

THE ECOLOGICAL NICHE OF STORM-PETRELS IN THE NORTH PACIFIC AND
A GLOBAL MODEL OF DIMETHYLSULFIDE CONCENTRATION

By


Grant R.W. Humphries

RECOMMENDED:




Clara Deal

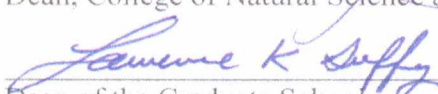

Advisory Committee Chair



P. BARBOSA
Chair, Wildlife Program
Department of Biology and Wildlife

APPROVED:



Dean, College of Natural Science and Mathematics


Dean of the Graduate School

March 31, 2010

Date

THE ECOLOGICAL NICHE OF STORM-PETRELS IN THE NORTH PACIFIC AND
A GLOBAL MODEL OF DIMETHYLSULFIDE CONCENTRATION

A
THESIS

Presented to the Faculty
of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

By

Grant R.W. Humphries, B.Sc

Fairbanks, Alaska

May 2010

BTJSCY
QL
G96
P64
H86
2010

BIOSCIENCES LIBRARY
UNIVERSITY OF ALASKA FAIRBANKS

~~UNIVERSITY OF ALASKA FAIRBANKS~~
UNIVERSITY OF ALASKA FAIRBANKS

Abstract

Ecological niche modeling techniques were used to create global, monthly predictions of sea surface dimethylsulfide (DMS) concentrations, and breeding season distribution of Leach's Storm-Petrel (*Oceanodroma leucorhoa*) and Fork-Tailed Storm-Petrel (*O. furcata*) in the North Pacific. This work represents the first attempt to model DMS concentrations on a global scale using ecological niche modeling, and the first models of Storm-Petrel distribution for the North Pacific. Storm-Petrels have been shown to be attracted to DMS, and it is therefore likely that a model of sea surface DMS concentration would help explain and predict Storm-Petrel distribution. We have successfully created the most accurate models of sea surface DMS concentrations that we are currently aware of with global correlation (r) values greater than 0.45. We also created Storm-Petrel models with area under the receiver operating characteristic curve (AUC) values of greater than 0.90. Using just DMS as a predictor variable we were also able to create models with AUC values upwards of 0.84. Future conservation efforts on pelagic seabird species may be dependent on models like the ones created here, and it is therefore important that these methods are improved upon to help seabird management on all scales (global, national, regional and local).

Table of Contents

	Page
Signature Page	i
Title Page.....	ii
Abstract.....	iii
Table of Contents	iv
List of Figures.....	viii
List of Tables.....	x
List of Appendices.....	xi
Acknowledgments	xii
General Introduction.....	1
Storm-Petrels of the North Pacific	2
Dimethylsulfide	3
Ecological niche modeling	4
Data mining (TreeNet)	5
Study goals	6
References	7
Chapter 1 Predicting monthly surface seawater dimethylsulfide (DMS) concentrations on a global scale using a machine learning algorithm (TreeNet).....	11
1.1 Introduction	12
1.1.1 Current DMS models.....	13
1.1.2 Spatial modeling with machine learning algorithms	14

1.1.3	Open access data.....	14
1.1.4	Study goal.....	15
1.2	Methods.....	16
1.2.1	Training data.....	16
1.2.2	Treenet algorithm.....	17
1.2.3	Scoring and output maps.....	18
1.2.4	Model assessment.....	19
1.3	Results.....	19
1.3.1	Ranking models.....	19
1.3.2	Predictor variable importance.....	20
1.3.3	Maps.....	20
1.4	Discussion.....	21
1.4.1	Towards a spatial ecology of DMS.....	22
1.4.2	Predictor variables.....	23
1.4.3	Model optimization.....	24
1.4.4	Seasonal variability.....	26
1.4.5	Considerations.....	29
1.5	Conclusions.....	30
	Acknowledgements.....	32
	References.....	33
Chapter 2 Predicted Distribution of Storm-Petrels (<i>Oceanodroma</i>) in the North Pacific using Geographic Information Systems (GIS), TreeNet and dimethylsulfide (DMS) concentrations.....		47

2.1	Introduction	48
2.2	Methods	51
2.2.1	Training data.....	51
2.2.2	Environmental layers.....	52
2.2.3	Model construction and assessment	52
2.2.4	Data flow	54
2.3	Results	54
2.3.1	Model ranks.....	54
2.3.2	Partial dependence plots	55
2.3.3	Distribution maps	55
2.3.4	Ground-truthing.....	56
2.4	Discussion.....	57
2.4.1	Spatial considerations.....	57
2.4.2	DMS as a predictor	59
2.4.3	Ground-truthing.....	60
2.4.4	Implications for Storm-Petrel management.....	61
2.4.5	Conclusions and future work.....	62
	Acknowledgements	63
	Literature Cited.....	65
	General Conclusions.....	78
	Dimethylsulfide	78
	Storm-Petrels	80

Final conclusions	81
References	83
Appendices	84

List of Figures

	Page
Figure 1.1: Plots of solar radiation dose, phosphates and salinity for all months. The x axis represents the unit value for each variable, and the y axis represents the partial dependence of each variable on sea surface DMS concentrations	39
Figure 1.2: Average Root Mean Squared Deviation (RMSD) of all months for Models 1 through 4 (subsets of 20, 60, 70 and 90% respectively).....	40
Figure 1.3: Monthly predictions of Dimethylsulfide concentrations (nM) created in ArcMap 9.3, Geographic Coordinate System WGS_1984, from TreeNet predictions based on public DMS and predictor data. Areas in white around coasts are areas of no data due to poor resolution of underlying datasets. Data are available from the author in ESRI grid format	41
Figure 1.4: Global annual surface concentration of DMS as created via an averaging of the best single run DMS models for each month. White regions around coastlines are areas of no data due to the coarseness of base predictor layers. Data are available from the author in ESRI grid format	42
Figure 1.5: Latitude time series plot of predicted sea surface concentration of DMS (nM).....	43
Figure 2.1: Partial dependence plots of Storm-Petrel distribution on DMS concentration (nM) for models using only DMS: a) Leach's Storm-Petrel model 2a b) Leach's Storm-Petrel model 2b c) Fork-Tailed Storm-Petrel model 2a d) Fork-Tailed Storm-Petrel model 2b	72

Figure 2.2: Maps of relative index of occurrence (RIO) of Leach's and Fork-Tailed Storm-Petrel as produced by TreeNet for the summer breeding season in the North Pacific using top ranked models with confirmed absence points.....	73
Figure 2.3: Maps of relative index of occurrence of Leach's Storm-Petrel (A) and Fork-Tailed Storm-Petrel (B) with confirmed presence and confirmed absence points from 2 opportunistic surveys performed in summer 2008	74
Figure 2.4: Mean relative index of occurrence (RIO) from predicted maps of confirmed presences and absences for both species of Storm-Petrel	75

List of Tables

	Page
Table 1.1: Highest ranking RMSD and corresponding R^2 values for each month. Models 3 and 4 refer to training subsets of 70 and 90% respectively. Letters a – e refer to one of 5 random permutations of each subset	44
Table 1.2: Average RMSD of 5 randomly drawn subset runs of models 1 – 4.....	45
Table 1.3: Relative importance of variables for models with lowest RMSD.....	46
Table 2.1: Area under the ROC curve (AUC) scores for default TreeNet settings (1), and altered TreeNet settings (2) using confirmed absences (a) and pseudo absences (b)	76
Table 2.2: Percent correctly classified presences for default TreeNet settings (1), and altered TreeNet settings (2) using confirmed absences (a) and pseudo absences (b)	77

List of Appendices

	Page
Appendix A: Data sources for all variables used for model development in Chapter 1 of thesis	84
Appendix B: Total number of data points used to train and assess each model for each month for Chapter 1 of thesis	85
Appendix C: Relative importance of predictor variables for highest ranking models (for both models a and b) as ranked by the AUC	86
Appendix D: Sources of predictor variables for Chapter 2 of thesis	87
Appendix E: Model automation for DMS model	88

Acknowledgements

This thesis was completed with the help and guidance of many individuals and institutions. I would like to acknowledge the following for their contributions in the execution of this study (in no particular order):

(1) Dr. Ian Jones for introducing me to Alaska in 2006 and for guiding my interests towards Marine Ornithology.

(1) Dr. Falk Huettmann for his support and guidance throughout the entirety of the thesis work, for instruction in using GIS, and for thoughtful and productive discussions in a variety of different fields. Dr. Huettmann has made a profound impact on my direction in life, and has been supportive of me since day one. This type of support and direction could only be given by someone passionate about his field of study, which has motivated me to be just as (if not more) passionate. Despite his accent and grammar, we managed to build a strong rapport throughout my time in Alaska, and will hopefully continue collaborations well into the future.

(2) Dr. Clara Deal for her patience and enthusiasm with this project. Her contributions have been key to the success of this thesis. Even with the prospects of moving on, she stuck things out with me and helped to make my time here amazing. Her support also sent me to sea with the Japanese and to Los Alamos, two experiences that have affected me greatly, allowing me to build strong connections with other scientists.

(3) Dr. David Atkinson for writing several scripts for figures and data processing, and for helping me make sense of programming languages. His feedback during this

work has been outstanding. I would also like to thank him for being Canadian and reminding me time and time again of my roots in Newfoundland. It has been important knowing that others share my love of my home-land.

(4) Jeff Williams of the US Fish and Wildlife Service for his help with logistics aboard the M/V Tigla. I would also like to thank Captain Billy Pepper and crew of the M/V Tigla for a wild ride and great times in the Aleutians, despite the “close call” with Mt. Cleveland in summer 2008.

(5) Dr. Sei-Ichi Saitoh of the University of Hokkaido for his help with logistics aboard the T/S Oshoro-Marui. I would also like to thank the Captain and crew of the T/S Oshoro-Marui for teaching me how to use chopsticks, and for allowing me to stand on the bow of the ship counting birds, even when the weather was less than ideal.

(6) Dr. Scott Elliot from the Los Alamos National Laboratory (LANL) for hosting me for the summer of 2009. The discussions that came from my time there led me to build a stronger understanding of my study system. I would also like to thank from LANL, Dr. Elizabeth Hunke, and Dr. Philip Jones.

(7) The Institute of Arctic Biology, the College of Natural Science and Mathematics, The Organization of Fish and Wildlife Information Managers and the Pacific Seabird Group for conference travel throughout my time “isolated” in Alaska.

(8) The International Arctic Research Center through the JAMSTEC-IARC Research Agreement, and IARC-NSF Cooperative Agreement, and Department of Energy EPSCoR Grant # DE-FG0208ER46502 for funding throughout my degree program, travel to Los Alamos, and funding my time aboard the T/S Oshoro-Marui.

- (9) All data contributors and data compilers for the North Pacific Pelagic Seabird Database, and the Pacific Marine Ecology Laboratory Dimethylsulfide database. Without the hard work of these contributors, none of this study would have been possible, and I hope to continue to see strong support for open access data worldwide so such studies may continue long into the future.
- (10) My parents, Kathleen (“Kate”) and Wayne (Chesley) Humphries, for constantly supporting me through all hard times, and for worrying about my safety in the “extreme” land of Alaska. I also want to thank my younger brother and sister for putting up with my distance from home, even though I am positive they enjoy having more space to themselves around the house.
- (11) Emily Weiser, who has supported me and believed in me from the beginning, despite my absent-mindedness. The impact she has made on my life has, and will continue, to affect me deeply. Emily’s thoughtful discussions and powers of observation have made a great difference to this work, and I shall forever be thankful.
- (12) Dr. Sergio Vallina, Gary Drew, Dr. Dave Verbyla, Dr. Hilmar Meier, Dr. Ian Jones, Dr. Mykhayl Golovnya, Michelle Wille, my labmates: Andy Baltensperger, Kim Jochum, Timothy Mullet, Susan Hazlett, and Micheal Lindgren, my friends, the ladies in the office, the international programs office, and a plethora of others who have accidentally been omitted from this list.

General introduction

The at-sea distribution of seabirds is a question important to scientists and managers. Studies have been performed correlating at-sea distributions of seabirds to certain environmental factors, but very few examine multivariate models as a predictive tool for testing hypotheses [*Elith et al.*, 2006; *Raymond and Woehler*, 2003].

Understanding and quantifying these distributions provides us with the ability to more accurately monitor and manage species, and to forecast anthropogenic or climate impacts.

Storm-Petrels (*Oceanodroma*) are a Genus of the family Hydrobatidae, of the Order Procellariiformes, which are tube-nosed, colonial seabirds. It is theorized that this group of birds uses their large, tubed noses to find food far out at sea where there are little to no visual cues for foraging [*Nevitt and Haberman*, 2003]. Dimethylsulfide (DMS) is a chemical that is released at the surface of the ocean, and is related to hotspots of primary productivity. Current models of global DMS exist, but are in contest with one another with respect to overall accuracy [*Bell et al.*, 2006; *Belviso et al.*, 2004]. In order to better understand the chemical composition of the oceanographic environment, a new model of DMS distribution using the best available science is required. The release of this chemical into the atmosphere from the ocean could act as an olfactory foraging cue for Storm-Petrels when visual clues are lacking in the open ocean. It is possible that a distribution model using DMS as a predictor variable will accurately classify Storm-Petrel distribution in the North Pacific. The proposed models will be developed using a type of regression tree modeling within a Geographical Information System (GIS) environment based on empirical data, in which no *a priori* assumptions are made

concerning which variables influence the target variable, allowing us to try a wide variety of predictors. The ultimate objective is the development of a model that will allow accurate predictions of DMS distribution that can be used to investigate the role DMS has on Storm-Petrel distribution.

Based on previous successful uses of the above modeling techniques [*Craig and Huettmann, 2009; Elith et al., 2006; Huettmann and Diamond, 2001; Ohse et al., 2009; Yen et al., 2004*], we can use algorithms that handle complex environmental interactions, enabling us to accurately model the spatial distribution of DMS, as well as the distribution of Storm-Petrels in the North Pacific. This would also allow us to capture the relationships between Storm-Petrels and DMS. These models can then be used for further analysis in determining effects of long term (climate change) and short term (oil spills, disturbance by ship traffic, etc...) factors which may alter the distribution of many different species.

Storm-Petrels of the North Pacific

Two species of Storm-Petrel breed in the North Pacific: Leach's Storm-Petrel (*Oceanodroma leucorhoa*) and Fork-Tailed Storm-Petrel (*Oceanodroma furcata*).

During the breeding season in the North Pacific they occupy deep burrows that can extend down to a meter in depth [*Boersma and Silva, 2001; Huntington et al., 1996*].

Both species are nocturnal and possess relatively poor eyesight which may have selected for enhanced development of other senses. These enhanced senses are important for inter and intra-species interactions. Both species leave their burrows at night to forage at sea for several days before returning to their colonies [*Boersma et al., 1980; Malakoff, 1999*;

Wilbur, 1969]. Like other Procellariiforms, Storm-Petrels have large olfactory bulbs, possibly because a well-developed chemical (olfactory) sense allows these birds to find these foraging areas as well as to find their breeding islands [*Grubb*, 1979].

The distribution of Fork-Tailed Storm-Petrels is limited to the Bering Sea, North Pacific and Sea of Okhotsk with breeding islands on all of the surrounding coasts. Winter and summer distributions of Fork-Tailed Storm-Petrels are essentially identical with recorded sightings at the ice edge during the boreal winter months [*Boersma and Silva*, 2001; *Onley and Scofield*, 2007]. Leach's Storm-Petrel have a more global distribution spreading from the Aleutian Islands and Sea of Okhotsk, southwards to central America in the Pacific Ocean, and from Norway to Brazil and Western Africa in the Atlantic Ocean. There is not much information on their winter distribution, though it is suggested there may be a southward migration during these months with an increase in Leach's Storm-Petrel sightings near Hawaii and Western Africa in the boreal winter [*Huntington et al.*, 1996; *Onley and Scofield*, 2007].

Dimethylsulfide

The Oceans are the primary influence on global climate and the mechanisms behind this link are poorly understood. DMS is a poorly studied biogenic compound that is the dominant source of sulfur to the atmosphere from the ocean and may act as a bridge between biology, oceans and the climate [*Andreae et al.*, 1985; *Lovelock et al.*, 1972]. DMS is also known to play a role as an olfactory cue in seabirds [*Nevitt and Bonadonna*, 2005] and even possibly in reef fish [*DeBose et al.*, 2008]. Currently it is believed that seabirds will “smell” DMS to locate foraging areas at sea as DMS is linked to areas of

high productivity where macro plankton such as Euphausiids may be located [*Nevitt and Bonadonna, 2005*]. DMS distribution has large implications for biological conservation management and species distribution modeling. We therefore need climatologies of DMS that will account for all the complexities involved in DMS formation.

Ecological niche modeling

Using GIS to build models has become very popular in ecology, and it has been shown that spatial variation in species is very important in determining how organisms use their environment [*Cushman, 2010*]. Environmental Systems Research Institute's (ESRI) ArcGIS is a widely used software package that can handle many of the functions required to build spatial models. ArcGIS can deal with datasets of a variety of different formats, and combined with the open access software Hawth's Tools (www.spataleecology.com), allows for processing of a large number of datasets. This sort of functionality means that users can overlay many predictor variables into a set of georeferenced data points, which can then be converted easily for use in machine learning software such as TreeNet. ArcGIS also has the added attraction of being easily programmed using the scripting language Python (www.python.org). Python can draw upon the statistical power of program R (www.r-project.org), and run programmable batch files, allowing for all geoprocessing and statistical analyses to be performed in one script. Such scripts allow for fast processing, and iterative testing of program settings in order to optimize model output.

Data mining (TreeNet)

Boosted regression trees (also known as Stochastic Gradient Boosting [*Friedman*, 2002a]) use an error minimization method to call upon an algorithm which creates a series of regression trees in an iterative fashion. Trees are created by boosting a classifier algorithm using a weighted subset of the training data. These trees have depths which are defined by the number of terminal nodes with the number of splits in the tree equaling the number of terminal nodes minus one [*Elith et al.*, 2008; *Friedman*, 2001; *Friedman*, 2002a]. Each split is computed based on the optimization (reduction) of the tree building criterion (that is, the minimization of the weighted least squares criterion). The error of each tree is estimated using v-fold cross validation, where the algorithm created by the tree is applied to the subset of the training data not used to build the tree. A loss function is then fitted to the data, and a new tree is calculated based on the weighting of the new subset [*Friedman et al.*, 2000]. This methodology allows us to avoid over fitting, and boosts prediction power significantly [*Breiman*, 2001; *Elith et al.*, 2006; *Friedman*, 2002b]. TreeNet (Salford Systems, San Diego, CA) is a graphical user interface that can implement this algorithm in either a UNIX or Windows environment. This program allows users to generate command codes in order to create batch files for running multiple models. TreeNet is resistant to over-learning, which can be detected by examining the divergence (or convergence) between the Mean Squared Error of the test and learning samples. Because of TreeNet's ability to handle "messy" data and large number of predictor variables, it has become increasingly popular with ecologists in order to predict species distributions [*Craig and Huettmann*, 2009; *Elith et al.*, 2008; *Ohse et*

al., 2009]. I therefore chose to use a TreeNet which does not require *a priori* assumptions about controllers in a system, allowing us to handle the non-linearity of ecological data [Breiman, 2001; Elith *et al.*, 2008].

Study goals

The development of a global DMS model that will be available for public use will establish a DMS dataset that will allow DMS to be included in a variety of spatial analyses, up to and including global circulation models. Such a model may also be used to help develop species distribution models (e.g. by being linked to prey such as Euphausiids). The development of a distribution model of Fork-Tailed and Leach's Storm-Petrel will help in conservation management of both species. It is hypothesized that DMS will play an important role in determining the distribution of both Storm-Petrel species. The goals of this thesis are to: (1) create a series of monthly DMS models using open access datasets, and to make some inferences on controlling factors in DMS production, and (2) to create models of Fork-Tailed and Leach's Storm-Petrel distribution in the North Pacific, and assess a possible link between Storm-Petrels and DMS.

References

- Andreae, M. O., et al. (1985), Dimethyl Sulfide in the Marine Atmosphere, *Journal of Geophysical Research-Atmospheres*, 90(D7), 2891-2900.
- Bell, T. G., et al. (2006), A Comparison of dimethylsulphide (DMS) data from the Atlantic Meridional Transect (AMT) programme with proposed algorithms for global surface DMS concentrations, *Deep-Sea Research II*, 53, 1720-1735.
- Belviso, S., et al. (2004), Comparison of global climatological maps of sea surface dimethyl sulfide, *Global Biogeochemical Cycles*, 18(3).
- Boersma, P. D., et al. (1980), The Breeding Biology of the Fork-Tailed Storm-Petrel (*Oceanodroma furcata*), *The Auk*, 97, 268-282.
- Boersma, P. D., and M. C. Silva (2001), Fork-tailed Storm-Petrel (*Oceanodroma furcata*), in *The Birds of North America*, No.569, edited by A. P. a. F. Gill, The Academy of Natural Sciences, Philidelphia, PA.
- Breiman, L. (2001), Statistical Modeling: The Two Cultures, *Statistical Science*, 16(3), 199-231.
- Craig, E., and F. Huettmann (2009), Using "Blackbox" Algorithms such as TreeNet and Random Forests for Data-Mining and for Finding Meaningful Patterns, Relationships, and Outliers in Complex Ecological Data: An Overview, an Example Using Golden Eagle Satellite Data and an Outlook for a Promising Future, in *Intelligent Data Analysis: Developing new Methodologies Through Pattern Discovery and Recovery*, edited by H.-F. Wang, Idea Group Inc, Hershey, PA, USA.

- Cushman, S. A. (2010), Space and Time in Ecology: Noise or Fundamental Driver?, in *Spatial Complexity, Informatics, and Wildlife Conservation*, edited by S. A. Cushman and F. Huettmann, Springer, New York, New York.
- DeBose, J. L., et al. (2008), Dimethylsulfoniopropionate as a foraging cue for reef fishes, *Science*, 319(5868), 1356-1356.
- Elith, J., et al. (2006), Novel methods improve prediction of species' distributions from occurrence data, *Ecography*, 29, 129-151.
- Elith, J., et al. (2008), A working guide to boosted regression trees, *Journal of Animal Ecology*, 77, 802-813.
- Friedman, J., et al. (2000), Additive Logistic Regression: A statistical view of boosting, *The Annals of Statistics*, 28(2), 337-407.
- Friedman, J. H. (2001), Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29(5), 1189-1232.
- Friedman, J. H. (2002a), Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38(4), 367-378.
- Friedman, J. H. (2002b), Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38, 367-378.
- Grubb, T. C. (1979), Olfactory guidance of Leach's Storm Petrel to the breeding island, *The Wilson Bulletin*, 91(1), 143-145.
- Huettmann, F., and A. W. Diamond (2001), Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic, *Ecological Modelling*, 141, 261-298.

- Huntington, C. E., et al. (1996), Leach's Storm Petrel (*Oceanodroma leucorhoa*), in *The Birds of North America*, No. 233, edited by A. P. a. F. Gill, The Academy of Natural Sciences, Philadelphia, PA.
- Lovelock, J. E., et al. (1972), Atmospheric dimethylsulfide and the natural sulfur cycle, *Nature*, 237, 452-453.
- Malakoff, D. (1999), Olfaction: Following the Scent of Avian Olfaction, *Science*, 22(5440), 704-705.
- Nevitt, G. A., and K. Haberman (2003), Behavioral attraction of Leach's storm-petrels (*Oceanodroma leucorhoa*) to dimethyl sulfide., *The Journal of Experimental Biology*, 206, 1497-1501.
- Nevitt, G. A., and F. Bonadonna (2005), Seeing the world through the nose of a bird: new developments in the sensory ecology of procellariiform seabirds, *Marine Ecology-Progress Series*, 287, 292-295.
- Ohse, B., et al. (2009), Modeling the distribution of white spruce (*Picea glauca*) for Alaska with high accuracy: an open access role-model for predicting tree species in last remaining wilderness areas, *Polar Biology*, 32(12), 1717-1729.
- Onley, D., and P. Scofield (2007), *Albatrosses, Petrels & Shearwaters of the World*, Princeton University Press, Princeton, New Jersey.
- Raymond, B., and E. J. Woehler (2003), Predicting Seabirds at sea in the Southern Indian Ocean, *Marine Ecology Progress Series*, 263, 275-285.
- Wilbur, H. M. (1969), The Breeding Biology of Leach's Petrel *Oceanodroma leucorhoa*, *The Auk*, 86, 433-442.

Yen, P. P. W., et al. (2004), A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (*Brachyramphus marmoratus*) during the breeding season in coastal British Columbia, Canada, *Ecological Modelling*, 171, 395-413.

Chapter 1. Predicting monthly surface seawater dimethylsulfide (DMS) concentrations on a global scale using a machine learning algorithm (TreeNet)¹

Abstract: In order to deal with the complexities of DMS, a machine learning algorithm (TreeNet) was combined with the framework of ArcGIS to make predictions of DMS concentrations on a global scale. The core of this method is an automated software code. Here we present monthly climatologies of DMS concentrations based on 15 environmental predictor variables downloaded from open access data sources, which is the first time DMS modeling has been based upon such a comprehensive set of input data. We also present the first use of spatial modeling for determining DMS concentrations at sea using a machine learning algorithm. Root Mean Squared Deviation (RMSD) and R squared values were used to determine model performance among a series of random subsets of data extracted from NOAA's Pacific Marine Ecological Laboratory ("Kettle") DMS database. R squared values, broken down by month, ranged from 0.21 to 0.69. Comparison with a global mean DMS climatology matched known hotspots. This research can act as a benchmark for other oceanographic models to further improve our understanding of global ocean systems and its predictions. The use of transparent, open access concepts conforms to best practices held highly by national science organizations such as the International Council for Science, International Polar Year, National Science Foundation and the European Union. The open access concepts, tools and data layers shown here may also be used for further hypothesis testing, and objectively quantify

¹ Humphries, G.R.W, F. Huettmann, C. Deal and D. Atkinson. 2010. Predicting monthly surface seawater dimethylsulfide (DMS) concentrations on a global scale using a machine learning algorithm (TreeNet). Prepared for submission to *Global Biogeochemical Cycles*.

spatial distribution of ocean compounds, allowing for improved global understanding of marine ecosystems and global sustainability.

Keywords: Dimethylsulfide, global, machine learning algorithms, TreeNet, GIS, Open access, model automation

1.1 Introduction

DMS is a marine biogenic compound that is the dominant source of natural sulfur to the atmosphere [Andreae *et al.*, 1985; Lovelock *et al.*, 1972]. The production of DMS begins in the cells of marine phytoplankton as Dimethylsulfoniopropionate (DMSP) which is released into the ocean upon cell senescence/grazing and transferred to the atmosphere where it forms sulfate aerosols via oxidation [Charlson *et al.*, 1987]. Control of the transfer of DMS into the atmosphere is a function of wind speed at the surface of the ocean, turbulence of ocean surface layers, gas diffusivity and seawater temperature [McGillis, 2000]. Once in the atmosphere, DMS oxidizes via reactions with OH and NO₃ radicals to form sulfur dioxide (SO₂), sulfate (SO₄²⁻) and methanesulfonic acid (MSA), which leads to the formation of non sea salt sulfates (NSS-SO₄²⁻) [Bardouki *et al.*, 2003; Yin *et al.*, 1990]. NSS-SO₄²⁻ are aerosols that are found in the marine atmosphere, and are hypothesized to be the primary source of atmospheric sulfur that contribute to cloud formation [Andreae and Crutzen, 1997; Charlson *et al.*, 1987]. Acting as cloud condensation nuclei, NSS-SO₄²⁻ enhance cloud formation and increase cloud albedo, which reduces incoming solar radiation. Cloud albedo can theoretically act as a brake on positive feedbacks that accelerate warming, such as the “ice-albedo” feedback [Charlson *et al.*, 1987]. Full details of the impact of DMS on the atmospheric radiation budget are

not yet well understood [*Charlson et al.*, 1987; *Watson and Liss*, 1998]. The effect of DMS on the radiation budget could link the atmosphere and its operation to those factors affecting marine biological productivity and relative abundance of phytoplankton [*Bopp et al.*, 2003; *Leck et al.*, 1990; *Malin and Kirst*, 1997]. The formation of cloud condensing nuclei is also important when examining the earth's annual rainfall budget because increases or decreases in these aerosols have been shown to have a strong effect on precipitation from clouds [*Nriagu et al.*, 1987]. The effect of human activity on global DMS concentrations could in turn alter trends in precipitation [*Nriagu et al.*, 1987; *Rosenfeld et al.*, 2008].

1.1.1 Current DMS models

Belviso et al. [2004] assessed a series of proposed DMS climatologies [*Anderson et al.*, 2001; *Aumont et al.*, 2002; *Belviso et al.*, 2004; *Chu et al.*, 2003; *Kettle et al.*, 1999; *Kettle and Andreae*, 2000; *Simo and Dachs*, 2002] and found that current DMS models were inaccurate and spatially variable. *Aumont et al.* [2002] was found to be best for the Atlantic Ocean, whereas *Simo and Dachs* [2002] and *Chu et al.* [2003] were better suited for the equatorial Pacific. It was found that none of the previously mentioned models could achieve global r^2 values greater than 0.06. Most of these models were calculated using strictly linear or deterministic techniques, and none have yet examined a truly multivariate or spatial approach that could better apply across the globe.

1.1.2 Spatial modeling with machine learning algorithms

Spatial modeling has been used widely in marine ecology to examine the relationships of environmental variables on the distribution of different species [Elith *et al.*, 2006; Huettmann and Diamond, 2001]. This type of digital science goes hand in hand with traditional *in situ* work via ground-truthing and data collection. Using presence-only data combined with novel methods of modeling such as boosted regression trees and Multivariate Adaptive Regression Splines (MARS), it is possible to improve model accuracy over methods such as linear, generalized additive models or general linear models [Elith *et al.*, 2006]. TreeNet by Salford Systems draws upon regression trees to create a series of predictions using stochastic gradient boosting [Craig and Huettmann, 2009; Friedman, 2002]. This method also uses a type of optimized error testing called v-fold cross validation in order to prevent over-fitting of the model [Friedman, 2002]. This means that one can use a large number of predictor variables to describe the patterns and processes in the system without having to make any *a priori* assumptions about potential importance of predictors [Breiman, 2001; Craig and Huettmann, 2009; Hochachka *et al.*, 2007]. This approach is a fresh and powerful way of obtaining good DMS predictions that are spatially and temporally explicit, and is a method that currently sees little use in oceanography.

1.1.3 Open access data

Open, free access to high quality datasets is essential to assure the repeatability of methodology and also encourages improvement of current analysis techniques, development of theoretical knowledge, and offers some protection against the faulty use

of data [*Fienberg et al.*, 1985]. We followed the Open Access policy, promoted by a variety of different organizations (International Polar Year (IPY), International Council for Science (ICSU), and National Science Foundation (NSF)) because the policy is becoming a best professional practice and a requirement for publication and funding [*Ohse et al.*, 2009].

1.1.4 Study goal

The objective of this study is to develop spatial patterns of monthly DMS concentration for the globe using recently available, online access data sources as applied to a novel, non-linear regression tree algorithm found in the software package, TreeNet. This study employs TreeNet to develop spatial patterns of DMS by relating observational DMS data to environmental predictor layers. The DMS data were obtained from the open access, online dataset available at the Pacific Marine Environmental Laboratory (PMEL) database [*Kettle et al.*, 1999]. Environmental predictor data sets (e.g. solar radiation dose) were selected for inclusion on the basis of current understanding of DMS formation/destruction processes. Uniform spatial overlays were developed from all input data sets (fields) (Table 1). These data fields were used to create monthly climatologies of DMS on a global scale based on the trained TreeNet algorithm. The output from TreeNet allowed us to make inferences on the controlling factors of DMS production/destruction and their interactions with DMS. The analysis tested the hypotheses that each variable plays a significant role in predicting DMS concentrations.

1.2 Methods

1.2.1 Training data

DMS measurements were obtained from the PMEL DMS database [Kettle *et al.*, 1999]. This database consists of approximately 40,500 mixed layer DMS measurements taken around the globe from 1972 to the present. Random subsets were extracted to form the training data by which the models were constructed. Data were filtered by month and projected to World Geodetic System (WGS) 1984.

Many observations in the DMS database were taken at the same location at different depths (down to a maximum of 20 meters). Only the records for the shallowest depths were used when multiple records were available at one location. DMS values greater than 100nM were also filtered out to account for extremely unusual values of DMS measured during algal blooms [Simo and Dachs, 2002]. Appendix B shows the filtered number of measurements for each month, as well as the number of data points used for training the algorithm and assessing the final outputs for each series of model runs.

Random subsets consisting of 20, 60, 70 and 90% (Models 1, 2, 3 and 4 respectively) of the total available data were removed from the PMEL database, which left the remaining 80, 40, 30 and 10% of the data for external assessment of the models. The random subsets were removed using the subset function with no replacement (to avoid pseudo-replication), a function in the R language. Using the freely available Hawth's Tools for ArcGIS 9.x (<http://www.spatial ecology.com/htools/>), spatial overlays of the datasets were performed by extracting the values of each environmental predictor

layer to the same shape file containing the data subset. This resulted in the creation of a comma separated values file with DMS as a response variable, and all the various environmental layers as predictors. All data, including the DMS measurements, were continuous variables. Input data did not require statistical transformations because the non-parametric nature of TreeNet does not require pre-conditioning.

A list of the environmental predictors used to construct the model and their spatial resolutions are listed in Appendix A.

1.2.2 TreeNet algorithm

The complexity of the ocean DMS cycle means that it is important to use an approach which does not require *a priori* assumptions about controllers in the system (boosted regression trees; [Breiman, 2001]). A regression algorithm creates a series of error-minimized regression trees in an iterative fashion to explain the variance in a dataset. This methodology avoids over fitting of data, boosts prediction power significantly, and can handle “messy” or missing data (via data imputation) [Craig and Huettmann, 2009; Elith *et al.*, 2006; Friedman, 2002]. TreeNet by Salford Systems uses the boosted regression tree algorithm, and allows users to generate command codes in order to create batch files for running multiple models. TreeNet can also easily allow users to select different options and settings to perform objective tests on data sets.

Testing was performed to determine which settings yielded the best results by altering parameters such as the number of trees and the number of terminal nodes. Testing focused on the months of January, March and July which represented a low, medium and high number of data points. The number of terminal nodes was varied

between 4 and 10 while keeping the number of trees constant at 500. The least squares error plots (created by cross validation testing) were examined, showing that in all cases, 500 trees were not enough to reach the minimum possible error in predictions. The same tests were performed with the number of trees set at 1000. We found that 10 terminal nodes and 1000 trees, using the Huber-M loss function [Huber, 1964] provided minimum prediction error in all cases.

1.2.3 Scoring and output maps

To create output maps of the models, a regular grid of empty data points was created over the surface of the ocean in ArcGIS 9.3 on a scale of $1^\circ \times 1^\circ$, to match the scale of the predictors that were used. Values for the environmental predictor variables were then calculated at the empty point locations via a spatial overlay in Hawth's Tools. We applied these data to the TreeNet algorithm that was trained in the previous steps ("scoring"), creating a regular grid of DMS predictions. Using the inverse distance weighted (IDW) interpolation tool in ArcGIS 9.3, the regular grid of predictions was smoothed across the surface of the ocean, creating output maps that could then be assessed using independent point measurements from the subset of data not used in the training process. In a similar manner a map of global average DMS concentrations was created. All of the maps were created with metadata in agreement with Federal Geographic Data Committee (FGDC) standards and are available for public access from the author.

1.2.4 Model assessment

Model assessment was performed by way of a “hold out test”, independent from the testing that is performed in the TreeNet software package. A spatial overlay of each model output was performed with the random subset of data not used to build the model, giving columns of predicted vs. observed values. To determine model performance, Root Mean Squared Deviation (RMSD) and R^2 values were calculated. RMSD is a metric that is used to show how well the model predicts the observed values based on a 1:1 slope drawn from the origin. RMSD is found to be one of the most effective methods for conducting an aggregate comparison of observed to predicted values on a continuous scale [Pineiro *et al.*, 2008] over an entire domain of interest.

1.3 Results

1.3.1 Ranking models

Models were ranked using RMSD scores. Subsets of model #4 contained the lowest single run RMSD for all months (except June and September) ranging from 1.24 in October to 19.999 in May. Subsets of model #3 contained the lowest single run RMSD values for June and September and were 15.44 and 2.353 respectively (Table 1.1). R squared values ranged from 0.2146 to 0.6935.

Average RMSD values for each month decreased slightly as we increased the size of the subset used to build the model (except for June and July) (Table 1.2). The highest RMSD values were found in May and June. Average RMSD remained relatively robust between models 1 through 4, which indicated that accuracy does not improve greatly by adding more measurements to build the model (Figure 1.1).

1.3.2 Predictor variable importance

The relative contributions of the various predictor variables for models with the lowest RMSD values, as determined by TreeNet, are listed in Table 1.3. Solar radiation dose (SRD) provided the highest contribution of any predictor variable with an average relative importance across months of 71.92. This was followed by phosphates and salinity (70.10 and 62.15 respectively). Euclidean distances to shore, standard deviation of sea surface temperature and mixed layer depth had a relatively minor contribution throughout all models (40.32, 44.25, and 47.45 respectively).

The partial dependence plots of SRD indicated that concentrations of DMS varied directly with SRD values (i.e. high SRD means high DMS). This data trend was also apparent with phosphates. The partial dependence plots of salinity did not follow an obvious pattern though certain months (i.e. February, March, July, October and November) suggest a range of salinities that were associated with high DMS concentrations (Figure 1.2)

1.3.3 Maps

Monthly maps predicted low concentrations of DMS in the open ocean gyres in all months. High concentrations of DMS in January were mostly located in the southern latitudes, with the highest values around the Antarctic. In February, relative high concentrations of DMS were found further north to mid-southern latitudes, which then decreased into March and April. May, June and July months showed an overall global increase in DMS concentrations, with hot spots of high DMS concentrations (>14 nM) in the northern latitudes, particularly in the Bering Sea, Labrador Sea, and Greenland Sea.

August, September and October showed decreases in global DMS concentrations with patchy hot spots ranging from 4 – 6 nM. In November and December, the model output predicted global increases in DMS which peaked with concentrations of about 12 – 16 nM (Figure 1.3).

An annual mean climatology of DMS shows areas with high concentrations (> 7nM) of DMS in equatorial upwelling regions, the Bering Sea, Grand Banks, west coast of Africa, west coast of Peru, Gulf of Alaska, Greenland Sea, the Falkland Islands, and the Southern Ocean. Open ocean gyres show average annual concentrations of DMS between 1.43 to approximately 2.5 nM (Figure 1.4).

The latitude time plot of sea surface DMS concentrations as created using the National Center for Atmospheric Research command language (NCL) (Figure 1.5) shows averaged concentrations of DMS for each month, by latitude. Average DMS concentrations are highest in the summer time in the northern and southern latitudes (5 – 7 nM), whereas the mid-range latitudes never increased above 4nM. This seems to show a lag of DMS production after the spring phytoplankton bloom where DMS concentrations increase after the peak productivity begins to decline.

1.4 Discussion

This study has developed for the first time, a spatial model of DMS on a global scale using machine learning algorithms. Our goals were to create a series of environmental layers that could be used openly and freely by the general public, to make inferences on important controlling processes in DMS production based on output from TreeNet, and to quantify how well the model results match observations.

1.4.1 Towards a Spatial Ecology of DMS

Within the GIS framework, there are many considerations that must be made when performing a spatial analysis. One of the first and most important issues is that of scale [Huettmann and Diamond, 2006]. The choice of scale is based on several factors including computational power available, input data available and the complexity of the system one wishes to examine. DMS is a globally relevant compound, playing roles in cloud formation [Ayers and Cainey, 2007; Charlson *et al.*, 1987; Johnson and Bell, 2008] and animal attraction [Cunningham *et al.*, 2008; DeBose *et al.*, 2008; Nevitt and Bonadonna, 2005]. This fact, combined with open access to global climatologies of predictor variables, and the access to fast and publicly available computing methods, led to the decision to model DMS distribution at a global scale. Another important issue to discuss is the choice of grain size (resolution). Resolution is important in determining the outcome of many spatial models in some ways, due to the possibility of autocorrelation. That is, if resolution is too coarse, pseudo-replication of data can occur in the process of the overlays, leading to results with no relevant ecological meaning. If the resolution is too fine, it is likely that point measurement errors (due to projection or GPS error) will cause an association with false environmental variables [Guisan *et al.*, 2007]. The ideal situation is for environmental layers of infinitely fine resolution and point data with perfect GPS locations, but this is a situation unlikely when dealing with real data. It is also important to note that climatologies of infinitely fine resolution in space and time are not yet available, and the common resolution used in ocean climatologies tends to be approximately 1° by 1°, and therefore limits the predictions of the DMS model.

1.4.2 Predictor variables

Though this model does not allow *per se* for mechanistic descriptions of how DMS is controlled in the ocean surface, it does allow us to test and examine (via partial dependence plots) possible inferences of the “oceanographic niche” (the specific oceanographic conditions in which DMS is produced). In deterministic models, it is often that variables are chosen *a priori* with a focus on the most parsimonious model, whereas with machine learning algorithms (such as TreeNet), the opposite approach is taken, where an algorithm is used to determine the relationships between predictors and response variables [Craig and Huettmann, 2009; Elith *et al.*, 2008]. The mechanisms of DMS formation are still uncertain [Steinke *et al.*, 2006; Vallina and Simo, 2007], and in fact, as suggested by these results, are not necessarily consistent (i.e. they change from month to month). Therefore it is advantageous to use methods that do not require prior assumptions to make predictions on DMS concentrations.

The results show that SRD, phosphates and salinity play important roles in determining concentrations of DMS. SRD was found to be positively correlated to DMS concentrations because high ultra-violet radiation inhibits DMS consumption and induces oxidative stress (DMS release) in phytoplankton [Vallina and Simo, 2007]. The partial dependence plots of SRD show that in general when SRD values are low, DMS concentrations are also low, thereby supporting the hypothesized link between SRD and DMS. Phosphates have been found to be linked to *Synechococcus* blooms, where good correlation existed between DMSP concentrations and number of cells [Wilson *et al.*, 1998]. Our results support this as well as our partial dependence plots for phosphate that

show that low DMS concentrations and low phosphate concentrations are correlated. Low salinity shock was suggested to affect DMS concentrations by increasing algal DMS contribution, and decreasing bacterial DMSP consumption [Niki *et al.*, 2007]. Our partial dependence plots for salinity seem to suggest a range of salinity in which DMS concentrations are highest. One notable feature is that neither average chlorophyll a or mixed layer depth were considered important predictor variables overall, contrary to a model suggested by *Simo and Dachs* [2002]. The SeaWiFS satellite can detect color changes in the surface of the ocean, which allows for an approximation of chlorophyll a concentrations. This satellite cannot distinguish chlorophyll a concentrations when turbidity in the ocean also produces color. This could possibly explain why chlorophyll a is not a strong predictor of DMS concentrations. SeaWiFS also has a limited range of coverage at any one time of the year. Though we dealt with this via TreeNet's ability to handle missing data via imputation, it is possible that the model has not accurately captured the relationship of chlorophyll a to DMS.

1.4.3 Model optimization

Performance of the output model is strongly affected by the settings used. Therefore for optimal model performance it is advantageous to run through all the different settings until the best model is determined. A full battery of tests on model settings were not performed in this case due to the high accuracy achieved. It would be of use in the future, with high performance computing capabilities, to fully automate batteries of tests in an effort to determine “the best” model settings.

Running models with different subsets of the data allowed us to determine the stability of the model by examining how model performance (RMSD) changed. This also highlights the minimum requirements for DMS observations at sea, feeding directly into possible future science or monitoring missions. It is also important to examine how RMSD changes when using random subsets of the different models to examine variability that may occur due to outliers in the evaluation data where in some cases evaluation data may contain more outliers than others. Figure 2 showed that average RMSD did not change substantially between different model runs. This was an exercise in the ability of TreeNet to remain robust even when the amount of data used to build a model was varied. This also spoke to the relative robustness of the natural relationships that define DMS production (i.e. the “oceanographic niche” of DMS).

The average RMSD within months seemed to remain relatively stable between all model runs, but it varied from month to month in general ranging from 1.39 in October to 28.01 in May. This sort of variation in RMSD was most likely due to the impact of unusually high measurements. The month of May contained a large number of high concentrations on the order of $\sim 80 - 100$ nM. When running a TreeNet model, iterative trees are boosted, that is, error is minimized between each tree by applying a loss function which down-weights outliers. This weighting causes the model to predict at a scale that eliminates such outliers. When performing a spatial overlay for the final model assessment, unusually high concentrations of DMS from the held out subset are still included, and are therefore overlaid on areas that were down-weighted by TreeNet.

The best single runs were found to be in model 4 in 10 of 12 cases, which, when compared to the stability of the mean RMSD values, indicates more variability in model 4. This variability is most likely due to the smaller subset of data (10%) used to evaluate this model. With such a small subset of data being taken randomly at every run, it is likely that the variability between each subset is high, leading to variability in the assessments. The best model runs most probably occur in this model because 90% of the data are being used to train TreeNet.

It is also of interest to examine r^2 values for best model runs to compare to other models that have been evaluated in the past. The r^2 values for our models range from 0.21 to 0.69. Currently, all climatologies of DMS concentration relative to the Kettle database have r^2 values less than 0.06. These comparisons were made to annual, global climatologies which can be difficult to assess when using discrete samples taken from monthly measurements. Comparing discrete measurements to annual climatologies may not accurately capture model performance as there is much seasonal variability in DMS that is not captured by such an analysis. To better examine overall accuracy, it is beneficial to examine monthly model performance to capture the seasonal variability in the data.

1.4.4 Seasonal variability

Monthly models were output in this case to examine monthly changes in DMS concentrations, and to examine if strong relationships between certain environmental predictor variables and DMS existed and remained through all months. This could allow

for a more thorough examination of mechanisms that control DMS concentrations on a global scale.

January shows high concentrations of DMS in the southern hemisphere with the highest concentrations existing along the Antarctic coast. Antarctic sea ice contains large and variable concentrations of DMSP. It is thought that the release of DMSP from the sea ice in the summer months (in this case, the austral summer), account for elevated concentrations in the ocean [Curran *et al.*, 1998; DiTullio *et al.*, 1998; Trevena and Jones, 2006]. It is interesting to note that though sea ice was not an environmental layer included in this model, high DMS concentrations were picked up in the Antarctic, where sea ice algae that contain high intracellular DMSP concentrations are found [DiTullio *et al.*, 1998]. The model for February has high DMS concentrations further north than in January, associated with the movement of solar activity (i.e. SRD), thought to be one controlling factor in DMS production [Vallina and Simo, 2007]. March and April months have overall lower DMS concentrations than any of the surrounding months, which may be associated with the summer paradox that was described in the Sargasso Sea, where higher concentrations of DMS are noted in the summertime, after the spring phytoplankton bloom [Toole *et al.*, 2003]. This could indicate that during the bloom seasons (Spring, Fall), DMS production is slowed on a global scale, perhaps through some chemical or physical forcing based on photolysis rates or perhaps due to sequestration of DMSP in phytoplankton in the spring. It is possible that due to a lag in the dynamics of phytoplankton and the organisms that prey upon them, DMS concentrations do not begin to peak until after the blooms, when DMSP has been released

into the oceans and then converted into DMS. This explanation seems to be supported by elevated concentrations of DMS in May, June and July in the output. May, June and July are also characterized by high concentrations of DMS in northern high-latitudes which persist through August. September, October and November once again show low DMS concentrations again possibly corresponding to the spring/fall bloom, followed by increases in DMS concentrations in December.

The latitude time series plot (Figure 1.5) shows monthly, global (by latitude) averages of DMS concentrations. This plot matches the northern hemisphere of other similar figures that have been created to illustrate seasonal variability in DMS [*Anderson et al.*, 2001; *Belviso et al.*, 2004; *Kettle and Andreae*, 2000; *Simo and Dachs*, 2002]. In all cases, there is an increase in DMS concentrations in the northern hemisphere during the boreal summer months relative to spring (>10 nM). Our model shows an increase in DMS during the austral summer months relative to the austral spring, which is only mirrored by the DMS database [*Kettle and Andreae*, 2000]. This indicates that our model matches the database better than other DMS climatologies.

Two aspects of the latitude time series plot that are not reflected in other models are high concentrations of DMS in both hemispheres during their respective winter months. The mechanisms behind such peaks in DMS concentration require further investigation. It is possible that these patterns are due to blooms during the autumn months, where DMS production lags until the following season when phytoplankton begin to senesce.

Concentrations of DMS in both polar regions are high ($> 7\text{nM}$) when examining the global annual climatology (Figure 4). Sea ice algae contain high concentrations of DMSP [DiTullio *et al.*, 1998] and even though sea ice was not included as a predictor variable, our model still reflects higher concentrations of DMS in polar regions.

1.4.5 Considerations

One of the major limitations of this study concerns the spatial distribution of the data from the “Kettle” database. In any type of data sampling, it is best to draw a random training sample from a parent distribution that is spatially uniform across the entire area of interest, but as DMS measurements in the Kettle database are from ships of opportunity, this analysis suffers from an irregular spatial distribution of measurements. Boreal summer months (May, June, and July) are heavily sampled in the northern seas (North Sea, Bering Sea, etc...) due to increased accessibility and better weather conditions. A similar pattern is found in the Austral summer (December, Jan, Feb), where the southern seas (Ross Sea, Weddell Sea, etc...) are highly sampled. Sample distribution drives the predictions of the model because the model is being trained based on parameters that are found only in those sampling areas that favor particular times of year. In other words, if the systems that control DMS are different from one region to another, we are over-generalizing due to the spatial bias in the data. The model also suffers from any errors that might be associated with the predictor variables, and any errors associated with DMS concentration measurements.

Due to the difficulty and cost of getting DMS measurements at sea and processing samples, and lack of awareness and coordination, it is unlikely that a perfect sampling

distribution will ever be achieved in the near future, and any studies or models will suffer from this bias. Consequently, the models over-generalize when predictions are made at global scales. Using a minimum number of sample points based on studies like ours, it may be possible to increase the effectiveness of expensive at-sea cruises. Faster and more accurate methods of measuring DMS concentration in water are being developed. Currently, a low-cost chemical ionization mass spectrometer has been produced for use in continuous measurements of DMS in seawater [Saltzman *et al.*, 2009]. Using such technologies placed aboard a plethora of different ships will increase sampling distributions, and in time, DMS predictions generated by models such as this will continue to improve.

It is also of importance to note that only 15 environmental predictor variables were used in this model to determine spatial distribution of DMS. It is likely that not all the factors involved in DMS formation were captured. Two main avenues for model improvement are: inclusion of more predictor variables and development of a three-dimensional modeling approach to more explicitly account for the depth aspect. Other model refinements could come via improvements in predictor variables (e.g. via satellite improvement). It would be of further benefit to perform a battery of tests by altering settings and removing certain predictor variables iteratively to determine the best settings for a model of this type.

1.5 Conclusions

DMS in many ways acts as a bridge between biology, oceanography and climatology. DMS, as released from phytoplankton, may act as a foraging cue for many

seabirds [Bonadonna *et al.*, 2006; Cunningham *et al.*, 2008; Nevitt and Bonadonna, 2005], may attract reef fishes [DeBose *et al.*, 2008], and in fact, can trigger search behavior in copepods [Steinke *et al.*, 2006]. DMS may also play a role in cloud formation, and therefore a role in global climate control [Charlson *et al.*, 1987]. The effect of DMS on the atmosphere may have implications in current climate scenarios where DMS has not been previously included.

These models have all been created using an open-access framework, allowing for full transparency, a concept currently adopted by organizations such as IPY, ICSU and NSF. Though TreeNet and ArcGIS do not offer freeware versions, there are various other alternatives that may be used to perform similar, adequate models. TreeNet versions are available for free trials, and ArcGIS follows OpenGIS Consortium formats. In addition, GRASS GIS is a free GIS program that can be run through program R, and packages such as *gbm* and *randomForests* in R offer algorithms that can deal with datasets as complex as the set used in our study.

The models created here seem to perform better than any current DMS climatology, and can form the basis for new environmental layers that can be considered in other oceanographic or climatic studies (e.g. global climate models). The creation of metadata and free access of these models allows for full transparency of science, and ease of import into GIS software and modeling programs. Such methodologies and concepts will help to build collaborations across a variety of fields, and give us a better understanding of global systems and how they interlink.

Acknowledgements

We would like to thank the following individuals and groups for support: The Pacific Marine Ecological Laboratory and its open access policies, the National Oceanographic and Atmospheric Administration, Salford Systems ltd, the EWHALE lab, Dr. Sergio Vallina, Dr. Dave Verbyla, Dr. Hilmar Meier, and Dr. Mykhaylo Golovnya. Funding for this study was provided by the University of Alaska Fairbanks, the International Arctic Research Center, and the Institute of Arctic Biology. Personal thanks to Emily Weiser for her support and thoughtful discussions, as well as to my family for support. Finally we would like to thank all data contributors to the datasets (NOAA, PMEL, etc...) used in this study, without their hard work these studies would not be possible.

References

- Anderson, T. R., et al. (2001), Global fields of sea surface dimethylsulfide predicted from chlorophyll, nutrients and light, *Journal of Marine Systems*, 30, 1 - 20.
- Andreae, M. O., et al. (1985), Dimethyl Sulfide in the Marine Atmosphere, *Journal of Geophysical Research-Atmospheres*, 90(D7), 2891-2900.
- Andreae, M. O., and P. J. Crutzen (1997), Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry, *Science*, 276(5315), 1052-1058.
- Aumont, O., et al. (2002), Dimethylsulfoniopropionate (DMSP) and dimethylsulfide (DMS) sea surface distributions simulated from a global three-dimensional ocean carbon cycle model, *Journal of Geophysical Research|Journal of Geophysical Research*, 107(C4), 4-1-4-21.
- Ayers, G. P., and J. M. Cainey (2007), The CLAW hypothesis: a review of the major developments, *Environmental Chemistry*, 4(6), 366-374.
- Bardouki, H., et al. (2003), Gaseous (DMS, MSA, SO₂, H₂SO₄ and DMSO) and particulate (sulfate and methanesulfonate) sulfur species over the northeastern coast of Crete, *Atmospheric Chemistry and Physics*, 3, 1871-1886.
- Belviso, S., et al. (2004), Assessment of a global climatology of oceanic dimethylsulfide (DMS) concentrations based on SeaWiFS imagery (1998-2001), *Canadian Journal of Fisheries and Aquatic Sciences*, 61(5), 804-816.
- Bonadonna, F., et al. (2006), Evidence that blue petrel, *Halobaena caerulea*, fledglings can detect and orient to dimethyl sulfide, *The Journal of Experimental Biology*, 209, 2165 - 2169.

- Bopp, L., et al. (2003), Potential impact of climate change on marine dimethyl sulfide emissions, *Tellus Series B-Chemical and Physical Meteorology*, 55(1), 11-22.
- Breiman, L. (2001), Statistical Modeling: The Two Cultures, *Statistical Science*, 16(3), 199-231.
- Charlson, R. J., et al. (1987), Oceanic Phytoplankton, Atmospheric Sulfur, Cloud Albedo and Climate, *Nature*, 326(6114), 655-661.
- Chu, S., et al. (2003), Global eddy permitting simulations of surface ocean nitrogen, iron, sulfur cycling, *Chemosphere*, 50, 223-235.
- Craig, E., and F. Huettmann (2009), Using "Blackbox" Algorithms such as TreeNet and Random Forests for Data-Mining and for Finding Meaningful Patterns, Relationships, and Outliers in Complex Ecological Data: An Overview, an Example Using Golden Eagle Satellite Data and an Outlook for a Promising Future, in *Intelligent Data Analysis: Developing new Methodologies Through Pattern Discovery and Recovery*, edited by H.-F. Wang, Idea Group Inc, Hershey, PA, USA.
- Cunningham, G. B., et al. (2008), African penguins (*Spheniscus demersus*) can detect dimethyl sulphide, a prey-related odour, *Journal of Experimental Biology*, 211, 3123 - 3127.
- Curran, M. A. J., et al. (1998), Spatial distribution of dimethylsulfide and dimethylsulfoniopropionate in the Australasian sector of the Southern Ocean., *Journal of Geophysical Research*, 103, 16677 - 16689.

- DeBose, J. L., et al. (2008), Dimethylsulfoniopropionate as a foraging cue for reef fishes, *Science*, 319(5868), 1356-1356.
- DiTullio, G. R., et al. (1998), Dimethylsulfoniopropionate in sea ice algae from the Ross Sea Polyna, *Antarctic Research Series*, 73, 139-146.
- Elith, J., et al. (2006), Novel methods improve prediction of species' distributions from occurrence data, *Ecography*, 29, 129-151.
- Elith, J., et al. (2008), A working guide to boosted regression trees, *Journal of Animal Ecology*, 77, 802-813.
- Fienberg, S. E., et al. (1985), *Sharing Research Data*, National Academies Press.
- Friedman, J. H. (2002), Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38, 367-378.
- Guisan, A., et al. (2007), Sensitivity of predictive species distribution models to change in grain size: insights from an international experiment across five continents, *Diversity and Distributions*, 13, 332-340.
- Halpern, B. S., et al. (2008), A Global Map of Human Impact on Marine Ecosystems, *Science*, 319(948).
- Hochachka, W. M., et al. (2007), Data-Mining Discovery of Pattern and Process in Ecological Systems, *The Journal of Wildlife Management*, 71(7), 2427-2437.
- Huber, P. J. (1964), Robust estimation of a location parameter, *Annals of Math and Statistics*, 35, 73-101.

- Huettmann, F., and A. W. Diamond (2001), Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic, *Ecological Modelling*, 141, 261-298.
- Huettmann, F., and A. W. Diamond (2006), Large-scale effects on the spatial distribution of seabirds in the Northwest Atlantic, *Landscape Ecology*, 21, 1089-1108.
- Johnson, M. T., and T. G. Bell (2008), Coupling between dimethylsulfide emissions and the ocean-atmosphere exchange of ammonia, *Environmental Chemistry*, 5(4), 259-267.
- Kettle, A. J., et al. (1999), A global database of sea surface dimethylsulfide(DMS) measurements and a procedure to predict sea surface DMS as a function of latitude, longitude and month, *Global Biogeochemical cycles*, 13(2), 399-444.
- Kettle, A. J., and M. O. Andreae (2000), Flux of dimethylsulfide from the oceans: A comparison of updated data sets and flux models, *Journal of Geophysical Research*, 106, 26,793 - 726,808.
- Leck, C., et al. (1990), Dimethyl Sulfide in the Baltic Sea - Annual Variability in Relation to Biological-Activity, *Journal of Geophysical Research-Oceans*, 95(C3), 3353-3363.
- Lovelock, J. E., et al. (1972), Atmospheric dimethylsulfide and the natural sulfur cycle, *Nature*, 237, 452-453.
- Malin, G., and G. O. Kirst (1997), Algal production of dimethyl sulfide and its atmospheric role, *Journal of Phycology*, 33(6), 889-896.

- McGillis, W. R., Dacey, J.W.H., Frew, N.M., Bock, E.J., and Nelson, R.K. (2000), Water-air flux of dimethylsulfide, *Journal of Geophysical Research*, 105, 1187-1193.
- Nevitt, G. A., and F. Bonadonna (2005), Seeing the world through the nose of a bird: new developments in the sensory ecology of procellariiform seabirds, *Marine Ecology-Progress Series*, 287, 292-295.
- Niki, T., et al. (2007), Effects of salinity downshock on dimethylsulfide production, *Journal of Oceanography*, 63, 873 - 877.
- Nriagu, J. O., et al. (1987), Biogenic Sulfur and the Acidity of Rainfall in Remote Areas of Canada, *Science*, 237(4819), 1189-1192.
- Ohse, B., et al. (2009), Modeling the distribution of white spruce (*Picea glauca*) for Alaska with high accuracy: an open access role-model for predicting tree species in last remaining wilderness areas, *Polar Biology*, 32(12), 1717-1729.
- Pineiro, G., et al. (2008), How to evaluate models: Observed vs. predicted or predicted vs. observed?, *Ecological Modelling*, 216, 316-322.
- Rosenfeld, D., et al. (2008), Flood or drought: How do aerosols affect precipitation?, *Science*, 321(5894), 1309-1313.
- Saltzman, E. S., et al. (2009), A chemical ionization mass spectrometer for continuous underway shipboard analysis of dimethylsulfide in near-surface seawater, *Ocean Science Discussions*, 6, 1569-1594.
- Simo, R., and J. Dachs (2002), Global ocean emission of dimethylsulfide predicted from biogeophysical data, *Global Biogeochemical Cycles*, 16(4).

- Steinke, M., et al. (2006), Dimethyl sulfide triggers search behavior in copepods, *Limnology and Oceanography*, 51(4), 1925-1930.
- Toole, D. A., et al. (2003), Photolysis and the dimethylsulfide (DMS) summer paradox in the Sargasso Sea, *Limnology and Oceanography*, 48(3), 1088-1100.
- Trevena, A. J., and G. B. Jones (2006), Dimethylsulphide and dimethylsulphoniopropionate in Antarctic sea ice and their release during sea ice melting, *Marine Chemistry*, 98, 210-222.
- Vallina, S. M., and R. Simo (2007), Strong relationship between DMS and the solar radiation dose over the global surface ocean, *Science*, 315, 506-508.
- Watson, A. J., and P. S. Liss (1998), Marine biological controls on climate via the carbon and sulphur geochemical cycles, *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 353(1365), 41-51.
- Wilson, W. H., et al. (1998), Population Dynamics of Phytoplankton and Viruses in a Phosphate-limited Mesocosm and their Effect on DMSP and DMS Production, *Estuarine, Coastal and Shelf Science*, 46, 49-59.
- Yin, F., et al. (1990), Photooxidation of Dimethyl Sulfide and Dimethyl Disulfide. I: Mechanism Development, *Journal of Atmospheric Chemistry*, 11, 309-364.

Figures

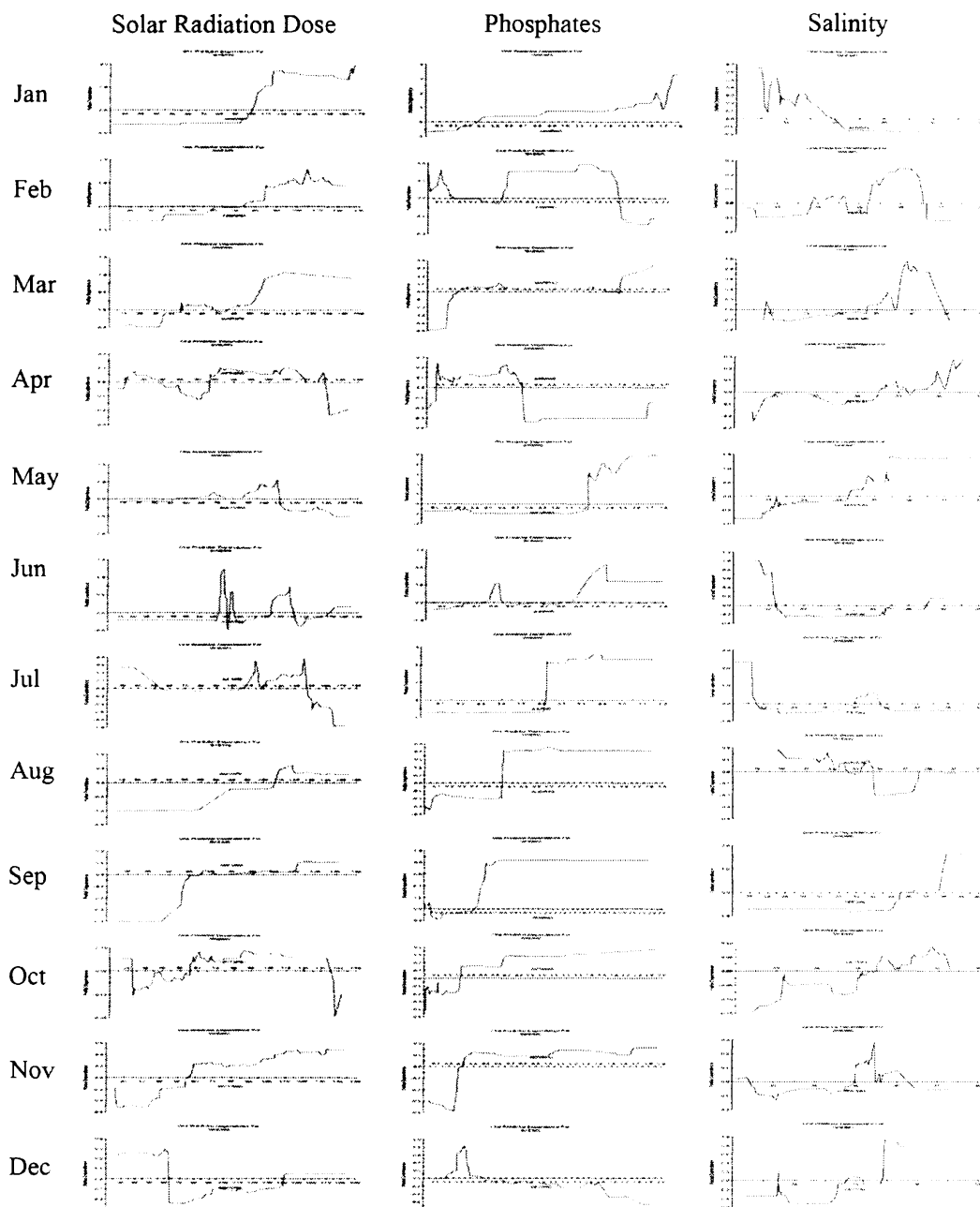


Figure 1.1: Plots of solar radiation dose, phosphates and salinity for all months. The x axis represents the unit value for each variable, and the y axis represents the partial dependence of each variable on sea surface DMS concentrations

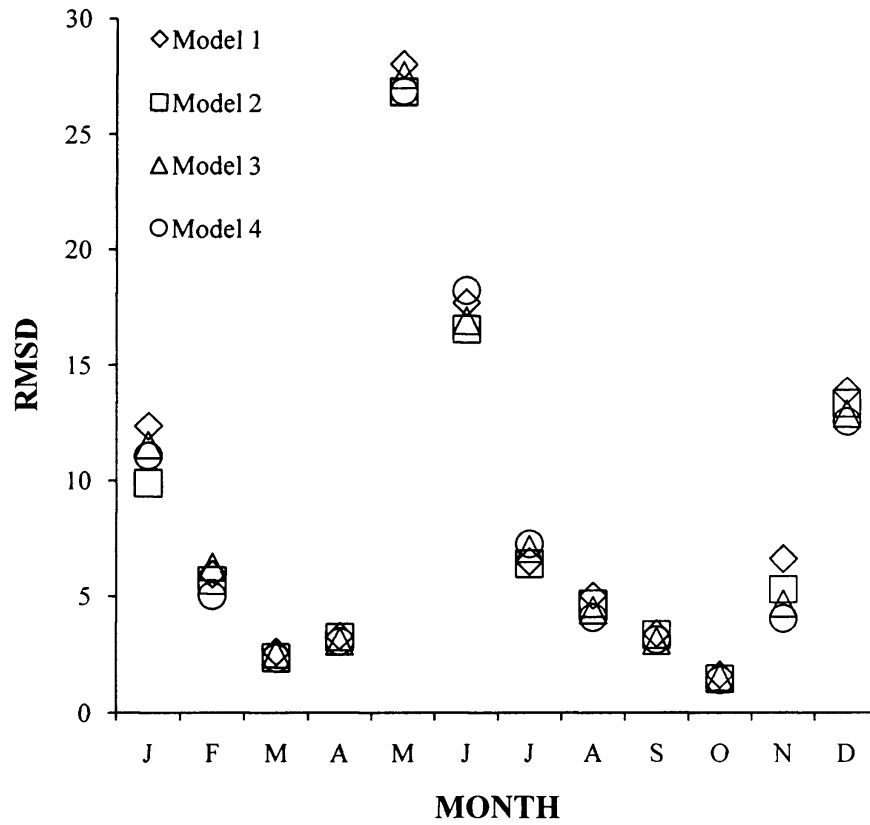


Figure 1.2: Average Root Mean Squared Deviation (RMSD) of all months for Models 1 through 4 (subsets of 20, 60, 70 and 90% respectively)

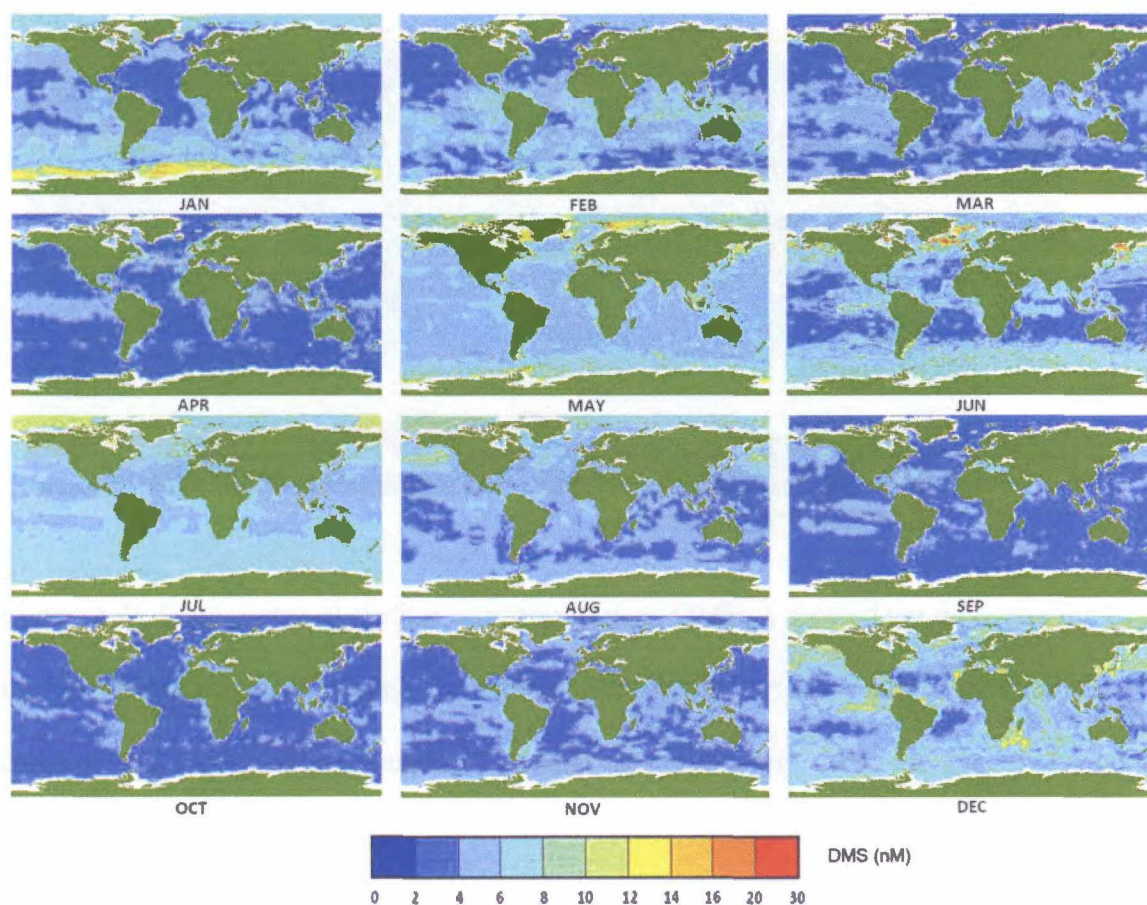


Figure 1.3: Monthly predictions of Dimethylsulfide concentrations (nM) created in ArcMap 9.3, Geographic Coordinate System WGS_1984, from TreeNet predictions based on public DMS and predictor data. Areas in white around coasts are areas of no data due to poor resolution of underlying datasets. Data are available from the author in ESRI grid format.

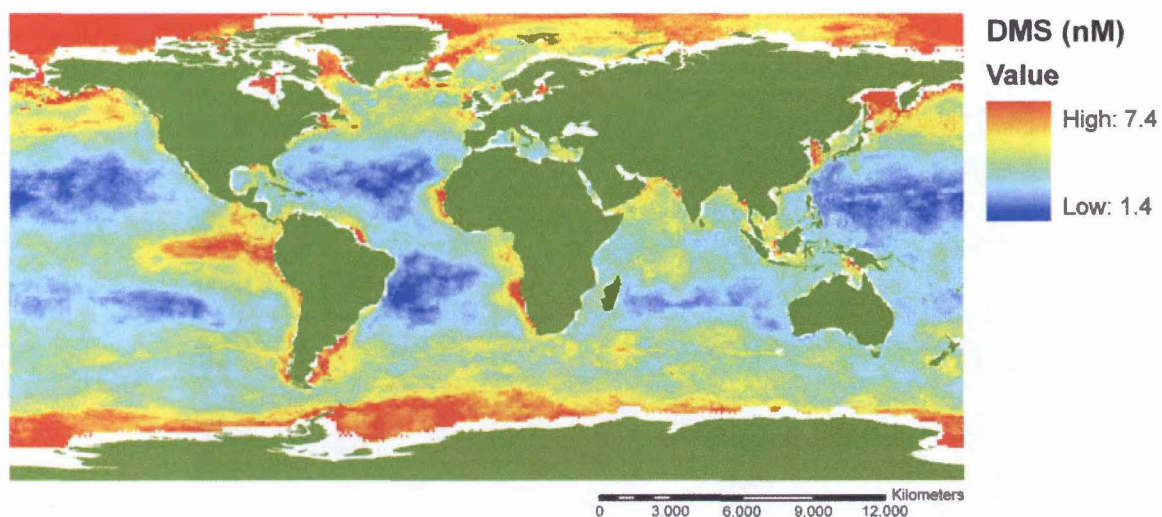


Figure 1.4: Global annual surface concentration of DMS as created via an averaging of the best single run DMS models for each month. White regions around coastlines are areas of no data due to the coarseness of base predictor layers. Data are available from the author in ESRI grid format.

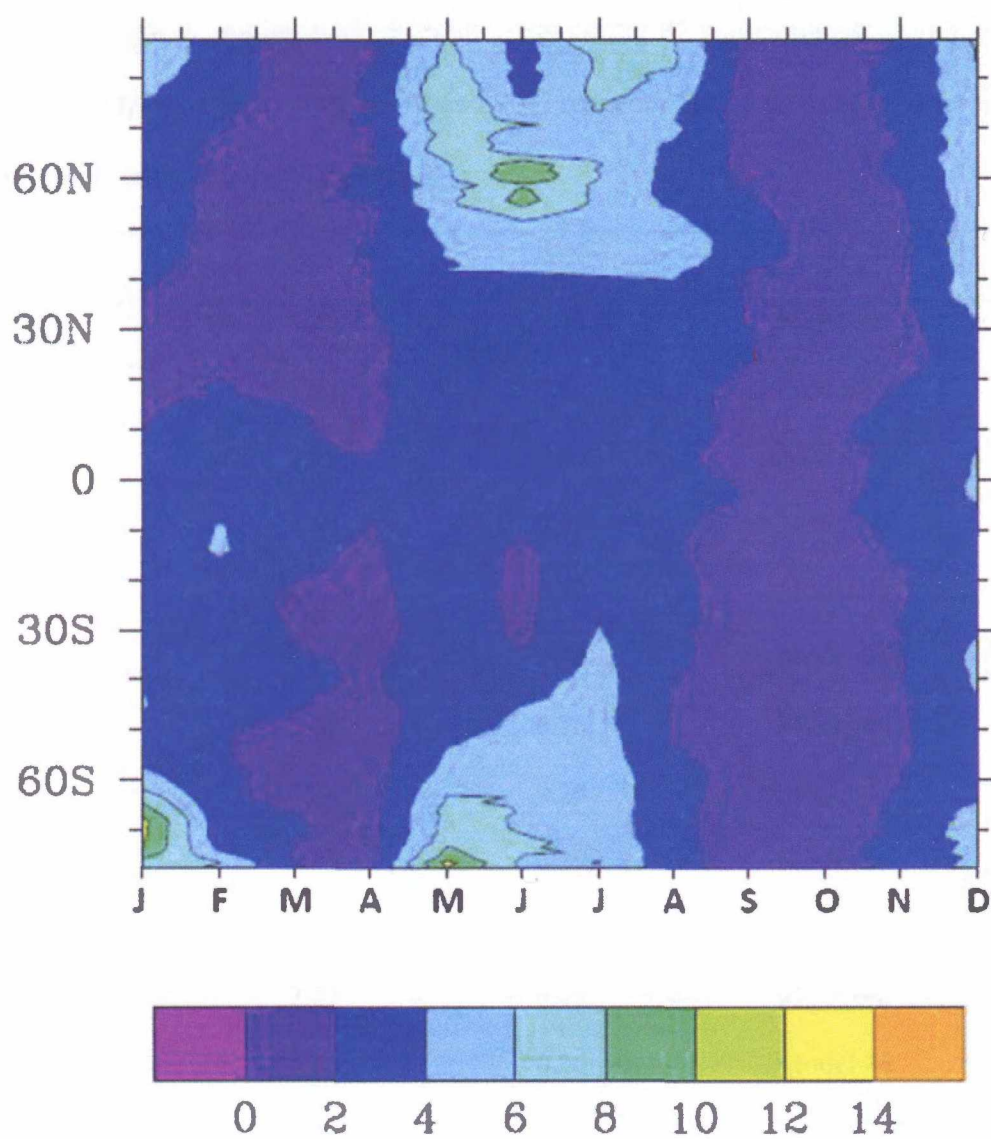


Figure 1.5: Latitude time series plot of predicted sea surface concentration of DMS (nM)

Tables

Table 1.1: Highest ranking RMSD and corresponding R^2 values for each month. Models 3 and 4 refer to training subsets of 70 and 90% respectively. Letters a – e refer to one of 5 random permutations of each subset

Month	RMSD	R^2	Model run
January	4.05	0.31	Model 4e
February	1.41	0.69	Model 4e
March	1.30	0.65	Model 4a
April	1.94	0.62	Model 4a
May	19.99	0.22	Model 4b
June	15.44	0.38	Model 3d
July	5.69	0.26	Model 4a
August	3.36	0.26	Model 4c
September	2.35	0.39	Model 3c
October	1.24	0.61	Model 4e
November	2.02	0.45	Model 4c
December	5.38	0.36	Model 4d

Table 1.2: Average RMSD of 5 randomly drawn subset runs of models 1 – 4

Month	Model 1	Model 2	Model 3	Model 4
January	12.35	9.88	11.52	11.06
February	5.95	5.67	6.26	5.02
March	2.59	2.35	2.50	2.32
April	3.28	3.24	3.06	3.04
May	28.01	26.83	27.57	26.86
June	17.68	16.55	16.91	18.21
July	6.49	6.39	7.02	7.24
August	5.01	4.66	4.38	4.07
September	3.39	3.34	3.07	3.12
October	1.63	1.44	1.52	1.40
November	6.62	5.30	4.67	4.04
December	13.87	13.33	12.88	12.55

Table 1.3: Relative importance of variables for models with lowest RMSD

Variable	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Solar Radiation Dose	100	100	100	61	69	46	56	100	25	65	100	41
Phosphates	97	36	73	100	77	46	100	63	38	88	68	55
Salinity	52	56	91	57	59	24	94	58	100	50	56	49
Dissolve Oxygen	46	40	57	86	44	41	97	71	31	95	70	29
Apparent O ₂ Utilization	58	37	91	66	86	49	36	82	46	57	54	26
Avg Sea Surface Temp	88	34	79	73	71	36	42	71	27	67	56	35
Avg Chlorophyll a	33	44	53	98	100	53	62	67	34	34	38	36
Stdev Chlorophyll a	45	29	59	74	75	53	46	66	39	66	42	36
Silicates	65	35	59	44	55	36	48	45	24	70	46	100
Bathymetry	44	21	55	82	99	55	47	61	17	41	36	31
Human Impact	40	25	61	60	86	49	46	39	25	100	32	27
Nitrates	52	39	43	64	41	100	52	61	29	46	36	25
Mixed Layer Depth	55	28	53	56	53	43	46	48	27	99	33	27
Stdev Sea Surface Temp	42	32	39	32	77	46	69	65	26	38	31	34
Distance to Shore	57	23	38	46	30	19	31	80	20	53	46	43

Chapter 2. Predicted Distribution of Storm-Petrels (*Oceanodroma*) in the North Pacific using Geographic Information Systems (GIS), TreeNet and dimethylsulfide (DMS) concentrations ¹

Abstract: Globally, scientists and managers still lack distribution models of Storm-Petrels despite the availability of large seabird databases (e.g. North Pacific Pelagic Seabird Database). We addressed this gap by using predictive modeling with machine learning software (TreeNet), and GIS to model storm-petrel distribution. Using a variety of environmental predictor variables that included detailed, newly available climatologies of sea surface DMS concentrations, we were able to construct species distribution maps for Fork-Tailed Storm-Petrel (*Oceanodroma furcata*) and Leach's Storm-Petrel (*O. leucorhoa*). We assessed accuracy of the models with area under the receiver operating characteristic (ROC) curve (AUC) values as well as a comparison of predicted distributions to presence-absence data from two opportunistic pelagic surveys performed in summer, 2008. Models including all predictor variables gave AUC values between 0.8 and 0.94, and including only DMS as a predictor gave models with AUC values between 0.75 and 0.84. Examination of the partial dependence plots led to the reinforcement of a possible large-scale Storm-Petrel – DMS link. Using DMS as a predictor variable for modeling still requires further research through additional ground truthing and model

¹ Humphries G, Huettmann F, Deal C, and Atkinson D. 2010. Predicted distribution of Storm-Petrels (*Oceanodroma*) in the North Pacific using Geographic Information Systems (GIS), TreeNet and dimethylsulfide (DMS). Prepared for submission in *Marine Ecology Progress Series*.

testing. The models presented here can be used as a scientific basis for management, and allow for a reduced impact in important wildlife areas.

Keywords: Leach’s Storm-Petrel, Fork-Tailed Storm-Petrel, Dimethylsulfide, machine learning algorithms, GIS, Open access, North Pacific, Bering Sea, seabird, predictive, TreeNet, modeling

2.1 Introduction

Ecological niche modeling has become a popular way of determining species distributions in terrestrial environments, and is important to conservation. The goal of ecological niche modeling is to predict species occurrence based on georeferenced “presence” and “absence” points that correlate to some environmental features. One of the major advantages of this type of modeling is that it enables the fast creation of large-scale models (e.g. global or regional). Traditionally, generalized linear models (GLMs) and generalized additive models (GAMs) have been used to analyze and predict species distributions, but more recently, a variety of more sophisticated algorithms have been developed and applied (Elith et al. 2006, Elith et al. 2008, Craig & Huettmann 2009). Many of these algorithms such as boosted regression trees or Random Forests “learn” the relationship between a target and many different predictor variables GLMs or GAMs require *a priori* assumptions of a data model (Breiman 2001, Elith et al. 2006, Elith et al. 2008). When we take a statistical approach with no *a priori* assumptions regarding what may control the distribution of these species at sea, the modeling process gains a greater degree of flexibility. That is, we can perform a multi-hypothesis test on a variety of predictor variables which define the region of interest, and can from there make

conclusions on the distribution of these species. This method also allows us to make inferences on the relationship between Storm-Petrel distribution and the various predictor variables used.

The at-sea distribution of pelagic seabirds is a difficult issue to address due to many unknowns with respect to the state of the open oceans, and the habits of these species at sea. Historical records of Storm-Petrels in the Bering Sea date back to 1779 during James Cook's expeditions which documented Fork-Tailed Storm-Petrels being taken from the ice edge in the Bering Sea, with no confirmed Leach's Storm-Petrel specimens (Stresemann 1948). During the mid 1900s, both species had been recorded at sea in high numbers east of the Kuril Islands to the Near Islands (Kuroda 1955). During the 1960s, recorded sightings of Fork-Tailed (*Oceanodroma furcata*) and Leach's Storm-Petrel (*O. leucorhoa*) at sea were limited to the central Pacific, Aleutian Islands, and the west coast of North America (Crossin 1974). The NPPSD (Drew & Piatt 2005) contains survey data from the 1970s to the early 2000s, and shows Fork-Tailed Storm-Petrel mostly distributed in the Bering Sea, the Gulf of Alaska and south along the western North American coast, with some sightings off the coast of Japan. The NPPSD shows Leach's Storm-Petrel extending south of, or around, the Aleutian islands, through the Gulf of Alaska, down the western coast of North America, and off the coast of Japan. Winter distribution of Fork-Tailed Storm-Petrel is essentially thought to be limited to its summer distribution in the North Pacific with recorded sightings on the ice edge during the winter time in the Bering Sea (Onley & Scofield 2007). Leach's Storm-Petrel are thought to move further south during the winter months, with the majority of sightings

occurring around Hawaii and in the Eastern Pacific (Huntington et al. 1996, Onley & Scofield 2007).

Fork – Tailed Storm-Petrel nest sympatrically with Leach’s Storm-Petrel during the breeding season in the North Pacific. Birds use deep burrows that can extend down to a meter in depth (Huntington et al. 1996, Boersma & Silva 2001). As they are both nocturnal, with relatively poor eyesight, the development of other senses is important for inter and intra-species interactions. Both species leave their burrows at night to forage at sea for several days before returning to their colonies (Wilbur 1969, Boersma et al. 1980, Malakoff 1999). Like other procellariiforms, these birds have large olfactory bulbs, possibly because a well-developed chemical (olfactory) sense allows these birds to find these foraging areas as well as to find their breeding islands (Grubb 1979).

DMS is a biogenic gas that is one of the dominant sources of sulfur to the atmosphere from the ocean (Lovelock et al. 1972, Andreae et al. 1985). DMS is produced in the cells of marine phytoplankton, and is released into the ocean upon cell senescence or grazing. DMS is then transferred to the atmosphere where it begins to form sulfate aerosols via oxidation and becomes climatically active (Charlson et al. 1987). DMS is linked to areas of high productivity where macroplankton (e.g. Euphausiids) may be located (Andreae & Raemdonck 1983). Currently, it is believed that Procellariids will “smell” DMS to locate foraging areas at sea (Nevitt et al. 1995, Nevitt & Bonadonna 2005). DMS has been shown to be a foraging cue for African penguins (*Spheniscus demersus*) (Cunningham et al. 2008), reef fishes (DeBose et al. 2008), and copepods (Steinke et al. 2006). Fork-Tailed Storm-Petrel and Leach’s Storm-Petrel feed

on Euphausiids (krill) and other planktonic organisms, so it is possible that Storm-Petrels may use DMS as an olfactory cue to find active foraging areas (Nevitt et al. 1995, Nevitt 1999)

In this study, we used ecological niche modeling with the program TreeNet combined with GIS in order to determine the at-sea distribution of Storm-petrels (*Oceanodroma*) in the North Pacific. We hypothesized that we could develop accurate models of Fork-Tailed and Leach's Storm-Petrel using ecological niche modeling techniques, and that this would be facilitated by correlating Storm-Petrels and DMS concentrations.

2.2 Methods

2.2.1 Training data

Training data were accessed from the NPPSD and consisted of 2803 presence and 19144 absence records for Leach's Storm-Petrel, and 6044 presence and 15354 absence records for Fork-Tailed Storm-Petrel. We only used data from large ship surveys in order to remain as consistent as possible with data collection methods. Data from the months of May through August were used in the analysis to coincide with the breeding season when Storm-Petrels are numerous in the North Pacific. Training data could only be obtained for 1974 through 2002 in version 1 of the NPPSD, and were then projected in the WGS 1984 geoid. We organized presence only points by species (Leach's Storm-Petrel and Fork-Tailed Storm-Petrel). The presence points were then paired with randomly generated pseudo-absence points (VanDerWal et al. 2009), and then with

confirmed absences points (Guisan & Zimmermann 2000) (from the NPPSD). This created four categories of training data that would be used to build the models.

2.2.2 Environmental layers

Currently, there are several global models of DMS that exist with r values less than 0.25 (Belviso et al. 2004). For this reason, a new monthly climatology of DMS was developed using spatial modeling techniques in chapter 1 of this thesis. The climatologies for May, June, July, and August ($r = 0.46, 0.61, 0.51, 0.51$ respectively) were averaged to create a summer climatology for use in prediction. DMS is thought to play an important role in many ecological processes (Nevitt & Bonadonna 2005, Bonadonna et al. 2006, Steinke et al. 2006, Stefels et al. 2007, Cunningham et al. 2008, DeBose et al. 2008), and it is therefore assumed that this predictor would be useful in spatial modeling situations, helping to further our understanding of species distributions, and their interactions with the environment.

All environmental layers were projected into WGS 1984, averaged for the months of May through August and clipped to the study area (36 to 66 degrees latitude in the North Pacific). A list of sources for the environmental data layers used in the modeling process is available in Appendix B.

2.2.3 Model construction and assessment

TreeNet is a program that uses boosted regression trees to derive the relationships between a series of predictor variables and a target (response) variable. This algorithm does not require any *a priori* assumptions about the relationships in the data, and therefore allows for great flexibility in model creation (Breiman 2001). Another

advantage of this tool is that over-fitting is avoided due to cross validation of the data, which also boosts prediction power (Friedman 2002, Elith et al. 2006).

The six sets of training data were modeled with all of the environmental variables, all of the environmental variables minus DMS, and then DMS alone. This design was established to facilitate a targeted assessment of the potential role of DMS as a predictor of Storm-Petrel distribution. This was first done by using the informed default settings in TreeNet, which is found to be useful in getting fast, accurate results (Craig & Huettmann 2009). Model settings were then tuned by iteratively increasing the number of trees grown, the number of terminal nodes and the learn-rate. The best models were those with the largest AUC values (Fielding & Bell 1997). Model performance was evaluated by examining the AUC plots generated in TreeNet. These plots are calculated from a subset of the data which is a preferred way to examine the accuracy of models generated by machine learning algorithms (Bradley 1997, Hegel et al. 2009).

For an external assessment using data independent of the NPPSD, two opportunistic surveys were performed in summer, 2008 (July and August). One survey was performed aboard the T/S Oshoro-Marui in the North-Eastern Bering Sea. The second survey was performed aboard the M/V Tiglax between Homer, Alaska and Adak Island, Alaska. All data were collected using distance sampling methods (Thomas et al. 2002) and processed (including metadata) in ArcGIS. These data are available for download from the author.

2.2.4 Data flow

Data processing was performed using ESRI's ArcGIS version 9.3, and Microsoft Excel 2007. The spatial analyst toolset and an open access tool set (Hawth's Tools) were used to perform spatial overlays. Python 2.5 was also used in some of the process to automate conversion from raw data to final maps. Metadata were created in ArcCatalog 9.3 using Federal Geospatial Data Committee (FGDC) standards (<http://www.fgdc.gov/standards>).

2.3 Results

2.3.1 Model ranks

The optimal settings for models using only DMS as a predictor were 1000 trees, 200 terminal nodes with a learn-rate of 0.0001. These settings were also optimal for 2 of the 12 remaining models. Default settings with 1000 trees were found to be the optimal settings for the other 10 models created.

AUC was highest in model 1b for Leach's Storm-Petrel using all predictors but DMS, and also Fork-Tailed Storm-Petrel using all predictor variables. AUC was lowest in model 1a using only DMS as a predictor for both species. AUC scores were similar (<0.069) when pseudo versus confirmed absence points were used. AUC was not substantially different between models with all predictor variables and models using only DMS as a predictor, but dropped dramatically when DMS was not included. Altering default TreeNet settings did not substantially change AUC values, except for models using DMS only as a predictor where AUC increased to between 0.78 and 0.87 (Table 2.1).

Percent of correctly classified presences (PCCP) was highest for Leach's Storm-Petrel in model 2a when using all predictors but DMS. For Fork-Tailed Storm-Petrel, PCCP was highest for model 1b when using all predictor variables. For all models PCCP was similar, when using either all predictor variables or all predictor variables except DMS. Using DMS only, PCCP is low (43 to 59 %) for Fork-Tailed Storm-Petrel in model 1, and higher for model 2 (76 – 80%). For Leach's Storm-Petrel, using DMS only, PCCP is highest in model 2a (Table 2.2).

2.3.2 Partial dependence plots

The partial dependence plots of presence of Fork-Tailed Storm-Petrel, and Leach's Storm-Petrel on DMS are relatively similar (Figure 2.1). For DMS concentrations ranging from 0 to ~10 nM the plots are variable, showing no discernable pattern. After approximately 10nM, the partial dependence plots begin to show increases (Figure 2.1b, Figure 2.1d) or a slow decline (Figure 2.1a, Figure 2.1c). In all cases, there is a sudden increase in the partial dependence plots after approximately 9 - 10 nM DMS representing a possible threshold value.

2.3.3 Distribution maps

Distribution maps for Fork-tailed and Leach's storm-petrels were produced at a resolution of 10km x 10km with an extent of 36 to 66 degrees latitude and 140 to -122 degrees longitude (Figure 2.2). The models show Fork-Tailed Storm-Petrel distribution extending much further north than that of Leach's Storm-Petrel. High relative index of occurrence (RIO) values for both species occurred along the Kuril islands, Aleutian archipelago, Gulf of Alaska and west coast of Canada (0.60 to 0.99 for both species).

The Sea of Okhotsk and most of the Bering Sea had lower RIO values for Leach's Storm-Petrel than Fork-Tailed Storm-Petrel (0.10 to 0.40 for Leach's Storm-Petrel, 0.30 to 0.60 for Fork-Tailed Storm-Petrel), while Leach's Storm-Petrel had higher RIO values between 36 and 45 degrees latitude.

2.3.4 Ground-truthing

Leach's Storm-Petrel sightings during the summer of 2008 were limited to only the M/V Tiglax, south of the Alaska Peninsula and occurred in areas where the model predicts high (> 0.80) RIO for this species. Surveys aboard the T/S Oshoro-maru were north of the Aleutians in the Bering Sea over the Bering shelf, in areas where the model predicts low (< 0.10) RIO. No Leach's Storm-Petrel sightings were recorded west of -164 degrees longitude, in disagreement with high RIO values found in the model (Figure 2.3A). This may be due to the fact that the ship was travelling relatively close to colonies where Storm-Petrels are not found during the day.

Fork-Tailed Storm-Petrel sightings occurred aboard both vessels and in regions where models predicted high (> 0.80) RIO values. Leach's and Fork-Tailed Storm-Petrel sightings south of the Alaska Peninsula overlapped greatly, occurring in the same transect lines. No Fork-Tailed Storm-Petrel sightings were recorded north of 56 degrees latitude, coinciding with areas that were predicted to have low RIO with the exception of 3 transects between 56 and 58 degrees latitude where high RIO was predicted (Figure 2.3B).

When we compare the RIO values of the predicted map to the confirmed presence or absence of both species we see that confirmed presences are associated with high

RIOs. For Leach's Storm-Petrel, we show that high RIO values (~ 0.80) are associated with confirmed presences, where low RIO values (~ 0.20) are associated with confirmed absences. We also show a similar trend for RIO values for Fork-Tailed Storm-Petrel with confirmed presences and confirmed absences being associated with RIO values of ~ 0.60 and ~ 0.45 respectively. There is heavy overlap between presence and absence of Fork-Tailed Storm-Petrel, and no overlap between presence and absence of Leach's Storm-Petrel RIO values (Figure 2.4).

2.4 Discussion

In this study, we were able to create accurate distribution models for both Leach's and Fork-Tailed Storm-Petrels in the North Pacific; furthermore, we were able to confirm a possible link between Storm-Petrels and DMS. By working in the framework of GIS, we were also able to create distribution maps of RIO for both species of Storm-Petrel that are available from the author.

2.4.1 Spatial considerations

In the field of landscape ecology, one of the top priorities for study is that of scale, for the reason that many species-environment associations can change based on the scale chosen (Schneider & Piatt 1986, Huettmann & Diamond 2006). The scale refers to the grain size and extent of the data being used in the analysis. For this study, we chose the extent to be between 36 degrees and 66 degrees North latitude, comprising the northern halves of the North Pacific Transition Zone Province and the Kuroshio Current Province, where 36 degrees lie between the Subtropical and Subarctic fronts, and 66 degrees is the Arctic circle (Longhurst 1998). The NPPSD was further clipped from this study extent

due to the coarse nature of some of the data layers (i.e. DMS and Salinity), which led the final models to have an extent stretching only to 63 degrees north. Because the data were clipped to this extent, there is some possibility that biases might exist by excluding these presence or absence points by limiting predictor variables such as distance to shore. Of the 8021 presence points for Fork-Tailed Storm-Petrel, 6055 data points occurred within the extent of the analysis. Most of the records not included in the analysis were in Prince William Sound, Cook Inlet and in the Shelikof Strait. With approximately 75% of the data still included in the analysis, it is likely the sample was still representative with no records from the Aleutians, or around Southeast Alaska being excluded. Of the 3138 presence points for Leach's Storm-Petrel, only 5 records were not included in the analysis. Biases in the data may occur as well due to biased sampling (sampling one area multiple times, and not sampling some areas at all), or due to resolution of the presented data.

The resolution of underlying datasets can also influence the extent of autocorrelation in the data. Autocorrelation occurs when ecological processes may be expressed as a function of spatial location, or time between samples (e.g. how closely samples are correlated to one another) (Cushman 2009). Spatial autocorrelation in a species dataset (for example, how closely birds flock together in space), can influence apparent relationships between environmental variables, (Huettmann & Diamond 2006). Though we did not correct for spatial autocorrelation (e.g. by binning, as per Huettmann and Diamond (2006)), TreeNet has the ability to deal with “messy” data and still produce accurate models (Friedman 2002, Elith et al. 2008, Craig & Huettmann 2009). It is

therefore assumed that our models are still representative. To address potential temporal autocorrelation in this study, we used summer averages for all predictor variables with the exception of the human impact layer, bathymetry, and distance to shore because they are static. Summer seabird observations from May, June, July, and August were also combined, and therefore our models do not take into account the possibility of monthly shifts in distribution. It is possible that this could affect the outcome of our analysis in that monthly (or possibly daily) shifts in predictors like DMS, may affect the location of seabirds.

2.4.2 DMS as a predictor

We ran these models using only DMS specifically to investigate its potential as a predictor variable. Above 10 nM of DMS, the RIO of Leach's and Fork-Tailed Storm-Petrel stabilizes or begins to increase. This indicates that even with temporally broad scaled models, a relationship between DMS concentrations and Storm-Petrel distribution is evident. Accuracies of models using just DMS as a predictor variable were also boosted to approximately 0.80 to 0.87 (based on AUC) with correctly classified presences between 76 and 92%.

According to TreeNet, DMS was not an important predictor variable in models where all predictor variables were included (Appendix A). The Random Forest algorithm is known to depress variable importance of correlated variables (Grömping 2009), and due to its relatedness to the TreeNet algorithm with respect to both building regression trees, it is possible TreeNet also exhibits the same trait. The DMS model was created

using all of the same predictor variables as used in this analysis; therefore, there may have been some effect in the algorithm pushing DMS to a lower importance.

Other possible sources of error could arise from the short lifetime of DMS (Yang & Tsunogai 2005), or because the model being used does not possess a fine-enough temporal resolution for patterns to be detected. Moreover, though this DMS model is more accurate than any other DMS climatologies, it still requires more fine-tuning.

2.4.3 Ground-truthing

Through AUC values, we can get some perspective on how well the model predicts the data, but an independent dataset is usually required for testing true model accuracy. We were limited in this study by only having 2 opportunistic surveys for one summer. Based on the data collected during these surveys, it was found that the model performed well (i.e. no overlap in error bars of mean RIO between presence and absence) for prediction of occurrence of Leach's Storm-Petrel. The majority of the data were taken north of the Aleutian Islands, where our models show Leach's Storm-Petrel distributed south of the Aleutians, which also agrees with other accounts (Huntington et al. 1996, Onley & Scofield 2007). Most of the survey data were taken within the known distribution of Fork-Tailed Storm-Petrel (Boersma & Silva 2001, Onley & Scofield 2007). Mean RIO for Fork-Tailed Storm-Petrel presences was higher than absences; however we found heavy overlap in the mean RIO values indicating that there were some areas of high RIO where we did not detect Fork-Tailed Storm-Petrel, and areas of low RIO where we did. It is of importance here to note that simply because we were in areas of high RIO (based on the model), it does not guarantee a Storm-Petrel sighting (e.g. low

detectability due to weather or fatigue). As well, in areas with low RIO, it is still possible to sight birds. The model may also be suffering in some areas due to temporal or spatial errors, which may also be reason for the patterns in mean RIO for Fork-Tailed Storm-Petrel.

2.4.4 Implications for Storm-Petrel management

Currently, both Leach's and Fork-Tailed Storm-Petrel have a widespread and abundant population, numbering in the millions, and are therefore not considered threatened species (Huntington et al. 1996, Boersma & Silva 2001). The most significant threats to these birds include introduced mammalian predators in breeding colonies (Boersma & Groom 1993), ingestion of plastics (Blight & Burger 1997), and collision with man-made structures (Bent 1922, Reed et al. 1985). Fragmentation of the species distributions may also occur due to large-scale climate events. The models presented in this study allow for further examination of how storm-petrels interact with areas of heavy human influence or areas known for high plastic concentrations. If we focus on applications of these models, it may be possible to buffer these species' populations against other, potentially more serious threats (e.g. climate change, shipping, and oil spills).

Wilson's Storm-Petrel (*Oceanites oceanicus*) are affected negatively by environmental conditions, and their populations are related to food (krill) availability (Quillfeldt 2001). Food sources (e.g. krill) for Leach's and Fork-Tailed Storm-Petrel may be heavily influenced by local or large-scale climatic events (Hays et al. 2005). Krill and copepods have been linked to DMS (Daly & DiTullio 1993, Steinke et al. 2006).

Because DMS can be linked to climate (Charlson et al. 1987), it follows that we can predict future DMS patterns from future climate scenarios (e.g. IPCC). Due to the possible Storm-Petrel - prey - DMS link, it may become feasible to model future Storm-Petrel distribution based on these DMS patterns.

The presented models of Storm-Petrel distribution show a seascape that is fairly ubiquitous. In terrestrial environments, it is of great interest to map and quantify movement corridors (landscape connectivity) in order to understand species distributions (Cushman et al. 2010). This type of landscape quantification may also be applied to seascapes. Programs like Circuitscape, and Fragstats may be used to analyze seascape features to show how Storm-Petrels currently use the ocean. Forecasting models can give managers and scientists some idea on how Storm-Petrel habitat utilization will change with ecological / environmental conditions.

2.4.5 Conclusions and future work

These predictive models are potentially valuable for policy and management decisions, we therefore encourage further work to improve them and models like them (e.g. ground-truthing). Once more recent data become available, the accuracy of these models could be further evaluated using the Boyce index (Boyce et al. 2002). Further efforts must also be undertaken to improve the accuracy of the underlying environmental models used to predict these distributions. Errors in the current seabird database must also be improved to obtain clean training data for model building. Currently, DMS predictions may be too coarse spatially and temporally to accurately define how Storm-Petrels are using this compound to locate foraging areas on a small scale. The

importance of individual predictive variables could be further assessed by removing predictor variables at random and testing for model improvement. Model improvement could be fine-tuned by altering all algorithm settings to determine which models perform best. These tests can be automated in the R (www.r-project.org/) or Python (www.python.org/) programming languages. Addressing autocorrelation in the models may also elucidate the relationships between the various predictor variables and the distributions of these species. The more accurately we can predict current seabird distributions, the more accurately we will be able to predict their responses to future climate change,

In this study, open access data and tools were used to construct our Storm-Petrel models in a GIS framework. In order for similar work to continue, we must advocate freely accessed and well described datasets and tools with high quality metadata. Freely accessible and certified software will also be critical to making these methods available for widespread use. Though we operated primarily in ArcGIS for spatial analyses, other free GIS packages exist (e.g. GRASS GIS, Open GIS consortium, Open Modeler), and open-access statistical languages such as R can be used for statistical analysis. As the tools to develop predictive models become more accessible, predictive models of species distribution will become more available and widespread in contributing to management and conservation decisions

Acknowledgements

We would like to thank the following individuals and groups for support: The Pacific Marine Ecological Laboratory and its open access policies, the North Pacific

Pelagic Seabird Database, the National Oceanographic and Atmospheric Administration, Salford Systems ltd, the EWHALE lab, Gary Drew, Dr. Sergio Vallina, Dr. Dave Verbyla, Dr. Hilmar Meier, and Dr. Mykhayl Golovnya. Funding for this study was provided by the University of Alaska Fairbanks, the International Arctic Research Center, and the Institute of Arctic Biology. Personal thanks to Emily Weiser for her support and thoughtful discussions, as well as to my family for support. Finally we would like to thank all data contributors to the datasets (e.g. NPPSD, PMEL, NOAA) used in this study, without their hard work these studies would not be possible.

Literature Cited

- Andreae MO, Ferek RJ, Bermond F, Byrd KP, Engstrom RT, Hardin S, Houmère PD, Lemarrec F, Raemdonck H, Chatfield RB (1985) Dimethyl Sulfide in the Marine Atmosphere. *Journal of Geophysical Research-Atmospheres* 90:2891-2900
- Andreae MO, Raemdonck H (1983) Dimethyl Sulfide in the Surface Ocean and the Marine Atmosphere - a Global View. *Science* 221:744-747
- Belviso S, Bopp L, Moulin C, Orr JC, Anderson TR, Aumont O, Chu S, Elliott S, Maltrud ME, Simo R (2004) Comparison of global climatological maps of sea surface dimethyl sulfide. *Global Biogeochemical Cycles* 18
- Bent AC (1922) Life histories of North American petrels and pelicans and their allies. US National Museum Bulletin 121
- Blight LK, Burger AE (1997) Occurrence of plastic particles in seabirds from the eastern north Pacific. *Marine Pollution Bulletin* 34:323-325
- Boersma PD, Groom MJ (1993) Conservation of storm-petrels in the North Pacific. In: Vermeer K, Briggs KT, Morgan KH, Siegel-Causey D (eds) The status, ecology, and conservation of marine birds in the North Pacific. Canadian Wildlife Service Special Publication, Ottawa, ON, p 112-121
- Boersma PD, Silva MC (2001) Fork-tailed Storm-Petrel (*Oceanodroma furcata*). In: Gill APaF (ed) The Birds of North America, No569. The Academy of Natural Sciences, Philadelphia, PA

- Boersma PD, Wheelwright NT, Nerini MK, Wheelwright ES (1980) The Breeding Biology of the Fork-Tailed Storm-Petrel (*Oceanodroma furcata*). *The Auk* 97:268-282
- Bonadonna F, Caro S, Jouventin P, Nevitt GA (2006) Evidence that blue petrel, *Halobaena caerulea*, fledglings can detect and orient to dimethyl sulfide. *The Journal of Experimental Biology* 209:2165 - 2169
- Boyce MS, Vernier PR, Nielson SE, Schmiegelow FKA (2002) Evaluating Resource Selection Functions. *Ecological Modelling* 157:281-300
- Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30:1145-1159
- Breiman L (2001) Statistical Modeling: The Two Cultures. *Statistical Science* 16:199-231
- Charlson RJ, Lovelock JE, Andreae MO, Warren SG (1987) Oceanic Phytoplankton, Atmospheric Sulfur, Cloud Albedo and Climate. *Nature* 326:655-661
- Craig E, Huettmann F (2009) Using "Blackbox" Algorithms such as TreeNet and Random Forests for Data-Mining and for Finding Meaningful Patterns, Relationships, and Outliers in Complex Ecological Data: An Overview, an Example Using Golden Eagle Satellite Data and an Outlook for a Promising Future. In: Wang H-F (ed) *Intelligent Data Analysis: Developping new Methodologies Through Pattern Discovery and Recovery*. Idea Group Inc, Hershey, PA, USA

- Crossin RS (1974) The Storm Petrels (Hydrobatidae). In: King WB (ed) Pelagic Studies of seabirds in the central and eastern Pacific Ocean, Vol 158. Smithsonian Contributions to Zoology
- Cunningham GB, Strauss V, Ryan PG (2008) African penguins (*Spheniscus demersus*) can detect dimethyl sulphide, a prey-related odour. *Journal of Experimental Biology* 211:3123 - 3127
- Cushman SA (2009) Animal Movement Data: GPS Telemetry, Autocorrelation and the Need for Path-Level Analysis. In: Cushman SA, Huettmann F (eds) *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer, New York, New York, p 131-150
- Cushman SA, Chase M, Griffen C (2010) Mapping Landscape Resistance to Identify Corridors and Barriers for Elephant Movement in Southern Africa. In: Cushman SA, Huettmann F (eds) *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer, New York, New York
- Daly KL, DiTullio GR (1993) Biogenic production of dimethylsulfide: Krill grazing. *Antarctic Journal of the United States* 28:140
- DeBose JL, Lema SC, Nevitt GA (2008) Dimethylsulfoniopropionate as a Foraging Cue for Reef Fishes. *Science* 319:1356
- Drew GS, Piatt JF (2005) North Pacific Pelagic Seabird Database) NPPSD: Compiling Datasets and Creating an Archive, Accessible Databas, and Pelagic Seabird Atlas, U.S. Geological Survey, Anchorage, Alaska

- Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Soberon J, Williams S, Wisz MS, Zimmermann NE (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129-151
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77:802-813
- Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absences models. *Environmental Conservation* 24:38-49
- Friedman JH (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38:367-378
- Grömping U (2009) Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician* 63:308-318
- Grubb TC (1979) Olfactory guidance of Leach's Storm Petrel to the breeding island. *The Wilson Bulletin* 91:143-145
- Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135:147-186
- Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, Bruno JF, Casey KS, Ebert C, Fox HE, Fujita R, Heinemann D, Lenihan HS, Madin EMP, Perry MT, Selig ER, Spalding M, Steneck R, Watson R (2008) A Global Map of Human Impact on Marine Ecosystems. *Science* 319

- Hays GC, Richardson AJ, Robinson C (2005) Climate change and marine plankton. *Trends in Ecology and Evolution* 20:337-344
- Hegel TM, Cushman SA, Evans J, Huettmann F (2009) Current State of the Art for Statistical Modelling of Species Distributions. In: Cushman SA, Huettmann F (eds) *Spatial Complexity, Informatics, and Wildlife Conservation*. Springer, New York, New York, p 273-309
- Huettmann F, Diamond AW (2006) Large-scale effects on the spatial distribution of seabirds in the Northwest Atlantic. *Landscape Ecology* 21:1089-1108
- Huntington CE, Butler RG, Mauck RA (1996) Leach's Storm Petrel (*Oceanodroma leucorhoa*). In: Gill APaF (ed) *The Birds of North America*, No 233. The Academy of Natural Sciences, Philadelphia, PA
- Kuroda N (1955) Observation on pelagic birds of the northwest Pacific. *Condor* 57:290-300
- Longhurst AR (1998) *Ecological Geography of the Sea*, Vol. Academic Press, San Diego, California
- Lovelock JE, Maggs RJ, Rasmussen RA (1972) Atmospheric dimethylsulfide and the natural sulfur cycle. *Nature* 237:452-453
- Malakoff D (1999) Olfaction: Following the Scent of Avian Olfaction. *Science* 22:704-705
- Nevitt G (1999) Olfactory foraging in Antarctic seabirds: a species-specific attraction to krill odors. *Marine Ecology Progress Series* 177:235-241

- Nevitt GA, Bonadonna F (2005) Seeing the world through the nose of a bird: new developments in the sensory ecology of procellariiform seabirds. *Marine Ecology-Progress Series* 287:292-295
- Nevitt GA, Veit RR, Kareiva P (1995) Dimethyl sulphide as a foraging cue for Antarctic procellariiform seabirds. *Nature* 376:680-682
- Onley D, Scofield P (2007) *Albatrosses, Petrels & Shearwaters of the World*, Vol. Princeton University Press, Princeton, New Jersey
- Quillfeldt P (2001) Variation in breeding success in Wilson's storm petrels: influence of environmental factors. *Antarctic Science* 13:400-409
- Reed JR, Sincock JL, Hailman JP (1985) Light Attraction in Endangered Procellariiform Birds: Reduction by Shielding Upward Radiation. *The Auk* 102:377-383
- Schneider DC, Piatt JF (1986) Scale-dependent correlation of seabirds with schooling fish in a coastal ecosystem. *Marine Ecology Progress Series* 32:237-246
- Stefels J, Steinke M, Turner S, Malin G, Belviso S (2007) Environmental constraints on the production and removal of the climatically active gas dimethylsulphide (DMS) and implications for ecosystem modelling. *Biogeochemistry* 83:245-274
- Steinke M, Stefels J, Stamhuis E (2006) Dimethyl sulfide triggers search behavior in copepods. *Limnology and Oceanography* 51:1925-1930
- Stresemann E (1948) Birds collected in the North Pacific area during Capt. James Cook's last voyage (1778 and 1779). *Ibis* 91:244-255

- Thomas L, Buckland ST, Burnham KP, Anderson DR, Laake JL, Borchers DL, Strindberg S (2002) Distance Sampling. *Encyclopedia of Environmetrics* 1:544-552
- VanDerWal J, Shoo LP, Graham C, Williams SE (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling* 220:589-594
- Wilbur HM (1969) The Breeding Biology of Leach's Petrel *Oceanodroma leucorhoa*. *The Auk* 86:433-442
- Yang G-P, Tsunogai S (2005) Biogeochemistry of dimethylsulfide (DMS) and dimethylsulfoniopropionate (DMSP) in the surface microlayer of the western North Pacific. *Deep-Sea Research I* 52:553-567

Figures

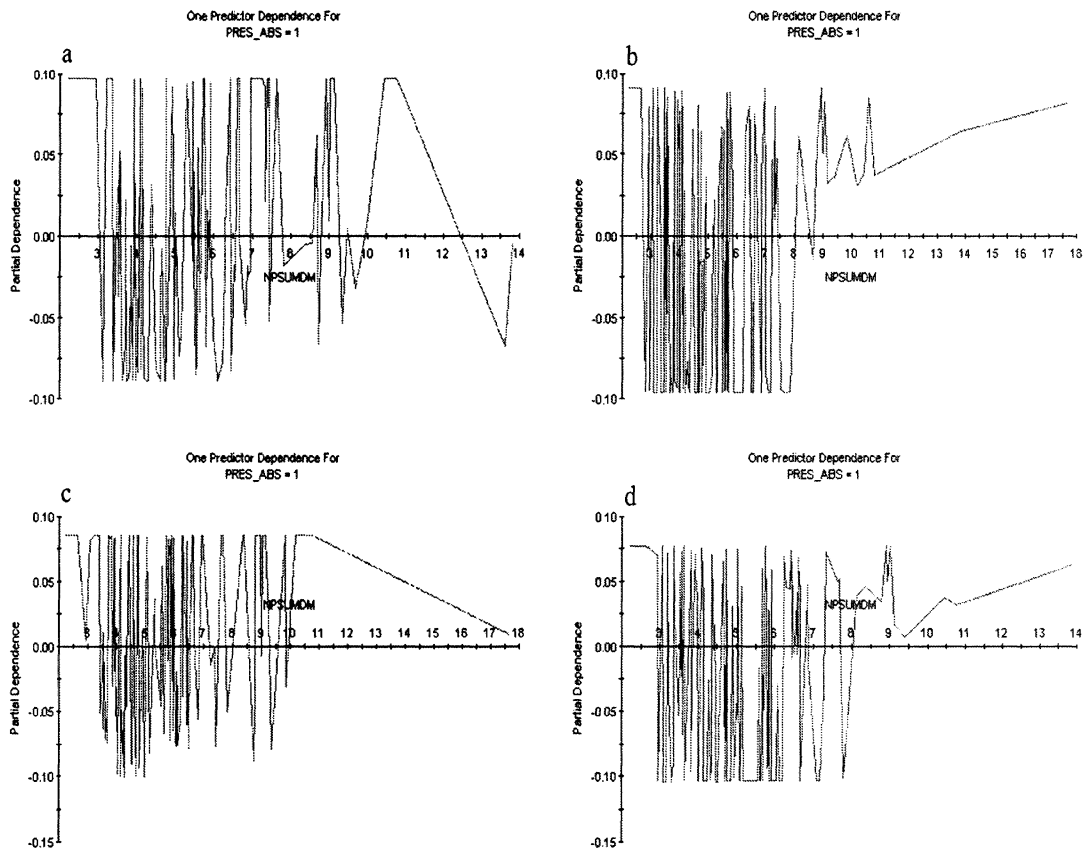


Figure 2.1: Partial dependence plots of Storm-Petrel distribution on DMS concentration (nM) for models using only DMS: a) Leach's Storm-Petrel model 2a b) Leach's Storm-Petrel model 2b c) Fork-Tailed Storm-Petrel model 2a d) Fork-Tailed Storm-Petrel model 2b

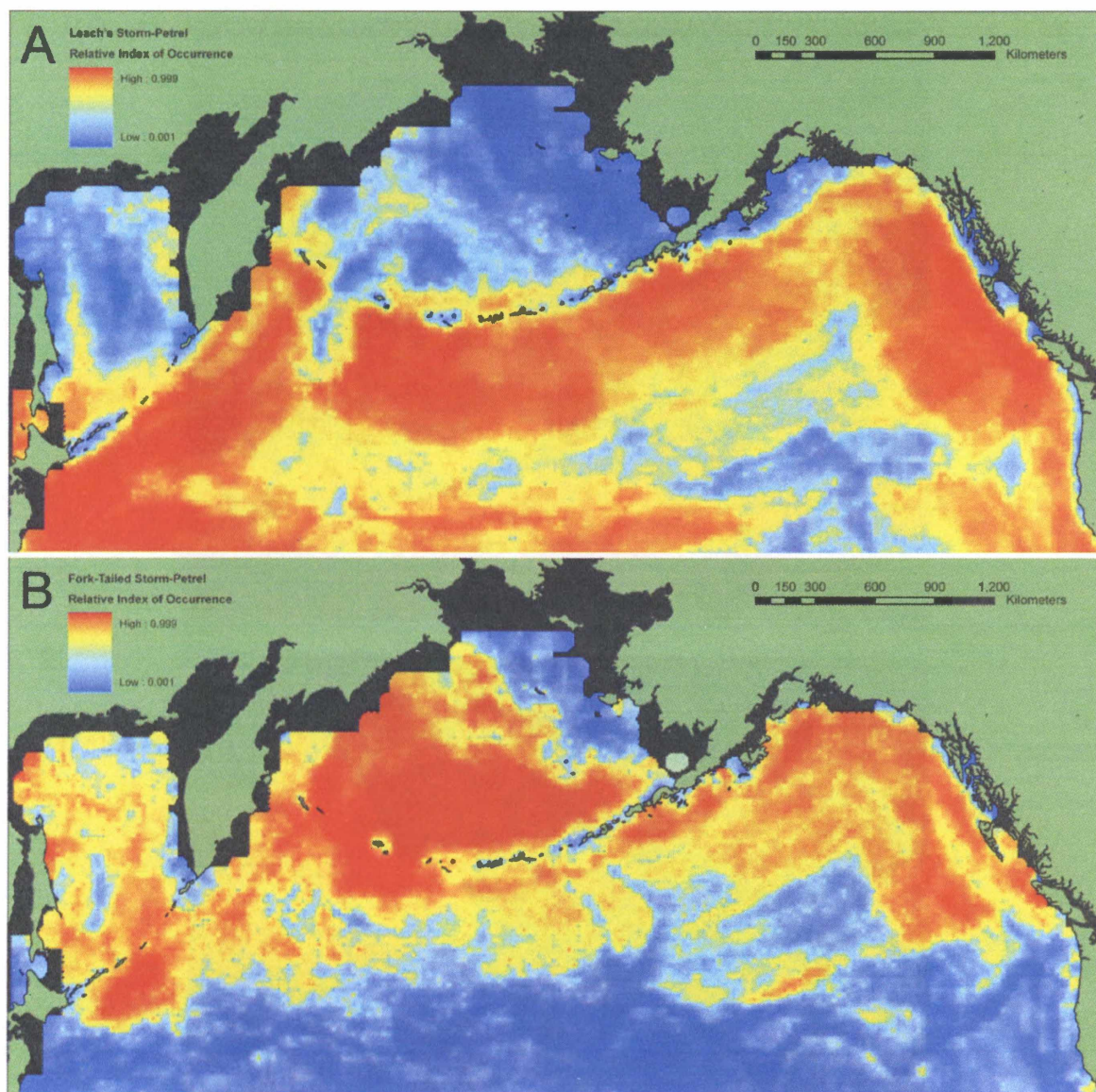


Figure 2.2: Maps of relative index of occurrence (RIO) of Leach's and Fork-Tailed Storm-Petrel as produced by TreeNet for the summer breeding season in the North Pacific using top ranked models with confirmed absence points

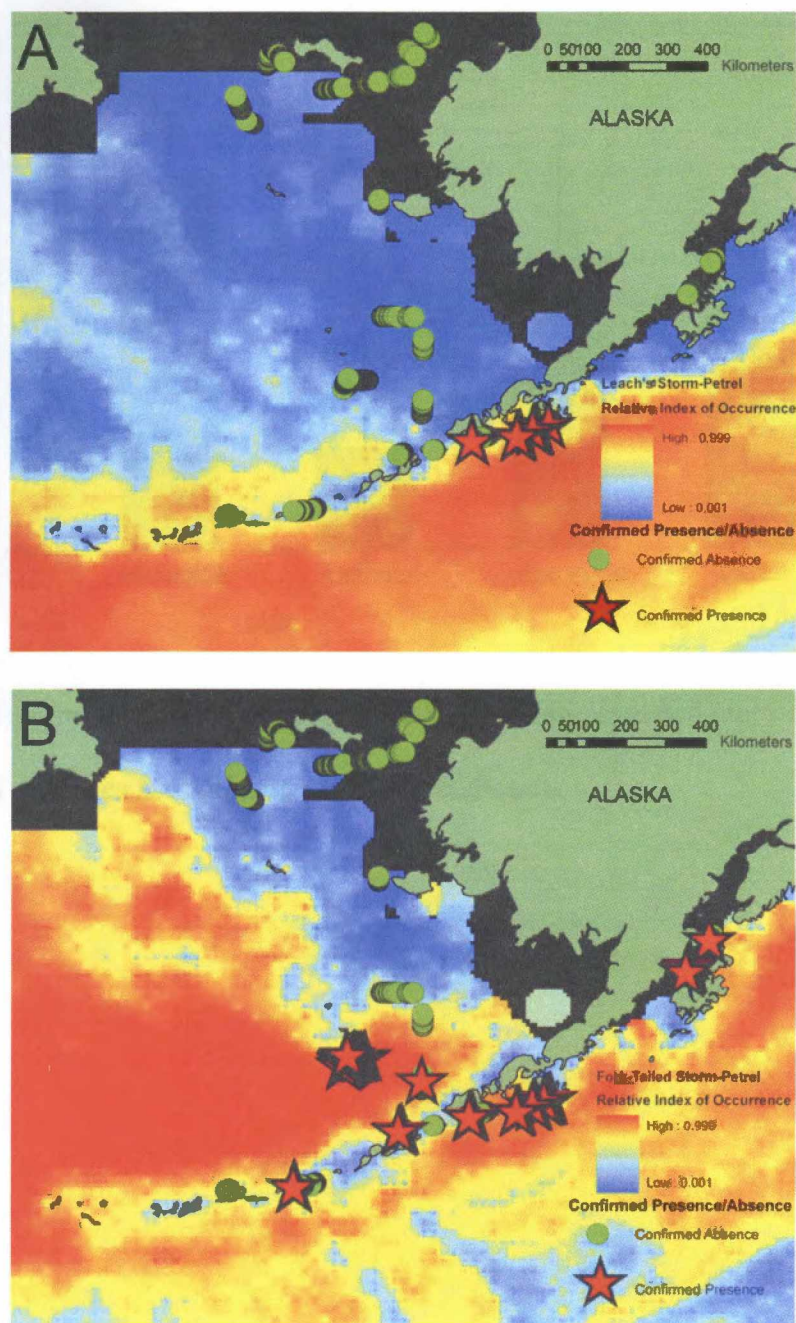


Figure 2.3: Maps of relative index of occurrence of Leach's Storm-Petrel (A) and Fork-Tailed Storm-Petrel (B) with confirmed presence and confirmed absence points from 2 opportunistic surveys performed in summer 2008

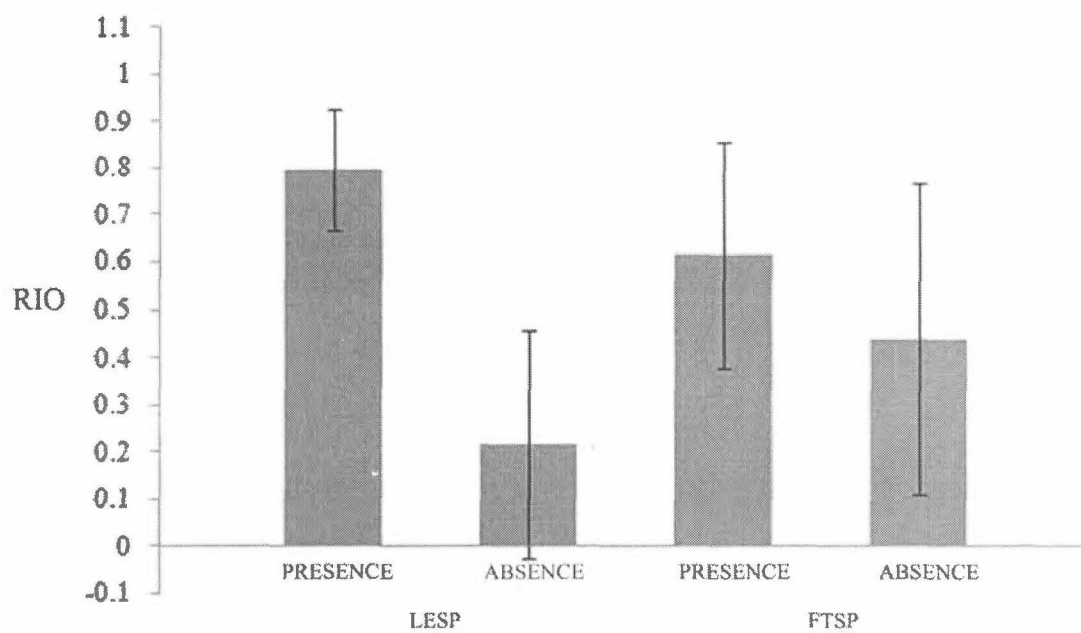


Figure 2.4: Mean relative index of occurrence (RIO) from predicted maps of confirmed presences and absences for both species of Storm-Petrel

Tables

Table 2.1: Area under the ROC curve (AUC) scores for default TreeNet settings (1), and altered TreeNet settings (2) using confirmed absences (a) and pseudo absences (b)

Species	Model											
	All predictor variables				All predictors but DMS				DMS only			
	1a	1b	2a	2b	1a	1b	2a	2b	1a	1b	2a	2b
Leach's Storm-Petrel	0.93	0.94	0.91	0.91	0.93	0.94	0.91	0.90	0.75	0.79	0.87	0.80
Fork-Tailed Storm-Petrel	0.87	0.94	0.87	0.92	0.87	0.94	0.86	0.92	0.63	0.70	0.80	0.84

Table 2.2: Percent correctly classified presences for default TreeNet settings (1), and altered TreeNet settings (2) using confirmed absences (a) and pseudo absences (b)

Species	Model											
	All predictor variables				All predictors but DMS				DMS only			
	1a	1b	2a	2b	1a	1b	2a	2b	1a	1b	2a	2b
Leach's Storm-Petrel	94.2	91.7	95.0	91.7	92.4	91.0	94.8	91.7	82.9	84.4	91.8	83.2
Fork-Tailed Storm-Petrel	84.6	93.5	83.7	80.7	84.6	93.5	83.4	80.5	42.8	59.0	80.4	76.0

General Conclusions

The goals of my thesis were to:

- (1)
 - a. create a series of monthly DMS models using open access datasets
 - b. makes some inferences on controlling factors in DMS production

- (2)
 - a. create models of Fork-Tailed and Leach's Storm-Petrel distribution in the North Pacific
 - b. assess a possible link between Storm-Petrels and DMS.

I hypothesized that by using DMS as a predictor within the framework of sophisticated ecological niche modeling techniques, accurate presence/absence models of Leach's and Fork-Tailed Storm-Petrel could be developed.

Dimethylsulfide

DMS models were run using various subsets of the data allowing me to determine the stability of the model by examining how model performance (RMSD) changes based on the percentage of the original data I use to build and evaluate the model. RMSD did not change substantially between different model runs which indicated that TreeNet remained robust even when the amount of data used to build the model was varied. This may also speak to the robustness of the natural relationships that are defining DMS production.

Currently all known climatologies of sea surface DMS concentration relative to the Kettle database have r^2 values less than 0.06 (Belviso et al. 2004). These comparisons

were made to annual, global climatologies which can be difficult to categorize when using discrete samples taken from monthly measurements. To better examine model performance, it is more beneficial to examine monthly model performance to capture seasonal variability to avoid over-generalization. R^2 values for my models range from 0.21 to 0.69. This work was a marked increase in performance over currently assessed models, but this may not be a fair comparison as those assessments were only for annual climatologies, and not for monthly output.

Accuracy assessments of data are included in the metadata for all layers created in this thesis and are available freely from the author.

Contrary to many current models (Belviso et al. 2004, Bell et al. 2006) this model did not allow *per se* for mechanistic descriptions of how DMS is controlled in the ocean surface, it did however, allow us to test predictor variables and make inferences (via partial dependence plots) on the oceanographic niche in which DMS is produced. The results showed that SRD, phosphates and salinity play important roles in determining concentrations of DMS. SRD and Phosphates were found to vary with DMS concentrations, while a range of salinity values were found to be related to higher DMS concentrations. Neither average chlorophyll a or mixed layer depth were considered important predictor variables overall, contrary to a model suggested by Simo and Dachs (2002). The predictor combination as selected by TreeNet allowed for a high predictive accuracy, and allowed for some inference on possible controllers in the DMS system.

Storm-Petrels

To create Storm-Petrel models, data from the NPPSD were used in a presence/absence or presence/pseudo-absence framework. Using ecological niche modeling techniques allowed us to accurately capture Storm-Petrel distribution in the North Pacific. AUC values for the Storm-Petrel models ranged from 0.63 to 0.94. The best models were created when using all predictor variables. I was also able to extend predictions of Storm-Petrel distribution in to the Sea of Okhotsk where no data exists in the NPPSD. The predictions of both species of Storm-Petrel in this region seemed to qualitatively match current known distribution of Fork-Tailed and Leach's Storm-Petrel (Arthukin & Burkanov 1999, Onley & Scofield 2007). Ground-truthing data performed in summer 2008 shows high model agreement. Predicted RIO of Storm-Petrel distribution seemed to match areas of confirmed presence and absences for both species. All of the survey data and distribution maps are available for download from the author with appropriate metadata.

I ran these models specifically using only DMS to investigate its potential as a predictor variable. Above 9 - 10 nM of DMS, the RIO of Leach's and Fork-Tailed Storm-Petrel stabilizes or begins to increase. This indicated that even with such broad scaled models, we were able to detect a pattern between DMS concentrations and Storm-Petrel distribution. Accuracies of models using just DMS as a predictor variable were also boosted to approximately 0.80 to 0.87 (based on AUC) with correctly classified presences between 76 and 92%.

Final conclusions

Throughout this thesis, I have investigated DMS on a global scale, and its use as a predictor in determining Storm-Petrel distribution. In Chapter 1, I found that a suite of 15 predictor variables could build a relatively accurate model of DMS available freely to the public. I also found through investigation that solar radiation dose, phosphates and salinity were major contributing factors in determining DMS concentrations at the surface of the ocean. In Chapter 2 I developed an accurate model of Storm-Petrel distribution in the North Pacific using a series of predictor variables including DMS. I confirmed the hypothesis here that DMS would be a good predictor of Storm-Petrel distribution. This however, does not confirm that Storm-Petrels “smell” DMS, and to fully address such a question using these techniques would involve dealing with a finer temporal scale model and directed experimentation.

Because these predictive models are potentially valuable for policy and management decisions, we encourage further work to improve them and models like them. For instance, further ground-truthing of such models is needed. Once more recent data become available, the accuracy of these models could be further evaluated with a method such as the Boyce index (Boyce et al. 2002). Further efforts must also be undertaken to improve the accuracy of the underlying environmental models used to predict these distributions. The seabird data must also be improved upon with respect to metadata, and data gaps. Currently, DMS predictions may be too coarse spatially and temporally to accurately define how Storm-Petrels are using this compound to locate foraging areas on a small scale. The importance of individual predictive variables could

be further assessed by removing predictor variables at random and testing for model improvement. Model improvement could be fine-tuned by altering all algorithm settings to determine which models perform best. These tests can be automated in programming languages such as R (www.r-project.org/) or Python (www.python.org/). Addressing autocorrelation in the models may also elucidate the relationships between the various predictor variables and the distributions of these species. The more accurately we can predict current seabird distributions, the more accurately we will be able to predict and manage their responses to future climate change and other impacts.

References

- Arthukin YB, Burkanov VN (1999) Seabirds and Sea mammals of the Russian Far East: A Field Guide., Vol. AST Publishing House, Moscow
- Bell TG, Malin G, McKee CM, Liss PS (2006) A Comparison of dimethylsulphide (DMS) data from the Atlantic Meridional Transect (AMT) programme with proposed algorithms for global surface DMS concentrations. *Deep-Sea Research II* 53:1720-1735
- Belviso S, Bopp L, Moulin C, Orr JC, Anderson TR, Aumont O, Chu S, Elliott S, Maltrud ME, Simo R (2004) Comparison of global climatological maps of sea surface dimethyl sulfide. *Global Biogeochemical Cycles* 18
- Boyce MS, Vernier PR, Nielson SE, Schmiegelow FKA (2002) Evaluating Resource Selection Functions. *Ecological Modelling* 157:281-300
- Onley D, Scofield P (2007) Albatrosses, Petrels & Shearwaters of the World, Vol. Princeton University Press, Princeton, New Jersey
- Simo R, Dachs J (2002) Global ocean emission of dimethylsulfide predicted from biogeophysical data. *Global Biogeochemical Cycles* 16

Appendices

Appendix A

Data sources for all variables used for model development in Chapter 1 of thesis

Dataset	Source	Resolution
Dimethylsulfide	Pacific Marine Ecological Laboratory (saga.pmel.noaa.gov)	Points
Salinity	World Ocean Atlas (www.nodc.noaa.gov)	1°
Dissolved Oxygen	World Ocean Atlas (www.nodc.noaa.gov)	1°
Apparent Oxygen Utilization	World Ocean Atlas (www.nodc.noaa.gov)	1°
Nitrates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Phosphates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Silicates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Sea Surface Temperature	Marine Conservation Biology Institute (distributed CD)	1°
Bathymetry	Marine Conservation Biology Institute (distributed CD)	1°
Solar Radiation Dose	Calculated as per Vallina and Simo (2007).	1°
- irradiance at top of atmosphere	Provided by Dr. Sergio M. Vallina	1°
- mixed layer depth	Provided by Dr. Sergio M. Vallina	1°
Human Impact	National Center for Ecological Analysis and Synthesis (www.nceas.ucsb.edu/GlobalMarine [<i>Halpern et al.</i> , 2008])	1 km
Euclidean Distance to shore	Calculated in arcGIS software from coastline polyline	0.833°
Chlorophyll a (SeaWIFS)	NASA - Oceancolor project (oceancolor.gsfc.nasa.gov)	1.1 km

Appendix B

Total number of data points used to train and assess each model for each month for
Chapter 1 of thesis

	Model 1		Model 2		Model 3		Model 4		
Month	Train	Assess	Train	Assess	Train	Assess	Train	Assess	Total
Jan	271	1026	778	519	908	389	1167	130	1297
Feb	613	2418	1819	1212	2122	909	2728	303	3031
Mar	1385	5063	3869	2579	4514	1934	5803	645	6448
Apr	1257	4484	3445	2296	4019	1722	5167	574	5741
May	768	2944	2227	1485	2598	1114	3341	371	3712
Jun	749	2873	2173	1449	2535	1087	3260	362	3622
Jul	871	3212	2450	1633	2858	1225	3675	408	4083
Aug	430	1491	1153	768	1345	576	1729	192	1921
Sep	452	1655	1264	843	1475	632	1896	211	2107
Oct	708	2375	1850	1233	2158	925	2775	308	3083
Nov	728	2699	2056	1371	2399	1028	3084	343	3427
Dec	402	1572	1184	790	1382	592	1777	197	1974

Appendix C

Relative importance of predictor variables for highest ranking models (for both models a and b) as ranked by the AUC

Variable	Model			
	Leach's Storm-Petrel		Fork-Tailed Storm-Petrel	
	All predictors	All predictors	All predictors but DMS	All predictors
	1a	1b	1a	1b
Bathymetry	100.0	57.6	89.8	45.6
Chlorophyll a	32.3	55.1	79.5	100.0
Distance to shore	72.0	86.5	100.0	49.2
Human Impact	31.5	55.9	66.3	47.9
Mixed Layer Depth	30.2	76.1	58.1	40.2
Dimethylsulfide	35.6	60.2	N/A	37.3
Dissolved O2	95.5	83.0	71.9	44.0
Nitrate	34.5	59.3	53.6	37.0
Apparent O2 Utilization	43.0	77.0	77.2	58.4
Phosphate	31.2	56.3	73.4	43.3
Salinity	36.1	100.0	66.0	41.8
Silicate	41.2	62.5	56.9	31.5
Sea surface temperature	40.1	67.6	54.8	39.3

Appendix D

Sources of predictor variables for Chapter 2 of thesis

Dataset	Source	Resolution
Dimethylsulfide	Thesis Chapter 1	1°
Salinity	World Ocean Atlas (www.nodc.noaa.gov)	1°
Dissolved Oxygen	World Ocean Atlas (www.nodc.noaa.gov)	1°
Apparent Oxygen Utilization	World Ocean Atlas (www.nodc.noaa.gov)	1°
Nitrates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Phosphates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Silicates	World Ocean Atlas (www.nodc.noaa.gov)	1°
Sea Surface Temperature	Marine Conservation Biology Institute (distributed CD)	1°
Bathymetry	Marine Conservation Biology Institute (distributed CD)	1°
Mixed layer depth	Provided by Dr. Sergio M. Vallina	1°
	National Center for Ecological Analysis and Synthesis	
Human Impact	(www.nceas.ucsb.edu/GlobalMarine)	1 km
	(Halpern et al. 2008)	
Euclidean Distance to shore	Calculated in ArcGIS software from coastline polyline	0.833°
Chlorophyll a (SeaWiFS)	NASA - Oceancolor project (oceancolor.gsfc.nasa.gov)	1.1 km

Appendix E

Model automation for DMS model

Automation was performed using Python 2.5.4 and program R version 7.0. Using approximately 1500 lines of code, a monthly DMS modeling tool was created which allows the user to go from raw DMS data in a comma separated values format to the completed model with smoothed maps and appropriate assessments performed. The model automation process is shown in Figure 1, and the code is available from the author.

