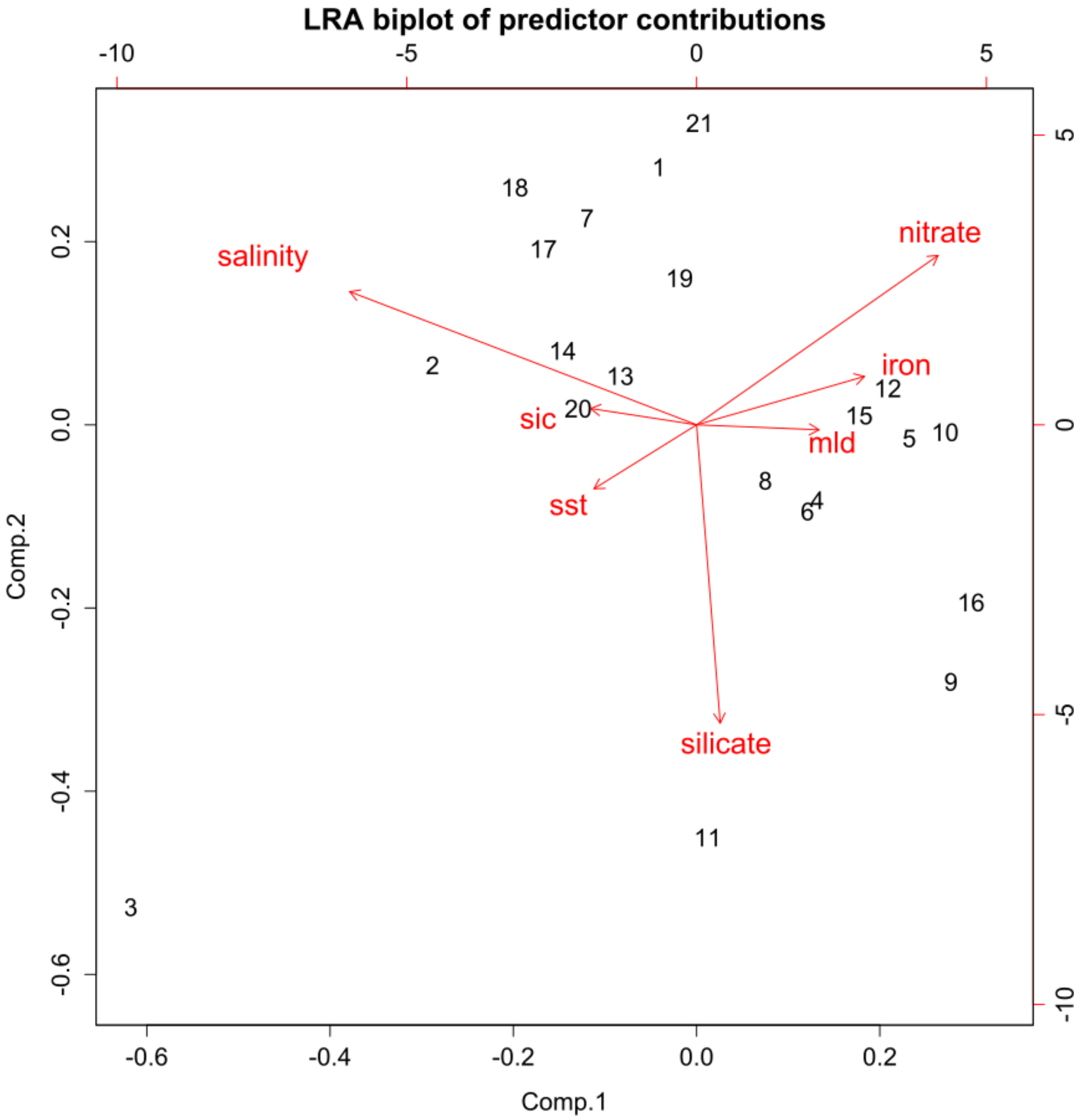

Figure 3.23: Model analysis using a biplot of a log ratio analysis of relative predictor contributions in the Maxent models. Variable loadings are represented by the red arrows. Black numbers indicate the component scores of the individual species. 1: *Fragilariopsis curta*, 2: *F. cylindrus*, 3: *F. linearis*, 4: *F. nana*, 5: *F. obliquecostata*, 6: *F. pseudonana*, 7: *F. rhombica*, 8: *F. ritscheri*, 9: *F. separanda*, 10: *F. sublinearis*, 11: *F. vanheurckii*, 12: *Asteromphalus roperianus*, 13: *A. heptacis*, 14: *A. hookeri*, 15: *A. hyalinus*, 16: *A. parvulus*, 17: *Corethron pennatum*, 18: *Dactyliosolen antarcticus*, 19: *Eucampia antarctica*, 20: *Azpeitia tabularis*, 21: *F. kerguelensis*

tribution in comparison. The relative predictor contributions were clustered, using a Manhattan distance matrix and a complete hierarchical clustering method (figure 3.22). In this case, the clusters, identified in the distribution pattern clustering and environmental conditions clustering, did not show up. This means, though using observation records from the same regions and similar resulting distribution patterns, the optimal model found by Maxent can be completely different.

The relative predictor contribution (see also fig. 3.21) of the 21 models was also analyzed by a log ratio analysis (LRA biplot in fig. 3.23). Nitrate and sea surface temperature are on the same axis, but in opposite directions. This means that nitrate and sea surface temperature can partly be replaced by each other, with just one of them playing the major role in a model. The second axis is built by salinity, followed by silicate on the third. Both predictors play an important role in the models, with each of them containing further independent information.

The distribution patterns correlate well with the most influential predictor (see fig. 3.21). Nitrate was most influential for the species that, according to their distribution of observation records, are endemic to the Southern Ocean. This includes models with and without a signal in the North. As an exception, nitrate had only a small influence in favor of silicate in three cases: *F. separanda*, *F. linearis*, and *F. vanheurckii*. These are also the three models with the least number of observation records. With a few exceptions, mixed layer depth plays a more important role in the group with a signal only in the Southern Ocean, whereas iron is more important in the models with (false) predictions in the North. In contrast, sea surface temperature was most influential in the models of the truly bipolar species. This is also reflected in fig. 3.23, where nitrate and sea surface temperature are on the same axis. Silicate can replace nitrate, and usually just one of them has a high importance. Salinity also shows a strong signal in that plot. With several exceptions, salinity plays a less important role for species endemic to the Southern Ocean than for bipolar species. These exceptions are the bipolar *F. pseudonana* with a low response on salinity (0.2%), and *F. curta* and *F. rhombica* with a strong response on salinity (7.5% and 27.2%).

Global distribution patterns can be partly identified in the model's input data. With the Maxent models, truly bipolar and, according to the distribution pattern, potentially bipolar species are distinguishable by the predictor influence.

3.4 Perturbation experiment

All cultures showed the same reactions when exposed to increased temperatures: slower growth, shorter chain length, as well as smaller and less colored chloroplasts. However, the temperature at which this was observed varied strongly among the cultures. The first culture (from station 301) already showed these signs at a temperature of 7°C. At 8°C, the cells of this culture were dead, whilst the other cultures still seemed to be vital. At a temperature of 9°C, shorter cell-chains were observed in two more cultures (from stations 364 and 404), but cells still seemed vital. After the next temperature rise to 10°C, one of these (station 364) contained only a few chains and many separated dead valves, whereas the other (station 404) still contained short chains of vital cells. Even at a temperature of 10°C, the fourth culture (station 374) still contained long chains of vital cells and was growing. At the temperature of 11°C, no vital cells remained in the first three cultures. They mostly contained single cells and only few cell chains. Only the last culture contained a few living cells but did not seem healthy. The experiment was stopped at this temperature.

4 Discussion

This chapter contains three main sections. The methodical aspects of this thesis are discussed in the first part, including a discussion of observation and environmental data quality as well as a discussion of the modeling process with its chances and limitations. Second, the findings of the species ecology and current biogeography are summarized and compared to previous knowledge. This part also includes a discussion of the main distribution patterns resulting from the models as well as the future projections. The last part contains a synthesis, a wrap-up of the research questions, and gives an outlook.

4.1 Methodical aspects

4.1.1 Data

Observation data

Increasing accessibility of species observation records, especially of the type presence-only data, allows a broad use for biogeographical modeling. Presence-only data contain the information that a species was observed at a certain location and time, but not on species abundance or absence. Data of higher quality, e.g., species abundance data, might allow more detailed models due to their higher information content. But the use of presence-only data in distribution models like Maxent might be more informative for phytoplankton studies, as their biomass varies widely and is not distributed homogeneously. Distribution areas, as predicted here, might even be more informative than the typically lower scaled population dynamics.

Data from diverse sources needs to be aggregated to achieve a decent amount of observation records for a species distribution model. In this process, all data need to be harmonized by conversion to the highest common denominator, which mostly turns out to be of the type presence-only. This is also the case in big data repositories, such as GBIF and OBIS, where all kinds of data are aggregated to large sets of presence-only data. Further, collections as typically found in natural history museums, often only allow deriving presence-only data.

The distribution models discussed in this study are all based on presence-only data, inferred from public databases, the Hustedt Diatom Collection, and literature. Various studies have shown that the quality of the observation data is crucial to build a reliable species distribution model. First results, published in *Pinkernell and Beszteri* (2014), showed that distribution models based on publicly available observation can give decent results. In that study, projections across time appeared to be very sensitive on the used observation dataset in the case of future scenarios, and an improved version of the observation dataset led to more robust distribution models.

Big data repositories provide access to aggregated data from various data providers by a single data portal and also make them available as web services. Organizations such as GBIF or the more marine oriented OBIS have established global networks for biodiversity data and became an invaluable resource for biodiversity and biogeographical studies. Harmonized data formats and access to a multitude of the many different sources are the main advantages of these systems, as many entries, aggregated from various sources, would certainly be much harder to find and to harvest if they were not included in such a network. These networks often are the first addresses for observation data in many studies. On the other hand, various problems arise using this data because of a lack of voucher images, sometimes questionable taxonomical classifications, and biased data due to huge gaps in the spatial and environmental coverage of samples. GBIF consists of more than 90 nodes, distributed worldwide. The GBIF-node for plants, algae, and protists, hosted at Botanischer Garten und Botanisches Museum Berlin-Dahlem at Freie Universität Berlin, is included in the global GBIF portal and provides some specialized data portals, e.g., for protists[11]. The great advantage of this GBIF subproject is the availability of voucher images, despite the currently still small amount of entries.

Many observation records used in the models for this study were gathered from the Hustedt Diatom Collection. The collection contains hundreds of samples from the Southern Ocean and the adjacent ocean basins, typically as permanent slides for light microscopy. In contrast to the use of observation data from the data repositories, voucher images can be made available for each observation. Deposited in the Hustedt Collections online database, these voucher images are accessible for further research, too. It is frequently suggested that all specimens sampled and used for biodiversity studies should be kept and made accessible as primary data in collections, as it is common in taxonomy and paleontology (*Schilthuizen et al.*, 2015).

With presence-only data, it does not matter for the models how many observations exist in a single grid cell, as, e.g., a sample with only a single diatom valve has the same value as a sample with thousands of valves. Thus, the valve density on the slides for light microscopy is not taken into account in this study. On some of the slides from the northern part of the Southern Ocean only a few valves were found, and just one valve in an extreme case. Whereas in other samples, especially in those from the core distribution area, sometimes up to hundreds of valves were found on the slides. This density information, though not usable as input data for this kind of distribution models, might be useful as a control of the model output.

In the Hustedt Collection, even remote regions like the Southern Ocean are spatially well covered with samples. Nevertheless, a temporal bias is still existing as the majority of the entries are from the summer season. Samples from the winter are quite rare in this region and even more important, entries from heavily sea-ice-covered sites are rare.

[11] http://protists.gbif.de/protists/

But even in regions that are well covered with sampling sites, poor species detectability can have an impact on data quality. *Monk* (2014) discusses this important aspect that is often neglected in studies, as a biased dataset due to low detection rates violates assumptions for SDM. *Cermeno et al.* (2014) showed that species richness in phytoplankton samples might often be strongly underestimated. They could double the number of detected species by a 10-fold increase in the sample volume. In an experiment with synthetic communities, *Rodriguez-Ramos et al.* (2013) found a 20-45% fraction of missing species in small volume samples. For field samples, they found a 1.5-fold increase in species numbers with an increased sampling effort. Most sampling campaigns probably suffer from this. Hence, most data repositories do, too. This just as well affects samples from the Hustedt Collection used for this study. To a certain degree, this issue can be countered by selecting an appropriate grid cell size in the environmental variables and adjusting the prevalence settings. Prevalence, the proportion of sampling sites (in terms of grid cells) where a species was observed, has a strong influence on the model's predictive power (*Santika*, 2011).

If images are available, taxonomic classification of the samples is under own control and responsibility, in contrast to entries from the data repositories, where, in the worst case, it is even unclear who is responsible for the classification and which taxonomy was used. About 150 different pelagic diatom species are known to occur in the Southern Ocean, with many of them indistinguishable by light microscopy. For a number of key taxa, identification up to the species level was possible, e.g., for the genera *Fragilariopsis* and *Asteromphalus*. We should, however, also bear in mind that taxonomic classification of diatoms, especially on the basis of light microscopy, is tough and error-prone.

In a few regions, especially in the north Atlantic and the northwestern Pacific, an unexpectedly high number of observation records are available for some of the taxa, e.g., for *Asteromphalus heptactis* and *A. hookeri* (see also maps in figures 3.19 A-D), *Corethron pennatum* (maps in figure 3.20 A+B) and *Dactyliosolen antarcticus* (map in figure 3.20 C+D). Most of these entries stem from plankton recorders. Such an accumulation of observation records in just one of the data sources might indicate problems in data quality, either due to misidentification or also under-sampling in the other resources.

Aside from the number of observations, their distribution is just as important. Most obvious is the spatial distribution of sampling sites. In case of *F. kerguelensis*, observation records from the Pacific sector of the Southern Ocean were strongly underrepresented in publicly available data. Compensating this obvious sampling bias by adding data from three transects in that area did not change the model output significantly. Later it turned out that a more subtle sampling bias existed towards the northern regions, which had a strong impact on the models future projections (*Pinkernell and Beszteri*, 2014). Further, the observation data can be biased in other dimensions. Most prominent in this study is the bias towards summer months and in missing observations from heavily covered sea ice regions. This less obvious bias has a huge impact on the calculation of the environmental space, which is used by the models. In some cases, existing observation data turn out to be unusable for distribution models, e.g., when a record in the repository lacks metadata. In GBIF and OBIS, most of these metadata are well-tended, but missing entries, especially missing sampling dates, led to some discarded observation records.

The need for (meta-) data standards, software, and work-flows for biodiversity research is beyond question and led to several projects working in this field. *Vos et al.* (2014) list several projects currently under development here. A brief summary of recommendations and requirements for the biodiversity informatics community is listed in *Hardisty et al.* (2013). The BioVeL[12] (Biodiversity Virtual e-Laboratory) project is also quite active, working on an IT environment for biodiversity science. In recent years, several workflows and tools were published, e.g., for biodiversity data management (*Mathew et al.*, 2014) and automated data curation in workflows (*Alper et al.*, 2013), but also publications about semantics and ontology of biological collections (*Walls et al.*, 2014), and environment ontology (*Buttigieg et al.*, 2013).

In conclusion, the number of observations matters for a good distribution model, but the representative distribution, especially in environmental space, is more important. For the Southern Ocean, this means that also the northern boundaries should be sampled, which seem to be frequently omitted in north-south sampling transects. A circumpolar sample coverage has less impact on model quality. Sea ice covered regions are massively underrepresented in the samples, forming a momentous bias. Generally spoken, a dataset with several north-south transects, covering different seasons and including regions affected by sea ice, would be ideal. Another aspect of observation data concerns quality and re-usability issues. Whereas public repositories provide observation data in high quantity, their quality level is not always clear. Thus, observation data documented by voucher images and annotated by meta-data shall be preferred if possible.

Environmental data

Increased use of remote sensing technologies allows worldwide observations of environmental variables. This is also true in the marine realm, where satellite data are supported by measured data of a huge fleet of autonomous buoys that can reach even remote regions such as the Southern Ocean. In recent years, many global ocean wide datasets became available and enabled new approaches such as the SDMs used here.

Bio-ORACLE (*Tyberghein et al.*, 2012) is a global marine dataset of 23 geophysical, biotic and climate variables in a spatial resolution of five arc-minutes. It was used in the first versions of the Southern Ocean diatom distribution models to get experience with promising predictors (data not shown). The Bio-ORACLE predictor-set compiles data of several years into one easy to use dataset. For some of the variables, just a mean value is given, e.g., mean pH or mean salinity, whereas others, e.g., sea surface temperature, are represented by four variables: minimum, mean, and maximum sea surface temperature and sea surface temperature range. This dataset proved its use in several marine biogeographical studies. Big advantages of this dataset are the high

[12] http://www.biovel.eu

spatial resolution and the high number of predictors, but unfortunately, Bio-ORACLE is not available at a monthly resolution.

As the Southern Ocean is subject to strong seasonal variation, data at a monthly resolution are used instead. Main environmental predictors, identified in previous models with Bio-ORACLE, are available in the World Ocean Atlas as monthly data. *F. kerguelensis* distribution models, also published in *Pinkernell and Beszteri* (2014), are based on a minimal set of four predictors: sea surface temperature and salinity, and silicate- and nitrate concentration. The models already showed seasonal distribution patterns and were also used for model projections on future environmental datasets to assess effects of climate change on this species' distribution. For this thesis, the models were extended by three predictors accounted to be important for Southern Ocean diatom biogeography: iron concentrations, mixed layer depth, and sea ice concentration.

In direct comparison, models based on monthly environmental data performed better than those based on yearly data (see discussion in chapter 4.1.2). It turned out that only a few predictors are necessary to predict the main distribution areas. Nitrate concentration on its own has the most predictive power but results in an overestimation of the distribution area. Complemented with silicate concentrations, water temperature and salinity good results are possible that even allow projection on future scenarios.

Some predictors show signs of correlation, e.g., in phosphate and nitrate concentrations in the Bio-ORACLE predictor set, and to a lesser degree also in sea surface temperature, salinity, silicate, and nitrate. It is regionally limited and not an ocean wide - and more importantly not even a Southern Ocean wide - phenomenon. This might lead to the phenomenon that one predictor can be exchanged by another one, e.g., nitrate by phosphate.

The predictors used in this study have a higher temporal resolution than the Bio-ORACLE dataset on the one hand, but also a much lower spatial resolution of just one degree on the other. Imperfect species detection, as mentioned before, is a strong issue in plankton observation data. A coarser spatial resolution is not a disadvantage, as it enhances the chance that several observations fall into a single grid cell and this way improves the chance to detect a species presence. The species biogeography and, at least for the open ocean, also the modeled habitat can be considered homogenous within a grid cell as well as within a whole region. Positions of the frontal systems, which are an important feature for orientation in that system, can also vary by several degrees.

The first four environmental predictors (*Pinkernell and Beszteri*, 2014), as well as the three additional ones, will be briefly discussed in the following in relation to the *F. kerguelensis* models.

Sea surface temperature was thought to be an important predictor for diatom distribution, supported by several studies (e.g., for Antarctic diatoms (*Fiala and Oriol*, 1990), for *Proboscia inermis* (*Boyd et al.*, 2013), and various phytoplankton groups (*Huertas et al.*, 2011)). *Thomas et al.* (2012) predict a poleward shift of phytoplankton of the low latitudes due to rising ocean surface temperatures based on mechanistic SDMs. However, for phytoplankton of the high latitudes, they expect a smaller impact, as their optimal temperatures typically are higher than the mean annual temperatures in that region. Indeed, in most of the models this predictor played an important role, e.g., in the *F. kerguelensis* models, it was the second most important predictor. Beside physiological requirements of the organisms, water temperature is also an important feature to distinguish the different water masses of the oceans, which is especially true for the Southern Ocean. As this parameter can be easily controlled in lab experiments, model runs were complemented by a series of eco-physiological experiments on temperature tolerance (see also discussion in chapter 4.2.3). Several studies on the effect of resource supply and ocean warming on phytoplankton productivity indicate that resource availability is more important than temperature, so warming of the ocean's surface might have a lesser impact than expected (*Marañón et al.*, 2012; *Maranon et al.*, 2014; *Peter and Sommer*, 2013).

Salinity on its own is not a good predictor; in the *F. kerguelensis* models its contribution is rather low, e.g., 2.5% in model 3. Used as the only predictor, it already resulted in a reasonable distribution pattern for *F. kerguelensis* in the Southern Ocean (see figure 3.4). Together with water temperature, salinity determines the density of seawater and therefore is an important property to characterize water masses in the ocean.

Silicate concentration was thought to have a strong influence on diatom biogeography, as it is needed to build their frustules. Some species in the Southern Ocean, e.g., *F. kerguelensis*, indeed are extremely thick shelled. In all models, silicate concentration has a lower influence than nitrate concentration. Silicate concentrations in the Southern Ocean decrease much stronger towards the North than nitrate concentrations, which is caused by silicate consumption by diatoms. This leads to very low silicate concentrations towards near the northern boundary of the ACC where most diatom species are still observed. Thus, the predictive power of silicate is relatively low, despite its importance for diatom growth. This fits well with the range of silicate concentrations *F. kerguelensis* requires, as published in *Jacques* (1983).

Nitrate concentration turned to be the most important predictor in most of the models. It is an important macronutrient, although not the only one. In the model runs using the Bio-ORACLE dataset, nitrate could be replaced by, e.g., phosphate. Nitrate concentrations in the surface waters of the Southern Ocean are decreasing towards the North. This distribution pattern makes it an ideal predictor in correlative species distribution models for the Southern Ocean.

Iron plays an import role for diatoms in the Southern Ocean, proved by several ocean fertilization experiments (*Smetacek et al.*, 2012; *De Baar et al.*, 2005; *van Creveld et al.*, 2016). However, exact iron measurements are costly and complex, and for the Southern Ocean, sampled areas and iron maps are patchy. For this reason, modeled iron data from the IPSL-CM5A model was used. This predictor was found useful, despite the poor quality of iron data, especially in the future predictions.

Sea ice concentration led to the most noticeable changes in the modeled distribution area compared to the models presented in *Pinkernell and Beszteri* (2014). Sea ice concentration data provided by satellites have good quality and are available since 1978/79. In *Pinkernell and Beszteri* (2014) it was claimed that further predictors had hardly any effect on the predicted distribution areas. On the basis of the models presented here, this statement has to be restricted to be valid just for the austral summer season. The new model versions reveal a strong influence on the predicted distribution area in sea-ice-covered regions (see discussion in chapter 4.2.4).

Including mixed layer depth did not have a big impact on the modeled distribution pattern. In the mixed layer, the upper part of the ocean, the density is nearly the same as on the surface, due to nearly identical physical properties such as temperature and salinity. This is the zone where the phytoplankton lives. Its depth has a strong impact on the average amount of light the phytoplankton are exposed to, and it is an important factor for phytoplankton blooms, especially in combination with light availability. Day length was not included as a parameter, though it could be calculated easily depending on latitude and sampling date. The weather also plays an important role, as, e.g., clouds and fog have a huge impact on the photosynthetically active radiation (PAR) that actually matters for the phytoplankton. Day length itself will not change in the future but PAR might. The samples are biased towards summer (more light) conditions and (at least in the southernmost sampling sites) cover a total range of 0-24 hours, so the explanatory power of this predictor is rather low.

Future scenarios for the end of the century are used for model projections to estimate consequences of a possibly changing environment for the modeled species distributions. They were developed for the fifth phase of the Coupled Model Intercomparison Project (CMIP5) (*Taylor et al.*, 2012). This project provides a series of coordinated climate change experiments, including simulations of the recent past for model evaluation, decadal runs, and long-term experiments. These simulations were also used for the fifth assessment report (AR5) of the Intergovernmental Panel on Climate Change (IPCC).

A set of four representative concentration pathways (RCP) is being used for long and near-term modeling experiments, replacing the earlier socio-economic and emission scenarios. They are named by the radiative forcing level for the year 2100 and include several factors, such as land use, emission of greenhouse gases and air pollutants. Developed independently by different work groups, they are based on published scenarios and shall give a plausible scenario for the future (*Moss et al.*, 2010; *van Vuuren et al.*, 2011).

With RCP4.5 and RCP8.5 two of these scenarios were selected for future model predictions in this study. The RCP4.5 scenario describes a pathway with stabilization without overshoot to 4.5 W/m^2 (~650 ppm CO_2 equivalent) after the year 2100. The RCP8.5 scenario describes a rising radiative forcing pathway leading to 8.5 W/m^2 (~1370 ppm CO_2 equivalent) by the year 2100. In the latter one, also called the "business as usual" scenario, the mean ocean temperature increased by +2.73° C (±0.72), and for the RCP4.5 scenario by 1.28° C (±0.56) from the 1990s to 2090s (*Bopp et al.*, 2013).

Unfortunately one of the predictors, mixed layer depth, was not available in the repositories for the HadGEM2-ES model, one of the GCMs used for the future projections. Models with and without this predictor are compared. For those models using MLD, the future predictions are based only on the remaining four of the five GCM outputs. Responses to MLD in the projections of the *F. kerguelensis* models were different for winter and summer. In the summer projections, the current distribution area did not significantly change when MLD was removed, but it decreased in the projection on future scenarios. This can be explained by changes in the stratification pattern in the future due to ocean warming, changing sea ice conditions and wind stress (*Petrou et al.*, 2016). In the winter, the distribution area strongly increased for the current projections, which shows the impact of this predictor. In the future projections, changes were only marginal, as the changed stratification pattern mostly affects summer season. On the one hand, the use of all available GCM makes the predictions more robust, but on the other hand, MLD has a positive impact on the distribution models predictive power, too. Finally, MLD was used as a predictor in the models for all of the other species, too. The impact on the model performance by using this predictor is considered higher than including modeled future environmental data of a fifth GCM.

Strong variations of modeled iron concentrations among the different GCMs used for the future scenarios turned out to be problematic. In the region of interest, the area between 40 and 70°S, the median of the iron concentrations reaches from 0.146 µmol l^{-1} in the IPSL model up to 0.829 µmol l^{-1} in the Hadley model. Within the IPSL model runs, future and current iron concentrations are similar. The projections on the future scenarios of the IPSL model match the ensemble projections of the models without iron. This has a huge impact on the projections on future scenarios (see figures 3.13 and 3.12), as including the iron predictor leads to a significantly decreased distribution area prediction. Similar to MLD, positive and negative effects of the iron predictor on overall model quality need to be compared. On the one hand, iron has a strong influence in the model, e.g., reaching 4.4% in *F. kerguelensis* model 3 (which is more than MLD, salinity, and silicate have). In contrast, strong variations in the data from future scenarios argue against using this predictor. Due to the use of MLD as a predictor, the Hadley models data cannot be used for future projections, so the most extreme variation is out (see figures 3.12 and 3.13). NorESM and MPI models still indicate higher iron concentrations than the IPSL and CESM model. Overall, the positive impact on the model performance using this predictor is considered higher than the increased uncertainty in future predictions caused by it.

In conclusion, the three additional predictors (sea ice concentration, mixed layer depth, and iron concentration) broaden the model's predictive power and can help to map the species distribution patterns better. Including sea ice concentration helps to visualize regions that are affected by a sampling bias, and otherwise, would lead to an unexplainable southern distribution boundary. Iron and MLD, both essential for diatoms, have a positive impact on the model quality. They, however, have just a small effect on the current distribution patterns, but a bigger impact on the modeled future distributions. With the RCP4.5 data, a scenario with moderate changes in the environmental conditions was chosen which only had a moderate impact on the model projections. The RCP8.5 scenario, in contrast, is the most extreme one of the four RCP scenarios with a much stronger impact on the model projections.

4.1.2 Model response and evaluation

The majority of the data records used for the models in this study are of the type presence-only data. All available data - even if of a higher quality level - were transformed into this type. This limited the choice of modeling methods.

A comparison of several different modeling approaches using the openModeller framework resulted in a similar distribution pattern among all models, forming a belt-shaped area around the Antarctic continent (see figure 3.1). In the northern regions of this belt, as well as in the Weddell and Ross seas, not all models agreed well about the distribution boundaries. These sensitive regions were identified to be important for further model improvements. First, additional predictors might lead to a better model performance in that area. Second, as these regions are not covered well in the public data repositories, further observation records from samples in the Hustedt diatom collection shall support the model's predictive power to reduce prediction uncertainties.

For further investigation, Maxent was chosen, as it is considered as one of the best methods for presence-only data. An absolute quantification, however, is not possible with these models (and data). They are not intended for use in, e.g., biogeochemical studies or for quantification of fluxes. Instead, these models are frequently used to study and map species biogeography and the impact of climate change on it. *Brun et al.* (2015) used such models to characterize the realized ecological niches of 133 phytoplankton taxa in the open ocean. *Brun et al.* (2016) systematically analyzed the predictive power of species distribution models for plankton in a changing climate by comparing several metrics and different modeling approaches based on data from continuous plankton recorders. They concluded that without intense model assessment even powerful models and extensive datasets are not a guarantee for reliable and robust climate change projections.

The quantity a species distribution model is able to predict is determined by the type of observation data, possible sampling bias, as well as the degree of imperfect species detection (*Guillera-Arroita et al.*, 2015). Presence-background data can only estimate a ranking or the relative likelihood, not the probability of occurrence, for which presence-absence data with a detection probability of one at presence sites would be necessary. Ranking means that regions that fit the needs of a species better than other ones get a higher value than those regions where, according to the model, species are less likely to occur. According to their schema, the SDM used for this study can predict a ranking. As the probability of detecting a species at a site is certainly smaller than one and not constant, and a sampling bias cannot be precluded, the interpretation as a relative likelihood is ruled out. The relative likelihood would result in a response proportional to the probability of occurrence, whereas ranking does not. Ranking instead leads to a sharper edge between regions with high and low estimated quantity, which might result in an underestimation of presence in regions with a low model output. This fits well to the low threshold of 0.2 used for this study.

A threshold is necessary to measure the predicted distribution area and helpful when current and future distributions are plotted on the same map for comparison (see fig. 3.16 to 3.20). The selection of the threshold value is arbitrary and has a strong impact on the measured values. Thus in most of the plots, the use of a threshold was avoided in favor of the original model output. In case of the area measurements and figures 3.16 to 3.20 a threshold of 0.2 was used for current and future distributions. On a global scale, the gradients at the distribution boundaries are quite steep in most of the models and regions. A small change in the threshold, however, can move the distribution boundary by a few degrees in latitude (see fig. 3.15), which equals several hundred kilometers. This affects both, the current and the future projections, so the trend will remain. These differences are in the same range of (interannual) variations of latitudinal frontal positions. The selected values (with a few exceptions) fit the distribution of observation records and could also be confirmed by independent observation data from a recent cruise (see chapter 4.2.2).

F. kerguelensis models

With more than 500 observation records, the *F. kerguelensis* models are based on a decent base of observation data. Several hundred model variants for this species were calculated and compared, of which of course only a few can be presented in this study. They revealed that the obvious spatial bias towards the Atlantic sector of the Southern Ocean has a negligible effect in the models environmental space. In contrast, the low number of observation records in public repositories in the belt north of the ACC was found to have an effect in the environmental space (see figure 3.2). In the context of presence-only data, this can be considered as a hidden bias, which had a strong impact on the predicted distribution areas in the future projections. Systematic assessment of samples from the Hustedt collection could close this gap in the observation data, especially in comparison to earlier models as described in *Pinkernell and Beszteri* (2014).

The improved observation dataset, at least locally, led to more credible model predictions, e.g., in the Weddell Sea. The anticipated model improvement in terms of clear distribution boundaries, however, could not be observed by just adding additional observation data. It is especially noticeable that several of the added observation records are outside the predicted distribution range (in models without and with the additional data).

The model quality is characterized by a good tradeoff between a close fit to the data and a high generalizing capacity. Validation is necessary to test how a model generalizes to an independent new dataset. It also can reveal a problematic data situation, e.g., due to bias.

Tests with independent observation data are a promising approach for model evaluation, which due to missing data often are not possible. The data used for this study can be separated into two independent groups: those from public data repositories (GBIF, OBIS, and GDD) and those from the Hustedt Diatom Collection. In fact, several entries in the Hustedt Collection were already included in the public databases, but the majority was not. Models, trained with one set and tested with the other, revealed strong discrepancies due to missing observation data at the northern edge of the ACC, especially in the public databases. Due to aggregation of observation records from the various data sources, this latitudinal bias could be eliminated. The Hustedt collection was checked for samples from locations far north of the predicted northern distribution boundary.

In case independent data are not available, the existent data can be separated into a test and a training dataset for which several strategies exist. In k-fold cross validation *k* bins of equal sample size are formed, and the models are iteratively build using *k-1* bins with the remaining one held for evaluation. This is repeated *k* times so that each bin is used for evaluation once. A disadvantage of this method is that the bias might remain in the calibration and evaluation dataset. Typically, but not necessarily, the background data is sampled from the entire study area.

It is obvious that the observation records are not distributed evenly over the ocean basins (see figure 3.2). To test the impact on model quality, k-fold cross-validation runs were used, where the observation records were separated into several bins depending on their longitude (in twelve steps of 30°) by data partitioning as described in *Muscarella et al.* (2014). Smaller steps in longitude led to worse test results, as not all bins contained data over the whole range of latitude anymore. Other tests confirmed that observations along North-South transects are important for good model quality.

Null models are often suggested to be used for model assessment, e.g., to detect spatial auto-correlation or sampling biases (*Raes and ter Steege*, 2007; *Merckx et al.*, 2011). In null models, occurrence sites are randomly selected from the spatial area of the species. Repeated several times, AUC values - or other metrics - are used to compare if these random models are better than the real SDM. It has already been shown that circumpolar distribution of observation data has only a limited effect on the model, whereas a distribution along north-south transect has, thus null-models, in this case, might have only a limited explanatory power.

Further, three additional predictors were included in this study. Their impact on the model performance was discussed already in section 4.1.1 on page 70 ff. at the example of various *F. kerguelensis* models. Including the sea ice predictor revealed the consequences of the bias in the observation data toward the summer.

Overall, the models seem to be of a decent quality, despite the bias in observation data towards summer season. The, of course slightly different, model projections of most of the models agree well. The model's response curves, indicating the model's internal behavior, are also of reasonable shape. For most predictors, the shape of the response curve of the single predictor model and full model are equal. This is not the case for silicate, which in this case might indicate a correlation with another predictor.

It is important to avoid projections on regions outside the range of the training conditions. Maxent's built-in function to mark these regions (so-called clamping) did not indicate any cases.

Several observation records are located far outside the predicted distribution area (see figure 3.2 C+D). Seasonal variations do not explain this, as these regions are not covered in any month. Compared to the total number of 2954 observations (table 2.1), which resulted in 712 distinct occupied grid cells and spread over 12 months, it is still a rather small number. The related observation records stem from different sources, including the Hustedt collection (e.g., on slide Hasle17-74 from October 1964 at 52.8°S, 38.7°W). As mentioned before on some of the slides only a small number of valves was found, and only one valve in a few extreme cases. Due to the presence only case such a sample on the one hand counts as much as a sample with hundreds of valves on it. But on the other hand, due to the small fraction and the fact that they are from a different environmental space, they appear as outliers in the Maxent model. Overall, it is quite likely that these regions do not belong to the core distribution area, especially for observations located far away from the species main distribution area. It has to be noted that the observation records were collected and aggregated over a period of several decades. A higher fraction of records outside the currently modeled distribution area certainly would have an impact on the model. Even though not included in the predicted distribution area, observations are real. Misidentification or problems in the record's metadata might have led to some of these dubious records though. Contamination, though certainly possible, can be considered unlikely in this high number of cases. Most probably the valves were transported northward by currents or eddies. As the environmental conditions are outside the suitable range, it is even likely that at the point of sampling the majority of cells were already dead.

Projections of the full monthly model (model 3) on a yearly averaged dataset were compared to a yearly projection of a model that was built on yearly minimum, maximum and mean values (model 6). Both models resulted in a similar spatial distribution, but with a stronger pronounced northern boundary for model 3, which also resulted in a less patchy distribution area. The yearly projection of model 3 is comparable to the projections of this model on (Austral) summer conditions (December to March in fig. 3.7).

On the one hand, yearly models result in plausible distribution predictions, comparable to a maximal distribution range. Information about the maximal range of a predictor seems suitable, and as an advantage, also include information that would be missing in temporally biased observation records. E.g., in a yearly model an observation record with information about the minimal and maximal temperature at a site in the course of a year is equivalent to at least two observation records in a monthly model - one from the warmest and one from the coldest month. On the other hand, this might lead to an underestimation of seasonality. The yearly projection of the *F. kerguelensis* model in figure 3.10 includes the regions that are heavily influenced by sea ice in the winter, whereas monthly models clearly indicate these regions as not suitable in the winter. In summary, monthly models are preferred, as seasonality effects are included. For partly biased data, these patterns might also be interpreted as limits of the model's scope (see discussion in chapter 4.2.4).

A first major aspect of this work was the evaluation of species distribution models for pelagic diatoms in the Southern Ocean. The species *F. kerguelensis* was selected because it is highly abundant in the Southern Ocean, relatively easy to recognize by light microscopy, and was also relatively well studied before. In conclusion, the model strongly benefits from the additional predictors and the additional observation data. The new model results in better spatial predictions, but also revealed a sampling bias towards the summer month and ice-free regions. It is not possible with the available data to reduce the impact of this bias, but being aware of it certainly helps to interpret the model predictions better. The updated spatial prediction of *F. kerguelensis* is discussed in chapter 4.2.1.

Further models

Distribution models for further 20 pelagic diatom species from the Southern Ocean were build based on the experience gained with the *F. kerguelensis* models. The basic setup was similar, as the same modeling method (Maxent) was used with the same predictors.

The main difference between these models is the varying data situation, ranging from only five distinct observation records for *F. vanheurkii* up to 3950 for *C. pennatum.* Models based on an only low number of observations are certainly of lesser explanatory power, and the distribution of the observation records has an even bigger impact on the model predictions. Two models for rare species, each based on less than 20 records, are the one for *F. linearis* (fig. 3.16E, 14 records) and *F. separanda* (fig. 3.18A, 19 records), which result in completely different distribution maps (see also discussion in section 4.2.1).

The models were built based on the experience from the *F. kerguelensis* models, and with the model settings found to be best for model 3. Models for abundant as well as rare species were built, so prevalence is certainly different, yet still unknown, for each of the species. For practical reasons, the default value of 0.5 was kept for all models presented here. Further, a threshold of 0.2 was defined and used for the measurement of the current and future distribution area in all of the models.

Another difference compared to the *F. kerguelensis* models are the species that are not endemic to the Southern Ocean but bipolar or even cosmopolitan instead. These models were built on the same predictor set and behaved similar to the models of endemic species. The resulting distribution patterns and ways to distinguish them are discussed in section 4.2.4.

Overall, quite reasonable distribution models could be built for most of the species, and evaluation runs using cross-validation did not indicate further issues. Nitrate is the most important variable (median: 34.7%), followed by SST (17.2%), silicate (11.65%), iron (7.85%), MLD (5%), SIC (4.9%), and salinity (1.7%). As an exception, the *F. cylindrus* model shows a high contribution of SST (60.2%) and SIC (24.2%). It is noticeable that in most models either nitrate or silicate plays the major role. This also indicates the correlation of these two predictors in at least some of the regions. An exception is the *F. ritscheri* model were nitrate explains 29.7% and silicate 29%.

The investigated species are closely related, and it is assumed that all of them are influenced by the same environmental factors that are used as predictors in this study. Two third of the species are considered as endemic to the Southern Ocean, just like *F. kerguelensis* is. The models, however, respond different on the predictors, visible, e.g., in the relative predictor contributions (figure 3.21, even in cases of similar distribution of observation records. The models resulted in plausible distribution projections, except for a few ones based on only a limited number of observation records, and model tests and evaluations did not indicate any further issues. So overall, experience gained with the *F. kerguelensis* models can be transferred on models for other pelagic diatoms.

4.2 Biogeography and ecology

The often quoted hypothesis that 'everything is everywhere, but, the environment selects' (*Baas-Becking*, 1934; *De Wit and Bouvier*, 2006) originally from the microbiology, is often stressed in the context of biogeographical studies for small species. All ocean regions indeed are connected and thus at least potentially allow a global dispersal of planktonic organisms. *Cermeno and Falkowski* (2009) analyzed fossil diatom assemblages from the world's oceans covering approximately the last 1.5 million years to separate effects of environmental selection and dispersal and showed that they are not limited by dispersal. Their results indicate that diatom community structure is influenced by environmental selection rather than by dispersal. The analysis of community compositions over wide latitudinal transect confirmed this (*Cermeno et al.*, 2010), leading to the conclusion that community structure is dramatically altered by changes in habitat conditions. Endemism in diatoms was frequently discussed, e.g. in *Vanormelingen et al.* (2008) contrasting two views in biogeography: the ubiquitous theory and species-specific endemism in Antarctic freshwater diatoms, further in *Mann and Droop* (1996) focusing especially on unidentified endemism due to coarse species classification, and in *Vyverman et al.* (2010) about several algal groups including diatoms. There is little doubt that endemism in diatoms exists. On the other hand, specific distribution areas for diatoms are rarely published.

Another important aspect, especially of marine pelagic diatoms is their typical bloom and bust behavior by which they are able to out-compete other species as soon as a phase of nutrient limitation finishes. Typical bloom and bust behavior is not of interest in this study, though, the diatom species analyzed by the models, of course, are subject to this. The models predict the so-called realized ecological niche of the species. Within this spatial area, blooms are possible but do not necessarily occur everywhere.

All models are projected on a global scale and can be grouped into bipolar, endemic to southern ocean and cosmopolitan. This study focuses especially on the Southern Ocean. As mentioned, most models show a similar distribution pattern by forming a belt-shaped area around the Antarctic continent. The species northern predicted boundaries vary strongly, whereas the southern boundary in most cases is defined by the sea ice zone. In the following, the species-specific distribution areas are discussed and compared to previous knowledge. Next, these findings will be set in relation to observations from expeditions in the Sothern Ocean and with observations from an experiment about temperature tolerance limits. Observed (global) patterns will be discussed, as well as the fate of the species regarding model projections on future environmental scenarios.

4.2.1 Diatom biogeography

In this section previous knowledge about diatom distribution is combined with new insights from the distribution models. Observation data from public repositories such as GBIF and OBIS is complemented with the findings of several studies about the distribution of the taxa of interest in plankton (*Cefarelli et al.*, 2010; *Hasle*, 1965, 1968, 1969, 1976), and sediment (*Armand et al.*, 2005, 2008; *Crosta et al.*, 2005; *Esper et al.*, 2010; *Zielinski and Gersonde*, 1997). This also includes studies about diatom biogeography (*Olguín et al.*, 2006; *Olguín and A. Alder*, 2011; *Olguín Salinas et al.*, 2015), and a book (*Semina*, 2003).

Most models predict a belt shaped distribution around the Antarctic continent with a southern distribution boundary related to the sea ice edge (see discussion in chapter 4.2.4). This is an improvement to previous models, as presented in *Pinkernell and Beszteri* (2014), where no clear boundary could be modeled. Unless stated otherwise, the southern distribution boundary in the models follows the sea ice edge, or the Antarctic continent during Austral summer respectively.

***Fragilariopsis* Hustedt, 1913** For the genus *Fragilariopsis* Hustedt, Algaebase currently lists 25 taxonomically accepted species names. Holotype species is *Fragilariopsis antarctica* (Castracane) Hustedt, a synonym for *Fragilariopsis kerguelensis* (O'Meara) Hustedt. Genus *Fragilariopsis* is marine and distributed worldwide. *Cefarelli et al.* (2010) analyzed species composition and abundance of phytoplankton samples from a transect covering the Argentine Sea, the Drake Passage, and the Weddell Sea. They focused on twelve *Fragilariopsis* species using light and electron microscopy. Though not fully consistent with the taxonomic classification on Algaebase, this paper is used as a reference for taxonomic classification throughout this thesis. *Hasle* (1965) analyzed species of the genus *Fragilariopsis* - mainly based on samples from the Brategg-Expedition - by light and electron microscopy, also summarizing information about their taxonomy and distribution.

***Fragilariopsis curta* (van Heurck) Hustedt, 1958** *Cefarelli et al.* (2010) report *F. curta* to be the most frequent *Fragilariopsis* species in their study area. It is present in the Argentine Sea, the Drake Passage, and the Weddell Sea. In the latter, it is the species with the highest relative abundance. It occurred at water temperatures from -1.6 to 13.35 °C and salinities from 33.10 to 34.24.

Based on sediment samples *Zielinski and Gersonde* (1997) in contrast report *F. curta* to be restricted to areas south of the Polar Front, where surface water does not exceed a temperature of 2 °C. Its distribution is linked to the presence of sea ice. The northern distribution boundary also marks the location of the winter sea ice edge. In contrast to Cefarelli, they report a temperature range of -2 to 2 °C. *Olguín and A. Alder* (2011) lists this species as sea ice related. *Semina* (2003) classified it to be ice-neritic in the high-antarctic region.

In total, 1106 observation records were gathered for *F. curta* (see table 2.1). Except for an observation record at the equator at the null-meridian (0°, 0°), which most probably is a data artifact, all observations are from the Southern Ocean, with the northernmost record at 44°S. The species is well covered by north-south transects, which seems a good quality criterion.

Only 180 of the 1106 presence records were used in the Maxent model, as just one presence record per grid cell and month is accounted for model training. These locations span a temperature range from -1.75 °C to 12.6 °C, and a salinity range of 32.9 to 34.4, which both are reflected by the model's response curves. The northern boundary roughly follows the Subantarctic Front and not the Polar Front as mentioned in *Zielinski and Gersonde* (1997).

***Fragilariopsis cylindrus* (Grunow ex Cleve) Frenguelli, 1958** According to *Cefarelli et al.* (2010), *F. nana* and *F. cylindrus* are not distinguishable by light microscopy. However, in their listing of morphometric data for *Fragilariopsis* spp. (page 1468, table 3), they report *F. nana* to be narrower. The transapical axis of *F. cylindrus* is 2.4 - 4µm and if *F. nana* 1.4 - 2.4µm. As not separated in their station list, the observation records from the paper are neglected for the distribution models. In case of the Hustedt data, *F. nana* and *F. cylindrus* are separated by their transapical length where possible.

Cefarelli et al. (2010) frequently found *F. cylindrus / F. nana* in the Drake Passage and the Weddell Sea, with high relative abundances in the latter region. They report a temperature range from -1.6 to 6.22 °C, and a salinity range from 33.10 to 34.24. According to *Lundholm and Hasle* (2008), *F. cylindrus* (maybe including *F. nana*) is a marine planktonic and sea ice species. It is present at north and south hemispheres. *Olguín and A. Alder* (2011) lists this species as sea ice related. *Semina* (2003) classified *F. curta* as bipolar and panthalassic.

F. cylindrus is considered as bipolar, which is also reflected in its pattern of presence records. In total 1542 presence records were gathered (see table 2.1), from which 132 were used for model training. Several suspicious observations were used for model training but are not covered by the models predicted distribution, e.g., a transect in the Argentine Sea up to 41°S, an observation in the South Pacific at 43°S, and observations south of Australia up to 47°S. In these cases, misidentification with, e.g., *F. nana* seems possible. Further suspicious observations were neglected for model training, as they fell in grid cells with missing environmental data, e.g., observations near Australia at 36°S, as well as in the eastern Mediterranean Sea. The temperature at observation sites ranges from -1.8 °C to 12.1 °C and it differs from the range mentioned by *Cefarelli et al.* (2010). Unlike in most of the other distribution models, sea surface temperature had a strong relative influence here reaching 60.2%, followed by sea ice concentration 24.2%. Nitrate, in this case, reached only 3.2%.

This species is known to be sea ice related. It has to be noted that for *F. cylindrus* 45% of the observation records are from regions with a sea ice concentration $\geq$ 15%, and many from regions with massive sea ice cover. Though the sampling bias towards ice-free regions seems less pronounced in regard to the observation records, it still has an impact on the predicted distribution in ice-covered regions.

Fragilariopsis kerguelensis **(O'Meara) Hustedt, 1952** *Cefarelli et al.* (2010) report *F. kerguelensis* to be found in each of their three studied areas, with the northernmost observation at 46.4°S in the Argentine Sea. Its highest frequency and relative abundance are reached in the Drake Passage. They found *F. kerguelensis* at water temperatures from -1.33 to 14.06 °C and salinities from 33.17 to 34.19.

Zielinski and Gersonde (1997) classify *F. kerguelensis* as an open ocean species, dominating the pelagic areas between the ACC and the winter sea ice edge. In this region, it is the main contributor to the Southern Ocean diatom ooze belt, a belt of well preserved diatom frustules in the surface sediments of the Southern Ocean, and can reach abundances up to 90% of the sediment diatom assemblages. North of the Subtropical front they report *F. kerguelensis* to decrease to less than 20% of the assemblages. The Weddell Sea and the Argentine Basin are mentioned as areas of lower abundance, the latter influenced by the input of neritic diatoms from the waters around the Falkland Plateau. *Zielinski and Gersonde* (1997) plotted abundances vs. surface water temperatures and report a temperature range from -1 to 18 °C, with a significant drop in abundance at temperatures above 13.5 °C.

Crosta et al. (2005) also correlated abundances with sea surface temperatures (February). They report a range from 0 to 20 °C, with greatest abundances between 1 and 8 °C. Further, they also found *F. kerguelensis* in sediment traps, which were covered by sea ice for up to eight months. The northern boundary of the distribution is reported to be the Subtropical front (*Semina*, 2003). *Hasle* (1976) located the northern distribution boundary of *Nitzschia kerguelensis* (=*F. kergulensis*) at "approximately 40 to 56°S, with the most frequent occurrence of the species in the open northerly waters". *Hart* (1942) already reported *F. kerguelensis* as the most abundant diatom in the Antarctic Seas. Other sources indicate a much broader distribution area: e.g., occurrence in surface water between 65°S and 30°N (*Van der Spoel et al.*, 1973), or records as far north as the Cape Verde Islands (*Heiden and Kolbe*, 1928). *Semina* (2003) classified this species as notal-antarctic and panthalassic.

In total, 2954 presence records were gathered for *F. kerguelensis* (see table 2.1), of which 576 were used for model training (model 3). Water temperature range from -1.8 °C to 16.5 °C. Six outliers are located in the Pacific at latitudes between 32°S and 10°S with significantly higher water temperatures from 22.7 °C to 29.2 °C. The median water temperature over all samples is 1.5 °C, with a skewed histogram towards higher temperatures and a strong drop at 5 °C. The salinity ranges from 32.7 to 35.7. In the model, nitrate was the most important predictor, reaching a relative contribution of 68.8%. Silicate, which was expected to have a strong influence due to the strong silicate frustules this species builds, just reached 0.8%.

The northern distribution boundary is predicted between the Subantarctic Front and the Subtropical front. In contrast to previous models (*Pinkernell and Beszteri*, 2014), the predicted latitudinal changes of the northern boundary of model 3 are much weaker in the course of the year.

F. kerguelensis might survive in ice-covered surface water, so the southern boundary might be the Antarctic continent instead of the sea ice edge.

***Fragilariopsis linearis* (Castracane) Frenguelli, 1943** According to *Semina* (2003), *F. linearis* is found only in the high latitudes and is designated as ice related. She classified it as high-antarctic and ice-neritic. Just 14 observation records of this species could be gained, of which 13 were used in the Maxent model. It is a very rare species that typically appears in low abundances. Most of the observations are from the western Ross Sea, collected in January 2006 and February/ March 2008. In this model, silicate was the most important predictor (49.5%), followed by SST (32%), SIC (14.7%) and salinity (3.8%). The remaining three predictors did not contribute to the models (0% each). Due to the small number of observations, the predictive power is considered as rather low.

***Fragilariopsis nana* (Steemann Nielsen) Paasche, 1961** *Cefarelli et al.* (2010) didn't separate *F. nana* from *F. cylindrus* by light microscopy (see the chapter about *F. cylindrus*). Further, in Algaebase *F. nana* is wrongly marked as a not valid taxon and listed as a synonym for *F. pseudonana* (Hasle) Hasle, and in GBIF as a synonym of *Fragilariopsis cylindrus* (Grunow) Krieger 1954. *Lundholm and Hasle* (2008) and *Cefarelli et al.* (2010) treat them as two distinct species, following *Hasle* (1965), who separated the elliptical-lanceotate morphotype as *F. pseudonana* (see also *Lundholm and Hasle* (2008)).

In total, just 55 presence records could be gained for this species (see table 2.1), of which 46 were used for model training. The temperature range at presence sites was from -1.7 °C to 9.8 °C, and the salinity range was 32.9 to 34.4. The model predicts a distribution with the Polar front as the northern boundary.

***Fragilariopsis obliquecostata* (van Heurck) Heiden, 1928** *Cefarelli et al.* (2010) found *F. obliquecostata* to be frequently abundant in the Weddell Sea, especially in the northern part, and at one station in the Drake Passage. The water temperatures of their observations were between -1.6 and 3.41 °C, and salinity between 33.10 and 34.24. They cite *Hasle* (1965), reporting *F. obliquecostata* to be present in samples from sea ice, and mention that this species was found at temperatures up to 18 °C in other studies. *Olguín and A. Alder* (2011) also lists this species as sea ice related. GBIF lists several entries around South Georgia Island and in the Indian sector of the Southern Ocean.

In total, 547 presence records were gained for this species (see table 2.1), of which 76 were used for model training. The temperature at the observation sites ranged from -1.7 °C to 4.3 °C, and the salinity from 32.9 to 34.4.

***Fragilariopsis pseudonana* (Hasle) Hasle, 1993** *Cefarelli et al.* (2010) found *F. pseudonana* in all three of their study areas: the Weddell Sea, the Drake Passage, and the Argentine Sea. They report a temperature range from -1.33 to 14.06 °C and salinities from 33.33 to 34.24. In *Tomas* (1997), this species is listed as cosmopolitan, and *Hasle* (1965) reports a "continuous distribution from arctic to antarctic waters". Further, she mentions this species to avoid coastal waters in the high latitudes.

In contrast, the observation records found for this species indicate a bipolar distribution (see also observations in figure 3.17 C+D). In total, 275 observation records were gained (see table 2.1), of which 59 were used for model training. The temperature at presence sites ranged from -1.7 °C to 13.6 °C, and the salinity from 32.5 to 35.2.

***Fragilariopsis rhombica* (O'Meara) Hustedt, 1952** In WoRMS and Algaebase, *Fragilariopsis rhombica* (O'Meara) is listed as a synonym of *Diatoma rhombica* O'Meara. *Cefarelli et al.* (2010) found this species in their complete study area, with high abundances in the Drake Passage and the Weddell Sea. They report water temperatures of -1.6 °C to 13.35 °C and a salinity range of 33.17 to 34.24 at observation sites. *Olguín and A. Alder* (2011) lists this species as sea ice related. *Semina* (2003) classified this species as notal-antarctic and panthalassic.

In total 483 observations were gained for this species, of which 115 were used for model training in Maxent. At observation sites, sea surface temperature ranged from -1.7 °C to 12.6 °C and salinity from 32.9 to 34.2.

All observation records are from the Southern Ocean, with a few sites further north up to 47°S. For the summer, the model predicts a belt around the Antarctic continent, with its northern boundary between the Polar front and the Subantarctic Front. During winter, this belt shrinks to a small band along the Subantarctic Front. Further, a weak signal in the north Pacific is present throughout the year.

***Fragilariopsis ritscheri* Hustedt, 1958** *F. ritscheri* is considered to be limited to southern cold water regions (*Tomas*, 1997). *Cefarelli et al.* (2010) found this species only in the Weddell Sea, at temperatures of -1.6 to -0.09 °C and a salinity range of 33.17 to 34.24. *Olguín et al.* (2006) report it to be present in spring phytoplankton in the Malvinas current, e.g., at a station at 37°15"S. *Olguín and A. Alder* (2011) lists this species as sea ice related. *Semina* (2003) classified this species as notal-antarctic and panthalassic.

In total, 39 records were gained for this species, of which 30 were used for model training in Maxent. At presence sites, sea surface temperature ranged from -1.4 °C to 3.8 °C, and salinity from 32.9 to 34.

All observation records are located in the Southern Ocean, which is reflected well by the model's summer projection. The northern boundary of the belt-shaped distribution is located between the Polar front and the Southern Antarctic Circumpolar Current Front (sACCf). During the winter, this belt disappears, except a few patchy spots, along with its northern summer distribution boundary. The strong decline in the south coincides well with sea ice cover.

***Fragilariopsis separanda* Hustedt, 1958** *F. separanda* is a rare species and considered to be limited to southern cold water regions (*Tomas*, 1997). *Cefarelli et al.* (2010) found this species in the Drake Passage and the Weddell Sea at water temperatures of -1.15 °C to 4.33 °C and a salinity range of 33.17 to 33.89. According to *Zielinski and Gersonde* (1997), this species is endemic to the Southern Ocean. They found it present in the sediment in the region between the Polar Front and the Permanent Open Ocean Zone (POOZ) (see also *Zielinski and Gersonde* (1997) and *Treguer et al.* (1995)) and related to surface temperatures of 0 °C to 4 °C.

In total, 23 records were gained for this species, of which 17 were used for model training in Maxent. At presence sites, sea surface temperature ranged from -1 °C to 6.2 °C, and salinity from 32.4 to 34.

All observations are from the Southern Ocean, except one south of Nova Scotia (Canada) at 43°N. The model predicts a belt in the Southern Ocean throughout the year, which is more pronounced during summer. The northern boundary follows the Subantarctic Front. This model also predicts occurrence in the north Pacific throughout the year, as well in the Arctic during (northern) summer.

***Fragilariopsis sublinearis* (Van Heurck) Heiden & Kolbe, 1943** *Hasle* (1965) summarized *F. sublinearis* as a neritic planktonic species and related to sea ice (*Jousé et al.*, 1962), and occurring "only in the immediate vicinity of dispersing pack-ice" (*Hart*, 1942). According to *Cefarelli et al.* (2010), this species can easily be confused with *F. obliquecostata.* They found it in the Drake Passage and in neritic and oceanic samples in the Weddell Sea. This species occurred only in low abundances, except for one open water station in the Weddell Sea. They report a temperature range of -1.6 °C to 4.33 °C and a salinity range of 33.1 to 34.24. *Semina* (2003) classified this species as notal-antarctic and ice-neritic.

In total, 396 records were gained for this species, of which 76 were used for model training in Maxent. At presence sites, sea surface temperature ranged from -1.7 °C to 12.8 °C, and salinity from 33.2 to 34.3.

All observations are from the Southern Ocean, except to one, located a bit more north at 43°S in the Pacific and at 44°S in the Atlantic. The predicted distribution covers all observation sites except for the latter two ones. The belt-shaped distribution has its northern boundary between the Polar Front and the Subantarctic Front. During winter, the belt is much less pronounced, with several gaps in the Indian Ocean sector of the Southern Ocean and south of Australia.

***Fragilariopsis vanheurckii* (Peragallo) Hustedt, 1958** *Hasle* (1965) reports the occurrence of *F. vanheurkii* close to the Antarctic continent inside the sea ice (brownish, under-surface sea ice). *Cefarelli et al.* (2010) found this species in the Weddell Sea at temperatures of -1.6 °C to -0.17 °C and a salinity range of 33.33 to 33.92. *Semina* (2003) classified this species as notal-antarctic and ice-neritic.

In total, just 5 records were gained for this species, of all were used for model training in Maxent. At presence sites, sea surface temperature ranged from -1.4 °C to 0.7 °C, and salinity from 33.7 to 33.8. The median sea ice concentration is 47%.

The observation records are from the Weddell Sea and from the sea ice. The model predicts a belt in the Southern Ocean, with a northern distribution boundary along the southern boundary of the ACC. The southern distribution boundary is the Antarctic continent. It is the only model that does not predict a retreat from sea-ice-covered regions. Of course, due to the small number of just five observation records, this model has just a limited explanatory power.

***Asteromphalus* Ehrenberg, 1844** For the genus *Asteromphalus* Ehrenberg, Algaebase currently lists 26 taxonomically accepted species names. The holotype is *Asteromphalus darwinii* Ehrenberg. Genus *Asteromphalus* is marine and distributed worldwide.

Hernández-Becerril (1991) analyzed *Asteromphalus'* morphology and taxonomy. He lists eight species to be cold water related but does not distinguish between northern and southern hemisphere. GBIF lists observations in the Southern Ocean for *A. hookeri*, *A. hyalinus*, and *A. parvulus*.

***Asteromphalus heptactis* (Brébisson) Ralfs, 1861** *A. heptactis* is distributed worldwide, with several observation records in the Southern Ocean. *Hernández-Becerril* (1991) and *Semina* (2003) classified this species as cosmopolitan. In AlgaeBase several observation sites are cited confirming a cosmopolitan distribution.

In total 3731 observation could be gained, of which 327 were used for training in Maxent. The temperature at presence sites ranges from -1 °C to 29.8 °C, with a median temperature of 20 °C. The salinity ranges from 27.5 to 38.1.

The model predicts a cosmopolitan distribution pattern with several huge gaps, e.g., in the Atlantic and the Pacific Ocean. The central Arctic is not covered. In the Southern Ocean, a belt around the Antarctic continent is visible throughout the year.

***Asteromphalus hookeri* Ehrenberg, 1944** According to GBIF, *A. hookeri* is distributed worldwide, with the majority of its observation records in the Southern Ocean. *Hernández-Becerril* (1991) mentioned (northern or southern) cold water regions for this species. *Semina* (2003) classified it as notal-antarctic.

In total 604 observation records could be gained for this species, of which 233 were used for training in Maxent. At the presence sites, the temperature ranges from -1.8 °C to 29.9 °C, with a mean temperature of 3.3 °C, and the salinity from 29.3 to 36.1.

Observation records are concentrated in the Southern Ocean and west Pacific, complemented by several observations at lower latitudes in all ocean basins in the southern hemisphere.

The model predicts the main distribution area in a belt around the Antarctic continent throughout the year, limited by the Subantarctic front in the north. Further regions that are covered throughout the year are the North Pacific, the region east of Canada, several regions in the Arctic, the eastern part of the Indian Ocean and the South China Sea.

***Asteromphalus hyalinus* Karsten, 1905** For *A. hyalinus*, GBIF and OBIS list observation records in the Southern Ocean. *Semina* (2003) classified this species as notal-antarctic and panthalassic. According to *Hernández-Becerril* (1991), it is limited to (northern or southern) cold water regions.

In total 499 observation records could be gained for this species, of which 68 were used for model training in Maxent. The temperature at the presence sites ranges from -1.7 °C to 6.9 °C, and the salinity from 32.9 to 34.3.

All records are from the Southern Ocean, and the model predicts a distribution range limited to this region. It shows a belt around the Antarctic continent throughout the year, which is less pronounced during the winter month. For the summer, the northern distribution boundary follows the Polar front.

***Asteromphalus parvulus* Karsten, 1905** For *A. parvulus*, GBIF and OBIS list observation records in the Southern Ocean. *Semina* (2003) classified this species as notal-antarctic. According to *Hernández-Becerril* (1991), it is found in (northern or southern) cold water regions.

In total 120 observation records could be gained for this species, of which 67 could be used for model training in Maxent. The temperature at the presence sites ranges from -1.5 °C to 6.7 °C, and the salinity from 33 to 34.4. All observations are from the Southern Ocean, except one, which is located in the Indian Ocean at a temperature of 29 °C.

Similar to *Asteromphalus hyalinus*, this model predicts a distribution area around the Antarctic continent. The northern boundary of that distribution cannot be clearly assigned to one of the main ocean fronts. Partially it follows the Polar front, in most regions it keeps south of it. In contrast to *A. hyalinus*, this model predicts a weak habitat suitability signal for *A. parvulus* in the north Pacific.

***Asteromphalus roperianus* (Greville) Ralfs, 1861** For *A. roperianus*, GBIF lists observations at the coast of Mexico and in the Atlantic Ocean. OBIS, in contrast, lists observations mainly in the Southern Ocean and at the coast of Tanzania. *Hernández-Becerril* (1991) lists this species as "world-wide warm-water and occasionally found in temperate regions". *Semina* (2003) classified this species as tropical and panthalassic.

In total 47 observation records could be gained for this species, of which 31 could be used for model training in Maxent. The temperature at the presence sites ranges from -1.8 °C to 22.2 °C, and a median temperature of -0.1 °C. The salinity ranges from 33.5 to 36.7. All observations are from the southern hemisphere, with the majority of the observations from the Southern Ocean. A few observations are from lower latitudes in all ocean basins.

The main predicted distribution area again is a belt around the Antarctic continent throughout the year. The distribution boundary in the north follows the Subantarctic Front. In the south, it is limited by the Antarctic continent or the sea ice edge in winter. Further, this model predicts *A. roperianus* to be present in the North Pacific throughout the year, but with a stronger intensity during (northern) winter.

***Azpeitia tabularis* (Grunow) G.Fryxell & P.A.Sims, 1986** For genus *Azpeitia* M. Peragallo 1912, Algaebase currently lists 11 taxonomically accepted species names.

Lectotype species is *Azpeitia temperi* M. Peragallo. The genus *Azpeitia* is marine and distributed worldwide.

The species *Azpeitia tabularis* is known to occur in southern cold water regions, especially in Subantarctic waters (*Tomas*, 1997). *Fryxell et al.* (1986) classified it as extremely cold-tolerant with a preference for the Southern Ocean, with high abundance in the Subantarctic zone and lesser abundance in the Antarctic zone (*Fenner et al.*, 1976). *Zielinski and Gersonde* (1997) located its southern distribution boundary at the winter sea ice edge based on sediment records. Further, they report highest concentrations between 10 °C and 20 °C, though it is also present at low water temperatures of 0 °C. *Semina* (2003) classified this species as notal-antarctic and panthalassic.

For *A. tabularis*, OBIS lists observations in the Southern Ocean, the Arctic, and the North Pacific. In contrast, records in GBIF indicate a cosmopolitan distribution. Algaebase further cites an observation in the Adriatic Sea (*Viličić et al.*, 2002).

In total, 373 observation records were gained (see table 2.1), of which 31 could be used for model training in Maxent. The observation records span a salinity range of 32.6 to 34.6 and a temperature range of -1 °C to 19.2 °C. Median temperature of all presence sites was 6.1 °C.

The model predicts the occurrence of *A. tabularis* in the North Pacific throughout the year. Both the High Arctic regions and the North Atlantic are only covered during summer. In the southern hemisphere, a belt around the Antarctic continent is predicted throughout the year. The northern distribution boundary follows the Subtropical Front in the Atlantic and Indian sector of the Southern Ocean, but not in the Pacific sector, where the boundary lies more northward between the Subantarctic and Subtropical Front.

***Corethron pennatum* (Grunow) Ostenfeld, 1909** For genus *Corethron* Castracane 1886, Algaebase currently lists 9 taxonomically accepted species names. The lectotype is *Corethron criophilum* Castracane (which is equal to *C. pennatum*). *Corethron* is reported to be marine and cosmopolitan. It occurs in high numbers, especially around the Antarctic continent. Species *Corethron pennatum* seems distributed worldwide. According to GBIF and OBIS, main areas are the Southern Ocean, the North Atlantic, and the North Pacific and the Philippine Sea. In contrast, *Crawford et al.* (1998) distinguish three *Corethron* species, of which *C. hystrix* was found in the north Atlantic and north Pacific, whereas *C. pennatum* and *C. inerme* were found in the Atlantic sector of the Southern Ocean. *Semina* (2003) classified *C. criophillum* (equal to *C. pennatum*) as cosmopolitan and panthalassic.

In total 13759 observation records could be gained for this species (see table 2.1), of which 1606 were used for training in Maxent. At observation sites, the salinity spans from 28.3 to 38.3, and the temperature from -1.8 °C to 29.6 °C, with a median of 10 °C.

The majority of the observations are located at latitudes of 50°S to 70°S and 40°N to 70°N, complemented by many observations at lower latitudes, often close to the coasts, as well as by several north-south transects in the western Pacific. It is quite likely that other species than *C. pennatum* are included in the dataset, as many database entries already existed before *Crawford et al.* (1998) clarified this genus.

Despite the worldwide distribution of observation records, the model projections show a bipolar spatial distribution pattern. Some coastal regions are covered as well, such as along north and south America, along with the African coast, etc. In the northern hemisphere it is predicted to occur throughout the year, but not in the high Arctic waters. In the southern hemisphere, a gap is visible in the Pacific sector, which is more pronounced during summer.

***Dactyliosolen antarcticus* Castracane, 1886** For genus *Dactyliosolen* Castracane 1886, Algaebase currently lists 6 taxonomically accepted species names. Holotype is *Dactyliosolen antarcticus* Castracane.

In GBIF, the marine species *D. antarcticus* Castracane has observation records for the North Atlantic, the Southern Ocean (mainly between Australia and the Antarctic continent), in the Philippine Sea, and along the coast of Mexico. OBIS shows a similar pattern of observation records. According to *Tomas* (1997), *D. antarcticus* has a cosmopolitan distribution and is especially important in the Southern Ocean. *Garrison* (1991) lists this species in relation to pack ice.

In total, 7961 observation records could be gained for this species (see table 2.1), of which 763 were used for training in Maxent. The majority of the observations are located in two belts in the northern and southern hemisphere, each at latitudes between 50° and 70°. The observations in the northern hemisphere are located mainly in the Atlantic Ocean; in the southern hemisphere, they are circumpolar but less frequent in the Pacific sector. Besides this bipolar pattern, several occurrences are located in the south Pacific: along a transect in the northwestern Pacific at 150°E, and along the east coast of the USA. The salinity at presence sites ranges from 32 to 35.7, and the temperature from -1.7 °C to 29.2 °C, with a median at 9.5 °C. In the histogram of the temperature values (figure 4.1), two peaks are visible between -1 °C and 4 °C, as well as between 9 °C and 14 °C. Further, the histogram is skewed to the right, up to 30 °C. The two peaks indicate that two or even more species might have been mixed in the occurrence records.

***Eucampia antarctica* (Castracane) Mangin, 1915** For genus *Eucampia* Ehrenberg 1839, Algaebase currently lists 5 taxonomically accepted species names. The holotype is *Eucampia zodiacus* Ehrenberg.

For species *E. antarctica*, GBIF lists several observation records in the Southern Ocean. In the Pacific, a few observations far north of the ACC are listed near the Cook Islands (9.9°S) and American Samoa (14.2°S). Further, the locations in the north Atlantic are listed. OBIS further lists observation records at the coast of Peru, South Africa, and Korea. *Semina* (2003) classified *Eucampia antarctica var. antarctica* as notal-antarctic and panthalassic.

In total, 750 observation records were gained for this species (see table 2.1), of which 176 were used for training in Maxent. The salinity ranges from 32.8 to 35.7. Most of

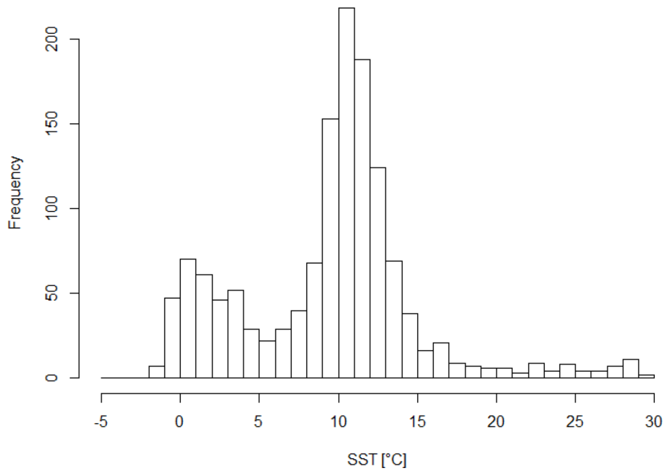

Figure 4.1: Histogram of SST at observation sites of *D. antarcticus*.

the observations are located at temperatures ranging from -1.8 °C to 6.6 °C, with eight outliers at temperatures up to 29.2 °C. The median temperature is 0.7 °C.

The model predicts a main distribution area in the Southern Ocean throughout the year, and in the northern hemisphere occurrence throughout the winter. Though a few observations are present in the northern hemisphere, this species is considered to be endemic to the Southern Ocean. The northern distribution boundary in the Southern Ocean is located between the Polar front and the Subantarctic Front.

4.2.2 Comparison with (recent) cruise data

The detailed models for *Fragilariopsis kerguelensis* revealed two regions with questionable observation data: along the northern distribution boundary and in relation to heavy sea ice coverage. The gaps in the first region could be filled by samples from the Hustedt collection, helping to find the northern distribution limit. While a spatial bias still can be observed regarding the circumpolar coverage, the sampling coverage over the latitudinal zones of the Southern Ocean is much better due to several north-south transects. These transects, however, barely cover massively ice-covered regions due to a strong bias towards Austral summer. Several winter campaigns were already carried out in the Southern Ocean (e.g., R/V Polarstern cruises ANT-V in 1986 and ANT-XXIX in 2013), but intense diatom samples have not found its way to the literature and observation repositories. The gaps are explained by missing samples from that regions, but it might still be possible that *F. kerguelensis* really avoids regions of massive ice coverage.

Bartsch (1989) investigated ice algae in the Weddell Sea, using data from the Winter Weddell Sea Project (Polarstern cruise ANT V/2+3 from July till December 1986) and cruise ANT III/3 (January and February 1985). She reports *Nitzschia cylindrus* (equivalent to *Fragilariopsis cylindrus*) to be one of the dominant ice species. *Nitzschia kerguelensis* (equivalent to *Fragilariopsis kerguelensis*), is frequently present between

70° S and the northern edge of the pack ice. She further cites *Rivkin and Voytek* (1987), who identified *F. kerguelensis* as one of the dominant species in water under the ice as well as waters near the ice edge.

Klöser (1990) analyzed plankton near the Antarctic Peninsula in late autumn (Polarstern cruise V/1 in May and June 1986). Southern Ocean diatom species developed survival strategies for the winter, which besides darkness also cover survival in (under) ice conditions. Cells from the ACC are transported under the ice but often die due to a lack of light. After the ice melt, those cells that overwintered in the ice act as seed cells.

Scharek (1991) compared plankton samples taken in the Weddell Sea from October to December 1986 before and after sea ice melting. Three regions were sampled: the northern sea ice edge zone (54°30'S - 60°S), the pack ice zone in the eastern Weddell gyre (60°S - 70°S), and the polynyas in the southeastern Weddell sea (70°S - 77°S). For *Nitzschia kerguelensis* (equivalent to *Fragilariopsis kerguelensis*), she reports occurrence in ACC area during winter in high numbers. In the pack ice regions, she distinguishes the northern part, where *F. kerguelensis* frequently occurs in winter and spring, but where several empty and broken valves were observed. For the southern region and the polynyas along the coast, she reports only sporadic observations, with most of the valves empty.

F. kerguelensis samples, taken on Polarstern cruise XXIX/5 (18.04. - 29.05.2013) on the way from the Falkland Islands towards South Africa, confirm northern distribution regions in the predicted distribution area.

During cruise PS103 of R/V Polarstern in December 2016 on the way from Capetown to Atka Bay (close to Neumayer III station), samples were taken in the region of the expected distribution boundary of *F. kerguelensis* for comparison with the model's predictions. As the first real station of that cruise was planned at a location far south of the region of interest, samples were taken from the seawater pipeline and concentrated by means of a net. This way, sampling could start at approximately 42°S. The last sample in which *F. kerguelensis* was not observed was taken at 42°29'S, 9°58'E. The first occurrence of *F. kerguelensis* was observed at 42°54.5'S, 9°34.9E, but this sample contained short chains of dead cells only. Living cells of *F. kerguelensis* first occurred at 43°22.2'S, 9°E. These sites are located at the predicted boundary of the *F. kerguelensis* model for December (Model 3). The first two sites are located in the same grid cell of the model projection. This cell has a model output value of 0.23. The 0.2 iso-line of the threshold used for these studies is passing directly through this cell. The third location, where living *F. kerguelensis* cells were observed first, is the grid cell directly south of the previous one, with a value of 0.3.

At 9°E *Orsi et al.* (1995) located the Subtropical front at approximately 38°S and the Subantarctic front at 48°S. In the model predictions, *F. kerguelensis'* northern distribution boundary is located between those two fronts. It has to be noted that these are average positions and that in the course of the year as well as over the years, these fronts can move hundreds of kilometers. In cruise PS103, the first CTD measurement was conducted far more south at 45°57.318'S, 6°17.241'E, which already showed characteristics of the Polar Frontal zone (this location according to Orsi is slightly north of the Subantarctic front).

Further, several net samples from the ice shelf near Neumayer Station contained several chains of living *Fragilariopsis* cells. A few samples were treated with hydrochloric acid to allow identification on a species level. In the sample the following *Fragilariopsis* species could be identified: *F. ritscheri*, *F. sublinearis*, *F. obliquecostata*, *F. curta*, *F. nana*, *F. cylindrus*, and maybe *F. rhombica*. The species *F. kerguelensis* was not present.

In summary, the first occurrence of *F. kerguelensis* on cruise PS103 in December 2016 felt exactly in the area predicted by model 3. Comparison to the location of the fronts described by Orsi does not fit well, as the fronts currently are shifted several hundred kilometers towards the north. It might be that *F. kerguelensis* really avoids ice-covered regions, whereas other (*Fragilariopsis*) species don't. Due to missing observation data, projections on ice-covered regions might still be considered out of the model's scope.

4.2.3 Upper temperature tolerance limits

Nitrate was the most influential predictor in the majority of the models, followed by sea surface temperature. Temperature is considered as a well suitable variable for lab experiments as it is technically well feasible in experiments on the one hand, and has a strong ecological impact on the species distribution as well as a high predictive capacity in the models on the other. A few earlier studies already determined upper growing temperature of *F. kerguelensis*: *Jacques* (1983) kept *F. kerguelensis* at temperatures of 3, 5, 7, 10, and 15° C. He observed the highest growth rate at a temperature of 5° C. Later, *Fiala and Oriol* (1990) estimated an upper growing temperature of 7° C. In contrast, the sea surface temperatures at the *F. kerguelensis* observation sites give a different picture (see figure 4.2). The upper temperature limit for *F. kerguelensis* in model 5f, based just on the temperature predictor, is at approximately 8° C (regarding a logistic value of 0.2). An experiment on temperature limits on diatom growth was conducted, to validate the upper temperatures from observation data and model predictions with lab data (see chapters 2.4 and 3.4).

This experiment was conducted to judge how realistic the model outputs are, given the enormous discrepancies between published temperature limits in literature and observation records (for which the temperature was taken from compiled data repositories such as, e.g., the World Ocean Atlas). Due to its simple setup, this experiment admittedly provides only limited insights to temperature tolerance of *F. kerguelensis*. But, however, it clearly extends the previously published temperature tolerance limits and potentially highlights an intraspecific variability in temperature tolerances. High temperatures estimated for of some of the observations cannot be confirmed by this experiment (the highest temperature of 29.2°C belonged to an observation record east of Samoa at 10°S, 166°W). The majority of the values are in a reasonable range (median = 2.57°C, third quartile = 3.74°C). This is also true for the model output that predicts an upper limit of 8° C.

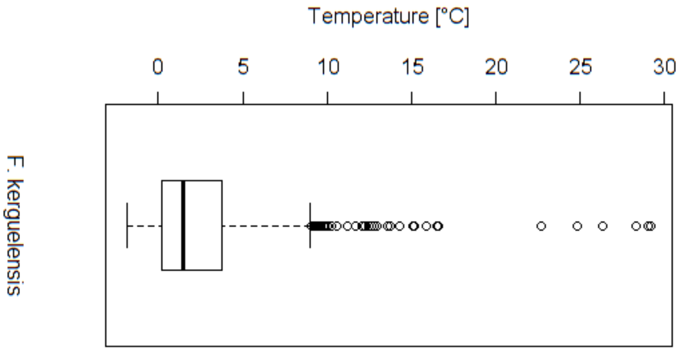

Figure 4.2: Sea surface temperature at *F. kerguelensis* observation sites.

In this experiment, biological interactions such as grazing, competition, etc. were ignored. Of course, tolerance limits estimated in such an experiment cannot be directly compared to the tolerance limits in nature (which are reflected in the observation data and in the model). Further, no samples were taken for an exact determination of growth rates by cells counts. Instead, thus not very accurate, vitality was estimated by observing the cultures under an inverted microscope. This worked best after inoculation of new culture flasks. The exact timing of the temperature rises was adapted to the inoculation dates, to give the cultures a few days for acclimation. As a useful addition, this experiment should be repeated with more replicates, longer adaption phases, and various light regimes, all in combination with exact cell counts to estimate growth rates.

4.2.4 Pattern analysis

Coarse distribution patterns are already visible in the global distribution maps (see figures 3.16 to 3.20), e.g., the belt shape areas around the Antarctic in the Southern Ocean. On the global scale, four main patterns could also be clearly identified with the clustering methods: (1) focus on the mid-latitudes, (2) cosmopolitan, (3) bipolar distribution, (4) focus on the Southern Ocean, with a small signal in the north Pacific. The clustering, however, could not separate the groups by their northern distribution boundaries in the Southern Ocean. Further, model parameters, as well as the model's input data, did not cluster the same way the distribution maps did. Models for different species that are resulting in the same distribution patterns can be based on the same model coefficients, but not all are.

These main patterns, however, correlate well with the most influential predictors, e.g., nitrate for the species endemic to the Southern Ocean and sea surface temperature for the bipolar ones. Further, values for nitrate, iron, and sea surface temperature at the observation sites cover a wider range in case of the truly bipolar species.

To model the species northern distribution boundaries in the Southern Ocean, nitrate and silicate are good predictors in correlative models due to their decreasing concentrations towards the North, just as sea surface temperature due to a northward temperature increase.

The models discussed in *Pinkernell and Beszteri* (2014) did not resolve a clear southern distribution boundary. The potential influence of sea ice coverage on the southern distribution boundary was expected to be covered by the sea surface temperature predictor. The new models, with the sea ice concentration predictor included, clearly contradicted this. They led to a sharp southern distribution boundary at the sea ice edge, e.g., in the *F. kerguelensis* models related to a sea ice concentration of 79%. In the models, this boundary is a consequence of missing samples of sea ice covered regions. The gaps in the model projections thus could be interpreted as regions outside the model's scope. Leaving out this predictor leads to unexplainable model outputs for these regions that have no observation records for the according environmental space. It is not clear, although *F. kerguelensis* is perceived to avoid sea ice (C. Klaas, pers. comm.), whether it might not be distributed till the ice shelf or continent as its southern border instead. The model's predictions might not be wrong as *F. kerguelensis*, as it does not immediately disappear from ice-covered regions, but only occurs in low abundances in those areas. Ice-covered regions appear as unsuitable for most of the species in the models, which, from an ecological point of view is doubtful.

Several approaches exist to group the world oceans into units of unique ecological properties. *Longhurst* (2010) partitioned the ocean into 56 bio-geochemical provinces. These partitions are frequently used as a reference in ecological studies but are also criticized for their static nature. *Reygondeau et al.* (2013) refined the partitions boundaries based on four predictors: sea surface temperature, sea surface salinity, chlorophyll a, and bathymetry. He provides monthly plots, but avoids predictions in regions affected by winter darkness and distinguishes four main biomes: polar, westerly wind, trade wind, and coastal, separated in the 56 bio-geochemical provinces Longhurst used. In contrast, the distribution models used here are based on seven predictors, which besides temperature and salinity to characterize water bodies, also take nutrients into account.

For many modeled species the northern distribution boundaries (for February) are located in the Subantarctic Water Ring province (SANT) in the westerly wind biome. Some also include regions of the coastal biome, such as the Humboldt current coast province (HUMB) and the South West Atlantic Shelves province (FKLD). A few other species, e.g., *Fragilariopsis curta* and *F. nana*, have their northern boundaries located closer towards the pole and just reach the Antarctic province (ANTA) south of the SANT in the polar biome. The ANTA biome, together with the Australian Polar biome (APLR) forms the main regions for most of the modeled species. Several models, even for species that are endemic to the Southern Ocean, show a signal in the north Pacific. This region belongs to the westerly winds biome and is separated in six provinces: Western Pacific subarctic gyres (PSAW), Kuroshio current (KURO), Northwest Pacific subtropical (NPSW), Eastern Pacific subarctic gyres (PSAE), North Pacific polar front (NPPF), and Northeast Pacific subtropical (NPSE). The analyzed species show individual distribution patterns that mostly fall in only a few ecological provinces. The pattern described by the provinces is too coarse, compared to the differences in the species distribution patterns. Of course, Longhurst partitions are not intended to comply with the distribution areas of individual species directly.

The southern ocean is characterized by a strong frontal system and consists of different water bodies with unique features, e.g., temperature, salinity, nutrient supply, etc. These frontal structures were believed to have a strong influence on the biogeography of the pelagic diatoms. Distribution boundaries and fronts were found to be related, but much less than expected. The fronts plotted according to *Orsi et al.* (1995) are average positions; whereas the real positions can vary by hundreds of kilometers. Further, the observation records were collected over many years, and also the environmental data was aggregated over decades. This might explain the loose relation of the modeled species distribution boundaries and average frontal positions. Subduction of the waters north of the Subantarctic Front might further result in a big fraction of the organisms disappear from the upper layer and this way form a natural border.

For ice related algae further model adaptions are needed, e.g., to resolve whether a species was detected in the ice or in the water column under the ice. This, currently, is not yet implemented in the model and thus also out of the model's scope and predictive power. Besides that, in many cases also the observation records are lacking metadata about a sample being from the ice or from the water column underneath.

In summary, despite the lack of barriers in the Southern Ocean, the models predict that the species have their individual northern distribution boundaries. These distribution boundaries might be determined by individual autecological features of the species. The models indicate this, especially for temperature. To clarify this, further experiments are needed, e.g., comparing studies of species-specific temperature tolerances.

4.2.5 Future distribution prognosis

The Southern Ocean is already affected by global change, but less than other regions, e.g., the Arctic. Until the end of this century, however, massive changes are expected for the Southern Ocean, including strengthening of westerly winds, warmer and fresher surface waters, a potential shift of wind and frontal systems towards the pole, an increased eddy activity, and increased stratification (*Constable et al.*, 2014).

For the measurements, only the Southern Ocean and adjacent ocean basins were considered, and for the bipolar species, the northern regions were neglected. Most models predict a decline of the species distribution area due to a poleward shift of the northern distribution boundary. The predicted decline varies among the species and depends on the future scenario. As expected, it is stronger for the RCP8.5 than for the RCP4.5 scenario, but a few exceptions exist. Overall, several common features to the Longhurst/Reygondeau provinces could be observed, e.g., the movement of the northern distribution boundary/ SANT province towards the pole and back in summer/winter.

During model building, nitrate appeared to be a very important predictor in most of the models, but this predictor is not necessarily the reason for the area loss in the future predictions. Iron was already identified as problematic for future projections (see the chapter on iron predictor in the discussion). In the projections on the NorESM1-ME and the MPI-ESM-LR models iron indeed had the strongest influence on the model output in the regions affected by area loss, followed by sea surface temperature and nitrate concentration. In the CESM1-BGC and IPSL-CM5A model, the effect can be explained mainly by sea surface temperature and to a lesser extent by nitrate concentration. Thus, in the resulting future predictions plotted in figures 3.16 to 3.20, the iron predictor probably has an exaggerated effect on the model output, resulting in an overestimated decline of the predicted distribution area. In case of *Fragilariopsis kerguelensis*, a comparison of a model with and a model without iron confirms this. For the RCP8.5 scenario, model 3 (with iron) predicts an area loss of 37.8% for the February 2100 projection in comparison to the current state, whereas model 4 (without iron) predicts just 33.6%.

Two models, those of *Fragilariopsis vanheurkii* and *Asteromphalus hookeri*, resulted in an increased predicted distribution area for the RCP8.5 scenario for the end of the century. These models had different model coefficients which means that this was caused by different predictors. Both models responded strongly on nutrients. For *A. hookeri* nitrate played the most important role (36.1%), followed by silicate (22.4%), whereas for *F. vanheurkii* only silicate was important (60.5%) and nitrate played a minor role (0.6%). In addition further two models, those of *Fragilariopsis cylindrus* and *F. pseudonana*, resulted in an increased area for the RCP4.5 scenario.

For the area measurements and future plots, a threshold of 0.2 was used. In case of the *Fragilariopsis linearis* model, this threshold is problematic. The model showed a very weak signal in the Southern Ocean and did not reach the threshold in the future projections, resulting in an area loss of 100%. As discussed earlier, this model suffers from a bad observation record coverage and thus has a limited explanatory power anyway.

4.3 Synthesis

As a synthesis, first the research questions raised in the introduction are answered, and a general conclusion and outlook are given.

4.3.1 Answers to research questions

Q1 - Evaluation of SDM methods exemplified by *F. kerguelensis* models

The first aspect of this thesis covers a comprehensive evaluation of species distribution models for pelagic marine diatoms at the example of *F. kerguelensis* using Maxent. This method was chosen because of its good performance for presence-only data, a common data type that can be derived from public observation data repositories and herbaria such as the Hustedt diatom collection. The latter was used extensively to generate a high-quality dataset, also supported by voucher images. Observation data quality (and quantity) have a huge influence on the model quality. A good north-south coverage over all water masses is necessary, whereas a circumpolar coverage is less important.

A strong temporal bias became visible, as winter samples, especially in connection with sea ice, are rare. Seven environmental predictors were tested. Including sea ice concentration helped to visualize the impact of the temporal bias. Overall, despite the partly poor data situation, the models result in robust and reliable current spatial predictions. Model tests and evaluations did not indicate further issues. In conclusion, species distribution models as used here can be considered suitable for modeling spatial distributions of Southern Ocean pelagic marine diatoms.

Q2 - Application on further species

Models for further 20 species, some of them endemic to the Southern Ocean, some bipolar, and a few even cosmopolitan, were built based on the experience gained in the *F. kerguelensis* models. For the majority of the species, plausible models could be built. The models suffer the same problems already known from the *F. kerguelensis* models, such as a partly poor data situation and a temporal bias avoiding winter samples, especially in relation to sea ice related coverage. Overall, a strong variation in the data situation was observed, ranging from five to more than 1000 observation records. Macro patterns, e.g., endemic to the southern ocean, bipolar, and cosmopolitan, could be found in the spatial model predictions but did not fit to the observation records and previous knowledge from literature in all cases. Several models for species that are considered endemic to the Southern Ocean overestimate the suitable regions by including regions in the north Pacific, an upwelling area with similar HNLC characteristics like the Southern Ocean. These models can be identified based on their model parameters, e.g., the most influential predictor. Distribution patterns and underlying model parameters of all 21 models (including the *F. kerguelensis* model) were compared by clustering, PCA, and LRA-methods. These methods led to different clusters which means that even if two models result in a similar spatial distribution, the underlying model can be completely different. The majority of the distribution patterns in the Southern Ocean cover the Longhurst provinces but do not strictly agree with their boundaries. The Southern Oceans has a pronounced frontal system which was expected to have an influence on the distribution boundaries. The identified boundaries indeed follow these fronts, but not as strict as expected. Observation records and environmental data are aggregated over decades, and also the frontal positions are moving over time, which complicates a clear assignment of the modeled species distribution boundaries to the averaged frontal positions.

Hasle plotted observation records to visualize distributions of several diatom species. The models presented here are going further by highlighting suitable areas for the species. According to their projected summer distribution, these regions fit well. A few species have observations far outside the predicted range, e.g., *F. cylindrus*, *A. roperianus*, *A. hookeri*, *D. antarcticus*, and *E. antarctica.*

Q3 - Prediction of potential future distribution patterns

The distribution models were projected on RCP4.5 and RCP8.5 scenarios for summer (February) and winter (August) 2100. For the plots, the median over four or five respectively different global biogeochemical ocean models was used. Area measurements require a threshold, for which 0.2 was defined for all models. The selection of a threshold is arbitrary, and different values lead to strong variations. The trend from current to future distribution areas, however, stays the same.

Various predictor influences on the predicted spatial future distribution were compared using *F. kerguelensis* models. The iron predictor led to strong variations within the projections on the different GCM outputs and thus result in a strong uncertainty. This affected only the future projections, so keeping this predictor was considered more important for model quality than removing it would benefit the future projections. As the mixed layer depth predictor was not available for the HadGEM2-ES model, only the remaining four GCMs were used for model projections. As with the iron predictor, keeping MLD in the model is considered more important for overall model quality than removing it to project on five GCMs. Despite the issues with the iron and MLD predictors, potential future distribution patterns were plotted and areas measured.

In summary, most models predict a moderate area loss and a poleward shift till the end of the century. Unsurprisingly, both features are more pronounced in the RCP8.5 scenarios in most of the cases. The area measurements are limited to the Southern Ocean, independent of the regarding macro distribution pattern.

4.3.2 Conclusion and outlook

Species distribution models have been widely used for terrestrial organisms during the last decades, and lately also in the marine world. Interest in these models for marine protists is growing, especially with regard to global climate change, but experience with it is still scarce (*Chust et al.*, 2017). In this thesis, SDM was first evaluated for marine pelagic diatoms of the Southern Ocean at the example of *Fragilariopsis kerguelensis*. Based on that experience, models for further selected species were built with a focus on current and potential future distribution patterns.

In conclusion, SDMs turned out to be well suitable to model the biogeography of marine pelagic diatoms in the Southern Ocean. False positive model signals in the north Pacific for species that are considered as endemic to the Southern Ocean frequently occurred but could be clearly distinguished from the truly bipolar ones by their model parameters. The sea ice edge determined the predicted southern distribution boundary in the Southern Ocean in most of the cases. Observation data from sea-ice-covered regions are rare, so predictions for these regions are considered as outside the model's scope. At least for some species, however, they still might be true, e.g., for *F. kerguelensis* which is known to avoid sea ice.

The model results for each species are discussed, also in the context of previous studies, e.g., by Hasle, and are used to update knowledge about the individual species biogeography. Model projections on future environmental scenarios for the end of this century are used to predict the fate of these species with regard to climate change. The predictions indicate a decrease in the species distribution area for most of the species due to a poleward shift of their northern distribution boundary. As expected, these changes are stronger in the RCP8.5 scenario, the so-called business as usual scenario, than in the RCP4.5 scenario. Future projections are limited due to strong uncertainties in the future scenarios, especially in case of the iron predictor.

Further observation records are most important for further improvements, especially from currently under-sampled regions and seasons. This includes the northern regions of the ACC, the sea ice covered regions in the South, as well as samples from the austral winter season in general. Voucher images are still rare in public observation data repositories and should be deposited - at least for new entries. In the future, species detection by molecular methods might also improve the observation datasets, especially for cryptic species. Iron is considered as an important predictor for distribution models of diatoms, but its data quality currently is still poor. Improved environmental data, especially for iron, certainly can improve model quality. In the same way, new generations of earth system models will allow more precise future scenarios. Further, lab experiments might help to validate the models better. Besides more sophisticated experiments on temperature tolerance also effects on resource supply need to be studied, ideally in combined experiments and also considering intraspecific variations. Presence-only data as used here limit the available modeling approaches. They can be derived from any other data type, thus resulting in the highest possible number of observation records and in consequence lead to reasonable models. Nevertheless, observation data of a higher quality level might lead to more informative models in the future.

Bibliography

Adl, S. M., et al., The revised classification of eukaryotes, *J Eukaryot Microbiol*, *59*(5), 429–93, 2012.

Alper, P., C. A. Goble, and K. Belhajjame, On assisting scientific data curation in collection-based dataflows using labels, in *Proceedings of the 8th Workshop on Workflows in Support of Large-Scale Science*, pp. 7–16, ACM, 2013.

Antonov, J. I., D. Seidov, T. P. Boyer, R. A. Locarnini, A. V. Mishonov, H. E. Garcia, O. K. Baranova, M. M. Zweng, and J. D. R., *WORLD OCEAN ATLAS 2009 Volume 2: Salinity*, NOAA Atlas NESDIS 69, 184 pp. pp., U.S. Government Printing Office, Washington D.C., 2010.

Armand, L. K., X. Crosta, O. Romero, and J.-J. Pichon, The biogeography of major diatom taxa in Southern Ocean sediments, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *223*(1-2), 93–126, 2005.

Armand, L. K., X. Crosta, B. Quéguiner, J. Mosseri, and N. Garcia, Diatoms preserved in surface sediments of the northeastern Kerguelen Plateau, *Deep Sea Research Part II: Topical Studies in Oceanography*, *55*(5), 677–692, 2008.

Assmy, P., et al., Thick-shelled, grazer-protected diatoms decouple ocean carbon and silicon cycles in the iron-limited Antarctic Circumpolar Current, *Proc Natl Acad Sci U S A*, *110*(51), 20,633–8, 2013.

Baas-Becking, L. G. M., Geobiologie; of inleiding tot de milieukunde, 1934.

Bartsch, A., Die Eisalgenflora des Weddelmeeres (Antarktis): Artenzusammensetzung und Biomasse sowie Ökophysiologie ausgewählter Arten= Sea ice algae of the Weddell Sea (Antarctica): species composition, biomass, and ecophysiology of selected species, *Berichte zur Polarforschung (Reports on Polar Research)*, *63*, 1989.

Belkin, I. M., and A. L. Gordon, Southern Ocean fronts from the Greenwich meridian to Tasmania, *Journal of Geophysical Research: Oceans*, *101*(C2), 3675–3696, 1996.

Bombosch, A., D. P. Zitterbart, I. Van Opzeeland, S. Frickenhaus, E. Burkhardt, M. S. Wisz, and O. Boebel, Predictive habitat modelling of humpback (Megaptera novaeangliae) and Antarctic minke (Balaenoptera bonaerensis) whales in the Southern Ocean as a planning tool for seismic surveys, *Deep Sea Research Part I: Oceanographic Research Papers*, *91*, 101–114, 2014.

Booth, T. H., H. A. Nix, J. R. Busby, M. F. Hutchinson, and J. Franklin, bioclim: the first species distribution modelling package, its early applications and relevance to most currentMaxEntstudies, *Diversity and Distributions*, *20*(1), 1–9, 2014.

Bopp, L., O. Aumont, P. Cadule, S. Alvain, and M. Gehlen, Response of diatoms distribution to global warming and potential implications: A global model study, *Geophysical Research Letters*, *32*(19), n/a–n/a, 2005.

Bopp, L., et al., Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, *Biogeosciences*, *10*(10), 6225–6245, 2013.

Boyd, P. W., et al., Mesoscale iron enrichment experiments 1993-2005: Synthesis and future directions, *science*, *315*(5812), 612–617, 2007.

Boyd, P. W., et al., Marine phytoplankton temperature versus growth responses from polar to tropical waters–outcome of a scientific community-wide study, *PLoS One*, *8*(5), e63,091, 2013.

Brun, P., M. Vogt, M. R. Payne, N. Gruber, C. J. O'Brien, E. T. Buitenhuis, C. Le Quéré, K. Leblanc, and Y.-W. Luo, Ecological niches of open ocean phytoplankton taxa, *Limnology and Oceanography*, *60*(3), 1020–1038, 2015.

Brun, P., T. Kiorboe, P. Licandro, and M. R. Payne, The predictive skill of species distribution models for plankton in a changing climate, *Glob Chang Biol*, *22*(9), 3170–81, 2016.

Buttigieg, P. L., N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, and E. Consortium, The environment ontology: contextualising biological and biomedical entities, *J Biomed Semantics*, *4*(1), 43, 2013.

Cefarelli, A. O., M. E. Ferrario, G. O. Almandoz, A. G. Atencio, R. Akselman, and M. Vernet, Diversity of the diatom genus Fragilariopsis in the Argentine Sea and Antarctic waters: morphology, distribution and abundance, *Polar Biology*, *33*(11), 1463–1484, 2010.

Cermeno, P., and P. G. Falkowski, Controls on diatom biogeography in the ocean, *Science*, *325*(5947), 1539–41, 2009.

Cermeno, P., C. de Vargas, F. Abrantes, and P. G. Falkowski, Phytoplankton biogeography and community stability in the ocean, *PLoS One*, *5*(4), e10,037, 2010.

Cermeno, P., I. G. Teixeira, M. Branco, F. G. Figueiras, and E. Maranon, Sampling the limits of species richness in marine phytoplankton communities, *Journal of Plankton Research*, *36*(4), 1135–1139, 2014.

Chase, J. M., and M. A. Leibold, *Ecological niches: linking classical and contemporary approaches*, University of Chicago Press, 2003.

Chust, G., et al., Mare Incognitum: A Glimpse into Future Plankton Diversity and Ecology Research, *Frontiers in Marine Science*, *4*, 2017.

Colwell, R. K., and T. F. Rangel, Hutchinson's duality: the once and future niche, *Proc Natl Acad Sci U S A*, *106 Suppl 2*, 19,651–8, 2009.

Constable, A. J., et al., Climate change and southern ocean ecosystems i: how changes in physical habitats directly affect marine biota, *Global Change Biology*, *20*(10), 3004–3025, doi:10.1111/gcb.12623, 2014.

Cortese, G., and R. Gersonde, Morphometric variability in the diatom Fragilariopsis kerguelensis: Implications for Southern Ocean paleoceanography, *Earth and Planetary Science Letters*, *257*(3-4), 526–544, 2007.

Cox, E. J., Morphological variation in widely distributed diatom taxa: taxonomic and ecological implications, in *Proceedings of the 13th international diatom symposium. Biopress, Bristol*, pp. 335–345, 1995.

Crawford, R. M., F. Hinz, and C. Honeywill, Three species of the diatom genus Corethron Castracane: structure, distribution and taxonomy, *Diatom Research*, *13*(1), 1–28, 1998.

Crosta, X., O. Romero, L. K. Armand, and J.-J. Pichon, The biogeography of major diatom taxa in Southern Ocean sediments: 2. Open ocean related species, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *223*(1-2), 66–92, 2005.

D'Alelio, D., A. Amato, W. H. Kooistra, G. Procaccini, R. Casotti, and M. Montresor, Internal transcribed spacer polymorphism in pseudo-nitzschia multistriata (bacillariophyceae) in the gulf of naples: recent divergence or intraspecific hybridization?, *Protist*, *160*(1), 9–20, 2009.

De Baar, H. J., et al., Synthesis of iron fertilization experiments: from the iron age in the age of enlightenment, *Journal of Geophysical Research: Oceans*, *110*(C9), 2005.

de Souza Muñoz, M. E., R. De Giovanni, M. F. de Siqueira, T. Sutton, P. Brewer, R. S. Pereira, D. A. L. Canhos, and V. P. Canhos, openModeller: a generic approach to species' potential distribution modelling, *GeoInformatica*, *15*(1), 111–135, 2009.

De Wit, R., and T. Bouvier, 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say?, *Environmental microbiology*, *8*(4), 755–758, 2006.

Deacon, G. E. R., Physical and biological zonation in the Southern Ocean, *Deep Sea Research Part A. Oceanographic Research Papers*, *29*(1), 1–15, 1982.

Dong, S., J. Sprintall, S. T. Gille, and L. Talley, Southern Ocean mixed-layer depth from Argo float profiles, *Journal of Geophysical Research*, *113*(C6), 2008.

Dufresne, J. L., et al., Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, *Climate Dynamics*, *40*(9-10), 2123–2165, 2013.

Dunne, J. P., J. L. Sarmiento, and A. Gnanadesikan, A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor, *Global Biogeochemical Cycles*, *21*(4), n/a–n/a, 2007.

Elith, J., and C. H. Graham, Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models, *Ecography*, *32*(1), 66–77, 2009.

Elith, J., M. Kearney, and S. Phillips, The art of modelling range-shifting species, *Methods in Ecology and Evolution*, *1*(4), 330–342, 2010.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates, A statistical explanation of MaxEnt for ecologists, *Diversity and Distributions*, *17*(1), 43–57, 2011.

Elton, C., Animal ecology. 207 pp, *Sidgwick & Jackson, LTD. London*, 1927.

Esper, O., and R. Gersonde, New tools for the reconstruction of Pleistocene Antarctic sea ice, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *399*, 260–283, 2014a.

Esper, O., and R. Gersonde, Quaternary surface water temperature estimations: New diatom transfer functions for the Southern Ocean, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *414*, 1–19, 2014b.

Esper, O., R. Gersonde, and N. Kadagies, Diatom distribution in southeastern Pacific surface sediments and their relationship to modern environmental variables, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *287*(1-4), 1–27, 2010.

Evans, K. M., A. H. Wortley, G. E. Simpson, V. A. Chepurnov, and D. G. Mann, A molecular systematic approach to explore diversity within the sellaphora pupula species complex (bacillariophyta)(1), *J Phycol*, *44*(1), 215–31, 2008.

Fenner, J., H. Schrader, and H. Wienigk, Diatom phytoplankton studies in the southern Pacific Ocean, composition and correlation to the Antarctic Convergence and its paleoecological significance, *Initial Reports of the Deep Sea Drilling Project*, *35*, 757–813, 1976.

Fiala, M., and L. Oriol, Light-temperature interactions on the growth of Antarctic diatoms, *Polar Biology*, *10*(8), 629–636, 1990.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski, Primary production of the biosphere: integrating terrestrial and oceanic components, *Science*, *281*(5374), 237–40, 1998.

Fitzpatrick, M. C., N. J. Gotelli, and A. M. Ellison, MaxEnt versus MaxLike: empirical comparisons with ant species distributions, *Ecosphere*, *4*(5), art55, 2013.

Fourcade, Y., J. O. Engler, D. Rodder, and J. Secondi, Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias, *PLoS One*, *9*(5), e97,122, 2014.

Franklin, J., *Mapping species distributions: spatial inference and prediction*, Cambridge University Press, 2010.

Fryxell, G. A., P. A. Sims, and T. Watkins, Azpeitia (Bacillariophyceae): related genera and promorphology, *Systematic botany monographs*, pp. 1–74, 1986.

Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, M. M. Zweng, O. K. Baranova, and D. R. Johnson, *WORLD OCEAN ATLAS 2009 Volume 4: Nutrients (phosphate, nitrate, silicate)*, p. 398 pp., U.S. Government Printing Office, Washington D.C., 2009.

Garrison, D. L., Antarctic Sea Ice Biota, *American Zoologist*, *31*(1), 17–34, 1991.

Gibson, L., B. Barrett, and A. Burbidge, Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot, *Diversity and Distributions*, *13*(6), 704–713, 2007.

Giorgetta, M. A., et al., Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597, 2013.

Grinnell, J., Barriers to Distribution as Regards Birds and Mammals, *The american naturalist*, *48*(558), 248–254, 1914.

Grinnell, J., Field Tests of Theories Concerning Distributional Control, *The American Naturalist*, *51*(602), 115–128, 1917a.

Grinnell, J., The Niche-Relationships of the California Thrasher, *The Auk*, *34*(4), 427–433, 1917b.

Guillard, R. R. L., and J. H. Ryther, Studies of marine planktonic diatoms: I. cyclotella nana hustedt, and detonula confervacea (cleve) gran, *Canadian Journal of Microbiology*, *8*(2), 229–239, 1962.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle, Is my species distribution model fit for purpose? Matching data and models to applications, *Global Ecology and Biogeography*, *24*(3), 276–292, 2015.

Guisan, A., and W. Thuiller, Predicting species distribution: offering more than simple habitat models, *Ecology Letters*, *8*(9), 993–1009, 2005.

Guisan, A., and N. E. Zimmermann, Predictive habitat distribution models in ecology, *Ecological Modelling*, *135*(2-3), 147–186, 2000.

Hardisty, A., et al., A decadal view of biodiversity informatics: challenges and priorities, *BMC Ecol*, *13*, 16, 2013.

Hart, T. J., *Phytoplankton periodicity in antarctic surface waters*, vol. 21, pp. 261–356, Cambridge Universtiy Press, 1942.

Hasle, G., Nitzschia and Fragilariopsis species studied in the light and electron microscopes. III. The genus Fragilariopsis, *Akad. Oslo, Mat.-Nat. Kl*, *21*, 1–49, 1965.

Hasle, G. R., *Primary productivity and benthic marine algae of the Antarctic and Subantarctic*, vol. 10, pp. 6–8, 1968.

Hasle, G. R., *An analysis of the phytoplankton of the Pacific Southern Ocean: abundance, composition, and distribution during the Brategg Expedition, 1947-1948*, vol. 52, Univ.-Forl., 1969.

Hasle, G. R., The Biogeography of Some Marine Planktonic Diatoms, *Deep-Sea Research*, *23*(4), 319–338, 1976.

Heiden, H., and R. W. Kolbe, *Die Marinen Diatomeen der Deutschen Südpolar-Expedition 1901-1903*, W. de Gruyter & Company, 1928.

Hernández-Becerril, D. U., *The morphology and taxonomy of species of the diatom genus Asteromphalus Ehr*, *Bibliotheca Diatomologica*, vol. 23, J. Cramer in der Gebrüder Bornträger Vertragsbuchhandlung, Berlin, Stuttgart, 1991.

Hirzel, A. H., and G. Le Lay, Habitat suitability modelling and niche theory, *Journal of Applied Ecology*, *45*(5), 1372–1381, 2008.

Huertas, I. E., M. Rouco, V. Lopez-Rodas, and E. Costas, Warming will affect phytoplankton differently: evidence through a mechanistic approach, *Proc Biol Sci*, *278*(1724), 3534–43, 2011.

Hutchins, D. A., and K. W. Bruland, Iron-limited diatom growth and Si: N uptake ratios in a coastal upwelling regime, *Nature*, *393*(6685), 561–564, 1998.

Hutchinson, G. E., Concluding Remarks, *Cold Spring Harbor Symposia on Quantitative Biology*, *22*(0), 415–427, 1957.

Irwin, A. J., A. M. Nelles, and Z. V. Finkel, Phytoplankton niches estimated from field data, *Limnology and Oceanography*, *57*(3), 787–797, 2012.

Jacques, G., Some ecophysiological aspects of the Antarctic phytoplankton, *Polar Biology*, *2*(1), 27–33, 1983.

Jones, C. D., et al., The HadGEM2-ES implementation of CMIP5 centennial simulations, *Geoscientific Model Development*, *4*(3), 543–570, 2011.

Jousé, A., G. Koroleva, and G. Nagaeva, Diatoms in the surface layer of sediment in the indian sector of the antarctic. (in russian - english summary), *Trudy Instituta Okeanologii Akademiya Nauk SSSR*, *61*, 20–91, 1962.

Kara, A. B., Mixed layer depth variability over the global ocean, *Journal of Geophysical Research*, *108*(C3), 2003.

Klaas, C., and D. E. Archer, Association of sinking organic matter with various types of mineral ballast in the deep sea: Implications for the rain ratio, *Global Biogeochemical Cycles*, *16*(4), 2002.

Klöser, H., Verteilung von Mikroplankton-Organismen nordwestlich der Antarktischen Halbinsel unter dem Einfluss sich ändernder Umweltbedingungen im Herbst= Distribution of microplankton organisms north and west of the Antarctic Peninsula according to changing ecological conditions in autumn, *Berichte zur Polarforschung (Reports on Polar Research)*, *77*, 1990.

Kramer-Schadt, S., et al., The importance of correcting for sampling bias in MaxEnt species distribution models, *Diversity and Distributions*, *19*(11), 1366–1379, 2013.

Kumar, S., S. A. Spaulding, T. J. Stohlgren, K. A. Hermann, T. S. Schmidt, and L. L. Bahls, Potential habitat distribution for the freshwater diatom Didymosphenia geminata in the continental US, *Frontiers in Ecology and the Environment*, *7*(8), 415–420, 2009.

Leblanc, K., et al., A global diatom database – abundance, biovolume and biomass in the world ocean, *Earth System Science Data Discussions*, *5*(1), 147–185, 2012.

Locarnini, R., A. V. Mishonov, J. I. Antonov, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, and D. R. Johnson, *WORLD OCEAN ATLAS 2009 Volume 1: Temparature*, p. 184 pp., U.S. Government Printing Office, Washington D.C., 2010.

Long, M. C., K. Lindsay, S. Peacock, J. K. Moore, and S. C. Doney, Twentieth-Century Oceanic Carbon Uptake and Storage in CESM1(BGC)*, *Journal of Climate*, *26*(18), 6775–6800, 2013.

Longhurst, A. R., *Ecological geography of the sea*, Academic Press, 2010.

Lundholm, N., and G. R. Hasle, Are Fragilariopsis cylindrus and Fragilariopsis nana bipolar diatoms? - Morphological and molecular analyses of two sympatic species, *Nova Hedwigia*, *133*, 231–250, 2008.

Mann, D. G., The species concept in diatoms, *Phycologia*, *38*(6), 437–495, 1999.

Mann, D. G., and S. J. M. Droop, 3. Biodiversity, biogeography and conservation of diatoms, *Hydrobiologia*, *336*(1-3), 19–32, 1996.

Mann, D. G., and P. Vanormelingen, An inordinate fondness? the number, distributions, and origins of diatom species, *J Eukaryot Microbiol*, *60*(4), 414–20, 2013.

Marañón, E., P. Cermeño, M. Latasa, and R. D. Tadonléké, Temperature, resources, and phytoplankton size structure in the ocean, *Limnology and Oceanography*, *57*(5), 1266–1278, 2012.

Maranon, E., P. Cermeno, M. Huete-Ortega, D. C. Lopez-Sandoval, B. Mourino-Carballido, and T. Rodriguez-Ramos, Resource supply overrides temperature as a controlling factor of marine phytoplankton growth, *PLoS One*, *9*(6), e99,312, 2014.

Martin, G., et al., The HadGEM2 family of Met Office Unified Model climate configurations, Geosci. Model Dev., 4, 723–757, doi: 10.5194, 2011.

Martin, J. H., Glacial-interglacial CO2 change: The iron hypothesis, *Paleoceanography*, *5*(1), 1–13, 1990.

Mathew, C., A. Guntsch, M. Obst, S. Vicario, R. Haines, A. R. Williams, Y. de Jong, and C. Goble, A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control, *Biodiversity Data Journal*, (2), e4221, 2014.

McInerny, G. J., and R. S. Etienne, Ditch the niche - is the niche a useful concept in ecology or species distribution modelling?, *Journal of Biogeography*, *39*(12), 2096–2102, 2012a.

McInerny, G. J., and R. S. Etienne, Pitch the niche - taking responsibility for the concepts we use in ecology and species distribution modelling, *Journal of Biogeography*, *39*(12), 2112–2118, 2012b.

McInerny, G. J., and R. S. Etienne, Stitch the niche - a practical philosophy and visual schematic for the niche concept, *Journal of Biogeography*, *39*(12), 2103–2111, 2012c.

Merckx, B., M. Steyaert, A. Vanreusel, M. Vincx, and J. Vanaverbeke, Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling, *Ecological Modelling*, *222*(3), 588–597, 2011.

Merow, C., M. J. Smith, and J. A. Silander, A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter, *Ecography*, *36*(10), 1058–1069, 2013.

Mohan, R., A. A. Quarshi, T. Meloth, and M. Sudhakar, Diatoms from the surface waters of the Southern Ocean during the austral summer of 2004, *Current Science*, *100*(9), 1323–1327, 2011.

Monk, J., How long should we ignore imperfect detection of species in the marine environment when modelling their distribution?, *Fish and Fisheries*, *15*(2), 352–358, 2014.

Morin, X., and W. Thuiller, Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change, *Ecology*, *90*(5), 1301–13, 2009.

Moss, R. H., et al., The next generation of scenarios for climate change research and assessment, *Nature*, *463*(7282), 747–56, 2010.

Muscarella, R., P. J. Galante, M. Soley-Guardia, R. A. Boria, J. M. Kass, M. Uriarte, R. P. Anderson, and J. McPherson, ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity forMaxentecological niche models, *Methods in Ecology and Evolution*, *5*(11), 1198–1205, 2014.

Neiva, J., J. Assis, F. Fernandes, G. A. Pearson, and E. A. Serrao, Species distribution models and mitochondrial DNA phylogeography suggest an extensive biogeographical shift in the high-intertidal seaweed Pelvetia canaliculata, *Journal of Biogeography*, *41*(6), 1137–1148, 2014.

Nelson, D. M., P. Tréguer, M. A. Brzezinski, A. Leynaert, and B. Quéguiner, Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation, *Global Biogeochemical Cycles*, *9*(3), 359–372, 1995.

Nix, H. A., A biogeographic analysis of Australian elapid snakes, *Atlas of elapid snakes of Australia*, *7*, 4–15, 1986.

Norton, T. A., M. Melkonian, and R. A. Andersen, Algal biodiversity*, *Phycologia*, *35*(4), 308–326, 1996.

Nowlin, W. D., and J. M. Klinck, The physics of the Antarctic Circumpolar Current, *Reviews of Geophysics*, *24*(3), 469, 1986.

Olguín, H. F., and V. A. Alder, Species composition and biogeography of diatoms in antarctic and subantarctic (Argentine shelf) waters (37–76°S), *Deep Sea Research Part II: Topical Studies in Oceanography*, *58*(1-2), 139–152, 2011.

Olguín, H. F., D. Boltovskoy, C. B. Lange, and F. Brandini, Distribution of spring phytoplankton (mainly diatoms) in the upper 50 m of the Southwestern Atlantic Ocean (30–61 S), *Journal of plankton research*, *28*(12), 1107–1128, 2006.

Olguín Salinas, H. F., V. A. Alder, A. Puig, and D. Boltovskoy, Latitudinal diversity patterns of diatoms in the Southwestern Atlantic and Antarctic waters, *Journal of Plankton Research*, *37*(4), 659–665, 2015.

Orsi, A. H., T. Whitworth, and W. D. Nowlin, On the meridional extent and fronts of the Antarctic Circumpolar Current, *Deep Sea Research Part I: Oceanographic Research Papers*, *42*(5), 641–673, 1995.

Peter, K. H., and U. Sommer, Phytoplankton cell size reduction in response to warming mediated by nutrient limitation, *PLoS One*, *8*(9), e71,528, 2013.

Peterson, A. T., J. Soberon, R. G. Pearson, R. P. Anderson, E. Martinez-Meyer, M. Nakamura, and M. B. Araujo, *Ecological Niches and Geographic Distributions*, Princeton University Press, 2011.

Petrou, K., S. A. Kranz, S. Trimborn, C. S. Hassler, S. B. Ameijeiras, O. Sackett, P. J. Ralph, and A. T. Davidson, Southern Ocean phytoplankton physiology in a changing climate, *J Plant Physiol*, *203*, 135–150, 2016.

Phillips, S. J., and M. Dudik, Modeling of species distributions with Maxent: new extension and a comprehensive evaluation, *Ecography*, *31*(1), 161–175, 2008.

Phillips, S. J., M. Dudik, and R. E. Schapire, A maximum entropy approach to species distribution modeling, *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 655–662, 2004.

Phillips, S. J., R. P. Anderson, and R. E. Schapire, Maximum entropy modeling of species geographic distributions, *Ecological Modelling*, *190*(3-4), 231–259, 2006.

Pinkernell, S., and B. Beszteri, Potential effects of climate change on the distribution range of the main silicate sinker of the Southern Ocean, *Ecol Evol*, *4*(16), 3147–61, 2014.

Pollard, R. T., M. I. Lucas, and J. F. Read, Physical controls on biogeochemical zonation in the Southern Ocean, *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(16), 3289–3305, 2002.

Radosavljevic, A., R. P. Anderson, and M. Araújo, Making better Maxentmodels of species distributions: complexity, overfitting and evaluation, *Journal of Biogeography*, *41*(4), 629–643, 2014.

Raes, N., and H. ter Steege, A null-model for significance testing of presence-only species distribution models, *Ecography*, *30*(5), 727–736, 2007.

Ragueneau, O., N. Dittert, P. Pondaven, P. Tréguer, and L. Corrin, Si/C decoupling in the world ocean: is the Southern Ocean different?, *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(16), 3127–3154, 2002.

Ragueneau, O., S. Schultes, K. Bidle, P. Claquin, and B. Moriceau, Si and C interactions in the world ocean: Importance of ecological processes and implications for the role of diatoms in the biological pump, *Global Biogeochemical Cycles*, *20*(4), n/a–n/a, 2006.

Raven, P. H., J. M. Scott, P. Heglund, and M. L. Morrison, *Predicting species occurrences: Issues of accuracy and scale*, Island Press, 2002.

Rayner, N. A., Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Journal of Geophysical Research*, *108*(D14), 2003.

Ren, J., R. Gersonde, O. Esper, and C. Sancetta, Diatom distributions in northern North Pacific surface sediments and their relationship to modern environmental variables, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *402*, 81–103, 2014.

Renner, I. W., and D. I. Warton, Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology, *Biometrics*, *69*(1), 274–81, 2013.

Reygondeau, G., A. Longhurst, E. Martinez, G. Beaugrand, D. Antoine, and O. Maury, Dynamic biogeochemical provinces in the global ocean, *Global Biogeochemical Cycles*, *27*(4), 1046–1058, 2013.

Rivkin, R. B., and M. A. Voytek, Photoadaptations of Photosynthesis and Carbon Metabolism by Phytoplankton from Mcmurdo Sound, Antarctica .1. Species-Specific and Community Responses to Reduced Irradiances, *Limnology and Oceanography*, *32*(1), 249–259, 1987.

Rodriguez-Ramos, T., M. Dornelas, E. Maranon, and P. Cermeno, Conventional sampling methods severely underestimate phytoplankton species richness, *Journal of Plankton Research*, *36*(2), 334–343, 2013.

Romero, O., and C. Hensen, Oceanographic control of biogenic opal and diatoms in surface sediments of the Southwestern Atlantic, *Marine Geology*, *186*(3), 263–280, 2002.

Roura-Pascual, N., L. Brotons, A. T. Peterson, and W. Thuiller, Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula, *Biological Invasions*, *11*(4), 1017–1031, 2009.

Rovira, L., R. Trobajo, S. Sato, C. Ibanez, and D. G. Mann, Genetic and physiological diversity in the diatom nitzschia inconspicua, *J Eukaryot Microbiol*, *62*(6), 815–32, 2015.

Santika, T., Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data, *Global Ecology and Biogeography*, *20*(1), 181–192, 2011.

Scharek, R., Die Entwicklung des Phytoplanktons im östlichen Weddellmeer (Antarktis) beim Übergang vom Spätwinter zum Frühjahr= Development of phytoplankton during the late-winter/spring transition in the eastern Weddell Sea (Antarctica), *Berichte zur Polarforschung (Reports on Polar Research)*, *94*, 1991.

Schilthuizen, M., C. S. Vairappan, E. M. Slade, D. J. Mann, and J. A. Miller, Specimens as primary data: museums and 'open science', *Trends Ecol Evol*, *30*(5), 237–8, 2015.

Semina, H. J., *SEM-studies Diatoms of Different Regions of the World Ocean*, *Iconographia Diatomologica: Annotated Diatom Micrographs.*, vol. 10, ARG Ganter Verlag, 2003.

Shukla, S. K., J. Crespin, and X. Crosta, Thalassiosira lentiginosa size variation and associated biogenic silica burial in the Southern Ocean over the last 42kyrs, *Marine Micropaleontology*, *127*, 74–85, 2016.

Smetacek, V., Diatoms and the ocean carbon cycle, *Protist*, *150*(1), 25–32, 1999.

Smetacek, V., et al., Deep carbon export from a Southern Ocean iron-fertilized diatom bloom, *Nature*, *487*(7407), 313–9, 2012.

Soberón, J., and S. Higgins, Commentary on Ditch, Stitch and Pitch: the niche is here to stay, *Journal of Biogeography*, *41*(2), 414–417, 2014.

Soberon, J., and M. Nakamura, Niches and distributional areas: Concepts, methods, and assumptions, *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 19,644–19,650, 2009.

Syfert, M. M., M. J. Smith, and D. A. Coomes, The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models, *Plos One*, *8*(2), 2013.

Takeda, S., Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters, *Nature*, *393*(6687), 774–777, 1998.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, An Overview of CMIP5 and the Experiment Design, *Bulletin of the American Meteorological Society*, *93*(4), 485–498, 2012.

Thomas, M. K., C. T. Kremer, C. A. Klausmeier, and E. Litchman, A global pattern of thermal adaptation in marine phytoplankton, *Science*, *338*(6110), 1085–8, 2012.

Timmermans, K. R., and B. Van Der Wagt, Variability in Cell Size, Nutrient Depletion, and Growth Rates of the Southern Ocean Diatom Fragilariopsis Kerguelensis (Bacillariophyceae) after Prolonged Iron Limitation1, *Journal of Phycology*, *46*(3), 497–506, 2010.

Tjiputra, J. F., C. Roelandt, M. Bentsen, D. M. Lawrence, T. Lorentzen, J. Schwinger, . Seland, and C. Heinze, Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM), *Geoscientific Model Development*, *6*(2), 301–325, 2013.

Tomas, C. R., *Identifying marine phytoplankton*, Academic press, 1997.

Townsend Peterson, A., M. Papeş, and M. Eaton, Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent, *Ecography*, *30*(4), 550–560, 2007.

Treguer, P., and G. Jacques, Dynamics of Nutrients and Phytoplankton, and Fluxes of Carbon, Nitrogen and Silicon in the Antarctic Ocean, *Polar Biology*, *12*(2), 149–162, 1992.

Treguer, P., D. M. Nelson, A. J. Van Bennekom, and D. J. DeMaster, The silica balance in the world ocean: a reestimate, *Science*, *268*(5209), 375, 1995.

Tyberghein, L., H. Verbruggen, K. Pauly, C. Troupin, F. Mineur, and O. De Clerck, Bio-ORACLE: a global environmental dataset for marine species distribution modelling, *Global Ecology and Biogeography*, *21*(2), 272–281, 2012.

van Creveld, S. G., S. Rosenwasser, Y. Levin, and A. Vardi, Chronic iron limitation confers transient resistance to oxidative stress in marine diatoms, *Plant Physiology*, *172*(2), 968–979, 2016.

Van der Spoel, S., G. M. Hallegraeff, and R. W. M. Van Soest, Notes on variation of diatoms and silicoflagellates in the South Atlantic Ocean, *Netherlands Journal of Sea Research*, *6*(4), 518–541, 1973.

van Vuuren, D. P., et al., The representative concentration pathways: an overview, *Climatic Change*, *109*(1-2), 5–31, 2011.

Vanormelingen, P., E. Verleyen, and W. Vyverman, The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism, *Biodiversity and Conservation*, *17*(2), 393–405, 2008.

Verbruggen, H., L. Tyberghein, G. S. Belton, F. Mineur, A. Jueterbock, G. Hoarau, C. F. Gurgel, and O. De Clerck, Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed, *PLoS One*, *8*(6), e68,337, 2013.

Viličić, D., I. Marasović, and D. Mioković, Checklist of phytoplankton in the eastern Adriatic Sea, *Acta Botanica Croatica*, *61*(1), 57–91, 2002.

Villarino, E., G. Chust, P. Licandro, M. Butenschön, L. Ibaibarriaga, A. Larrañaga, and X. Irigoien, Modelling the future biogeography of North Atlantic zooplankton communities in response to climate change, *Marine Ecology Progress Series*, *531*, 121–142, 2015.

Vos, R. A., et al., Enriched biodiversity data as a resource and service, *Biodiversity Data Journal*, (2), e1125, 2014.

Vyverman, W., et al., Evidence for widespread endemism among Antarctic microorganisms, *Polar Science*, *4*(2), 103–113, 2010.

Walls, R. L., et al., Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies, *PLoS One*, *9*(3), e89,606, 2014.

Warren, D. L., In defense of 'niche modeling', *Trends in Ecology & Evolution*, *27*(9), 497–500, 2012.

Weinmann, A. E., D. Rödder, S. Lötters, and M. R. Langer, Traveling through time: The past, present and future biogeographic range of the invasive foraminifera Amphistegina spp. in the Mediterranean Sea, *Marine Micropaleontology*, *105*, 30–39, 2013.

Whitworth, T., and W. D. Nowlin, Water masses and currents of the Southern Ocean at the Greenwich Meridian, *Journal of Geophysical Research*, *92*(C6), 6462, 1987.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan, Effects of sample size on the performance of species distribution models, *Diversity and Distributions*, *14*(5), 763–773, 2008.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, S. Veran, and R. B. O'Hara, Presence-only modelling using MAXENT: when can we trust the inferences?, *Methods in Ecology and Evolution*, *4*(3), 236–243, 2013.

Zielinski, U., and R. Gersonde, Diatom distribution in Southern Ocean surface sediments (Atlantic sector): Implications for paleoenvironmental reconstructions, *Palaeogeography, Palaeoclimatology, Palaeoecology*, *129*(3-4), 213–250, 1997.