Physical Properties of Marine Sediments and their Application Toward Climate Change Studies

Tania Lado Insua
University of Rhode Island, ladoinsuat@egr.uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation
https://digitalcommons.uri.edu/oa_diss/47

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.
PHYSICAL PROPERTIES OF MARINE SEDIMENTS
AND THEIR APPLICATION TOWARD CLIMATE
CHANGE STUDIES

BY

TANIA LADO INSUA

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
OCEAN ENGINEERING

UNIVERSITY OF RHODE ISLAND
2013
DOCTOR OF PHILOSOPHY DISSERTATION

OF

TANIA LADO INSUA

APPROVED:

Dissertation Committee:

Major Professor  Kathryn Moran

Annette R. Grilli

Arthur J. Spivack

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2013
ABSTRACT

The Integrated Ocean Drilling Program (IODP) is the world’s largest collaborative research initiative for the study of the ocean seafloor, subseafloor, and their application to earth system science. The processing of recovered cores from the subseafloor by IODP has been standardized to maintain constant, reproducible results and to make the measurements available to the scientific community using a world-wide-web based database. The study of physical properties is a basic requirement for the study of marine sediments as they form a foundation data set for most of earth system dynamic studies.

This thesis focuses on solving four different scientific problems using analyses and modeling of marine sediments based on physical properties among other IODP data. The four manuscripts presented here are applied cases that use marine sediment physical properties to understand various aspects of past and future climate change. The main objectives of this dissertation are: validation of the design and development steps of a new seafloor observatory that will measure physical properties over time in IODP boreholes; dimensioning of this system based on the physical properties previously measured at the deployment site; development of a classification system for the major types of carbonate sediments used to reconstruct past sea level rise; and reconstructing past Pacific Ocean maximum bottom water salinities during the Last Glaciation Maximum (LGM) based on physical properties of marine sediments. These objectives were addressed by using multivariate statistics, machine learning techniques and a one dimensional diffusion model.
A new modular borehole observatory, named SCIMPI (Simple Cabled Instrument for Measuring Parameters In situ) was successfully developed and tested for deployment. This system is equipped to take time series measurements of temperature, pressure and electrical resistivity in the geological formation where it is emplaced. The characteristics of the system and the testing process are presented in Chapter 1.

In Chapter 2, the modular design of the first SCIMPI prototype is dimensioned using a clustering approach based on physical properties from the deployment site. The multivariate statistics applied show direct relation with the geological formation and gas hydrate dynamics at the site.

In the third Chapter of this dissertation, physical properties are related to characteristics of marine carbonates that, in turn, are indicative of their lithological type. Three different models were tested: Linear Discriminant Analysis, Random Forest and, Support Vector Machines. This study demonstrates a strong nonlinear relationship between physical properties and carbonate lithotypes. The results show that machine learning models can help identify lithologies, thus aiding the selection of sample locations, core-log correlations, and processing of these rare core materials.

Finally, pore fluid chloride concentration profiles from six Pacific Ocean deep sea sites were used to reconstruct past salinities. This study expands the spatial coverage of salinity reconstructions during the Last Glacial Maximum (LGM) to the equatorial Pacific and North and South Pacific Gyres. These reconstructed LGM chloride concentrations of deep Pacific bottom water are ~4% greater than today’s values. This is consistent with the view that the deep ocean density structure was primarily controlled by salinity variations during the LGM.
ACKNOWLEDGMENTS

“From here I see a path, but I don’t know where it leads; not knowing is the reason why I would like to walk it”

Rosalía de Castro, (1837 - 1885)

The Department of Ocean Engineering at the University of Rhode Island (URI), made this work possible, particularly my advisor, Dr. Kate Moran, who gave me some opportunities that I would never have dreamt of and the freedom to choose my path. Kate thanks for giving me the chance of working with you and for your infinite patience with me. Thanks for being there in the hard times. I really admire you as a researcher and as a person and I am proud to be your student.

I would like to thank my committee; this dissertation wouldn’t be possible without all of you. Dr. Spivack, I really enjoyed our science discussions and our advancements in the world of paleosalinity. Thanks for all you taught me, for your patience, your reviews and advice. Dr. Annette Grilli, thanks for your constant support, advice, kindness, the coffees and overall, thanks for your friendship. Dr. Hamel, thanks for sharing your knowledge and for the work you put on this thesis. Dr. Baxter, thanks for all your help during the hard-bureaucracy times, for your advice, your help with technical knowledge and your availability.

This work was supported by a three years fellowship from Barrié de La Maza that gave me the opportunity of my life. Later on I was funded by research assistantships from NSF and IODP-MI, I am extremely grateful for their support of this work.
Thanks to the crew, scientists, technical and land support of the Expeditions IODP 310, 325, 336 and 329, Knorr 195-3 and the Endeavor SCIMPI cruise. The staff of the Pell library, in particular Joyce A. Downey, has been instrumental in the access to the literature used in this work. So have been the personnel at shipping and receiving in Bay Campus. Thanks to Dr. Albert Gerheim for his quick reviews of formatting.

Thanks to the SCIMPI team, Kate Moran, Lucy Hurlbut, Ian and Tori Kulin, Jim Newman and Steve Farrington. Thanks for all that you taught me, for the great discussions and the good times. Working with you has been a pleasure. Ian, thanks for always being there for me, not only as a manager, as a coach and a friend, I have really missed you on the last months of this thesis.

Dennis Graham, Louise Anderson, Sally Morgan, Adam Klaus, Gary Acton and Jody Webster provided, lab, writing, graphs and data analysis assistance or training, reviews and great moments together. I would also like to thank Dr. Steven D’Hondt, Nelly and the girls and the Grilli and Sharma families for their academic and personal support. Jane Eaton, Kathy Beck, Gail Paolino, Dania Brandford-Calvo, Ami Limon, Melissa de Jesus and the staff at the Office of International Students and Scholars at URI, and at the Institute of International Education (IIE) for their assistance on administrative work.

Thanks to Dave Tarasevic, Fred Pease, Cornelia and Sandee Lineham, Mark Brasil, Linda Palazzo, Dennis Chase and Dennis Bernier, for being the heart of the Bay Campus, for always having smile and a good advice on a stormy day.

Thanks to my Rhody family: Ursula Hebinck, Myriam El Bettah, Anabela M. Resende da Maia, Ping Li, Alesandra Morales, Erin Papa, Eric Rulison, Ravi Sharma
and Jeff Harris. I have no words for you guys; I will never forget some of our crazy stories… Amir Banari, Christian Janssen, Kevin Hopkins and Antigone thanks for being my “soul sisters” and for Bonnet-summer! Rocio Escalante, Heather Price, Elias Shalhoub, Sarah Wojciechowski, Corey Richard, George Dib, Lou Licate, Alexis Ortiz, Elizabeth McLean, Marina Meier, Thorsten Fink, George Konstantinidis, Georges Dossot, Kevin Bollinger, Taylor Asher, Kevin Turco and Amanda Michael for all the fun! Anne and Joe Ilacqua, thanks for such a nice stay in Maple and for all the home dinners that made me feel in family. Thanks to Pamela Hurst for her advice and friendship. Thanks to all of you, my Rhode Islanders that helped me make this place my home during the last years.

Special thanks to Matthew C. Smith, my soul mate, for always being beside me even in the distance. Thanks for never failing me, for the laughs together, for the hours of flight, for your care, all the cooking in desperate times and for all that you teach me as a scientist but overall as a person. Thanks to Janice and Bill Smith for treating me like a daughter. Thanks to Olivia, Wilbur, Greg and Sarah Smith, for the paintings and the fun family time together.

Thanks to my family for being always there. To my parents, Hortensia Insua e M.J. Lado, thanks for the long hours of work that brought me where I am and for your support. Mum, thanks for believing in me and for being always there. To my brother, Germán, thanks for being always beside me, having a brother is invaluable. Thanks to Manuel Insua, my grandfather, for the computer you gifted be when I left for the United States and that I used so much for this dissertation and for your endless support. Thanks to my aunt Carmen Insua, for always making time to see me, for the
“torta” and the “tortilla” and to my uncle Pepe Insua for always having a coffee ready and a good talk when I pass by the bar. Fina, thanks for the “filloas”! And thanks to the ones in Valcobo for the BBQs! Ana Rama and Diego Rodríguez-Guisán, you count as siblings, I don’t know what I would do without you. Thanks for the support 24/7 in skype and for always having time for me without demanding my time. Maneiro, my “father in law”, thanks for the sailor advice. Thanks to “the miller’s family”: Pepe, Carmen, Sebas, Mar, Sonia and Alex Insua, Jano, Martín and Susu Muñoz, Sergio Mejuto and the “Boliños” (Alberto Insua an Silvia García). The few weekends I’m at home are amazing thanks to you. Pili Rebón, Cesar Pedreira, Loly Piñeiro and Carmen Seijal, thanks to you too!

To Alicia Abalo and Isa Lago, thanks a thousand for your friendship. To Fernando Cela, Xabier Boullón and the “rebebos” thanks so much for your support and all that you taught me these years. Thanks to the gang in Arteixo: Estela, Marcos, Lobo, Ra, Sindi, Javi, Susana, Anlli, Marta, Laura, Vitru, Juancho, Pablo… and the rest! To my other Galicians: David Raposo, Lyndsay Kyker, Nacho Barrio, Laura Murillo, Carlos Sierra, Enrique González (Henry), Almu Janeiro, Ana Cruz, Dani Casas, Angeles Guerra and Patty Piñeiro.

Thanks to the colleagues and professors in Vigo that left a mark in me and helped me to choose this path: Alejandro C. Collazo, Mariano Lastra, Soledad Gil, David Posada, Angel P. Diz, Fernándo Cruz, Castor Muñoz and Esther de Blas.

Thanks to all of you for believing in me. I hope I didn’t forget many of you and that those that I forgot can forgive me. This dissertation wouldn’t have been possible if I lack any of you.
AGRADECIENTOS

“Dende aquí vexo un camiño, que non sei adónde vai;

polo mesmo que n’o sei quixera o poder andar”

Rosalía de Castro, (1837 - 1885)

Ó Departamento de Enxeñería Oceánica da Universidade de Rhode Island (URI), fixo posible este traballo, en particular a miña directora, a Dr. Kate Moran, que me deu oportunidades cas que nunca tería soñado e a liberdade de escoller o meu camiño. Kate, gracias por darme a oportunidade de traballar contigo e pola túa infinita paciencia comigo. Gracias por estar ái nos tempos difíciles. Tes a mina admiración como investigadora e como persoa. Estou orgullosa de ser estudante túa.

Quero agradecer ó meu comité; esta tese non sería posible sen vós. Dr. Spivack, desfrutei moito das nosas discusións e avances no mundo das paleosalinidades. Gracias por ensinarme, pola súa paciencia, polas revisións e polos consellos. Dr. Annette Grilli, gracias polo teu apoio constante, consellos, amabilidad, os cafés, e sobre todo pola túa amistade. Dr. Hamel, gracias por compartir os seus coñecementos e polo seu traballo nesta tese. Dr. Baxter, gracias por toda a axuda durante os tempos de complicacións burocráticas, polos consellos, os coñecementos técnicos e a súa disponibilidade.

Este traballo foi financiado por unha beca de tres anos da Fundación Pedro Barrié de La Maza que me deu a oportunidade da miña vida. Despois fun financiada como axudante de investigación por iso estou moi agradecida a NSF e IODP-MI. Gracias á
tripulación, científicos, e apoio técnico e de terra das Expedicións IODP 310, 325, 336 e 329, Knorr 195-3 e a campaña SCIMPI no Endeavor. O persoal da Biblioteca Pell, en particular Joyce A. Downey, foi instrumental no acceso ás referencias bibliográficas usadas neste traballo. Tamén foi imprescindible o persoal na sección de transporte do Bay Campus. Gracias ó Dr. Albert Gerheim pola rapidez nas revisións.

Gracias ó equipo SCIMPI, Kate Moran, Lucy Hurlbut, Ian and Tori Kulin, Jim Newman e Steve Farrington. Gracias por todo o que me ensinaches, polas grandes discusións e os bos momentos. Traballar convosco foi un pracer. Ian, gracias por estar sempre aí para min, non so coma director do proxecto, tamén coma conselleiro e amigo, metínche moito de menos os últimos meses desta tese.

Dennis Graham, Louise Anderson, Sally Morgan, Adam Klaus, Gary Acton e Jody Webster axudaron no laboratorio, análises de datos, revisións, mellorando a miña capacidade gráfica e de escritura e con grandes momentos xuntos. Gracias ó Dr. Steven D’Hondt, Nelly e as nenas, e ás familias Grilli e Sharma, polo apoio académico e persoal. Jane Eaton, Kathy Beck, Gail Paolino, Dania Brandford-Calvo, Ami Limon, Melissa de Jesus e o persoal da Oficina de Estudiantes e Escolares Internacionales de URI e o Instituto de Educación Internacional (IIE) pola axuda administrativa.

Estou moi agradecida a Dave Tarasevic, Fred Pease, Cornelia e Sandee Lineham, Mark Brasil, Linda Palazzo, Dennis Chase e Dennis Bernier, por ser a alma do Bay Campus, por ter sempre un sorriso e un bo consello nun día de tormenta.

Gracias á miña familia en Rhode Island: Ursula Hebinck, Myriam El Bettah, Anabela M. Resende da Maia, Ping Li, Alesandra Morales, Erin Papa, Eric Rulison, Ravi Sharma e Jeff Harris. Non teño palabras suficientes para vós; nunca vou esquecer
alguna das nosas historias de tolemia... Non vos preocupedes, non vou comentalas en público nos agradecementos dunha tese! Amir Banari, Christian Janßen, Kevin Hopkins e Antigone grazias por ser os meus “soul sisters” e polo verán de Bonnet! Rocío Escalante, Heather Price, Elias Shalhoub, Sarah Wojciechowski, Corey Richard, George Dib, Lou Licate, Alexis Ortiz, Elizabeth McLean, George Konstantinidis, Marina Meier, Thorsten Fink, Georges Dossot, Kevin Bollinger, Taylor Asher, Yaurel Guadalupe, Kevin Turco e Amanda Michael por toda a festa! Anne e Joe Ilacqua, grazias por unha estancia marabillosa en Maple e por facerme sentir en familia. Gracias a Pamela Hurst polos seus consellos e amistade. Gracias a todos vós, os meus “Rhode Islanders” que axudaches a facer deste sitio a miña casa durante os últimos anos.

Gracias especialmente a Matthew C. Smith, a miña alma xemelga, por estar sempre o meu lado, incluso na distancia. Gracias por non fallarme nunca, polas risas xuntos, por todas as horas de voo, polo teu cariño, polas cociñadas en tempos desesperados e por todo o que me ensinas como científico pero sobre todo como persoa. Gracias a Janice e Bill Smith por tratarme coma unha filla. Gracias a Olivia, Wilbur, Greg e Sarah Smith, polos debuxos e a diversión en familia xuntos.

Gracias á miña familia por estar sempre aí. Gracias a meus pais, Hortensia Insua e M.J. Lado polas duras horas de traballo que me axudaron a chegar onde estou e polo voso apoio. A miña nai, Hortensia Insua, polo teu apoio, por estar sempre comigo. Gemán, grazias por estar aí o meu carón, ter un irmán e algo de valor incalculable. Gracias a Manuel Insua, meu avó, polo ordenador que me regalaches cando marchei para Estados Unidos e que tanto usei nesta tese e polo teu apoio incondicional. Gracias
a miña tía Carmen Insua, por sempre facer o tempo para verme e pola torta e a tortilla, 
e a meu tío Pepe Insua por terme sempre un cafetiño listo e unha boa charla cando 
paso polo bar. Fina, gracias polas filloas! E gracias ós de Valcobo polas churrascadas!
A meus tíos e primos, por ser vós e por estar aí. Ana Rama e Diego Rodríguez-Guisán, 
vós contades coma irmáns, sen vós non sei que faría. Gracias polo apoio 24h en skype 
e por sempre sacar tempo para min e nunca demandalo. Maneiro, “meu sogro”, gracias 
polos consellos de mariñeiro. Gracias “ós do muiñeiro”: Pepe, Carmen, Sebas, Mar, 
Sonia e Alex Insua, Jano, Martín e Susu Muñoz, Sergio Mejuto e os Boliños (Alberto 
Insua e Silvia García). Os poucos domingos que estou na casa son unha pasada gracias 
a vós. Pili Rebón, Cesar Pedreira, Loly Piñeiro e Carmen Seijal, gracias a vós tamén!

A Alicia Abalo e Isa Lago, mil gracias pola vosa amistade. Fernando Cela e 
Xabier Boullón e os rebebos moitas gracias polo voso apoio e por todo o que me 
ensinaches estes anos. Graciñas á tropa de Arteixo: Estela, Marcos, Lobo, Ra, Sindi, 
Javi, Susana, Anlli, Marta, Laura, Vitru, Juancho, Pablo… e todo o resto! Os meus 
outros galegos: David Raposo, Lyndsay Kyker, Nacho Barrio, Laura Murillo, Carlos 
Sierra, Enrique González (Henry), Almu Janeiro, Ana Cruz, Dani Casas, Angeles 
Guerra e Patty Piñeiro.

Gracias ós profesores e compañeiros de Vigo que deixaches marca en min e que 
me animaches a tomar este camiño: Alejandro C. Collazo, Mariano Lastra, Soledad 
Gil, David Posada, Angel P. Diz, Fernándo Cruz, Castor Muñoz e Esther de Blas.

Gracias por crer en min. Espero non terme esquecido de moitos de vós e espero 
que ós que esquecén, me saibades perdoar. Esta tese non sería posible se algún de vós 
me faltase.
A meu avó, Manuel Insua, que me aconsellou en terra de lobos a ouvear coma todos.

En memoria de mina avoa, Hortensia Figueroa.

[to my grandfather, Manuel Insua, who advised me in the land of the wolf, just howl.

In loving memory of my grandmother, Hortensia Figueroa]
PREFACE

The focus of this thesis is the application of engineering approaches, including sediment and rock physical properties and modeling, to enable a better understanding of the climate system by making new observations of the present-day ocean and developing techniques for extracting past climate records. Four major publications are presented in this dissertation in manuscript format. All of these contribute to cutting edge issues in the study of climate change from an engineering perspective by providing innovative solutions to measure, analyze and integrate data.

In the first chapter of this thesis, I present a new borehole observatory (SCIMPI). SCIMPI is designed to observe sub-seafloor gas hydrate dynamics. Methane reservoirs are an important, yet poorly known aspect of the global carbon cycle and their effect on climate change is currently of major science focus. The warming potential of methane as a greenhouse gas is 25 times the potential of CO$_2$ over a 100 year period (Foster et al. 2007, reviewed in Bridgham et al. 2013). Global methane emissions are between 500 and 600 Tg CH$_4$/yr (Dlugokencky et al. 2011). An unknown amount of this emission is due to destabilization of sub-seafloor gas hydrates; most of the uncertainty is due to the challenge on assessing the gas hydrate reservoirs (Ruppel 2011). SCIMPI represents a new approach for generating sub-seafloor time series for the study of these gas hydrate systems. A team of engineers led by my major professor designed SCIMPI. As part of that team, I led sensor calibrations, testing, planning of offshore tests, participation in offshore testing, and all aspects of site selection and sensor positioning. This manuscript was submitted to the journal Scientific Drilling.
In the second chapter the dimensioning of this sub-seafloor observatory is addressed, which optimizes the positioning of SCIMPI sensors with depth. Optimization of sensor position is critical for observing change at the depths where fluid and methane hydrate dynamics occur. The optimization code that I developed and applied uses a new technique for this application – a k-mean cluster on the previously collected data obtained for the proposed study area, the Cascadia margin, offshore western Canada. SCIMPI is scheduled to be deployed at a site on the Cascadia Margin, called “Bullseye Vent” from the Drillship (D/V) JOIDES Resolution in May, 2013. I analyzed the seismic and borehole data from this site and configured the SCIMPI sensors using this new technique. This manuscript will be submitted to the journal *Geosphere*.

The third and fourth chapters are separate studies that use ocean sediment and rock cores to reconstruct Earth’s past climate, specifically sea level rise (Chapter 3) and deep sea salinity (Chapter 4).

The instrumental and historical human records of climate cover only a few hundred years of the planet climate history (Ruddiman, 2001). Sediment and ice cores host records of Earth’s past climate over thousands and millions of years. The cores contain indicators of past climate referred to as climate proxies that can be biotic, geological and geochemical.

In chapter 3, cores recovered from coral reefs were studied for reconstruction of past sea level because they provide direct evidence of sea level, based on coral type. Coupled with accurate dating, using mass spectrometry, the timing of deglaciation events contributes to our understanding of the mechanisms that drove past glacial-
interglacial cycles. Furthermore, scleractinian coral colonies can monitor and record past sea-surface temperatures (SSTs) (Camoin et al. 2007). The niche of the species of coral is known, and several proxies such as U/Th, $^{231}$Pa/$^{235}$U and $^{14}$C dating measured on their skeleton can also be tied to past sea level heights (e.g. Thomas et al. 2009, Yokoyama et al. 2000).

Ocean corals have $^{234}$U and $^{238}$U but not $^{230}$Th. When the corals die, the $^{238}$U decays and produces $^{230}$Th that is also radioactive and decays. Gradually, the amount of $^{230}$Th increases to a level that reflects a balance between the slow decay of the parent U and the faster loss of daughter $^{230}$Th. U/Th ratios can be obtained by mass spectrometry (Schulz and Zabel 2006). Other isotopic ratios can also be used in the study of sea level changes with corals, such as $^{231}$Pa/$^{235}$U and $^{14}$C dating that can be tied to past sea level heights. The species Acropora palmata is of special importance for sea level record because it grows in the shallowest five meters, thus providing one of the most reliable sea level indicators.

The reconstruction of an accurate sea level curve is of vital importance for climate change modeling and reconstruction of paleoclimate (e.g. Adkins and Schrag 2003). IODP Expedition 310 Tahiti Sea Level and IODP Expedition 325 Great Barrier Reef Environmental Changes are the most recent contributions to the study of past sea levels.

It is quite difficult to drill and recover high quality coral cores because of their brittle nature and large void spaces that naturally occur in these animals. Therefore, it is crucial to extract as much information as possible from the paucity of recovered
cores and to take extreme care during core processing so as not to damage and degrade the climate signals represented by the species of coral.

One of the primary instruments for initially analyzing all cores, including coral cores is the non-destructive core measurement system, the Multi-sensor Core Logger (MSCL), which was primarily developed to analyze sediment cores. The MSCL has also been used less frequently for hard rock or corals. The performance of the MSCL in hard rock, fossil corals and some carbonate sediments has been questioned because the data are adversely affected by several parameters related to the quality of the core sample, including under filled liners, bad contact due to a gap between the core liner and the core and differences in the saturation of the sediment. In the third chapter, a completely new and innovative methodology for analyzing MSCL data to identify fossil corals and other carbonate lithologies based on machine learning techniques. I was selected to be a member of the science party for IODP Expedition 325 and applied this technique to the MSCL data from the recovered cores for this expedition and its sister one (Expedition 310). This manuscript was submitted as part of a special issue in the journal *Sedimentology*, titled “*Reef response to sea-level and environmental changes: reading the past, informing the future*”.

McDuff (1984) pointed out the importance of some non-reactive species in the sediment pore fluids, including chloride, which diffusively moves within sediments. Isotopic composition information is stored in the calcite shells of planktic organisms and benthic foraminifera buried on deep sea sediments. Oxygen isotopes are used to study changes in past global ice and global temperature and carbon isotopes can be used for studies of movement of organic materials among reservoirs and can help to
detect changes in deep ocean circulation. Similarly, the pore water trapped in the sediment gives us information on past climate and variations of conservative properties in the ocean, e.g. chloride concentration or the changing ratio of strontium to calcium sea water.

Adkins and Schrag (2001, 2003), Adkins et al. (2002) and Schrag et al. (1996, 2002), described a one-dimensional Eulerian numerical model based on the diffusion/advection tracer equation that reproduces paleo-chloride concentration and $\delta^{18}O$ measured in deep-sea sediments. They analyzed two sites in the boundaries of the Pacific ocean showing that saltier waters (compared with the present) were present during the Last Glacial Maximum (LGM), suggesting higher stratification of deep and intermediate waters and a deepening of nutrient and carbon rich waters in the eastern North Pacific. These studies are important contributors to efforts to reconstruct past deep ocean circulation — much needed to test climate models used for forecasting future climate on the planet. This paleo-circulation reconstruction has not yet been achieved partially due to a lack of proxy records and salinity reconstructions, particularly for the Pacific Ocean (MARGO 2009). Thus, further research that improves and provides more paleosalinity reconstructions is important for improved understanding of overturning circulation during times of different climate forcing, such as the LGM (Stocker 1998, Lynch-Stieglitz et al 2007, MARGO 2009).

In the fourth chapter, I present results from my one-dimensional diffusion model that uses sediment formation factor (derived from physical properties) as a proxy for tortuosity. Tortuosity contributes to the ability for species, such as chloride, to diffuse within sediments. This new procedure greatly reduces the errors of the modeled
diffusion system, generating more reliable outputs of paleo-chloride concentration. Using this improved approach, a reconstruction of the paleo deep-water salinity was competed, adding six new locations in the Pacific Ocean for reconstructing deep water circulation during the LGM. This manuscript will be submitted to the journal *Geophysical Research Letters*.

REFERENCES


# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>AGRADECEMENTOS</td>
<td>viii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>xii</td>
</tr>
<tr>
<td>PREFACE</td>
<td>xiii</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>xviii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>xxi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xxv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xxvii</td>
</tr>
</tbody>
</table>

## CHAPTER 1 - SCIMPI: A NEW BOREHOLE OBSERVATORY

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>3</td>
</tr>
<tr>
<td>2. SYSTEM CONFIGURATION AND SPECIFICATIONS</td>
<td>4</td>
</tr>
<tr>
<td>3. TESTING AND DEPLOYMENT</td>
<td>10</td>
</tr>
<tr>
<td>4. FLEXIBILITY IN THE CONFIGURATION</td>
<td>11</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>12</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>13</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>14</td>
</tr>
<tr>
<td>HIGHLIGHTS</td>
<td>14</td>
</tr>
</tbody>
</table>

## CHAPTER 2 - A MULTIVARIATE APPROACH TO OPTIMIZE

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>15</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>16</td>
</tr>
</tbody>
</table>
1. INTRODUCTION .................................................................................................................. 17
  1.1. Physical properties and pore water in gas hydrates .................................................... 18
  1.2. Subseafloor observatories ............................................................................................ 20
  1.3. SCIMPI deployment in Cascadia ................................................................................. 22
  1.4. Preliminary information about the sites ...................................................................... 23
  1.5. Location of the sensors .............................................................................................. 24
2. MATERIALS AND METHODS ............................................................................................ 25
  2.1. Data from previous expeditions .................................................................................. 25
  2.2. Statistical analysis ....................................................................................................... 28
3. RESULTS AND DISCUSSION ............................................................................................. 29
  3.1. Cluster results and interpretation for CAS05 and CAS10 ....................................... 29
  3.2. Cable lengths and distribution of modules for CAS05 and CAS10 ..................... 36
4. CONCLUSIONS .................................................................................................................. 39
ACKNOWLEDGMENTS ....................................................................................................... 40
REFERENCES ...................................................................................................................... 40

CHAPTER 3 - ADVANCED CLASSIFICATION OF CARBONATE SEDIMENTS BASED ON PHYSICAL PROPERTIES .................................................. 46

ABSTRACT ........................................................................................................................... 47
1. INTRODUCTION ............................................................................................................... 49
2. MATERIALS AND METHODS .......................................................................................... 52
  2.1. Sample selection ......................................................................................................... 53
  2.2. Linear Discriminant Analysis .................................................................................... 56
  2.3. Random Forest ........................................................................................................... 56
2.4. Support Vector Machines ................................................................. 58
2.5. Evaluating the accuracy of the models ........................................... 59
3. RESULTS AND DISCUSSION .............................................................. 63
  3.1. Total accuracy of the models ...................................................... 63
  3.2. Accuracy for each lithology ......................................................... 65
    3.2.1. Expedition 325: GBREC ......................................................... 65
    3.2.2. Expedition 310: Tahiti ......................................................... 66
    3.2.3. Comparison GBREC vs. Tahiti: Re-saturate or directly process? ... 69
  3.3. Comparison of the lithologies: Confusion matrices ....................... 69
4. DISCUSSION AND CONCLUSIONS ................................................... 76
ACKNOWLEDGMENTS ............................................................................. 80
REFERENCES .......................................................................................... 80

CHAPTER 4 - RECONSTRUCTION OF PACIFIC OCEAN BOTTOM
WATER SALINITY DURING THE LAST GLACIATION MAXIMUM .... 85

ABSTRACT ............................................................................................. 86
1. INTRODUCTION ................................................................................... 87
2. MATERIALS AND METHODS: RECONSTRUCTING
   PALEOSALINITIES .................................................................................. 90
3. RESULTS AND DISCUSSION ........................................................... 94
4. CONCLUSIONS .................................................................................. 98
ACKNOWLEDGMENTS ........................................................................... 99
REFERENCES ........................................................................................ 100
APPENDICES .......................................................................................................................... 107

APPENDIX I .......................................................................................................................... 107
  Selection of the number of clusters .................................................................................. 107

APPENDIX II .......................................................................................................................... 108
  R code for cluster analysis ............................................................................................... 108

APPENDIX III .......................................................................................................................... 116
  Histograms of the physical properties ............................................................................. 116

APPENDIX IV .......................................................................................................................... 118
  R code for advanced classification of carbonate sediments ........................................ 118

APPENDIX V .......................................................................................................................... 128
  MATLAB for finite difference diffusion model ............................................................. 128

APPENDIX VI .......................................................................................................................... 133
  WOCE lines and locations used for modern salinity ................................................. 133
LIST OF TABLES

TABLE PAGE

CHAPTER 1
Table 1. Specifications of the SCIMPI prototype .......................................................... 9

CHAPTER 2
Table 1. Variables used in the analyses with the tool and method of measurement used, and the geographic location and depth below seafloor locations for Sites CAS05-CORK and CAS10-SCIMPI, based on Site U1328 ......................... 27
Table 2. Depth intervals for each cluster and observations based on the lithology and gas hydrates present. Major cluster areas are marked in dark colors, while intercalated areas are marked in pale ................................................................. 33
Table 3. Distribution of the SCIMPI modules for Sites CAS05 and CAS10 and major features expected for that depth ................................................................. 37

CHAPTER 3
Table 1. Number of samples for each lithology on each expedition ......................... 54
Table 2. Total accuracy range for both expeditions calculated for all models (LDA: Linear Discriminant Analysis, RF: Random Forest, SVM: Support Vector Machines) and datasets (up-balanced, down-balanced and unbalanced) ...... 65
Table 3. Accuracy range for each lithology on both expeditions for all models (LDA: Linear Discriminant Analysis, RF: Random Forest, SVM: Support Vector Machines) and datasets (up-balanced, down-balanced and unbalanced) ...... 65
Machines) and datasets (up-balanced, down-balanced and unbalanced) .... 68

Table 4. Confusion matrix for the RF up-sampled dataset for GBREC ................. 72
Table 5. Confusion matrix for the RF up-sampled dataset for Tahiti ..................... 73
Table 6. Confusion matrix for the LDA unbalanced dataset for GBREC ................. 74
Table 7. Confusion matrix for SVM down-sampled dataset for Tahiti ................. 75

CHAPTER 4

Table 1. Site information, modern and reconstructed LGM values for chloride concentration and salinity and relative difference of LGM salinity compared to modern salinity ................................................................. 97
LIST OF FIGURES

FIGURE       PAGE

CHAPTER 1

Figure 1. SCIMPI schematic showing the modules connected to each other to form a 
    single string that is deployed in an open borehole to depths up to 300 mbsf ... 4

Figure 2. Cutaway view of: PTR (Pressure-Temperature-Resistivity) measurement 
    module, TRB (Temperature-Resistivity-Batteries) measurement module and 
    PTRB (Pressure-Temperature-Resistivity-Batteries) measurement module .... 7

CHAPTER 2

Figure 1. Location of the SCIMPI deployment Sites ................................................. 22

Figure 2. Major clusters identified in the area analyzed and their relation in depth to 
    the variables analyzed..................................................................................... 31

Figure 3. Depth locations selected for the nine SCIMPI modules (not to scale) based 
    on the PCA and CA analyses shown on the seismic reflection record from the 
    Cascadia margin site. Images of the actual SCIMPI sensor module and 
    command module are also shown...................................................................... 38

CHAPTER 3

Figure 1. Location of Expedition 310, Tahiti Sea Level ........................................... 50

Figure 2. Expedition 325, Great Barrier Reef Environmental Changes .................. 51
Figure 3. Process to combine the descriptions and physical properties data into the database used for this study................................. 55

Figure 4. Data analysis diagram for all models................................................. 62

CHAPTER 4

Figure 1. Bathymetric map of the eight sites used for this study...................... 89

Figure 2. Salinity profiles plotted with depth below seafloor for all the sites......... 96
CHAPTER 1

SCIMPI: A NEW BOREHOLE OBSERVATORY

Tania Lado Insua\textsuperscript{1, *}, Kathryn Moran\textsuperscript{1, 2, 3, 4}, Ian Kulin\textsuperscript{2, 4}, Stephen Farrington\textsuperscript{5}
and James B. Newman\textsuperscript{1, 6},

Manuscript submitted to Scientific Drilling

\textsuperscript{1} Department of Ocean Engineering, University of Rhode Island, Bay Campus, Narragansett, RI, 02882
\textsuperscript{2} Graduate School of Oceanography, University of Rhode Island, Bay Campus, Narragansett, RI 02882
\textsuperscript{3} Ocean Networks Canada, University of Victoria, PO BOX 1700 STN CSC, Victoria, BC V8W 2Y2 Canada
\textsuperscript{4} NEPTUNE Canada, PO Box 1700 STN CSC, Victoria, BC V8W 2Y2
\textsuperscript{5} Transcend Engineering and Technology, LLC, P.O. Box 222, Gaysville, VT 05746
\textsuperscript{6} Woods Hole Marine Systems Inc., PO Box 164, Woods Hole, MA 02543

*Correspondence author: T. Lado-Insua (e-mail: ladoinsuat@egr.uri.edu)
ABSTRACT

Recently, the scientific community has begun to incorporate long-term ocean and subseafloor observatory data for studying the dynamics of the earth system.

The Simple Cabled Instrument for Measuring Parameters In-Situ (SCIMPI) is a new observatory instrument designed to study dynamic processes below the seafloor. SCIMPI performs time series measurements of temperature, pressure and electrical resistivity at a series of depths, tailored for site-specific scientific objectives. SCIMPI’s modular design enables tailoring of the type, depth distribution, and frequency of measurements based on the study goals and sediment characteristics. The instrument can be configured for autonomous or cabled observatory deployments and has successfully undergone a number of tests, including pressure, communications, battery life, and interfacing with other drill ship equipment. Here we discuss the design of the instrument, its capabilities, and the testing process it has passed during four years of development. SCIMPI currently awaits its first deployment on the Cascadia Margin within the NEPTUNE Canada observatory network on IODP Expedition 341S, Scheduled for May 2013.

Key words: borehole observatory, gas hydrate, time series, logging, drilling
1. INTRODUCTION

A full understanding of earth system science requires the study of elemental (e.g., carbon and nitrogen) and heat fluxes across the seafloor-ocean boundary. Polar areas are currently responding the greatest to global warming and are of special interest for answering scientific questions about the processes of destabilization of permafrost and hydrates, which may accelerate the release of methane into the ocean, and potentially the atmosphere. The amount of methane released from warming and its impact as a positive feedback in climate change remain undetermined.

Methane reservoirs are also ubiquitous in other areas of the deep ocean in hydrate form and are an important, little-known component of the carbon cycle whose effect on global warming as greenhouse gases is a major component of this problem (Kvenvolden 1995, Beerling et al. 2009).

SCIMPI (Figure 1) is a borehole observatory instrument for placement in unconsolidated sediments. It will operate for two to four years on internal batteries that can be replenished via remotely-operated vehicle (ROV) providing high spatio-temporal resolution measurements of the physical properties in the sediment, or connected to a cabled observatory system for real-time data acquisition. With either periodic battery replacement or connection to a cabled observatory, SCIMPI provides long-term observations for understanding sub-seafloor dynamics such as changes in seafloor and sub-seafloor gas hydrate systems. The main advantages of SCIMPI are its configurability, comparatively low equipment cost, and simple operational requirements, making it an economical and versatile system for scientific research.
Figure 1. SCIMPI schematic showing the modules connected to each other to form a single string that is deployed in an open borehole to depths up to 300 mbsf.

2. SYSTEM CONFIGURATION AND SPECIFICATIONS

SCIMPI is designed for dynamic geotechnical conditions in which the borehole closes in on the device once the drill string through which it is emplaced is withdrawn. Borehole relaxation occurs because of two different processes: slower, creep-dominated deformation in fine-grained clays and shales, and immediate collapse in uncemented coarse-dominated sediments.

A SCIMPI string consists of multiple measurement modules and a command module connected by varying lengths of cable, with a ballast weight at the bottom of
the string and flotation distributed along the string to keep the cable taut during deployment. It is powered and controlled via an underwater-mateable connection to either a ROV-replaceable command module or cabled observatory infrastructure. The modules and cable segments are physically daisy-chained while power and communications are on a multi-drop bus, a combination that provides several benefits. Each SCIMPI module is composed of interchangeable housing components that, with cables of varying length, allow any combination and distribution of sensors to be formed. Sensors currently included in SCIMPI modules are a Seabird Electronics SBE-38 precision oceanographic thermometer and a Paroscientific Digiquartz series 4000 pressure sensor, both commonly used in ocean sciences to characterize sub-seabed environments having dynamic fluid flow. Resistivity is measured by an Electrical Resistivity Smart Sensor (ERSS) from Transcend Engineering and Technology which acquires measurements from any Wenner-style (e.g., galvanically coupled, 4-electrode) resistivity array. SCIMPI measurement modules each incorporate a Wenner-style array for measuring electrical resistivity of borehole sediments. The ERSS configuration can be tailored to the specific requirements of the deployment with the SCIMPI default set for a range of 0.1-100 ohm-m using bi-polar excitation at 100 Hz. The ERSS does not correct internally for ambient temperature, as thermal effects are media-dependent and should be considered in relation to the deployment medium during data analysis. All SCIMPI sensors are factory calibrated by their manufacturers. Additional 5 to 15 VDC sensors that communicate via RS-485 can be added without requiring software or circuit modifications.
Other benefits of the design are efficient power management and flexible communications. Each measurement module contains a Texas Instruments MSP430-based process responsible for powering and communicating with up to four sensors on an internal two-wire RS-485 bus (Figure 2). Each sensor's power is isolated and individually switchable. A low voltage signal from the command module activates all measurement modules when needed, causing them to latch on to the external power bus, boot up, and await instructions.

Inactive Measurement Modules consume a few microwatts of power. Active modules respond to a Modbus RTU-compliant protocol over an external two-wire RS-485 bus that connects all modules. The protocol encapsulates each sensor's command set, so software revisions are not necessary to support new sensors. Power is supplied to all modules via a two-wire bus that allows battery packs to be distributed anywhere in the SCIMPI system string. Internal battery packs deliver 16,500-mAh at 14v using primary lithium thionyl chloride cells. Distributing the battery power throughout the modules limits the required module size to allow a complete autonomous SCIMPI to be deployed through the drill pipe without needing an ROV. Back-charging is prevented so that power initially provided by modules down in the borehole can be replaced by power from a module containing fresh batteries at the seafloor.
Figure 2. Cutaway view of: PTR (Pressure-Temperature-Resistivity) measurement module (above), TRB (Temperature-Resistivity-Batteries) measurement module (center) and PTRB (Pressure-Temperature-Resistivity-Batteries) measurement module (below).
The Command Module, which remains above the seafloor, contains power conversion and seawater isolation circuitry, a programmable datalogger, and can be extended to hold additional battery packs. The datalogger is a single-board computer (SBC) based on an ARM9 processor running Microsoft's dot-NET Micro Framework. To conserve power, the SBC is supervised by a TI MSP-430 microcontroller that includes a real time clock, keeps track of the datalogger's status, and powers up the dot-NET system to acquire data according to the user-defined schedule. The system records data to log files in ASCII format with UTC time stamps on a 32-GB SD memory card.

SCIMPI is configured via a Windows application, "SCIMPI Config", provided by Transcend Engineering and Technology. The program enables the user to specify sensor commands, sequences of sensor power switching and polling, and the overall polling interval. The command module automatically switches between different operational modes depending on the status of its "up-hole" and "down-hole" ends. Cabled mode is automatically entered whenever power (9 to 80VDC) is applied to the up-hole end, such as through a wireline umbilical or cabled network. In cabled mode, SCIMPI acquires measurements continually and writes time-stamped ASCII data to the up-hole RS-485 bus in addition to the internal SD memory card. When up-hole power is removed, SCIMPI enters autonomous mode. In this mode, the command module electrically isolates the up-hole connector to avoid any connection with seawater, and enters an ultra-low-power sleep mode that is interrupted at a user-specified interval to acquire and record measurements. SCIMPI will be lowered into its borehole in cabled mode while powered via the drill ship's wireline cable. Upon
release, it will automatically switch to *autonomous* mode powered by the internal batteries. If connected via ROV to cabled observatory power, SCIMPI will again enter *cabled* mode and produce data continually.

SCIMPI is designed for a water depth of 6000 mbsl and a sediment depth up to ~300m below seafloor. Although the first SCIMPI prototype is limited by its flotation material to 1500 mbsl, the flotation of the system can be adapted to different site depths. The flotation keeps the SCIMPI cable taut during installation with the underwater-mateable connector and command module accessible above the seafloor during installation.

The first SCIMPI prototype is equipped to measure temperature, electrical resistivity, and pressure (Table 1). The design allows other sensors to be easily added, and we expect to incorporate other types of measurements in the future.

Table 1. Specifications of the SCIMPI prototype.

<table>
<thead>
<tr>
<th>Temperature (Modified Seabird SBE-38)</th>
<th>Electrical Resistivity Smart Sensor (ERSS)</th>
<th>Pressure (Paroscientific Model 410K-101)</th>
<th>Housing (Standard Measurement Module)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range: -5 to +35°C</td>
<td>Range: 0.1 to 98 ohm-m (default reconfigurable)</td>
<td>Range: 0 to 10000 psi (other ranges available)</td>
<td>Diameter: 76.2 mm</td>
</tr>
<tr>
<td>Resolution: 0.00025°C</td>
<td>Absolute accuracy: 0.025 ohm-m</td>
<td>Absolute accuracy: 0.01%</td>
<td>Length: 870mm</td>
</tr>
<tr>
<td>Absolute accuracy: +/-0.001°C</td>
<td>Drift: +/- 0.0001°C per 6 months</td>
<td>Resolution: 0.0001%</td>
<td>Weight: 16.4 kg</td>
</tr>
<tr>
<td>Deployment accuracy (3 years): +/-0.007°C</td>
<td>Resolution: 0.00001 ohm-m (reconfigurable)</td>
<td>Design pressure: 8760 psi</td>
<td>Material: 17-4PH Stainless Steel, condition H1025</td>
</tr>
</tbody>
</table>
3. TESTING AND DEPLOYMENT

All SCIMPI housings were designed and pressure tested to 60 MPa and all sensors are factory calibrated by their manufacturers and inspection tested prior to integration. The integrated system was tested both on land and at sea, with final pre-deployment testing completed in November 2011. We are now preparing SCIMPI for its first deployment in an Integrated Ocean Drilling Program borehole. SCIMPI was specifically designed to be deployed through the drill pipe of the D/V JOIDES Resolution so no reentry cone or casing is required. The borehole will be drilled to the target depth using a bit with a diameter of 9-7/8”. The bit will be dropped and SCIMPI will be lowered down the drill string, suspending it when needed in the Schlumberger C-plate as it is fed into the drill pipe. During deployment, the multi-function telemetry module (MFTM) designed by Lamont Doherty Earth Observatory Borehole Research Group enables continuous serial communication with SCIMPI through the Schlumberger wireline cable. Once lowered to target depth in the borehole, SCIMPI is released using the electrical release system (ERS) designed and built by Stress Engineering, the wireline cable, ERS, and MFTM are retrieved, and the drill pipe is stripped.

When deployed as an autonomous instrument, SCIMPI requires servicing by an ROV every two to four years to swap out the command module. If SCIMPI is connected to a network for electrical power and real time data reporting, then an ROV is required only to complete the initial connection.
4. FLEXIBILITY IN THE CONFIGURATION

The adjustable spacing between SCIMPI modules is one of the major advantages of the instrument. This spacing can be flexibly adjusted during a drilling expedition using core recovery and log data from the deployment site to determine optimal module positioning. To ensure this flexibility, spare cables with varying lengths will be carried on board allowing the science team to configure the instrument based on sub-seafloor characteristics determined from the interpretation of the first holes drilled. Lado-Insua et al (2012) has developed a methodology to calculate the number and optimal distribution of modules so that long-term data from SCIMPI captures the sediment intervals with most potential for determining dynamic fluid flow processes.

For example, the formation of gas hydrates requires particular conditions of pressure and temperature. SCIMPI’s sensors make it ideal for the study of gas hydrate dynamics. The stability of gas hydrates is generally assessed based on geothermal gradients (Gorman & Senger 2010). These models could be better determined by using several SCIMPI installations in an area. Spatial and temporal gradients in pressure and temperature can be determined to study lateral changes in the geothermal setting. The use of SCIMPI in these environments can provide unique information on gas hydrate changes in electrical resistivity and overpressure due, for example, to processes related to changes in pore pressures. We expect SCIMPI to provide insight into the dynamics of gas hydrates in areas affected by climate change (e.g. in the Arctic where warming is amplified), fluid flow (e.g. subduction zones), and methane release (e.g. outer continental shelves and slopes).
SUMMARY

SCIMPI is a new seafloor observatory in its final stages of development. A prototype has undergone thorough testing and is currently ready for deployment. SCIMPI can measure long-term time series of temperature, pressure and electrical resistivity spatially at multiple depths below the seafloor. The instrument is highly modular and customizable to different environments, different spatial distributions of sensors, and autonomous as well as cabled operation. SCIMPI is designed for deployment into soft sediments directly through the drill string to minimize equipment and operational costs and can operate autonomously for several years on battery power.

Science applications of SCIMPI include studies of fluid flow, hydrate formation, and seismically induced pore pressure changes. The Arctic Ocean is an area particularly suited for SCIMPI deployment. Kitidis et al. (2010) demonstrated the existence of water and sediment sources of methane and nitrous oxide in the Arctic Ocean, that indicate that future sea-ice retreat may increase the flux of these gases from the sea to the atmosphere. They pointed out the importance of spatio-temporal studies of these systems. The use of SCIMPI to study the variations over long-time series is critical for understanding their dynamics.

SCIMPI's comparatively low cost will enable time series measurements to become more commonplace, thereby improving our temporal and spatial knowledge of sub-seafloor gas, fluid, and pore pressure activity. Most notable among potential deployment targets for SCIMPI are sites with sub-seafloor gas hydrate and those with biogenic methane. Understanding the dynamics of methane’s role in the oceans as
climate change proceeds will contribute to a better understanding of the Earth’s carbon budget.

ACKNOWLEDGMENTS

SCIMPI is a collaborative effort funded by IODP-MI and led by the University of Rhode Island with major contributions from Transcend Engineering and Technology LLC and Woods Hole Marine Systems Inc. (WHMSI). Stress Engineering and Lamont Doherty Earth Observatories (LDEO) provided technical input and the deployment equipment. The authors would like to thank the R/V Endeavor crew. We are also thankful to Tori Kulin and Lucy Hurlbut for their support during the development of this project. Tania Lado Insua was funded during this project by Fundación Pedro Barrié de La Maza, NSF and IODP-MI.

REFERENCES


HIGHLIGHTS

SCIMPI is a new borehole observatory designed to record physical properties time series.

The first prototype is able to measure temperature, electrical resistivity and pressure.

SCIMPI can run on batteries that require refreshment after two years, or it can be powered from a cabled network to provide real time measurements.

SCIMPI is ideal for studies of dynamic environments such as gas hydrates
CHAPTER 2

A MULTIVARIATE APPROACH TO OPTIMIZE BOREHOLE OBSERVATORY DESIGNS

Tania Lado Insua¹, *, Kathryn Moran¹, ², ³, ⁴, Ian Kulin², ⁴, Sally Morgan⁵, Stephen Farrington⁶ and James B. Newman¹, ⁷,

Manuscript to be submitted to Geosphere

¹ Department of Ocean Engineering, University of Rhode Island, Bay Campus, Narragansett, RI, 02882
² Graduate School of Oceanography, University of Rhode Island, Bay Campus, Narragansett, RI 02882
³ Ocean Networks Canada, University of Victoria, PO BOX 1700 STN CSC, Victoria, BC V8W 2Y2 Canada
⁴ NEPTUNE Canada, PO Box 1700 STN CSC, Victoria, BC V8W 2Y2
⁵ Borehole Research Group Department of Geology, University of Leicester, University Road, Leicester, LE1 7RH, UK
⁶ Transcend Engineering and Technology, LLC, P.O. Box 222, Gaysville, VT 05746
⁷ Woods Hole Marine Systems Inc., PO Box 164, Woods Hole, MA 02543

*Correspondence author: T. Lado-Insua (e-mail: ladoinsuat@egr.uri.edu)
ABSTRACT

Long-term monitoring of the subseafloor has seen increased application in the past few decades. Systems such as the Circulation Obviation Retrofit Kit (CORK) have been used since the 1970s providing the scientific community with time series measurements of geophysical properties below the seafloor and in the latest versions with temporal pore water sampling. The Simple Cabled Instrument for Measuring Parameters In-Situ (SCIMPI) is a new subseafloor observatory to be deployed in May 2013 in Cascadia as part of the NEPTUNE Canada observatory. SCIMPI is designed to measure physical properties over time with a design that simplifies the deployment and reduces the costs of operations.

The subseafloor sensor depth positions within an observatory borehole have been determined in the past by identifying prominent features, i.e. the major lithologies or major flux areas based on individual analysis of the physical properties and downhole logging measurements of the site. The separation distance of sensor modules for SCIMPI has been selected using a cluster approach for identifying the most significant depth intervals. Two potential deployment sites close to a cold vent area on the Cascadia margin are presented. Where borehole data are available in the surrounding area, this approach can optimize the ideal locations using an unbiased procedure that takes into account relationships among the data types that are not obvious when each value is analyzed independently. SCIMPI’s first deployment will use the configuration explained here.

Key words: borehole observatory, time series, cluster, downhole logging, gas hydrates
1. INTRODUCTION

Many geological oceanographic research objectives require a dynamic approach to fully understand systems, such as the carbon cycle, that changes over relatively short time-scales of months to years (e.g. IPCC 2007, Skinner et al. 2010, Zickfeld et al. 2012). Subseafloor dynamic systems, e.g. methane hydrates, are dependent upon in situ hydrogeology and variations in temperature and pressure, which all remain poorly understood in the subseafloor. Methane is a powerful greenhouse gas that can produce 25 times the potential warming of CO$_2$ for a 100 year period (Forster et al. 2007, reviewed in Bridgham et al. 2013). The most recent global methane emissions estimates indicate that between 500 and 600 Tg of methane are released to the atmosphere every year (Dlugokencky et al. 2011).

Gas hydrates are ice-like solid structures in which gas (mostly methane) is trapped in cavities formed by water molecules in solid phase. There is no chemical bonding between the gas and water molecules and when they dissociate from the solid into a mixed fluid phase the volume increases with severe effects on the fluid pressure, volumetric deformation and effective stress and strength of the sediment (Waite et al. 2009). Gas hydrates are studied by industry as a new energy source with the potential for production at commercial scales by 2016 (Koh et al. 2012, Waite et al. 2009). The increased awareness of climate change and the still unknown role of gas hydrates in global warming (Ruppel 2011, IPCC 2007, Buffett and Archer, 2004) indicate the need to monitor gas hydrates and include them as a major point in the debate on climate change (Krey et al. 2009).
1.1. Physical properties and pore water in gas hydrates

Due to the destabilization that the gas hydrates suffer when they are retrieved from the seafloor, most of our knowledge about their physical properties and characteristics has been achieved in laboratory experiments (e.g. Ren et al. 2010), theoretical calculations and models (e.g. Clennell et al. 1999) and thanks to the use of wireline logging techniques and other detection methods (e.g. Riedel et al. 2006, Koh et al. 2012, Waite et al. 2009). The velocity of compression waves (P-waves) through gas hydrates is commonly used for detection of gas hydrates with seismic methods (e.g. Riedel et al. 2002). Clathrate hydrates usually appear as a strong reflector due to the change in impedance between the solid hydrate and the sediments with free gas in their pores. This reflector is named the bottom-simulating-reflector (BSR) because it usually mimics the shape of the seafloor reflector because hydrate stability in deep water is primarily a function of temperature, which has increasing vertical gradients with depth. P-wave velocities in methane hydrates reach 3600 m/s and S-wave velocities are on the order of 1900 m/s with a density of 0.912 g/cc (Dvorkin and Nur 2007). These results show high variability depending on the sediment surrounding the gas hydrates and significant differences have also been reported among different laboratories (Waite et al. 2011). When free gas is present (e.g. BSR) these values drop substantially (MacKay et al. 1995, Riedel et al. 2006).

Gas hydrates also have a higher electrical resistivity than oceanic sediments (e.g. Ren et al. 2010, Hyndman et al. 1999). This is due to the displacement of the salt in the pore fluid during their formation. This distinctively high electrical resistivity makes gas hydrates easily detected in wireline logs. The type of hydrate formed
determines the presence or absence of higher concentration of salt in a brine solution during the process surrounding the hydrate formation (Judge 2007 reviewed in Gabitto and Tsouris 2010). Gas hydrate content can be accurately estimated based on electrical resistivity and P-wave velocity measurements (Koh et al. 2012, Riedel et al. 2006).

Generally, the wireline logging measurements correlate well with the equivalent measurements obtained from the recovered cores. The quality of the samples recovered strongly depends on the use of pressurized systems to bring the sample to surface without suffering major dissociation (e.g. Santamarina et al. 2012, Riedel et al. 2006). However, there is still a limitation on the transference of laboratory experiments to the field and on the large variations among laboratories (Waite et al. 2011).

Another physical property measurement that can be useful in the detection of gas hydrates is low values of magnetic susceptibility. Novosel (2005) and Riedel et al. (2006) reported lower magnetic susceptibility values on the upper 30 meters below seafloor (mbsf) at Site U1328 related to gas hydrate formation.

Natural gamma ray variations reflect major changes in lithology with higher values related to more clay-rich units and lower values associated with intervals with lower clay content (Westbrook et al. 1994). In general natural gamma ray can help to identify lithofacies. Gas hydrate formation in the sites studied occurs predominantly in sandy formations (Riedel et al. 2006) that should be easily identifiable by magnetic susceptibility and natural gamma radiation.

Pore fluid chloride concentrations show a strong relationship with the presence of gas hydrates (Westbrook et al. 1995). High dilutions of chloride correspond to high
gas hydrate concentrations (up to ~36%). The dissociation of the gas hydrates, adding fresh water to the pore fluids prior to analyses can also cause this decrease in chlorinity (Riedel et al. 2006).

A full understanding of gas hydrate dynamics in situ can only be achieved with long term observations of changes in these systems.

1.2. Subseafloor observatories

An important science facility that has advanced our understanding of marine geodynamics is the Integrated Ocean Drilling Program (IODP) that, together with the Ocean Drilling Program (ODP) and the Deep Sea Drilling Program (DSDP), has been advancing sub-seafloor research since the 1960s. Thousands of sites have been drilled since these programs started, making data and samples available for several scientific disciplines. Here we use data from the IODP to optimize the selection of long-term observatory sensor depth intervals for dynamic studies of gas hydrate. Of primary importance to dynamic seafloor systems are sediment physical properties, which constrain the rates and volumes of fluid and gas flow. Measurements of physical properties of recovered sediments and rocks comprise one of many datasets collected during scientific ocean drilling.

The latest advances in scientific ocean drilling include the installation of long-term observatories in the subseafloor. CORKs (Circulation Obviation Retrofit Kit) have been used for several decades to study in situ fluid flow dynamics (Becker and Davis, 2005) primarily in hard rock formations. They comprise a reentry cone, a mud skirt and a casing hanger for running up to four sizes of nested casing that get
sequentially smaller deeper into the hole. The CORK body seals the borehole with strings of sensors inside of it. Pressure gauges and thermistor cables have been used since the first models, while osmo-samplers used to sample pore water flowing through screens located in the casing are a later addition to CORK II. The different spaces in the CORK structure are separated by packers that avoid flow between the different sections of the system (Becker and Davis 2005).

The new modular borehole observatory, SCIMPI, was developed to expand sub-seafloor observatories into softer sediments, increase depth intervals, simplify deployment and reduce costs. The first SCIMPI prototype is equipped with temperature, pressure and electrical resistivity sensors designed to record time series of measurements. Its battery life is 2 years after which batteries can be replaced together with the recovery of the command module where the data are recorded, or the system can be connected to a land network that can act as a source of power and a channel for real time data transmission. A more detailed description of SCIMPI can be found in Lado-Insua et al. (submitted, see Chapter 1).

As in any engineering development project the location and number of sensors in the observatory is a compromise between the scientific objectives and cost. Here we present an analysis based on physical property data from ODP and IODP to determine the subseafloor depth placement of the sensors to optimize SCIMPI for long-term observations.
1.3. SCIMPI deployment in Cascadia

Three priority Sites were selected in Cascadia, all of them located within the area of the NEPTUNE Canada observatory (Barnes et al., 2011). Here we present the dimensioning for the first priority and second priority sites. All the selected sites are located in an area of cold venting where gas hydrates are the source. The most prominent venting in the area is named the Bullseye Vent, an active venting feature that has been previously drilled (IODP Expedition 311, Site U1328). This site, located offshore the west coast of North America (Figure 1) constitutes one of the best-documented examples of dynamic gas hydrate environments.

Figure 1. Location of the SCIMPI deployment Sites (Image courtesy of M. Riedel)
The priority site for deployment, CAS05-CORK, is located northeast of the Bullseye Vent in an area of vigorous degassing. CAS05-CORK was located 400 m away from Site U1328 sampled during IODP Expedition 311. Site CAS10-SCIMPI is the second option for deployment. This Site is located close to the priority site and only 100m away from Site U1328. Water depth in all the sites is similar (1257m for CAS05-CORK and 1262m for CAS10-SCIMPI). All the sites proposed present deep-sea marine mud interbedded with sand and silt rich turbidites (Riedel et al. 2006).

1.4. Preliminary information about the sites

A major drilling expedition took place on the Cascadia continental margin, IODP Expedition 311. This expedition drilled, recovered core samples, and conducted wireline logging (WL) and/or logging while drilling (LWD) at sites in close proximity to the selected location (IODP Site U1328) close to the Bullseye Vent (Figure 1). During the expedition, pressure coring was used, which aided in achieving many of the objectives of this ambitious gas hydrate scientific drilling expeditions. Site U1328 has a water depth of ~1267.9 meters below sea level (mbsl). It is a location with fluid flow that presents a geothermal gradient of ~54°C/km. The bottom simulating reflector (BSR) depth varies slightly at each of the borehole locations, and small variations depend also on the variables studied for its location. The Bullseye Vent shows carbonate sheets covering the seafloor area and a shallow gas hydrate cap that can be visualized from 3D seismics (Riedel et al. 2006). The estimated hydrate concentration in the sediment column at this site varied from 0.7% to 38% and geochemical evidence of gas hydrate formation is higher in the upper few meters of
the boreholes (Riedel et al. 2006).

The physical properties measured at Site U1328 include bulk density, acoustic compressional wave velocity (P-wave), magnetic susceptibility, electrical resistivity, and natural gamma radiation. In this study, physical property data were supplemented with some pore fluid chemistry for the analyses. The tool of choice for the non-destructive measurement of physical properties on recovered cores is the multi-sensor core logger (MSCL), a standard instrument used in IODP. Physical properties are also measured and complemented by downhole techniques including logging while drilling and wireline logging of the open borehole. These logging approaches provide a better measure of the actual in situ characteristics of the sediment and rock because they are not influenced by sample disturbance.

The logging while drilling density measurements were of good quality for Hole U1328A. The wireline logging quality for Hole U1328C was compromised for the formation microscanner tool due to poor hole conditions. But both density and acoustic data quality for the hole are good. Three logging units were identified that matched the lithological units (Riedel et al. 2006).

1.5. Location of the sensors

The methodology presented here has been applied for the configuration of the current SCIMPI prototype. At the time of designing, the distribution of the prototype sensors was limited by the depths where the modules were located. Due to budget and time constraints, the cable lengths were dimensioned and purchased for a different site. The deployment site was afterwards relocated to the Cascadia margin, based on
the route of the D/V JOIDES Resolution and its proximity to NEPTUNE Canada. The length optimization approach was applied to data from existing sites on the Cascadia margin. The cable lengths, which are interchangeable in depth, were re-organized based on the output of this cluster approach. Initially 7 modules were going to be deployed, but the change in location and the interest to cross the BSR in this first deployment meant that 9 modules will be used in the first deployment to sufficiently span the depth below seafloor.

This dimensioning approach is not limited to this SCIMPI application and can be used for CORK designs or other observatory installations. It can also aid in identification of depth intervals of significance in dynamic fluid flow subseafloor settings and in this case it shows the major features using a wide range of datasets (wireline logging, logging while drilling, physical properties and pore water). Similar approaches are becoming more common in the literature (Shepherd et al. 1987, Pelling et al. 1991, Tudge et al. 2009).

2. MATERIALS AND METHODS

2.1. Data from previous expeditions

Measurements of bulk density (from MSCL and logging data), electrical resistivity, magnetic susceptibility, natural gamma, and pore water calcium and chlorinity from the recovered cores were used in our analyses in combination with wireline logging and logging while drilling data (Table 1). When different logging tools were used for the same measurement, they were included as different variables due to the differences in calibration that the logs presented. SCIMPI is equipped with
sensors for measuring electrical resistivity, pressure and temperature, making the system ideal for study of dynamic flow systems because changes in sub-seafloor fluid flow would be reflected to varying degrees in these measurements. The borehole variables used in these analyses are also sensitive to fluid flow changes, thus the degree of their collective variability will identify the depth intervals of hydrogeological significance.

In order to dimension the priority site CAS05-CORK and the secondary site CAS10, datasets from U1328 were used. We combined data from different boreholes to create a unique dataset that integrated the values obtained for the different boreholes as an average of the values with depth. These new variables were considered to be representative of an average environment in the study location, since they capture the variability in closely located boreholes. In order to relate records from the borehole and the recovery, a core-log correlation of the variables used for this study was performed before their integration. The core-log integration was based on densities since no magnetic susceptibility was available for the wireline logging and logging while drilling data and no natural gamma radiation was measured on the core. Thanks to the high recovery (>85%) the core-log integration was not a limitation in this study. All variables were interpolated each 10 cm with depth since precision in deployment would make it complicated to work with smaller intervals. Data affected by the pipe on the logging datasets was not used for the analysis since all the variables need to be present in order to perform it and we considered that an extrapolation would produce unreasonable results. The same applies for the depths where logging while drilling or wireline logging data were absent, in this case sonic P-wave velocity. Therefore the
cluster analysis for U1328 only covers depths below 61.1 mbsf (pipe depth during wireline logging operations) down to 227.9 mbsf (end of the sonic data record).

Table 1. Variables used in the analyses with the tool and method of measurement used, and the geographic location and depth below seafloor locations for Sites CAS05-CORK and CAS10-SCIMPI, based on Site U1328

<table>
<thead>
<tr>
<th>Variable obtained and units</th>
<th>Tool and measurement</th>
<th>Hole</th>
<th>Instrument</th>
<th>Depths (mbsf)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Bulk Density (g/cc)</td>
<td>MAD discrete samples (BD)</td>
<td>B</td>
<td>MAD discrete samples</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>MAD discrete samples (BD)</td>
<td>C</td>
<td>MAD discrete samples</td>
<td>56.7</td>
</tr>
<tr>
<td>Magnetic Susceptibility (System Units)</td>
<td>MSCL (MS)</td>
<td>B</td>
<td>MSCL</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>MSCL (MS)</td>
<td>C</td>
<td>MSCL</td>
<td>56.52</td>
</tr>
<tr>
<td>Chlorinity (mM)</td>
<td>Chlorinity</td>
<td>B</td>
<td>Interstitial water</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Chlorinity</td>
<td>C</td>
<td>Interstitial water</td>
<td>60.68</td>
</tr>
<tr>
<td>Pore water calcium (mM)</td>
<td>Calcium</td>
<td>B</td>
<td>Interstitial water</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Calcium</td>
<td>C</td>
<td>Interstitial water</td>
<td>60.68</td>
</tr>
<tr>
<td>Bulk density (g/cc)</td>
<td>Hostile Environment Lithodensity Tool (RHOM)</td>
<td>C</td>
<td>Wireline Logging</td>
<td>59.14</td>
</tr>
<tr>
<td></td>
<td>LWD Density Log-adVISION Tool (RHOB)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LWD Density EcoScope (RHOB)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>14.05</td>
</tr>
<tr>
<td>Electrical Resistivity</td>
<td>Dual Induction Tool (IDPH)</td>
<td>C</td>
<td>Wireline logging</td>
<td>61.12</td>
</tr>
<tr>
<td></td>
<td>LWD Phase shift Resistivity EcoScope (average all variables)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>12.06</td>
</tr>
<tr>
<td></td>
<td>LWD Attenuation Resistivity EcoScope (average all variables)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>12.06</td>
</tr>
<tr>
<td>Gamma Ray</td>
<td>Hostile Environment Gamma Ray Spectrometry Sonde-Uplog (HSGR)</td>
<td>C</td>
<td>Wireline Logging</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LWD Gamma Ray GeoVISION Tool (GR)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LWD Gamma Ray EcoScope (GRMA)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>14.05</td>
</tr>
<tr>
<td>P-wave velocity</td>
<td>Dipole Sonic Imager Pass 1 (VCO)</td>
<td>C</td>
<td>Wireline Logging</td>
<td>50.14</td>
</tr>
<tr>
<td></td>
<td>LWD Sonic Velocity (VELP)</td>
<td>A</td>
<td>Logging while drilling</td>
<td>7.5</td>
</tr>
</tbody>
</table>
2.2. Statistical analysis

A principal components analysis (PCA) and a k-means cluster analysis (CA) were applied as a multivariate approach to the data. Depth below seafloor was used separately from the other variables and not included in the PCA or CA process since it would create dependences and bias the analysis. But, results are plotted against depth to make their visualization easier.

Multivariate statistics are an important tool in the analysis of datasets with several variables. The PCA (Mardia 1979) consists of transforming a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components reveal the internal structure of the data explaining their variance. The use of PCA before CA has been proven to be a good tool to reduce the dataset and standardize it (Ding and He 2004). In our case, 12 principal components were used which represented 98.36% of the variance in the dataset.

We used a k-means clustering based on the Hartigan and Wong (1979) algorithm. This cluster analysis was applied on the principal components space to find similar trends in the data, grouping the values that present similar values. Clusters always provide grouping structure, but it is the task of the scientist to identify if those clusters have a physical meaning. They are a type of unsupervised learning because no predefined classes are assigned. This methodology groups data into clusters according to the their common characteristics. The use of \( k \)-means clusters requires the input of a
$k$ value that determines the number of clusters. A good quality cluster will define groups with high internal similarity but high dissimilarity with other groups. Different types of variables can be used in clustering (binary, nominal interval and ratio variables). The CA constructs various hierarchical partitions that define centroids and the optimal number of clusters and then evaluates them by some criterion, in this case, minimizing the within-cluster sum of squares. First, the number of clusters should have a physical meaning in the context where they are analyzed; however, a bootstrap can be run on the dataset by columns, and an assessment of the sum of squares performed within the cluster can be done for each one of the new datasets and compared to the values obtained for the original dataset (Peeples 2011). We generated 250 new datasets and compared them with the original one (Appendix I and II) using the “k.1” function programmed by Peeples (2011). We also investigated the physical meaning and distribution in depth of the different clusters. Clusters that are intercalating in depth may present characteristics of the sediment that are subtler, but such centimeter-scale variations cannot be addressed on a sub-seafloor deployment. The lateral variation in the site and the inability to physically deploy sensors in such small depth intervals mean these higher clusters are not useful for our purpose.

3. RESULTS AND DISCUSSION

3.1. Cluster results and interpretation for CAS05 and CAS10

The results for CAS05 and CAS10 are presented together since they both are based on the same dataset from previously drilled site U1328. A preliminary analysis of the number of clusters indicates the presence of 3 major areas ($k=3$, Appendix I).
We also discuss here the differences for $k=4$ and $k=5$.

Three major clusters were differentiated in this site (Figure 2, Table 2). No points in depth belong to two different clusters at the same time, and therefore, areas where two clusters intercalate indicate prominent features that are more similar to another cluster or they can indicate an area of change from one cluster to the next.

Cluster 1 corresponds to the upper depths analyzed going from 61 mbsf down to 123.5 mbsf, with two gaps corresponding to cluster 2 from 71.7 to 73.9 mbsf and 96.8 to 107 mbsf. Between 122.9 and 129.8 mbsf the three clusters show alteration with depth. A single point (123.0 mbsf) and small interval of ~40cm (123.6 to ~124 mbsf) were identified as cluster 3. Clusters 1 (124.1 to 127.9 and 128.2 to 129.8 mbsf) and 2 (128 to 128.1 mbsf) intercalate again before an extended interval in depth identified as cluster 2. This is consistent with a change in characteristics of the sediment from one cluster to the next. Since the model has to identify points as one single cluster, in areas where the difference is subtle, clusters may alternate in depth. Some small areas from cluster 1 extend below these depths (Figure 2, Table 2), appearing in the middle of clusters 2 and 3. A major unit (I) was identified from downhole logging data that broadly matches the distribution of Lithostratigraphic Unit I (determined by core description) from 0 to ~128 mbsf. This unit largely corresponds with our cluster 1. In Unit I (0 to 128 mbsf) the density increases with depth (1.7 to 2 g/cc), the P-wave velocity average is ~1550 m/s, and the electrical resistivity shows higher values between 0 and 46 mbsf alternated with lower values, indicating the presence of gas hydrates. This lithostratigraphic unit is identified as being composed of dark gray and dark greenish gray clay and silty clay often interbedded with clayey silt, silt, sandy-
silt-clay, silty sand and sand layers. Soupy texture was found in this unit indicating the presence of gas hydrates (<100 mbsf). More than 90% of the gas hydrates were found in sandy layers at this location. Sandy layers correspond with the lower values found for the natural gamma emission from the site in cluster 1.

Figure 2. Major clusters identified in the area analyzed and their relation in depth to the variables analyzed.
Cluster 2 covers most of the sediment column below cluster 1, from 129.9 to 190.9 mbsf, except for two single points in depth (154.9 and 155.9 mbsf) and three small intervals (160.1 to 160.7, 182.3 to 181.6 and 183.5 to 184.7 mbsf). This area corresponds to the second unit delineated by the downhole logging data (Unit II) (128 to 200 mbsf), which corresponds to the Lithostratigraphic Unit II. This unit has a decrease in density with depth (2 to 1.75 g/cc), constant electrical resistivity values (~1Ωm) except for a peak at 160 mbsf. The porosity is ~50% and the P-wave velocity is ~1600 m/s. Chlorinity and salinity have constant values with no presence of gas hydrate from 60 to 150 mbsf. However, from 150 to 250 mbsf discrete excursions of fresher water indicate the presence of gas hydrates that dissociate during the processing of the cores, these mostly correspond to clusters 1 and 3 that intercalate with cluster 2 in this analysis (Table 2). This unit, equivalent to Lithostratigraphic Unit II, is composed of dark greenish gray and dark gray clay with nanofossils, silty clay, with different presences of diatoms and diatom ooze, interbedded with sand and sandy silt layers.

Cluster 3 appears mostly from ~191 mbsf to the bottom of the hole. The Unit III identified from the in situ data (200 to 300 mbsf) has higher variability in electrical resistivity (1 to 4 Ωm) and P-wave velocities (1500 m/s to 1800 m/s). The density and porosity in this unit are similar to Unit II. Our analysis covers only the area drilled and therefore no clusters are shown below 227.9 mbsf. This cluster corresponds to Lithostratigraphic Unit III and is composed of dark greenish gray, dark gray and dark olive-gray clay and silty clay with varying presence of diatoms. At the bottom of this unit more indurated sediments were found. There are mousselike textures related to
gas hydrates in both Units II and III. Authigenic carbonates appear in the three lithostratigraphic units (and clusters) indicating formation related to the gas hydrates close to soupy and mousselike textures.

Table 2. Depth intervals for each cluster and observations based on the lithology and gas hydrates present. Major cluster areas are marked in dark colors, while intercalated areas are marked in pale.

<table>
<thead>
<tr>
<th>Depth interval (mbsf)</th>
<th>Cl.</th>
<th>Main cl. area</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.1-71.6</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations: the density increases with depth (1.7 to 2g/cc), the P-wave velocity average is ~1550 m/s, and the electrical resistivity shows higher values between 0 and 46 mbsf alternated with low values indicating the presence of gas hydrates</td>
</tr>
<tr>
<td>71.7-73.4</td>
<td>2</td>
<td>1</td>
<td>Low bulk density (MAD) values combined with low magnetic susceptibility and high gamma ray values from the Hostile Environment tool (WL) and the Geovision tool (LWD)</td>
</tr>
<tr>
<td>73.5-73.7</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>73.8-73.9</td>
<td>2</td>
<td>1</td>
<td>Low bulk density (MAD) values combined with low magnetic susceptibility and high gamma ray values from the Hostile Environment tool (WL) and the Geovision tool (LWD)</td>
</tr>
<tr>
<td>74.0-96.7</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>96.8</td>
<td>2</td>
<td>1</td>
<td>Low bulk density (MAD) values combined with low magnetic susceptibility and high gamma ray values from the Hostile Environment tool (WL) and the Geovision tool (LWD). Decrease in calcium and bigger range of values for the Gamma Ray EcoScope tool</td>
</tr>
<tr>
<td>96.9-97.1</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>97.2-105.8</td>
<td>2</td>
<td>1</td>
<td>Low bulk density (MAD) values combined with low magnetic susceptibility and high gamma ray values from the Hostile Environment tool (WL) and the Geovision tool (LWD). Decrease in calcium and bigger range of values for the Gamma Ray EcoScope tool. Lower electrical resistivity than general values from cluster 1</td>
</tr>
<tr>
<td>105.9-106.4</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>106.5-107.0</td>
<td>2</td>
<td>1</td>
<td>Low bulk density (MAD) values combined with low magnetic susceptibility and high gamma ray values from the Hostile Environment tool (WL) and the Geovision tool (LWD). Decrease in calcium and bigger range of values for the Gamma Ray EcoScope tool. Lower electrical resistivity than general values from cluster 1</td>
</tr>
</tbody>
</table>
Table 2. (Continuation).

<table>
<thead>
<tr>
<th>Depth interval (mbsf)</th>
<th>Cluster</th>
<th>Major cluster area</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>107.1-122.9</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>123</td>
<td>3</td>
<td>1</td>
<td>High values of P-wave velocity combined with intermediate to high values of electrical resistivity, and high bulk density values.</td>
</tr>
<tr>
<td>123.1-123.5</td>
<td>1</td>
<td>1</td>
<td>Cluster 1 general observations</td>
</tr>
<tr>
<td>123.6-124.0</td>
<td>3</td>
<td>1</td>
<td>High values of P-wave velocity combined with intermediate to high values of electrical resistivity, and high bulk density values.</td>
</tr>
<tr>
<td>124.1-127.9</td>
<td>1</td>
<td>1</td>
<td>LWD peak in electrical resistivity</td>
</tr>
<tr>
<td>128-128.1</td>
<td>2</td>
<td>1</td>
<td>LWD peak in electrical resistivity combined with discrete fresh water excursion. The Mg/Ca ratio presents a maximum close to this depth suggesting the formation of autigenic carbonate</td>
</tr>
<tr>
<td>128.2-129.8</td>
<td>1</td>
<td>1</td>
<td>LWD peak in electrical resistivity</td>
</tr>
<tr>
<td>129.9-154.8</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations: decrease in density with depth (2 to 1.75 g/cc), constant electrical resistivity values (~1Ωm). The porosity is ~50% and the P-wave velocity is ~1600 m/s.</td>
</tr>
<tr>
<td>154.9</td>
<td>1</td>
<td>2</td>
<td>LWD peak in electrical resistivity combined with discrete fresh water excursion</td>
</tr>
<tr>
<td>155-155.8</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations</td>
</tr>
<tr>
<td>155.9</td>
<td>1</td>
<td>2</td>
<td>LWD peak in electrical resistivity combined with discrete fresh water excursion</td>
</tr>
<tr>
<td>156-160</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations</td>
</tr>
<tr>
<td>160.1-160.7</td>
<td>1</td>
<td>2</td>
<td>LWD peak in electrical resistivity</td>
</tr>
<tr>
<td>160.8-181.2</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations</td>
</tr>
<tr>
<td>181.3-181.6</td>
<td>1</td>
<td>2</td>
<td>Slight increase in electrical resistivity</td>
</tr>
<tr>
<td>181.7-183.4</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations</td>
</tr>
<tr>
<td>183.5-184.7</td>
<td>1</td>
<td>2</td>
<td>Slight increase in electrical resistivity</td>
</tr>
<tr>
<td>184.8-190.9</td>
<td>2</td>
<td>2</td>
<td>Cluster 2 general observations</td>
</tr>
<tr>
<td>191-227.9</td>
<td>3</td>
<td>3</td>
<td>Cluster 3 general observations: higher variability in the electrical resistivity (1 to 4 Ωm) and P-wave velocities (1500 m/s to 1800 m/s). The porosity is ~50% and the density is ~2 g/cc. Low chlorinity and low calcium in the pore waters. Low magnetic susceptibility</td>
</tr>
</tbody>
</table>

In summary, if we consider the gas hydrate distribution, cluster 1 is identified as an area of significant gas hydrate occurrence with active gas hydrate formation in some areas. Cluster 2 reflects areas where gas hydrates are scarce or not present at all. Cluster 3 represents areas with gas hydrate present and free gas surrounding the methane stability zone.
Intercalation of clusters is common in the areas where clusters converge or where features more similar to other clusters occur. Working with sediment and excluding depth as a variable simply highlights that we are working in an area where the lithology, or the properties of the sediment, are alternating. The fine-scale (< 3 m) intercalation of these intervals makes it difficult to consider placement of a sensor in an individual layer, particularly because of the lateral differences encountered between sites that are located within only a few meters of each other (Sites U1327 and CAS05 and CAS10).

When extra clusters are analyzed, we can see that the most important information is in the first 3 clusters, in agreement with the calculations made to determine the $k$ value to use (Appendix I and II). A fourth cluster ($k=4$) only adds three small areas located at ~68.6-68.8 mbsf, ~93.5-95.6 mbsf and 217.1-217.5 mbsf. The first two areas overlap with cluster 1 (in $k=3$ analysis) and the last one with cluster 3. These areas are small and cannot be considered as features that will be present in the site where the instrument will be deployed. However, they seem to have a signal strong enough to differentiate them. The first and second areas at ~68.7 and ~93 mbsf seem to correspond with maximums on electrical resistivity. The first one represents a local maximum and the second area is a global maximum that seems to correspond with an area of high gas hydrate content evidenced by high electrical resistivity. The highest concentrations were found at ~92 mbsf for site U1328E, corresponding to a high resistivity interval in U1328A. This feature seems to be absent from Site U1328C confirming the lateral heterogeneity across the different holes. The presence of this feature only in certain holes explains why it does not appear in the first three clusters.
The last area at ~217.5 mbsf is very close to the reported BSR (at 219 mbsf, according to the seismics) and shows a maximum in electrical resistivity values downhole and a maximum in the Mg/Ca ratio at ~220 mbsf.

When a fifth cluster (k=5) is requested, it divides cluster 2 into two sections, an upper and a lower one and a few small areas from cluster 3 are redistributed into these new two clusters. The upper section is generally characterized by slightly higher densities (MAD and EcoScope tool) and magnetic susceptibilities than the lower section. In addition, the upper section has a lower calcium content and smaller range of gamma ray emission than the lower section.

The higher k values produced extremely mixed intervals and as such have not been considered for the purpose of dimensioning SCIMPI. The mineralogy (e.g. glauconite) of the sediment has been described as the reason for changes in the counts of natural gamma ray records (Ellis and Singer, 2007). Since we are not interested in this type of detail for the deployment we have not considered a finer scale.

3.2. Cable lengths and distribution of modules for CAS05 and CAS10

Due to time constraints during the project, the cable lengths for SCIMPI were fixed before the deployment location was decided. The current cable lengths (63, 44, 35.5, 32.5, 20, 10, 10, 9 and 5 m) are interchangeable and the clusters obtained from these analyses identify the cable lengths that best match the fluid flow characteristics of the different lithologies encountered at the sites.

In the distribution of the modules in depth we also need to consider the area that has not been covered by the cluster analysis due to the paucity of data or the presence
of drillpipe in the wireline logging data. In these areas the best information available is from the previous literature.

Table 3. Distribution of the SCIMPI modules for Sites CAS05 and CAS10 and major features expected for that depth.

<table>
<thead>
<tr>
<th>Module</th>
<th>Depth (mbsf)</th>
<th>Cable length above (m)</th>
<th>Cluster</th>
<th>Feature expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>10</td>
<td>-</td>
<td>Massive gas hydrates. Active gas hydrate formation</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>5</td>
<td>-</td>
<td>Gas hydrate formation</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>44</td>
<td>-</td>
<td>Gas hydrate formation</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>10</td>
<td>1</td>
<td>Gas hydrate presence</td>
</tr>
<tr>
<td>5</td>
<td>99.5</td>
<td>32.5</td>
<td>2</td>
<td>Gas hydrates absence</td>
</tr>
<tr>
<td>6</td>
<td>120.5</td>
<td>20</td>
<td>1</td>
<td>Gas hydrate presence</td>
</tr>
<tr>
<td>7</td>
<td>157</td>
<td>35.5</td>
<td>2</td>
<td>Gas hydrates absence</td>
</tr>
<tr>
<td>8</td>
<td>221</td>
<td>63</td>
<td>3</td>
<td>BSR, free gas</td>
</tr>
<tr>
<td>9</td>
<td>231</td>
<td>9</td>
<td>-</td>
<td>Methane hydrate stability</td>
</tr>
</tbody>
</table>

In the upper ~30 meters of Site U1328 gas hydrates are prominent. Massive gas hydrates have been recovered in this area from 1 to 8 mbsf and its presence is detected in the sediment recovered and the electrical resistivity records from borehole logging from 0 to 15 mbsf (Riedel et al. 2006). The methane source to maintain the gas hydrate formation on this site is probably a fracture zone laterally located between 5 and 25 mbsf (Riedel et al. 2006). The interstitial water geochemistry for this site (Riedel et al. 2006) indicates that gas hydrate formation controls the upper 30 mbsf and diffusion controls the chemical profiles below 30 mbsf down to 250 mbsf. Chlorinity and salinity values are low in the upper 5 mbsf and between 20 and 60 mbsf. The high values between 5 and 20 mbsf indicate a constant formation of gas hydrates producing an accumulation of brine during the formation in a diffusion-
controlled system (Riedel et al. 2006). Seismics locate the BSR at U1328 at a depth of ~219 mbsf. The temperature gradient at this site predicts the base of the methane hydrate stability at 220 to 245 mbsf.

Figure 3. Depth locations selected for the nine SCIMPI modules (not to scale) based on the PCA and CA analyses shown on the seismic reflection record from the Cascadia margin site. Images of the actual SCIMPI sensor module and command module are also shown. Flotation is not shown in this diagram.
4. CONCLUSIONS

A new sub-seafloor observatory, SCIMPI, will be deployed on the Cascadia margin to form part of the NEPTUNE Canada observatory. SCIMPI has a modular design that allows tailoring the instrument for the deployment site. To distribute the sensors we used a multivariate statistics approach that identified the optimal depth intervals for long-term gas and fluid flow observations. The optimal configuration selected for the current SCIMPI prototype at this site locates the modules at depths of 4, 10, 55, 66, 99.5, 120.5, 157, 221 and 231 mbsf (Figure 3).

Cluster methodologies have previously been used to identify unique lithologies based on wireline logging data (Anderson et al. 2012, Lovell et al. 1993, Pelling et al. 1991, Shepherd et al. 1987). The different areas identified by the clusters correspond to the major areas interpreted by scientists from Expedition 327 as areas of changes in lithology or gas hydrate presence. $K$-means clustering presents an advantage in situations where the number of modules is predetermined and this number is smaller than the best $k$. Since the $k$ can be set up to fit the number of modules the most important features will directly show up in the analysis. Applying this methodology to complex geological formations identifies critical intervals that should be monitored. In cases where a new site is explored in which no previous drilling has been done, different cable lengths could be carried on board and the separation of the modules optimized in situ when the first data from logging and core physical properties are measured. This methodology presents an unbiased and automatic procedure to identify the best position for sensors in seafloor observatories. Using this approach can help to accelerate the process of determining the location of seafloor observatories directly based on the first data input from the core process. Modular tools such as SCIMPI
present an advantage due to their adaptability to the site and this analysis adds to their flexibility to specify the separation between modules on site.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Louise Anderson for her help on the initial steps of this project. SCIMPI is a collaborative effort funded by IODP-MI and led by the University of Rhode Island with major contributions from Transcendev and Woods Hole Marine Systems Inc. (WHMSI). Stress Engineering and Lamont Doherty Earth Observatories (LDEO) provided technical input and the deployment instrumentation. The authors would like to thank the D/V JOIDES Resolution crew and the science parties and technical support of IODP Expedition 311. We are also grateful to Matthew Smith, Louise Anderson, Tori Kulin and Lucy Hurlbut for their support during the development of this project. Tania Lado Insua was funded during this project by Fundación Pedro Barrié de La Maza, NSF and IODP-MI.

REFERENCES


Technologies 1-7 at

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5774134

the Ocean Drilling Program. In Fisher, A. T., T. Urabe, A. Klaus, and the
Program, Volume 301, Integrated Ocean Drilling Program Management
International, Inc., Washington, DC.

emissions from wetlands: biogeochemical, microbial, and modeling
perspectives from local to global scales. Global Change Biology. doi:
10.1111/gcb.12131

to changes in the deep ocean, Earth and Planetary Science Letters 227, 185

Formation of natural gas hydrates in marine sediments: 1. Conceptual model of
gas hydrate growth conditioned by host sediment properties. Journal of

Proceedings of the twenty-first international conference on Machine learning
(p. 29). ACM.

methane: budget, changes and dangers. Philosophical Transactions of the


CHAPTER 3

ADVANCED CLASSIFICATION OF CARBONATE SEDIMENTS BASED ON PHYSICAL PROPERTIES

Tania Lado-Insua1,*, Lutz Hamel2, Kathryn Moran1,3,4, Louise M. Anderson5 and Jody M. Webster6

Article to be submitted to a special issue Sedimentology titled “Reef response to sea-level and environmental changes: reading the past, informing the future”

1Department of Ocean Engineering, University of Rhode Island, Bay Campus, Narragansett, 02882 RI, USA
2Department of Computer Science, University of Rhode Island, Main Campus, Kingston, 02881 RI, USA
3Graduate School of Oceanography, University of Rhode Island, Bay Campus, Narragansett, 02882 RI, USA
4Ocean Networks Canada and NEPTUNE Canada, University of Victoria, PO BOX 1700 STN CSC, Victoria, BC V8W 2Y2 Canada
5Department of Geology, University of Leicester. University Road, Leicester, LE1 7RH, UK
6Geocoastal Research Group, School of Geosciences, The University of Sydney, Sydney NSW 2006 Australia

*Correspondence author: T. Lado-Insua (e-mail: ladoinsuat@egr.uri.edu)
ABSTRACT

The Integrated Ocean Drilling Program (IODP) Expeditions 310 and 325 are the two more recent contributions to improving our understanding of sea level change and coral reef responses. Physical properties such as bulk density (gamma ray attenuation), P-wave velocity, electrical resistivity, and magnetic susceptibility are related to characteristics of the marine sediments that, in turn, are indicative of the lithologic type. Non-destructive physical properties are measured routinely using a Multi-Sensor Core Logger (MSCL) on whole cores. This study presents a nonlinear relationship between MSCL physical properties and different carbonate sediment types. The database and model presented here integrate sedimentology with physical properties data. All data were analyzed with R software using three different techniques: Linear Discriminant Analysis (LDA), Random Forest (RF) and, Support Vector Machines (SVM). The models that best describe the nature of the data are Random Forest and Support Vector Machines. The similarities between models for both expeditions indicate that the additional P-wave data collected on Expedition 310 cores does not significantly improve the dataset capability for differentiating lithologies.

Here we present the first application of machine learning as a tool for classifying sediment types. This tool was developed specifically for assisting in the challenging identification of the lithologies in poor quality coral cores. But, this technique has broader applications. The tool can be used for describing cores prior to their opening, enabling the pre-identification of critical intervals for special analyses and study. These methods of data analysis can also assist with sample interval selection for specific studies. This technique can also be applied to the interpretation of lithotypes
from wireline geophysical logging data, particularly in boreholes where sampling is
limited to cuttings.

KEY WORDS: Physical properties, core descriptions, carbonate sediment, machine
learning, coral.
1. INTRODUCTION

The Integrated Ocean Drilling Program (IODP) Expedition 310 (Tahiti Sea Level) took place on board the DP *Hunter* during the fall of 2005 (Figure 1) (Camoin et al. 2007). A sister IODP Expedition, 325, Great Barrier Reef Environmental Changes (GBREC) took place on board the *Greatship Maya* during the spring of 2010 (Figure 2) (Webster et al. 2011). These expeditions provided new information on past sea-level changes and sea surface temperature variations during the postglacial interval and a better understanding of the impact of sea level changes on reef growth and geometry. The samples recovered are extremely useful for understanding coral reef responses to environmental changes and for investigating how physical properties can be used in the identification of carbonate lithologies.

Physical properties, such as bulk density, electrical resistivity, and magnetic susceptibility are related to characteristics of sediments that, in turn, are indicative of their depositional environments. Changes in the physical properties in carbonate sedimentary rocks and fossil corals can reflect density changes in the coral composition, different connectivity between the pores, textural properties of the coralg framework, lack of burial compaction or even diagenesis (Camoin et al. 2007; De’ath et al. 2009).

The use of advanced classifiers allows the learning from several variables to identify different types of lithology. These techniques have been applied to different fields of study and they are of common use in both the natural sciences and medical research (e.g. Ambroise and McLachlan 2002, Svetnik et al. 2003, Cutler et al. 2007, Seiler et al. 2012, Lind and Maltseva 2003). They are not new to geological studies,
random forest techniques have been successfully applied to understand mechanisms in gully erosion (Kuhnert et al. 2010), on identifying driving factors of landslides (Vorpahl et al. 2012) and to predict the characteristics of sediment based on seismic data (Huang et al. 2012). However, they have never been applied to identify carbonate sediments using their physical properties.

Figure 1. Location of Expedition 310, Tahiti Sea Level. (Copyright: © 2010 MDA Information Systems, USGS, NASA).

Our measurements are taken non-destructively with a Multi-Sensor Core Logger (MSCL) on whole core round samples. MSCL data collected from hard rock, fossil corals and carbonate sediments can be compromised because of poor core quality. Physical properties measurements with this system are affected by several parameters (e.g. degree of saturation, under filled liners, etc.), making it difficult to relate the
MSCL values to the physical properties measured on discrete samples (Blum 1997; Jarrard and Kerneklian, 2007). However, the use of this instrument is a standard procedure and usually the initial step in the core lab.

Figure 2. Expedition 325, Great Barrier Reef Environmental Changes. (Copyright: © 2010 MDA Information Systems, USGS, NASA).
The main objective of this study was to model the relationship between physical properties and lithology in well-preserved samples to assess whether the MSCL can be used to generate useful information in reef sediments. A second objective was to establish whether this approach could be used as a new rapid and accurate tool for classifying sediment types. With this approach, relationships between physical properties and lithology can be identified prior to opening cores so that a careful subsampling strategy can be developed for the limited material recovered. It can also assist in correlation of core data with wireline logging datasets. This approach not only could accelerate the process of identification but also supplement identification of samples when a visual analysis alone is not sufficient. Here we report on the results, using this approach, from data collected during IODP Expeditions 310 and 325. This first application demonstrates the validity of these techniques for identifying coral lithotypes.

2. MATERIALS AND METHODS

Samples from all sites except M0058A (fine grain lithology) from GBREC and from Maraa West Transit Site M0005 (Holes A to E) from Tahiti (smaller terrestrial influence), were used for this study (Figures 1 and 2). The whole cores recovered from Tahiti were re-saturated with surface seawater and left for a couple of hours to re-equilibrate at 20°C previous to the measurements (Expedition 310 Scientists 2007). This step enabled P-wave velocity measurements to be collected and these were used in this study. In GBREC, the cores were processed without re-saturation to avoid the loss of fine materials that could be washed out. The scarce P-wave measurements obtained by the MSCL for GBREC were not used in this study. Both datasets were analyzed.
independently since their values were not comparable (Appendix III). They presented not only differences in the variables measured but also in the values of these variables due to the saturation level of the samples. However, the comparison of models helps us to understand the relationship between physical properties and lithology and clarifies the best procedures to follow in future coral reef drilling studies.

Other physical properties were measured on the cores recovered, such as color reflectance. The color reflectance data were purposely not included in the present study because they were measured after splitting the cores, and we intend to identify lithologies before core splitting.

2.1. Sample selection

An initial filtering of the data from the MSCL was done to avoid areas where any sensor failed to record data (i.e. missing values). Well preserved, mostly homogeneous lithologies were then visually selected from the core images and labeled according to lithology, as described by the sedimentologists (Figure 3). Rubble samples were discarded. Samples from cores where two different lithologies were found together at the same depth were avoided when possible. The corresponding measurements for bulk density, porosity, magnetic susceptibility, P-wave velocity (only in the Tahiti dataset) and electrical resistivity from the MSCL log (1 cm sample interval) were selected for data analysis. The values of each one of these measurements for the same depth in the core is referred to as a sample hereafter.

All the variables were analyzed with the R 2.15.3 software by three different supervised machine-learning techniques: Linear Discriminant Analysis (LDA),
Random Forest (RF) and Support Vector Machines (SVMs). The final matrices of data presented four or five quantitative variables (respectively for GBREC and Tahiti) and an identification variable or label indicating the lithology of each sample.

A total of 3852 (Expedition 325) and 1586 (Expedition 310) samples corresponding to a total of six different lithologies (Table 1) were selected for the study: sand, silt, microbialite (Microb), coral algae (CorAlg), massive coral and coral framework (MassCoral) and encrusting coral (EncrCoral) (Camoin et al. 2007, Webster et al. 2011). In both expeditions the number of samples for each lithology was different, and therefore we had a relatively unbalanced dataset.

Table 1. Number of samples for each lithology on each expedition.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Total samples</th>
<th>Lithology</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CorAlg</td>
<td>EncrCoral</td>
<td>MassCoral</td>
<td>Microb.</td>
<td>Sand</td>
</tr>
<tr>
<td>GBREC</td>
<td>3852</td>
<td>305</td>
<td>112</td>
<td>680</td>
<td>463</td>
<td>2004</td>
</tr>
<tr>
<td>Tahiti</td>
<td>1586</td>
<td>343</td>
<td>232</td>
<td>148</td>
<td>409</td>
<td>262</td>
</tr>
</tbody>
</table>

The dataset was especially unbalanced in the case of Expedition 325 where more than half of the samples were sand. To correct for this, three datasets were analyzed for each expedition: 1) the unbalanced original dataset, 2) a down-sampled dataset, in which a re-sampling with replacement was done to bring all the lithologies to the same number of samples from the minority class (the lithology with less number of samples) and, 3) an up-sampled dataset, in which a re-sampling with replacement was done to bring all the lithologies to the same number of samples from the majority class (the lithology with higher number of samples). All these processes are commonly used techniques for machine learning with unbalanced datasets (Hamel 2009).
Figure 3. Process to combine the descriptions and physical properties data into the database used for this study.

In the case of the database from GBREC, since we were more interested on the model being able to predict corals than sand, we initially corrected the unbalance by randomly down-sampling the sand, this is, we randomly reduced the number of sand samples to the amount of samples of the next majority class (MassCoral). This was a preliminary step before up-sampling for creating the dataset to be analyzed. The process of balancing the samples was integrated in the algorithm used to generate training and testing datasets (see section 2.5, Figure 3).
Here we present the results of the data balanced by random up- and down-sampling and the unbalanced data for each of the expeditions (GBREC and Tahiti).

2.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a multivariate statistical technique that seeks a linear combination of variables optimizing the separation of the classes’ (lithologies) by looking for a combination that maximizes the difference between the classes and minimizes the differences within the classes (reviewed in Venables and Ripley, 1998). In summary, this technique looks for a linear relationship in the data that will optimize the parameters of a linear equation based on the physical properties measurements so that they show the maximum separation between different lithologies. This approach has the advantage of simple computation. LDA requires non-missing data and Gaussian distribution of the data. This analysis gives as a result a linear combination of the variables. In cases where the solution of the problem is not linear or there is noise in the data, this method is not capable of good predictions.

2.3. Random Forest

Random Forest (RF) (Breiman, 2001) is a machine learning technique based on decision trees. The model builds a large collection of de-correlated trees, and then averages them (Hastie et al. 2009) combining the predictions of all of them. Decision trees are able to handle high dimensional data ignoring irrelevant variables, but they can be inaccurate when used independently. Creating ensembles of trees or “forests” helps to increase the accuracy. Our RF is composed of 500 trees \{T_1(X)… T_{500}(X)\}
where $X=\{x_1, \ldots, x_n\}$ is an n-dimensional vector of physical properties of the sediment associated to a lithology. Each one of our trees will have a prediction for each lithology. The outputs of all trees are aggregated to produce a final prediction, this is the class predicted by the majority of trees. At each node the RF algorithm decides between two physical properties of the formation that are selected at random (Svetnik et al. 2003). The final products, or the “leaves” of the tree, are the lithologies identified after deciding between values of the different variables in a binary manner. The models are simple to train and tune, and able to capture complex interactions in the data. RF does not require any particular distribution of the data and is able to handle missing data and qualitative variables. The solution proposed by RF is non-linear, but it is a powerful tool to classify the variables using linear and non-linear solutions. In this study, the number of trees generated in random forest to be averaged was set at 500, well above the 200 recommended as a minimum in the literature (Hastie et al. 2009). The number of variables randomly sampled at each split was set as 2 for the model. This value is usually optimized as the square root of the number of variables (Svetnik et al. 2003), which in our case are four for Expedition 325 and five for Expedition 310. The minimum node size was set as the default value (one), which makes the trees fully develop or ‘grow’. This parameter determines the minimum size of nodes below which no split is performed. This technique is not affected by the importance of the descriptors (physical properties) used, even if some of them are correlated as in the case of density, porosity and P-wave velocity. Random forest does not need an optimal dataset of descriptors, information can be redundant without affecting the analysis. Random forest suffers no overfitting. In overfitting situations
the “test” dataset error increases after the “training” error has reached zero (Svetnik et al. 2003). In our datasets the training error never reaches zero showing no overfitting. Random forest has been proved to be among the best performing models in several studies (Meyer et al. 2003, Vorpahl et al. 2012). Additional details on how random forest works is provided in Svetnik et al. (2003).

### 2.4. Support Vector Machines

Support Vector Machines (SVMs) are based on nonlinear binary boundaries (Hastie et al. 2009). The optimal separating hyperplane between two classes is obtained by maximizing the margin between the different classes’ closest points. The points lying on the boundaries are called support vectors (Hamel 2009). This means that the points in the borderlines for different lithologies for all the different physical properties, considering each one as a different dimension of the data, are selected as support vectors. Margins are created around the line (hyperplane when separating several dimensions) that separates them to limit the boundaries between the six lithologies or classes. The optimal hyperplane that separates the classes maximizes the distance to the closest points of the different lithologies (these points are the support vectors).

SVMs are in general easier to interpret than neural networks and they are based on optimizing a convex function. This, unlike neural networks prevents SVMs from getting false minima. The tuning of a SVM is also simpler than a neural network, since fewer parameters need to be optimized (cost, kernel and gamma in our case). A kernel is a function used in SVMs to transform the data, similar to a change of coordinates.
The use of different kernels on SVM allows them to be efficient at finding linear and nonlinear decision surfaces. Different kernels were tested to find the one performing at higher accuracies (linear, sigmoidal, radial and different degrees of polynomial), based on this the radial kernel was selected. The cost parameter of the model was optimized to 1000 for both expeditions. The cost is a way of penalizing the model when it misclassifies a sample. Higher costs imply higher penalization and hence, longer processing time. Another parameter to optimize a SVM model is gamma. Gamma modifies how broad or narrow the kernel used is; the smaller the gamma value, the wider the kernel. The gamma value was optimized to 0.45 for Expedition 310 (Tahiti) and to 0.55 for Expedition 325 (GBREC). The decision surface is not affected by the repetition of samples in these types of models, as it would not affect the points used as support vectors to separate the classes to have a point repeated. The reader can find more detailed information on SVM in Hamel (2009).

2.5. Evaluating the accuracy of the models

All the models presented here have their own methods of error evaluation in place.

In the case of the LDA, a jackknifed prediction that follows the Leave-One-Out (LOO) cross-validation methodology is implemented. In this case, one sample is systematically left out of the model during training and then the model prediction is tested on the sample left out (Venables and Ripley 1998, Ripley 2013).

The RF algorithm in R has Out-Of-Bag (OOB) error implemented. During the training of the model each tree is grown using a bootstrap dataset, in which some
samples are not presented to the model (Svetnik et al. 2003). These are OOB samples on which the model is tested afterwards. Calculating the error by leaving samples out of the bag has been proved problematic in some cases (Svetnik et al. 2003).

A cross-validation is performed in the SVM to calculate the accuracy of the model. The $X$-fold cross validation splits the dataset into $X$ partitions of which $X-1$ are used as a training dataset, and the one left is used for testing the accuracy of the model (Hamel 2009).

All these different error calculation methods (ECM) available in the models would give optimistic accuracies for them, especially in the case of the up-sampled balanced datasets, this is due to the repetition of points for testing and training in these ECM. Since these accuracies are unrealistic we do not report them here.

To avoid repetition of points, being able to compare the models, and in order to be realistic about the capability of prediction by different models, we followed a hold-out methodology (i.e. only part of the dataset was presented to the model for testing). The process was repeated in a 200 samples bootstrap to get intervals up to the 95% level of confidence for all models. This means that sampling with replacement was done in the original dataset to create 200 new datasets. The algorithm used for the bootstrapping was set up to sample with replacement a sample that was 2/3 the size of the original dataset, this was the “training” dataset. Then, the samples present in this dataset were eliminated from the original dataset, leaving us with a “testing” dataset (>1/3 of the original size due to the replacement in the “training” set). The up-sampling or down-sampling were then applied only to the “training” dataset (Fig. 3).
The model accuracy was tested in the “testing” dataset, which was composed of samples that the model had never seen before during training. This was done 200 times generating 200 different models, and analyzing the accuracy of each one of them individually to generate confidence intervals. The 5th lower accuracy value and the 195th higher accuracy value were taken, representing the 95% level confidence intervals where the accuracy of our model would sit. The process of bootstrapping and analyzing accuracy was done externally to the model, meaning that it was done independently of the model cross-validation (LDA, SVM) or out-of-bag (RF) method to calculate accuracy. The importance of an external bootstrap has been demonstrated in the past. Internal cross-validation errors are usually underestimated due to the bias in the data produced during the prediction (Ambrose and McLachlan, 2002). This happens in particular where attributes (physical properties in this study) are selected as a first step in the modeling. In our case the reduced number of physical properties available for the analysis was not preselected, despite the correlation between some of them (i.e. bulk density, porosity, and P-wave velocity). We decided to perform an external calculation of the accuracy to determine how confident we could be of the calculation of our error, and also to be able to compare the different models with the same methodology (Figure 4).

The use of confidence intervals allows us to differentiate between models. When two intervals overlap for two different models it means that the performance of those models are undistinguishable (Hamel 2009).
Figure 4. Data analysis diagram for all models. Code provided in Appendix IV.

The models presented in this study can also indicate which lithologies present more similar physical properties. Confusion matrices give a better understanding of the data indicating which lithologies are being misclassified. In a confusion matrix we have the real values from the dataset we are trying to predict in the rows, this is, if we add all the values of a row we will get the total of samples for the lithology indicated in that row. The columns indicate the values that the model is predicting. The total
accuracy of the model is determined by the number of samples in the diagonal of the confusion matrix, which are the samples that the model has identified correctly, divided by the total number of samples in the dataset. The accuracy for each lithology is the percentage of samples for that lithology that are identified correctly and therefore located for both, row and column under the same lithology.

3. RESULTS

3.1. Total accuracy of the models

In this first approach to the identification of carbonate sediments using physical properties alone, we tested three different methodologies (LDA, RF and SVMs). Data from Tahiti presented in all cases exhibits less total accuracy than GBREC. The higher accuracies for the unbalanced datasets in all the analysis (Table 2) are due to differences in sample number for each class (Table 1) and ease to predict some of the classes versus others (e.g. Sand is easier to predict than CorAlg or EncrCoral). The model will assign more samples to the majority classes underestimating or totally neglecting the minority classes (e.g. 0% accuracy for LDA unbalanced GBREC for CorAlg and EncrCoral, Tables 3 and 6). The models will also perform better in classes that present more similar values for all the samples (e.g. Sand).

The results obtained by the LDA (Table 2) indicate that linear relations between the physical properties studied are able to capture part of the required information in the data to identify some of the lithologies studied. The total accuracy of this methodology is surprisingly good, with a range of ~61 to ~58% for GBREC and ~50 to ~44% for Tahiti for the unbalanced datasets. The up- and down-sampled datasets
for GBREC have accuracies that overlap with the unbalanced dataset for the LDA model making these models undistinguishable. In the Tahiti dataset only the down-sampled LDA accuracy overlaps with the unbalanced LDA.

The RF in the unbalanced dataset, with total accuracy values of ~79 to ~75% for GBREC and ~65 to ~60% for Tahiti, is the best performing model. The up- and down-sampled datasets for Tahiti overlap with these values matching the performance of the unbalanced dataset. No other models present as higher accuracy as the unbalanced dataset for RF for GBREC.

The SVM also performed better in the unbalanced datasets with ~74 to ~71% total accuracy for GBREC and ~58 to 52% for Tahiti. The unbalanced dataset model overlaps with the up-sampled model for the Tahiti dataset making them undistinguishable for a 95% confidence interval.

The total accuracies of the models highlights the model that predicts the lithology of more samples accurately, however that does not mean it works well for all lithologies. It is important to examine the accuracies for each lithology and not just the total accuracies. When we look at total accuracies, the models performing better in all cases are unbalanced datasets, which is an artifact caused by the more easily predicted classes (e.g. sand and silt) and a bias toward majority classes versus those with small sample numbers. The model selected by the user in most cases is driven by the lithology that we are interested in predicting.
Table 2. Total accuracy range for both expeditions calculated for all models (LDA: Linear Discriminant Analysis, RF: Random Forest, SVM: Support Vector Machines) and datasets (up-balanced, down-balanced and unbalanced). Dark blue indicates the model with higher accuracy for each expedition. Pale blue indicates models that overlap with the accuracy range of the best model (for a 95% interval of confidence) for each expedition.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data structure</th>
<th>GBREC Exp. 325</th>
<th>Tahiti Exp. 310</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>Up-sampled</td>
<td>59.77-49.14</td>
<td>43.97-38.51</td>
</tr>
<tr>
<td></td>
<td>Down- sampled</td>
<td>59.79-44.39</td>
<td>44.88-37.82</td>
</tr>
<tr>
<td></td>
<td>Unbalanced</td>
<td>60.92-57.60</td>
<td>50.00-44.46</td>
</tr>
<tr>
<td>RF</td>
<td>Up-sampled</td>
<td>75.38-70.79</td>
<td>62.95-56.38</td>
</tr>
<tr>
<td></td>
<td>Down- sampled</td>
<td>67.70-60.00</td>
<td>60.27-52.77</td>
</tr>
<tr>
<td></td>
<td>Unbalanced</td>
<td>78.96-75.65</td>
<td>65.29-59.66</td>
</tr>
<tr>
<td>SVM</td>
<td>Up-sampled</td>
<td>66.07-59.36</td>
<td>54.92-47.90</td>
</tr>
<tr>
<td></td>
<td>Down- sampled</td>
<td>61.24-52.56</td>
<td>51.98-43.68</td>
</tr>
<tr>
<td></td>
<td>Unbalanced</td>
<td>74.48-70.93</td>
<td>58.05-52.05</td>
</tr>
</tbody>
</table>

3.2. Accuracy for each lithology

3.2.1. Expedition 325: GBREC

In the case of GBREC the best model for Sand prediction is the LDA applied to the unbalanced dataset (up to ~99% accuracy, Table 3). However, there were no other cases for any lithology on GBREC where LDA was the preferred model (Table 3). This is mainly due to the extremely high number of samples (2004, Table 1) for the Sand in this dataset. The model ignores completely two of the classes (CorAlg and EncrCoral have 0% accuracy) and strongly neglects another class (Silt, ~7% accuracy), prioritizing the predictions for Sand over all the others. It also implies a strong linear relationship between the variables analyzed for this lithology. This
model’s accuracy is not comparable to any other for the Sand, since no models overlap with the accuracy interval of the unbalanced LDA (Table 3).

The RF presents the best models for all the other lithologies (CorAlg, EncrCoral, MassCoral, Microbialite and Silt). Up-sampled RF works best for CorAlg (~68% to ~52% accuracy) and Microbialite (~70 to ~55% accuracy), down-sampled RF for MassCoral (up to ~55% accuracy) and Silt (~99 to ~86% accuracy) and unbalanced RF for MassCoral (up to ~69% accuracy, Table 3). However, most of the RF models fall within the interval of accuracy of the best model (pale blue, Table 3). The intervals also overlap with most of the SVM ranges of accuracy indicating that either RF or SVM are the best performers for all lithologies except Sand.

The SVM models present high accuracies, especially for the balanced datasets; overlapping with the RF best models, however, their upper limit of the range for accuracy is slightly lower than RF (Table 3). The best performing model for all the lithologies of GBREC, except Sand, is the up-balanced RF.

3.2.2. Expedition 310: Tahiti

In the case of Tahiti, where samples were re-saturated, a slightly different pattern is apparent. A small effect due to the re-saturation is present for some lithologies in the accuracy of the models.

The LDA is the model which best performs for EncrCoral (down-sampled, with a small accuracy, ~40 to ~18%, Table 3) and Microbialite (unbalanced, ~82 to ~61% accuracy, Table 3). Microbialite is the most abundant class (409 samples, Table 1). This indicates that the relationships between the physical properties studied for these
lithologies are mostly linear, and the non-linear information in the data is not adding any value with regards to the identification of encrusting corals and microbialites. However, this model accuracy overlaps with most of the RF and SVM models (pale blue, Table 3), making them as accurate as LDA for these lithologies.

The class with higher accuracy for RF is Silt (up to ~100%, Table 3), with a rather small number of samples (192 samples, Table 1). All the RF models for Silt perform best with a nearly identical model, indicating that this dataset is highly homogeneous. RF also performs well for CorAlg unbalanced (~70 to 56% accuracy), MassCoral down-balanced (~63 to ~39%) and Sand up-balanced (~70 to ~51%). All models for RF (up-sampled, down-sampled and unbalanced) overlap in accuracy intervals, making their results extremely similar.

The SVM results for all lithologies overlap with the best model except the up-and down-sampled models for the Microbialite, which slightly underperform (Table 3).

The best performing model for the Tahiti expedition for all lithologies would be any of the RF models.
Table 3. Accuracy range for each lithology on both expeditions for all models (LDA: Linear Discriminant Analysis, RF: Random Forest, SVM: Support Vector Machines) and datasets (up-balanced, down-balanced and unbalanced). Dark blue indicates the model with higher accuracy. Pale blue indicates models that overlap with the accuracy range of the best model (for a 95% interval of confidence).

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Model</th>
<th>Data structure</th>
<th>CorAlg</th>
<th>EncrCoral</th>
<th>MassCoral</th>
<th>Microb</th>
<th>Sand</th>
<th>Silt</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBREC</td>
<td>LDA</td>
<td>Up-bal.</td>
<td>16.43-3.18</td>
<td>32.30-3.63</td>
<td>33.62-10.54</td>
<td>45.04-17.21</td>
<td>88.45-68.76</td>
<td>68.28-31.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>20.26-1.43</td>
<td>38.64-3.45</td>
<td>34.87-4.21</td>
<td>46.67-11.16</td>
<td>89.97-58.60</td>
<td>91.41-9.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>0.00-0.00</td>
<td>0.00-0.00</td>
<td>26.57-15.95</td>
<td>38.16-25.86</td>
<td>99.41-98.06</td>
<td>6.92-2.07</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>Up-bal.</td>
<td>68.39-52.53</td>
<td>40.00-16.13</td>
<td>67.26-54.29</td>
<td>69.78-55.23</td>
<td>85.07-77.76</td>
<td>98.03-87.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>68.21-49.67</td>
<td>55.17-24.53</td>
<td>57.14-39.30</td>
<td>64.68-48.31</td>
<td>75.62-60.50</td>
<td>98.64-85.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>52.23-34.81</td>
<td>22.45-3.77</td>
<td>69.46-57.34</td>
<td>67.49-51.93</td>
<td>94.79-91.19</td>
<td>93.70-81.13</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>Up-bal.</td>
<td>64.43-47.71</td>
<td>42.37-17.19</td>
<td>51.18-35.17</td>
<td>62.66-45.73</td>
<td>76.50-64.60</td>
<td>97.16-87.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>62.89-42.77</td>
<td>43.14-15.38</td>
<td>44.21-23.86</td>
<td>61.70-39.48</td>
<td>71.41-55.71</td>
<td>96.64-86.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>38.36-9.80</td>
<td>11.32-0.00</td>
<td>62.50-50.44</td>
<td>57.87-42.67</td>
<td>95.50-89.54</td>
<td>91.39-77.78</td>
</tr>
<tr>
<td>Tahiti</td>
<td>LDA</td>
<td>Up-bal.</td>
<td>42.52-24.41</td>
<td>40.34-17.50</td>
<td>55.88-34.67</td>
<td>34.91-12.27</td>
<td>56.12-34.31</td>
<td>99.07-95.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>45.03-23.21</td>
<td>42.37-15.93</td>
<td>56.96-33.33</td>
<td>39.23-13.27</td>
<td>55.30-34.31</td>
<td>99.07-95.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>55.00-29.10</td>
<td>5.41-0.00</td>
<td>36.36-21.13</td>
<td>81.55-60.75</td>
<td>44.44-19.33</td>
<td>99.07-95.19</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>Up-bal.</td>
<td>66.86-50.58</td>
<td>35.59-17.36</td>
<td>59.46-37.65</td>
<td>74.40-54.98</td>
<td>70.45-51.08</td>
<td>100.00-92.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>62.36-45.14</td>
<td>34.51-17.32</td>
<td>62.82-38.89</td>
<td>68.40-46.63</td>
<td>69.05-46.03</td>
<td>100.00-89.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>70.37-56.32</td>
<td>30.51-13.68</td>
<td>57.75-35.37</td>
<td>82.63-66.67</td>
<td>70.08-48.34</td>
<td>100.00-92.22</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>Up-bal.</td>
<td>64.8-44.85</td>
<td>28.97-13.22</td>
<td>55.26-29.87</td>
<td>60.39-42.65</td>
<td>54.96-31.92</td>
<td>98.08-89.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down-bal.</td>
<td>61.62-39.56</td>
<td>31.71-13.33</td>
<td>54.55-31.71</td>
<td>57.07-34.31</td>
<td>50.74-30.66</td>
<td>98.95-86.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbal.</td>
<td>67.78-47.64</td>
<td>25.45-9.23</td>
<td>51.95-26.67</td>
<td>76.61-55.87</td>
<td>58.73-35.21</td>
<td>99.04-89.22</td>
</tr>
</tbody>
</table>
3.2.3. *Comparison GBREC vs. Tahiti: Re-saturate or directly process?*

The Tahiti samples show higher accuracy for three lithologies (CorAlg, Microbialite and Silt) compared with the GBREC samples, which show higher accuracies in the other lithologies (EncrCoral, MassCoral and Sand). However, all the accuracy ranges for all lithologies from both expeditions overlap except for Sand, which has higher accuracy for GBREC. Therefore, for all lithologies other than sand, the models performances are indistinguishable and highlight that the time-consuming and “smelly” process of re-saturation does not provide significant improvements in the datasets produced for this sort of analysis. The re-saturation process can wash out fine-grained material in the formation, therefore voiding any preservation improvements.

3.3. *Comparison of the lithologies: Confusion matrices*

We present here some examples of the confusion matrices obtained for the best models. We randomly selected one of the 200 bootstrap samples from the models to show the relations between lithologies for different scenarios. Two of the best performing models, RF up-balanced dataset from GBREC and Tahiti are presented in Tables 4 and 5. The values located in the diagonal of the matrix indicate samples that have been well classified. By looking at the rows we can see which class was erroneously assigned to each lithology, and by looking at the columns we can see which lithologies are predicted.

The EncrCoral is the class worse predicted for both GBREC and Tahiti in all the models (accuracy goes as low as ~18% for Tahiti and ~25% for GBREC, Table 3). This is expected if we consider the nature of the material. Encrusted corals grow
around other materials and therefore when physical properties are measured they will rarely have a single lithology signal. The number of other associated lithologies is the factor causing encrusting corals being associated with other classes. If we look at the confusion matrix for RF up-sampled for Tahiti we can see that 20% of EncrCoral have been classified as CorAlg, 28% as Microbialite, 12% as MassCoral and 12% as Sand (Table 5). Similar values are observed for GBREC (Table 4), where 22% of the samples are identified again as Microbialite, 16% as MassCoral, 13% as CorAlg and 13% as Sand. In all the other lithologies the misclassifications are lower than 20% (Tables 4 and 5).

The next two lithologies with poor predictions are CorAlg (59% accuracy, Table 4; 56%, Table 5) and MassCorals (59%, Table 4; 60%, Table 5).

CorAlg gets confused with Sand (15%, Table 4; 12%, Table 5) no matter what the core curation method (Tahiti vs. GBREC). However, the confusion with other classes such as EncrCoral and Microbialite is more marked for the Tahiti dataset (13% and 14%, respectively) than for the GBREC (8% and 6%, respectively). This may be due to the most commonly recovered lithologies being different for each expedition. The higher presence of microbialite and coralline algae in the Tahiti cores (Camoin et al. 2012), and the distribution of these three lithologies (CorAlg, EncrCoral and Microbialites) around other structures makes it easier to have mixed lithologies in the samples or adjacent to them in the area measured by the sensors of the MSCL.

Massive corals are the most important lithology in these cores because of their importance in the study of sea level changes. However, they are not easy to predict.
They are commonly confused with Microbialite (13%, Table 4; 17% Table 5) and Sand (13% Tables 4 and 5).

Microbialites are commonly misclassified as MassCorals (19%) in the dataset from GBREC, while in the case of the Tahiti dataset they are mostly misclassified as EncrCoral (18%). The similarity in these three classes may again be due to the presence of microbialite and encrusting corals that grow attached to massive corals. This causes the physical properties’ values to appear similar for the samples studied.

Sand samples are better identified in the GBREC dataset (82%) than in Tahiti (67%). In Tahiti these samples get misclassified mostly as CorAlg (14%).

In both tables we can see that Silt is classified with an accuracy of 95%. The samples of Silt that are misclassified are identified by the model as a CorAlg (2%) or Sand (3%), and occasionally as MassCoral (1%) or Microbialite (1%).
Table 4. Confusion matrix for the RF up-sampled dataset for GBREC. Total values per lithology are indicated as an extra feature in the last column. Percentages per lithology in parenthesis. Color code marks the higher percentage of samples from a lithology in steps of 20% (dark blue: 100-80%, pale blue: 0-20%).

<table>
<thead>
<tr>
<th>Real lithology</th>
<th>CorAlg</th>
<th>EncrCoral</th>
<th>MassCoral</th>
<th>Microb</th>
<th>Sand</th>
<th>Silt</th>
<th>Total real litho.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorAlg</td>
<td>91 (59%)</td>
<td>12 (8%)</td>
<td>18 (12%)</td>
<td>9 (6%)</td>
<td>23 (15%)</td>
<td>0 (0%)</td>
<td>153</td>
</tr>
<tr>
<td>EncrCoral</td>
<td>8 (13%)</td>
<td>23 (37%)</td>
<td>10 (16%)</td>
<td>14 (22%)</td>
<td>8 (13%)</td>
<td>0 (0%)</td>
<td>63</td>
</tr>
<tr>
<td>MassCoral</td>
<td>24 (7%)</td>
<td>29 (8%)</td>
<td>207 (59%)</td>
<td>47 (13%)</td>
<td>44 (13%)</td>
<td>0 (0%)</td>
<td>351</td>
</tr>
<tr>
<td>Microbialite</td>
<td>13 (6%)</td>
<td>17 (7%)</td>
<td>44 (19%)</td>
<td>138 (61%)</td>
<td>15 (7%)</td>
<td>0 (0%)</td>
<td>227</td>
</tr>
<tr>
<td>Sand</td>
<td>37 (4%)</td>
<td>24 (2%)</td>
<td>69 (7%)</td>
<td>26 (3%)</td>
<td>847 (82%)</td>
<td>31 (3%)</td>
<td>1034</td>
</tr>
<tr>
<td>Silt</td>
<td>3 (2%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>5 (3%)</td>
<td>139 (95%)</td>
<td>147</td>
</tr>
<tr>
<td>Tot. pred. litho.</td>
<td>176</td>
<td>105</td>
<td>348</td>
<td>234</td>
<td>942</td>
<td>170</td>
<td>1975</td>
</tr>
</tbody>
</table>
Table 5. Confusion matrix for the RF up-sampled dataset for Tahiti. Total values per lithology are indicated as an extra feature in the last column. Percentages per lithology in parenthesis. Color code marks the higher percentage of samples from a lithology in steps of 20% (dark blue: 100-80%, pale blue: 0-20%).

<table>
<thead>
<tr>
<th>RF 310 Up-sampled</th>
<th>Real lithology</th>
<th>Predicted values</th>
<th>Total real litho.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CorAlg</td>
<td>EncrCoral</td>
<td>MassCoral</td>
</tr>
<tr>
<td>CorAlg</td>
<td>96 (56%)</td>
<td>23 (13%)</td>
<td>8 (5%)</td>
</tr>
<tr>
<td>EncrCoral</td>
<td>23 (20%)</td>
<td>31 (27%)</td>
<td>14 (12%)</td>
</tr>
<tr>
<td>MassCoral</td>
<td>4 (6%)</td>
<td>6 (8%)</td>
<td>41 (60%)</td>
</tr>
<tr>
<td>Microbialite</td>
<td>19 (9%)</td>
<td>38 (18%)</td>
<td>6 (3%)</td>
</tr>
<tr>
<td>Sand</td>
<td>18 (14%)</td>
<td>10 (8%)</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Silt</td>
<td>2 (2%)</td>
<td>0 (0%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Tot. pred. litho.</td>
<td>162</td>
<td>108</td>
<td>73</td>
</tr>
</tbody>
</table>
Table 6. Confusion matrix for the LDA unbalanced dataset for GBREC. Total values per lithology are indicated as an extra feature in the last column. Percentages per lithology in parenthesis. Color code marks the higher percentage of samples from a lithology in steps of 20% (dark blue: 100-80%, pale blue: 0-20%).

<table>
<thead>
<tr>
<th>Real lithology</th>
<th>LDA 325 Unbalanced</th>
<th>Predicted values</th>
<th>Total real litho.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CorAlg</td>
<td>EncrCoral</td>
<td>MassCoral</td>
</tr>
<tr>
<td>CorAlg</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>18 (12%)</td>
</tr>
<tr>
<td>EncrCoral</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>9 (14%)</td>
</tr>
<tr>
<td>MassCoral</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>82 (23%)</td>
</tr>
<tr>
<td>Microbialite</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>26 (11%)</td>
</tr>
<tr>
<td>Sand</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>5 (0%)</td>
</tr>
<tr>
<td>Silt</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Tot. pred. litho.</td>
<td>0</td>
<td>0</td>
<td>141</td>
</tr>
</tbody>
</table>
Table 7. Confusion matrix for SVM down-sampled dataset for Tahiti. Total values per lithology are indicated as an extra feature in the last column. Percentages per lithology in parenthesis. Color code marks the higher percentage of samples from a lithology in steps of 20% (dark blue: 100-80%, pale blue: 0-20%).

<table>
<thead>
<tr>
<th>RF 310</th>
<th>Predicted values</th>
<th>Total real litho.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-sampled</td>
<td>CorAlg</td>
<td>EncrCoral</td>
</tr>
<tr>
<td>CorAlg</td>
<td>77 (45%)</td>
<td>28 (16%)</td>
</tr>
<tr>
<td>EncrCoral</td>
<td>11 (10%)</td>
<td>31 (27%)</td>
</tr>
<tr>
<td>MassCoral</td>
<td>4 (6%)</td>
<td>11 (15%)</td>
</tr>
<tr>
<td>Microbialite</td>
<td>24 (11%)</td>
<td>34 (16%)</td>
</tr>
<tr>
<td>Sand</td>
<td>14 (11%)</td>
<td>11 (9%)</td>
</tr>
<tr>
<td>Silt</td>
<td>1 (1%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Tot. pred. litho.</td>
<td>131</td>
<td>116</td>
</tr>
</tbody>
</table>
In Table 6 we present the unbalanced LDA model for GBREC since this was the best performing model for Sand. This way the reader can appreciate the weaknesses of the model and understand the importance of a model with more balanced data. If we look into the columns in Table 6, we see that in this model no samples are predicted as CorAlg or EncrCoral and only 16 samples are predicted as Silt. The model completely neglects these classes and believes most of the samples to be Sand despite that they are 82% of the CorAlg, 70% of the EncrCoral, 60% of the MassCoral, 59% of the Microbialite and 94% of the Silt samples. Basically, this model assumes nearly everything is Sand. Only 23% of the MassCoral samples and 30% of the Microbialite samples are identified correctly.

Finally, we also chose to present the up-balanced SVM for Tahiti (Table 7) for comparison with the RF model (Table 5). Since both of these models’ confidence intervals for the accuracy overlap, it is easy to notice that the accuracies for most of the lithologies are smaller than for the RF model. The percentage of CorAlg and MassCoral misclassified as Microbialite increases by ~10% for all these samples. This increase is smaller (~6%) in the case of the EncrCoral samples misclassified as Microbialite. Increases in the percentage of misclassified samples for the rest of the lithologies was generally from ~1 to ~8%.

4. DISCUSSION AND CONCLUSIONS

Three different mathematical models were used to predict the lithologies of carbonate sediment cores recovered during IODP Expeditions 310 and 325. The non-linear models proved more accurate than the linear approach. The high accuracies
obtained by two completely different mathematical models (Random Forest and Support Vector Machines) indicate a non-linear component in the relationship between the different physical properties measured, and a clear signal to differentiate types of sediment. Despite the presence of underfilled core liners and non-saturated samples, the model performs well when it is provided with several variables and a reasonable sample size. This indicates that MSCL data from carbonate sediments can provide useful information despite not presenting a direct and easy association with the discrete measurements.

The combination of the physical properties allows the models to find hidden relationships in the hyperspace (mainly radial as indicated by the optimized SVM) between variables that are crucial to the classification. Verwer and Braaksma (2009) demonstrated that P-wave velocity measurements present a complex relation with other physical properties of samples from IODP Expedition 310 due to mixed-mineralogy composition, diagenesis, rock texture, and cementation. The non-linearity in the model could be due to two reasons: 1) the combination of samples in different states of preservation that would slightly alter the relationship between the variables and, 2) non-linear relationships related to the magnetic susceptibility and P-wave velocities due to mixed lithology and cementation of the sample. Therefore, the data obtained by the MSCL can be used for classification of reefal carbonate sediments. The accuracies obtained for the datasets presented are particularly high for a real-world multiclass dataset. The accuracy of the model for each one of the lithologies is variable depending on the approach taken, but higher than the random \( \sim 17\% \) value that we would get for 6 different lithologies in all classes but EncrCoral. The accuracy
of prediction for EncrCoral goes as low as ~17.5% (Tahiti, Table 3). Since encrusting corals usually grow surrounding other lithologies their physical properties are strongly influenced by these adjacent lithologies due to the resolution of the MSCL sensors. With such a high error this class should not be considered in further studies, since the information on the physical properties is not sufficient to identify them as a separate class.

The use of P-wave velocity from Tahiti cores produced no significant improvement on classification for most of the lithologies. A ~2% increase in accuracy was noticed for CorAlg and Silt, and the lithology with major improvements was Microbialite with an ~11% increase in accuracy. However, all the intervals here reported for the 95% level of confidence overlapped, suggesting that the models with re-saturated samples were not significantly better than the models without. Considering this, physical properties experts may want to evaluate the cost in time versus results obtained with the saturation of the samples. Despite of being re-saturated, the dataset from Tahiti presented more samples of the classes that were more difficult to identify, in particular microbialites were highly abundant in the cores from this expedition. We believe that the more mixed lithology of these cores can be the cause for the lower performance of the models for this expedition. If the data obtained by the physical properties are primarily used for identification of lithologies and core-log relations, cores can be run on the MSCL without a previous re-saturation process. This would avoid the loss of fines that may be washed out during the re-saturation process.
Despite that SVMs are one of the most effective classifiers for multivariate data, sometimes they can be overtaken by other techniques, in particular, Random Forest has been proved to reduce the error for some datasets (Meyer et al. 2003, Liam and Wiener 2002). This is in agreement with the “no free lunch theorem” (Wolpert 1992, Wolpert and Macready 1997, Wolpert 2001) that states that any two optimization algorithms perform similarly if their performance is averaged for all possible datasets. This indicates that despite our initial belief that SVM will perform better than RF, this is not the case. The overlapping of the 95% confidence intervals calculated for the models presented show that most of the SVM models perform as well as the RF models.

Both, random forest and support vector machines are known to perform better than the linear discriminant analysis (e.g. Cutler et al. 2007, Meyer et al. 2003) due to their capability to address non-linear relationships between the measured variables.

So far, MSCL data is not commonly published on from hard rock and carbonate sediments, other than the IODP reports. Despite this study only being a first step, the implementation of machine-learning models such as the ones presented here could provide new insight in ocean sediment and hard rock studies. In order to run supervised machine-learning models the data available needs to be labeled based on the classification need. This part of the process has been, so far, the most time consuming. The addition of artificial intelligence models for analysis of physical properties as part of lithologic interpretations would require training of the personnel in the use of the techniques applied in this study to better understand how the models work, and how the label assigned in them helps analyze the dataset.
The addition of other variables measured on the split core, such as color reflectance, could improve the identification, however, it would not allow a first ‘look’ at whole rounds.

The use of supervised machine learning models to identify carbonate sediments could be used in the near future to better relate the core recovery of carbonate sediments to downhole logging data, and therefore provide valuable information on the study of sea-level and past environmental changes. This information could be implemented real-time during core processing on board, and may provide new insight in the selection of subsequent drilling and coring locations.

ACKNOWLEDGMENTS

The authors want to thank the crew, technical support and science party of IODP expeditions 310 and 325. This project has been funded by IODP as part of the Post Expedition 325 Award to Kate Moran and Tania Lado Insua. Jody Webster was supported by an Australian Research Council Discover Grant (DP1094001). We would like to thank Fernando Cela Diaz and Annette Grilli for their help and support on the initial steps of this project. Special thanks to Matthew C. Smith, Ian and Tori Kulin and Lucy Hurlbut for their support during the development of this project.

REFERENCES


CHAPTER 4

RECONSTRUCTION OF PACIFIC OCEAN BOTTOM WATER SALINITY DURING THE LAST GLACIAL MAXIMUM

Tania Lado-Insua¹,*, Arthur J. Spivack², Dennis Graham³, Steven D’Hondt³ and Kathryn Moran¹,²,³

Article to be submitted to Geophysical Research Letters

¹Department of Ocean Engineering, University of Rhode Island, Bay Campus, Narragansett, 02882 RI, USA
²Graduate School of Oceanography, University of Rhode Island, Bay Campus, Narragansett, 02882 RI, USA
³Ocean Networks Canada and NEPTUNE Canada, University of Victoria, PO BOX 1700 STN CSC, Victoria, BC V8W 2Y2 Canada

*Correspondence author: T. Lado-Insua (e-mail: ladoinsuat@egr.uri.edu)
ABSTRACT

Understanding salinity in the deep ocean is important for reconstructing past ocean circulation and climate. Based on sediment pore fluid chloride measurements on a limited number of samples, Adkins et al. (2002) presented evidence, that during the last glacial maximum (LGM), the Pacific was more stratified than it is today.

Here we present chloride concentration profiles from six Pacific Ocean sites. This study expands the spatial coverage of salinity reconstructions for the LGM in the South, Equatorial and North Pacific Ocean.

These reconstructed LGM chloride concentrations of deep Pacific bottom water were ~3.54 to 4.64% greater than today’s values. Pacific Ocean bottom water salinity was also homogeneous across the wide range of latitudes studied here. These LGM salinity reconstructions are consistent with sea level at the time, which was ~120 to ~145 m higher than today.
1. INTRODUCTION

The Meridional Overturning Circulation (MOC) directly impacts the climate system via its heat flux (e.g. Srokosz et al. 2012) and the sequestration of CO$_2$ (e.g. Ito and Follows 2005, IPCC 2007, Zickfeld et al. 2012, Skinner et al. 2010). For these reasons, identifying its variation and understanding its driving mechanisms are areas of active research. Different hypothesis about the MOC future and past strength are under discussion (McManus et al. 2004, Meissner 2007, Okazaki et al. 2010, Lippold et al. 2012, Ritz et al. 2013). Deep ocean circulation can reach many different states of equilibrium but uncertainty remains about what states will emerge in response to climate change (Stommel 1961, reviewed in Kuhlbrodt et al. 2007). Variations in the MOC are strongly affected by changes in density generated by temperature and salinity variations (e.g. Bryden et al. 2005).

The reconstruction of past deep ocean conditions enable a better understanding of how ocean circulation has changed in the past and could change in future. Ocean circulation is understood by the direct measure and proxies of salinity and temperature because they affect the geostrophic flow and are used to identify water masses and circulation patterns (e.g. Macdonald et al. 2009, Adkins et al. 2005). Knowledge of the salinity and temperature of past oceans has increased since McDuff (1984, 1985) pointed out that the salinity of the last glacial maximum (LGM) is preserved in marine sediments. Schrag and DePaolo (1993) and later on, Adkins and Schrag (2001) successfully built upon this idea and reconstructed both salinity and temperature of the LGM ocean at a limited number of geographic locations. Based on these results (Schrag et al. 1996, Adkins and Schrag 2001, 2003, Adkins et al. 2002, Schrag et al.
2002), they argue that the glacial-age deep ocean was dominated by density differences due to salinity variations rather than temperature variations and that the ocean was strongly stratified.

Presently, Pacific bottom waters are renewed by the flow of Lower Circumpolar Deep Water (LCDW) flowing from the Antarctic Circumpolar Current System (ACCS). This water enters the Pacific as a deep western boundary current, in the southwestern part of the basin, east of New Zealand. The LCDW is the major water source for the entire Pacific basin and it shows a small salinity decrease going from south to north due to diapycnal mixing. Repeated circulation in the ACCS, where the North Atlantic Deep Water (NADW) and the Antarctic Bottom Water (AABW) mix to become LCDW makes this water mass homogeneous (Schmitz 1996). The nearly uniform composition of Pacific bottom water is due to the uniformity of this source. No bottom waters are presently forming in the Pacific; however, it is unclear that the location where water masses formed during the LGM was the same as today (Ganopolski and Rhamstorf 2001, Lynch-Stieglitz et al. 2007, Toggweiler and Russell 2008, Okazaki et al. 2010, Skinner et al. 2010, Kwon et al. 2012).

Despite that a saltier and more stratified deep glacial ocean has been reported (Lea et al. 2000, Adkins and Schrag 2001, 2003, Adkins and Pasquero 2004, Lynch-Stieglitz et al. 2007, Herguera et al. 2010) reconstruction of global ocean salinity during the LGM has not yet been realized. The results to date are important (Adkins and Schrag 2001, 2003, Schrag et al. 1996, 2002), but represent a small area of the Pacific, specifically two sites: Ocean Drilling Program (ODP) Sites 1123 and 1239 (Figure 1). These sites are located on the margin of the Southern Ocean and the
eastern boundary of the Pacific. Site 1123 is located in the region where a deep western boundary current presently supplies the bottom waters of the Pacific with LCDW.

Figure 1. Bathymetric map of the eight sites used for this study. Water depth of the site is indicated in parenthesis. Blue indicates expedition KN-195 (III), orange indicates ODP Leg 201, green indicates IODP Expedition 329 and white indicates Sites 1123 (ODP Leg 202) and 1239 (ODP Leg 181), previously reported by Adkins et al. (2002) and Adkins and Schrag (2003).

If the data from Site 1123 are reliable and the bottom waters of the Pacific circulated in a similar manner during the LGM we would predict that the LGM bottom waters would have a salinity similar to that inferred for Site 1123. Here we expand upon these limited geographic studies by using sediment pore fluid measurement
techniques, coupled with diffusion modeling, to reconstruct salinity during the LGM in six locations that span the bottom waters of the equatorial Pacific, the south Pacific gyre and the north Pacific gyre (Figure 1).

2. MATERIALS AND METHODS: RECONSTRUCTING PALEOSALINITIES

Pore water samples from six sites were collected during three separate cruises: R/V Knorr 195 (III) (Sites EQP10 and EQP11), ODP 201 (Site 1225) and IODP Expedition 329 (Sites U1365, U1370 and U1371). Sites EQP10 and EQP11 were collected with long piston cores, while all the other samples were drilled and cored. These samples cover deep areas (>3760 m water depth) of the Pacific basin in the eastern central Pacific gyre, the north Pacific and the central south Pacific gyre (Figure 1). Core recovery was high in all cases (>90%) and the sediment recovered was dominated by red clays and carbonate ooze. In this study, three sites (1225, U1370 and U1371) were drilled to basement and one (U1365) contained an impermeable chert layer. For the sites where basement was not reached (EQP10 and EQP11 from KN 195 (III)), the numerical model for reconstructing salinity was extrapolated to 100 meters below seafloor (mbsf), the depth to basalt is estimated from seismic reflection profiles. At sites EQP10 and EQP11 chloride concentration measurements only reach a depth of ~27 mbsf, yet the column of sediment (>50 m) was thick enough to model chloride concentration for the LGM.

Sediment interstitial waters (IW) were collected by squeezing of whole core rounds using Manheim squeezers (Manheim and Sayles 1974). Chloride concentration was either determined by AgNO3 titration (Site 1225) or with an ion chromatography
procedure optimized for chloride concentration measurement (all other sites). The measured relative standard deviation for the titrations was 0.12% and 0.09% for ion chromatography. For all analyses, IAPSO Standard Seawater was used as the standard (D’Hondt et al. 2003, 2011). Measurements were made within three days of core recovery to minimize the potential effect of evaporation. Measured chloride concentrations for the Pacific are shown as equivalent salinities in all graphs. Chloride concentration describes a curve with a maximum in depth that corresponds to the signal from the Last Glacial Maximum (LGM) as it diffuses through the sediment. This maximum is achieved at approximately 40 mbsf for all of our study sites. This means that sites must have a sediment thicknesses of 50 m or greater and core recovery of ~25 meters or greater to fulfill the requirement for analyses to span at least 90% of the maximum (Adkins and Schrag 2002).

Diffusivity depends on the molecular paths within sediment characterized by tortuosity (Boudreau, 1997). This value is often estimated from measured porosity based on different variants of Archie’s Law (Shen and Chen 2007). The porosity of the sediment is the ratio of the volume of voids to the total volume. To obtain porosities, wet and dry weight of a sediment sample and either its wet or dry volume are measured. For sites EQP10 and EQP11 wet volume was measured with a cut syringe (method C, Blum 1997). For the rest of the samples dry volume was measured with a pycnometer (Method A, Blum 1997). Unlike previous studies (e.g. Adkins and Schrag 2001), we use measured formation factors to infer tortuosity for all sites except Site 1225; for this site we used Archie’s Law (Archie, 1947) following Adkins and Schrag (2003), since tortuosity was not measured. The tortuosity ($\theta^2$) can be calculated as the
formation factor \( (f) \) multiplied by the porosity \( (\phi) \):

\[
\theta^2 = f\phi
\]

The formation factor is the ratio of the electrical conductivity measured in the sediment in relation to the electrical conductivity of a known standard (IAPSO water in our case). We measured conductivity with a Metrohm 712 meter with platinum electrodes by inserting them and a temperature probe into the split core. Changes in diffusivity due to temperature variations were taken into account following the Stokes-Einstein relation. We use a one-dimensional, time-dependent, numerical diffusion model to find the magnitude of the LGM salinity that best fits the measured data:

\[
\frac{\partial C}{\partial t} = \frac{\partial}{\partial z} \left( \phi D \frac{\partial C}{\partial z} \right) = \frac{\partial}{\partial z} \left( f \frac{\partial C}{\partial z} \right)
\]

Where \( C \) is the chloride concentration, \( t \) is time, \( z \) is depth below sediment-water interface, \( (\phi) \) is the porosity in the sediment, and \( f \) is formation factor. The diffusion equation was implemented in MATLAB by using finite differences with a second order Adams-Bashforth-Moulton method (Appendix V). All of the sites are dominated by low sedimentation rates and Peclet numbers were too small to produce significant advection. No reaction term was added to the equation because our sites are dominated by red clays and carbonate ooze far from volcanic areas and therefore free of ash that could react and alter the chloride values (Adkins and Schrag 2003).

The chloride concentration of the bottom water as a function of time, (bottom water boundary condition, BWBC), is the optimized parameter. Its relative variation with time is based on reconstructed sea level variation, but its magnitude is optimized by a factor, \( a \), to best fit to the measured pore water profile. An \( a \) value of 0 implies
that the salinity difference predicted solely from the sea level curve produces the best-fit of the data while a value of $x$ infers that the best fit is produced by multiplying the salinity predicted by the sea level curve by the constant $\alpha = x$. We used the sea-level curve described by Adkins and Schrag (2003) based on the coral sea-level data for the last 30,000 yr (Fairbanks 1989, Bard et al. 1990, Edwards et al. 1993) and the deep water record of Chappell and Shackleton (1986) from 115 ky to the present. We optimized the BWBC by minimizing the mean square error between the modeled and measured data. We show a variation of the optimal value, $\alpha \pm 0.1$ for all models except EQP10 and EQP11 where $\alpha \pm 0.2$ is shown. The three most important parameters that affect the accuracy of this reconstruction are diffusivity (improved here with the use of formation factors), scatter in measured chloride concentration, and the shape and amplitude of the BWBC (Adkins and Schrag 2003).

The numerical analysis also requires initial conditions and an additional boundary condition: 1) an initial chloride concentration profile, and 2) a boundary condition at the bottom of the profile. For consistency, an iterative method was used, where the initial condition of the model was the average of chloride concentration for the last 115 ky based on the optimized sea level curve and the $\alpha$ value, but the model is insensitive to the initial condition for the depths where the LGM signal is captured (Adkins and Schrag 2003).

Three different conditions for the bottom of the sediment column were used: a concentration boundary condition at Sites 1225, EQP10, EQP11 and U1370 (equal to the initial condition of the model), a bottom boundary condition equal to the BWBC
for Site U1371 (where diffusion from the basement was observed), and a no-flux bottom boundary condition for Site U1365 where the chert layer occurs.

3. RESULTS AND DISCUSSION

Chloride concentrations reconstructed from the one dimensional diffusion model are consistent with measured values (Figure 2). The reconstructed LGM bottom water salinities ranged from 35.91±0.1 to 36.25±0.1 mg/kg (Table 1). This is equivalent to salinity values between ~3.54% and ~4.64% higher than today for the deep Pacific Ocean during the LGM (Table 1). The salinity of the bottom water for these sites today is ~34.70±0.1 mg/kg (WOCE database, Table 1).

A range of average global ocean sea levels during the LGM have been reported (e.g., Gehrels 2010). These levels vary from ~120 m higher than today in the most conservative records (Fairbanks 1989, Bard et al. 1990, Lambeck et al. 2002, Peltier and Fairbanks 2006) to ~135 m higher or more, as reported by Colonna et al. (1996) and Yokoyama et al. (2000). Milne and Mitrovica (2008) indicate that these results have a large range because of changes in mantle viscosity that can generate uncertainties of 15 to 20 m. By applying a sea level of 120 to 135 m higher than today for the LGM, current salinity values for the deep Pacific (WOCE database), and average depth for the modern ocean of 3682 m (Charette and Smith 2010), the LGM salinity range expected, if the salinity only varied due to sea level changes, is 35.83 to 35.97 mg/kg. The reconstructed LGM values for U1371 and U1370 are within this range. The LGM salinity values for ODP Site 1225 (36.07±0.1 mg/kg) and EQP11 (36.06±0.2 mg/kg) are close to the higher end of this range, implying a slightly greater
sea level change of ~145 m, which is consistent with some of the LGM sea level literature (e.g., Colonna et al. 1996, Yokoyama et al. 2000).

Sites EQP10 (36.21±0.2 mg/kg) and IODP U1365 (36.25±0.1 mg/kg) show higher reconstructed values. The sea level change necessary to fulfill these reconstructed salinities would be approximately ~160 to 164 m, which is outside the range of all the reported values. Both of these sites present some limitations. Site U1365 has a chert impermeable layer close to the depth where the LGM signal would reach its maximum. This could have altered the diffusion pattern generating ~ 0.3 mg/kg higher salinity than the other sites in the South Pacific (U1370, U1371). The results for this site should therefore be used with caution. The reduced number of measurements in depth for sites EQP10 and EQP11 limits the precision on the prediction for these sites. However, despite a higher range, the values for salinity at these sites are accurate.

The mean value of the salinity for our data, 36.08±0.06 mg/kg is slightly lower that from the Site 1123 (Chatham Rise, 3290 m water depth) data reconstruction of Adkins and Schrag (2002), 36.19±0.07, although the uncertainties overlap.

Adkins et al. (2002) suggests that a mass of fresh water in the glacial ocean or the sequestration of fresh water in groundwater and ice occurring simultaneously with an ~135 m sea level high stand (Adkins et al. 2002) could explain these higher values.
Figure 2. Salinity profiles plotted with depth below seafloor for all the sites. Optimized ($\alpha$) modeled salinity curve (black line), salinity curve error (grey dashed lines) produced by $\alpha \pm 0.1$ for all sites except EQP10 and EQP11 where $\alpha \pm 0.2$, and measured salinity based on chloride concentration (red dots).
Table 1. Site information, modern and reconstructed LGM values for chloride concentration and salinity and relative difference of LGM salinity compared to modern salinity.

<table>
<thead>
<tr>
<th>Site</th>
<th>Lat.</th>
<th>Long.</th>
<th>Water depth (m)</th>
<th>Modern values</th>
<th>LGM values</th>
<th>Relative difference LGM salinity compared to modern</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chloride conc. (mM/L)</td>
<td>Salinity (mg/kg)*</td>
<td>Chloride conc. (mM/L)</td>
</tr>
<tr>
<td>EQP11</td>
<td>30.355</td>
<td>-157.871</td>
<td>5813</td>
<td>551.26</td>
<td>34.69</td>
<td>576.37±2.93</td>
</tr>
<tr>
<td>EQP10</td>
<td>20.683</td>
<td>-143.357</td>
<td>5412</td>
<td>551.42</td>
<td>34.70</td>
<td>578.86±2.93</td>
</tr>
<tr>
<td>ODP 1225</td>
<td>2.771</td>
<td>-110.571</td>
<td>3760</td>
<td>551.26</td>
<td>34.69</td>
<td>576.66±1.46</td>
</tr>
<tr>
<td>IODP U1365</td>
<td>-23.851</td>
<td>-165.644</td>
<td>5695</td>
<td>551.42</td>
<td>34.70</td>
<td>579.52±1.46</td>
</tr>
<tr>
<td>IODP U1370</td>
<td>-41.852</td>
<td>-153.106</td>
<td>5074</td>
<td>551.58</td>
<td>34.71</td>
<td>574.03±1.46</td>
</tr>
<tr>
<td>IODP U1371</td>
<td>-45.964</td>
<td>-163.184</td>
<td>5312</td>
<td>551.42</td>
<td>34.70</td>
<td>574.61±1.46</td>
</tr>
</tbody>
</table>

*Salinity obtained from the WOCE atlas closest line (Appendix VI)
The chloride concentration at Site U1371 could only be reconstructed when the same boundary condition was applied to the top and bottom of the sediment column, indicating similar changes over time in the bottom water and basement pore fluid. For example, Site U1371 suggests diffusion from the basement based on the measured chloride concentration values. This effect does not impact the salinity signal for the LGM at this site (Figure 2) and at other sites (Table 1). The model is consequently insensitive in the upper 40 m of the sediment column to the boundary condition at the basement for sites with more than 100 m sediment thickness. This suggests that a change in chloride concentration values in basement fluids co-vary with changes in the overlying deep bottom water, and, therefore, with sea level changes. It also implies that there may be a link between basement fluids and deep ocean bottom waters. Similar fluid connections have been recorded in the North Pond results obtained by IODP Exp. 336 (Lado-Insua et al. in preparation).

These bottom water salinity values presented here suggest strong stratification in Pacific Ocean waters during the LGM (Lynch-Stieglitz et al. 2007, Herguera et al. 2010).

4. CONCLUSIONS

The salinity of the bottom waters of the Pacific during the Last Glacial Maximum is better constrained with data and model results from six new sites presented in this study. Our results indicate a consistent increase in the salinity over a wide geographic region of the Pacific Ocean, significantly higher than today and predicted based solely on sea level. We found no major changes in latitude for the reconstructed salinities.
They are all in agreement with a nearly homogeneous deep (>3500 m) Pacific Ocean during the LGM.

Our results suggest a change in Pacific bottom water salinity of 4.02±0.16, consistent or slightly higher than the global salinity change, 3.6±0.3 %, expected for a sea level maximum of ~120 to ~145 m during the LGM. The mean value of the salinity for our data, 36.08±0.06 mg/kg, is slightly lower that from the Site 1123 data reconstruction of Adkins and Schrag (2002), 36.19±0.07, although the uncertainties overlap. Thus our results support the accuracy of the Site 1123 results and are consistent with the bottom waters of the Pacific being supplied with a homogenous water mass as it is today. This implies that the primary mechanism that drives Pacific bottom water uniformity, the ACCS, was active during the LGM. In addition, our results are in agreement with the Pacific Ocean having been stratified and its bottom water slightly saltier than the average ocean as compared to today (Lea et al. 2000, Adkins and Schrag 2001, 2003, Adkins and Pasquero 2004, Lynch-Stieglitz et al. 2007, Herguera et al. 2010).

ACKNOWLEDGMENTS

The authors want to thank the Barrié de La Maza Foundation, IODP-MI and NSF for its financial support to Tania Lado Insua during the development of this study. We would like to thank Dr. Christian Janssen and Dr. Jeffrey Harris for their useful code reviews. We thank the crew, science party and technical staff of IODP Expedition 329 and Knorr 195 (III).
REFERENCES


Nature, 342, 637-642


APPENDICES

APPENDIX I

SELECTION OF THE NUMBER OF CLUSTERS (Chapter 2)

Figure 1. Cluster solution against logarithmic of sum of squared error minus the sum of squared error of the randomized samples. This technique helps us to determine the number of clusters.
APPENDIX II

R CODE FOR CLUSTER ANALYSIS (Chapter 2)

# k-means cluster for Cascadia data U1328

setwd("~/Desktop/SCIMPL_kmeans/U1328")
library("akima")
library(RColorBrewer)
library(ggplot2)
library(lattice)
library(reshape2)
k<-3

## load each dataset independently because they have different lengths

MAD_U1328<-read.csv(file="u1328_bd_dm.csv")

MS_U1328<-read.csv(file="u1328_ms_dm.csv")
MS_U1328$MS[MS_U1328$MS<0]=NA

CO3_U1328<-read.csv(file="CO3_U1328.csv")
#CO3 Site 889A

Ch_U1328<-read.csv(file="u1328_ch_dm.csv")
Ca_U1328_1<-read.csv(file="u1328_ca_dm.csv")

BD_U1328C<-read.csv(file="WL_HostileEnv_Lithodensity_U1328C.csv")
BD_U1328C$RHOM[BD_U1328C$RHOM<1.1]=NA

BD_U1328Aadn<-read.csv(file="LWD_Density_adn_U1328A.csv")
BD_U1328Aadn$RHOB[BD_U1328Aadn$RHOB<1.1]=NA

BD_U1328Aeco<-read.csv(file="LWD_Density_Eco_U1328A.csv")
BD_U1328Aeco$RHOB[BD_U1328Aeco$RHOB<1.1]=NA
## Electrical resistivity

- **LWD attenuation resistivity ecoscope (Average all variables but depth=Mean)**
  
  ```r
  Res_U1328Aatt <- read.csv(file="LWD_Attenuation_Resist_Eco.csv")
  ```

- **LWD phase resistivity ecoScope (Average all variables but depth=Mean)**
  
  ```r
  Res_U1328Apha <- read.csv(file="LWD_Phase_Resist_Eco_U1328A.csv")
  ```

- **Dual induction resistivity (vble=IDPH)**
  
  ```r
  Res_U1328C <- read.csv(file="WL_Dual_Induct_resistivity_U1328C.csv")
  ```

## Gamma Ray logging

- **Hostile Environment Lithodensity tool (RHOM)**
  
  ```r
  GR_U1328C <- read.csv(file="WL_HostileEnv_Gamma_U1328C.csv")
  ```

- **Gamma Ray Spectrometry Dual Induction Tool String (use SGR only)**
  
  ```r
  GR_U1328C$HSGR[GR_U1328C$HSGR<0]=NA
  ```

- **LWD Gamma EcoScope (GRMA)**
  
  ```r
  GR_U1328Aeco <- read.csv(file="LWD_Gamma_Eco_U1328A.csv")
  ```

- **Gamma Ray Spectrometry Dual Induction Tool String (CGR only)**
  
  ```r
  GR_U1328Aeco$GRMA[as.numeric(GR_U1328Aeco$GRMA)<0]=NA
  ```

- **LWD Gamma geovision (GR)**
  
  ```r
  GR_U1328Ageo <- read.csv(file="LWD Gamma_geovis_U1328A.csv")
  ```

## Sonic data

- **LWD Velocity sonic (VELP)**
  
  ```r
  PW_U1328A <- read.csv(file="LWD_Velocity_Sonic_U1328A.csv")
  ```

- **Dipole sonic Pass 1 (pass 2 and repeated not used) (vble=VCO)**
  
  ```r
  PW_U1328C <- read.csv(file="WL_Dipole_Sonic_Pass1_U1328C.csv")
  ```

### ELIMINATE DATA FROM PIPE ON LOGGING

# Pipe located down to 61.12 m for site 1328C and no pipe on drilling while logging
# values over 74m were eliminated

dep<-61.12

## minimum maximum depth recorded 227.69 (Dipole sonic imager)

```r
mind<-228
```
## Prepare the data: INTERPOLATION

### Vectors for interpolation

```r
xout <- seq(dep, mind, 0.1)  # depths limited by WL sonic Pwave
```

### INTERPOLATION FOR MAD

```r
BDm_U1328 <- data.frame(approx(MAD_U1328$Depth, MAD_U1328$BD, xout, method = "linear", rule = 2, ties = mean))
coltitlesBDm <- c("Depth", "BD")
names(BDm_U1328) <- as.vector(unlist(coltitlesBDm))
```

### Combining MAD values

```r
BD <- BDm_U1328$BD
```

### INTERPOLATION FOR MS

```r
MSm_U1328 <- data.frame(approx(MS_U1328$Depth, MS_U1328$MS, xout, method = "linear", rule = 2, ties = mean))
coltitlesMS <- c("Depth", "MS")
names(MSm_U1328) <- as.vector(unlist(coltitlesMS))
MS <- MSm_U1328$MS
```

### INTERPOLATION FOR Cl

```r
Cl_U1328 <- data.frame(approx(Ch_U1328$Depth, Ch_U1328$Chlorinity, xout, method = "linear", rule = 2, ties = mean))
coltitlesCl <- c("Depth", "Cl")
names(Cl_U1328) <- as.vector(unlist(coltitlesCl))
Cl <- Cl_U1328$Cl
```

### INTERPOLATION FOR Ca

```r
Ca_U1328 <- data.frame(approx(Ca_U1328_1$Depth, Ca_U1328_1$Calcium, xout, method = "linear", rule = 2, ties = mean))
coltitlesCa <- c("Depth", "Ca")
names(Ca_U1328) <- as.vector(unlist(coltitlesCa))
Ca <- Ca_U1328$Ca
```

### INTERPOLATION FOR Bulk density logging

```r
BDla_U1328 <- data.frame(approx(BD_U1328Aadn$Depth, BD_U1328Aadn$RHOB, xout, method = "linear", rule = 2, ties = mean))
coltitlesBDl <- c("Depth", "BDla")
names(BDla_U1328) <- as.vector(unlist(coltitlesBDl))
```

```r
BDle_U1328 <- data.frame(approx(BD_U1328Aeco$Depth, BD_U1328Aeco$RHOB, xout, method = "linear", rule = 2, ties = mean))
names(BDle_U1328) <- as.vector(unlist(coltitlesBDl))
```
BDl_U1328C<-data.frame(approx(BD_U1328C$Depth,BD_U1328C$RHOM, xout,method="linear",rule=2,ties=mean))
names(BDl_U1328C)<-as.vector(unlist(coltitlesBDl))

BDl<-BDl_U1328C$BDl
Bdla<-BDla_U1328$BDl
BDle<-BDle_U1328$BDl

## INTERPOLATION ELECTRICAL RESISTIVITY
Resl_U1328C<-data.frame(approx(Res_U1328C$Depth,Res_U1328C$IDPH, xout, method="linear",rule = 2,ties=mean))
coltitlesRes<-c("Depth","Res")
names(Resl_U1328C)<-as.vector(unlist(coltitlesRes))

Resla_U1328<-data.frame(approx(Res_U1328Aatt$Depth,Res_U1328Aatt$Mean, xout,method="linear",rule=2,ties=mean))
names(Resla_U1328)<-as.vector(unlist(coltitlesRes))

Reslp_U1328<-data.frame(approx(Res_U1328Apha$Depth, Res_U1328Apha$Mean,xout,method="linear",rule=2,ties=mean))
names(Reslp_U1328)<-as.vector(unlist(coltitlesRes))

Resl<-Resl_U1328C$Res
Resla<-Resla_U1328$Res
Reslp<-Reslp_U1328$Res

## INTERPOLATION GAMMA RAY
GRle_U1328<-data.frame(approx(GR_U1328Aeco$Depth,GR_U1328Aeco$GRMA, xout, method="linear",rule = 2,ties=mean))
coltitlesGR<-c("Depth","GR")
names(GRle_U1328)<-as.vector(unlist(coltitlesGR))

GRlg_U1328<-data.frame(approx(GR_U1328Ageo$Depth,GR_U1328Ageo$GR, xout, method="linear", rule = 2,ties=mean))
names(GRlg_U1328)<-as.vector(unlist(coltitlesGR))

GRI_U1328C<-data.frame(approx(GR_U1328C$Depth,GR_U1328C$HSGR, xout, method="linear",rule = 2,ties=mean))
names(GRI_U1328C)<-as.vector(unlist(coltitlesGR))

#Combining Gamma Ray values
GRle<-GRle_U1328$GR
GRlg<-GRlg_U1328$GR
GRI<-GRI_U1328C$GR
### INTERPOLATION PWave

```r
PWl_U1328A <- data.frame(approx(PW_U1328A$Depth, PW_U1328A$VELP, xout, method = "linear", rule = 2, ties = mean))
coltitlesPW <- c("Depth", "PW")
names(PWl_U1328A) <- as.vector(unlist(coltitlesPW))

PWl_U1328C <- data.frame(approx(PW_U1328C$Depth, PW_U1328C$VCO, xout, method = "linear", rule = 2, ties = mean))
names(PWl_U1328C) <- as.vector(unlist(coltitlesPW))

# Combining Pwave values
PWlA <- PWl_U1328A$PW
PWlC <- PWl_U1328C$PW
```

### COMBINING DATA IN ONE MATRIX

```r
DATA <- data.frame(MS, Cl, Ca, BDm, BDla, BDle, Resl, Resla, Reslp, GRl, GRle, GRlg, PWlA, PWlC)
coltitlesDATA <- c("MS", "Cl", "Ca", "BDm", "BDla", "BDle", "Resl", "Resla", "Reslp", "GRl", "GRle", "GRlg", "PWlA", "PWlC")
Depth <- (xout)
names(Depth) <- "Depth"
names(DATA) <- as.vector(unlist(coltitlesDATA))
```

### PCA

```r
PCA <- princomp(~., DATA, cor = TRUE)
# standard deviation explained by each component
screepplot(PCA)
# Identify the variables that we want to take from the PCA
# In this example we selected 5 out of 7 principal component variables.
# The variance explained by each one of them and the loading of each variable is provided
# Percentage of variance explained (Fall)
per <- as.data.frame(PCA$sdev)^2/sum(PCA$sdev^2)
print(per)
# Percentage explained by the first 5 components:
per12 <- sum(per[1:12,])
print(per12)
# Loading of each variable (Fall)
print(PCA$loadings)
```
#### selection of PC variables for cluster

**#data from PCA**

nv<-12 #number of variables in DATA
c<-data.frame(PCA$scores[,1:nv])

pc<-data.frame(PCA$scores[,1:nv])

c<-	ext{c}("PC1","PC2","PC3","PC4","PC5","PC6","PC7","PC8","PC9","PC10","PC11","PC12")
names(pc)<-as.vector(unlist(c))

**# From Peeples 2011 code: HELP TO SELECT NUMBER OF CLUSTERS**

**#number of k to test**

n.lev<-10
c<-as.matrix(pc)

wss <- rnorm(10)

while (prod(wss==sort(wss,decreasing=T))==0) {
    wss <- (nrow(c)-1)*sum(apply(c,2,var))
    for (i in 2:n.lev) wss[i] <- sum(kmeans(c, centers=i, iter.max=50, nstart=10)$withinss)}

**# Calculate the within groups sum of squared error (SSE) for the number of cluster solutions selected by the user**

wss <- rnorm(10)

while (prod(wss==sort(wss,decreasing=T))==0) {
    wss <- (nrow(c)-1)*sum(apply(c,2,var))
    for (i in 2:n.lev) wss[i] <- sum(kmeans(c, centers=i, iter.max=50, nstart=10)$withinss)}

**# Calculate the within groups SSE for 250 randomized data sets (based on the original input data)**

k.rand <- function(x){
    km.rand <- matrix(sample(x),dim(x)[1],dim(x)[2])
    rand.wss <- as.matrix(dim(x)[1]-1)*sum(apply(km.rand,2,var))
    for (i in 2:n.lev) rand.wss[i] <- sum(kmeans(km.rand, centers=i, iter.max=50, nstart=10)$withinss)
    rand.wss <- as.matrix(rand.wss)
    return(rand.wss)
}

rand.mat <- matrix(0,n.lev,250)

k.1 <- function(x) {
    for (i in 1:250) {
        r.mat <- as.matrix(suppressWarnings(k.rand(kdata)))
        rand.mat[,i] <- r.mat
    }
    return(rand.mat)
}

rand.mat <- k.1(kdata)
# Plot within groups SSE against all tested cluster solutions for actual and randomized data - 1st: Log scale, 2nd: Normal scale

```r
xrange <- range(1:n.lev)
yrange <- range(log(rand.mat),log(wss))
plot(xrange,yrange, type='n', xlab='Cluster Solution',
    ylab='Log of Within Group SSE',
    main='Cluster Solutions against Log of SSE')
for (i in 1:250) lines(log(rand.mat[,i]),type='l',col='red')
lines(log(wss), type="b", col='blue')
legend('bottomleft',c('Actual Data', '250 Random Runs'), col=c('blue', 'red'), lty=1)
yrange <- range(rand.mat,wss)
plot(xrange,yrange, type='n', xlab="Cluster Solution",
    ylab="Within Groups SSE", main="Cluster Solutions against SSE")
for (i in 1:250) lines(rand.mat[,i],type='l',col='red')
lines(1:n.lev, wss, type="b", col='blue')
legend('topright',c('Actual Data', '250 Random Runs'), col=c('blue', 'red'), lty=1)
```

# Calculate the mean and standard deviation of difference between SSE of actual data and SSE of 250 randomized datasets

```r
r.sse <- matrix(0,dim(rand.mat)[1],dim(rand.mat)[2] )
wss.1 <- as.matrix(wss)
for (i in 1:dim(r.sse)[2]) {
    r.temp <- abs(rand.mat[,i]-wss.1[,1])
    r.sse[,i] <- r.temp
}
r.sse.m <- apply(r.sse,1,mean)
r.sse.sd <- apply(r.sse,1,sd)
r.sse.plus <- r.sse.m + r.sse.sd
r.sse.min <- r.sse.m - r.sse.sd
```

# Plot difference between actual SSE mean SSE from 250 randomized datasets - 1st: Log scale, 2nd: Normal scale

```r
xrange <- range(1:n.lev)
yrange <- range(log(r.sse.plus),log(r.sse.min))
plot(xrange,yrange, type='n', xlab='Cluster Solution',
    ylab='Log of SSE - Random SSE',
    main='Cluster Solutions against (Log of SSE - Random SSE)')
lines(log(r.sse.m), type="b", col="blue")
lines(log(r.sse.plus), type="l", col="red")
lines(log(r.sse.min), type="l", col="red")
legend('bottomright',c('SSE - random SSE', 'SD of SSE-random SSE'),
       col=c('blue', 'red'), lty=1)
```

```r
xrange <- range(1:n.lev)
yrange <- range(r.sse.plus,r.sse.min)
plot(xrange,yrange, type='n', xlab='Cluster Solution',
     ylab='SSE - Random SSE',
     main='Cluster Solutions against (SSE - Random SSE)')
```
lines(r.sse.m, type="b", col='blue')
lines(r.sse.plus, type='l', col='red')
lines(r.sse.min, type='l', col='red')
legend('bottomright',c('SSE - random SSE', 'SD of SSE-random SSE'),
col=c('blue', 'red'), lty=1)

###################################################
## K-MEANS CLUSTER
###################################################

cl<-kmeans(pc,k,nstart=25)
#evaluate k
print(c(cl$totss,cl$tot.withinss,cl$betweenss))

###################################################
## INITIAL PLOTS
###################################################

#plot results of cluster
#in relation to coordinates
plot(Depth, col = cl$cluster)
#cluster values in the same matrix as coordinates
clz<-data.frame(cbind(as.numeric(Depth),cl$cluster))

#number of points in each cluster
cl$size
Cluster<-as.factor(as.numeric(cl$cluster))

pc1<-data.frame(cbind(Depth,pc))
df <- melt(pc1, id = 'Depth', variable_name = 'variable')
ggplot(df, aes(Depth,value)) + geom_line(aes(colour = variable))

qplot(Cluster,Depth,colour=Cluster,binwidth=0.1, ylim=c(mind,dep))
APPENDIX III

HISTOGRAMS OF THE PHYSICAL PROPERTIES (Chapter 3)

Figure 1. Histograms of the variables analyzed for both expeditions (Tahiti and GBREC) for advanced classification of carbonate sediments.
Figure 1. (Continuation)
APPENDIX IV

R CODE FOR ADVANCED CLASSIFICATION OF CARBONATE SEDIMENTS (Chapter 3)

LOADING DATA AND LIBRARIES

########################################################################
## LOAD DATA AND LIBRARIES
########################################################################

## Load the data; make sure you set the working directory to the
## directory where the data resides, either via setwd("/path/to/directory")
## or via the user interface of your choice

DATA<-read.csv(file="Unbalanced_325.csv") #Unbalanced_310.csv #for Exp. 310

library(MASS)
library(ROCR)
library(RColorBrewer)
require("randomForest")
library(ggplot2)
library(lattice)
library(e1071)
UNBALANCED DATASET SAMPLING FUNCTION

###########################################################
## SAMPLING FUNCTION FOR UNBALANCED DATASET
###########################################################

splitdf <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(2*length(index)/3), replace=TRUE)
  trainset<-dataframe[trainindex, ]
  testset <- dataframe[-trainindex, ]
  list(trainset=trainset,testset=testset)
}
UP-SAMPLING FUNCTION

```r
splitdf <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(2*length(index)/3), replace=TRUE)
  DATAU <- dataframe[trainindex, ]
  CorAlgU=subset(DATAU, Species=='CorAlg')
  MicU=subset(DATAU, Species=='Microbialite')
  EncCU=subset(DATAU, Species=='EncrCoral')
  MassCU=subset(DATAU, Species=='MassCoral')
  SandU=subset(DATAU, Species=='Sand')
  SiltU=subset(DATAU, Species=='Silt')
  maxi<-max(length(CorAlgU$Species),length(MicU$Species),
           length(EncCU$Species),length(MassCU$Species),
           length(SiltU$Species)) #sand out because of being too many in 325
  val<-maxi #or maxi
  #sampling up with replacement
  CorAlg<-CorAlgU[sample(1:nrow(CorAlgU),val,replace=TRUE),]
  Mic<-MicU[sample(1:nrow(MicU),val,replace=TRUE),]
  EncC<-EncCU[sample(1:nrow(EncCU),val,replace=TRUE),]
  MassC<-MassCU[sample(1:nrow(MassCU),val,replace=TRUE),]
  Sand<-SandU[sample(1:nrow(SandU),val,replace=TRUE),]
  Silt<-SiltU[sample(1:nrow(SiltU),val,replace=TRUE),]
  #bind after sampling
  trainset<-data.frame(rbind(CorAlg,Mic,EncC,MassC,Sand,Silt))
  testset <- dataframe[-trainindex, ]
  list(trainset=trainset,testset=testset)
}
```
splitdf <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(2*length(index)/3), replace=TRUE)
  DATAU <- dataframe[trainindex,]
  CorAlgU=subset(DATAU, Species=='CorAlg')
  MicU=subset(DATAU, Species=='Microbialite')
  EncCU=subset(DATAU, Species=='EncrCoral')
  MassCU=subset(DATAU, Species=='MassCoral')
  SandU=subset(DATAU, Species=='Sand')
  SiltU=subset(DATAU, Species=='Silt')

  mini<-min(length(CorAlgU$Species),length(MicU$Species),
    length(EncCU$Species),length(MassCU$Species),
    length(SandU$Species),length(SiltU$Species))
  val<-mini
  #sampling up with replacement
  CorAlg<-CorAlgU[sample(1:nrow(CorAlgU),val,replace=TRUE),]
  Mic<-MicU[sample(1:nrow(MicU),val,replace=TRUE),]
  EncC<-EncCU[sample(1:nrow(EncCU),val,replace=TRUE),]
  MassC<-MassCU[sample(1:nrow(MassCU),val,replace=TRUE),]
  Sand<-SandU[sample(1:nrow(SandU),val,replace=TRUE),]
  Silt<-SiltU[sample(1:nrow(SiltU),val,replace=TRUE),]

  #bind after sampling
  trainset<-data.frame(rbind(CorAlg,Mic,EncC,MassC,Sand,Silt))
  testset <- dataframe[-trainindex,]
  list(trainset=trainset,testset=testset)
}
LINEAR DISCRIMINANT ANALYSIS MODEL

# real values
real.val<-DATA$Species

# bootstrap
n<-c(1:200)
resB<-c(1:200)
dia<-data.frame(matrix(nrow=length(n),ncol=6))
names(dia)=namesSp

for(j in n) {
    # apply the function
    splits <- splitdf(DATA, seed=593+j)
lapply(splits,nrow)
    training <- splits$trainset
    testing <- splits$testset

    model.ldaB <- lda(Species~., training, CV=FALSE)
pred.ldaB <- predict(model.ldaB, testing)
    # Assess the accuracy of the prediction
    # percent correct for each category of Species
    ctB <- table(testing$Species, pred.ldaB$class)
    resB[j]<-sum(diag(prop.table(ctB)))
dia[j,]<-diag(prop.table(ctB, 1))
    print(resB)
}

aB<-sort(resB)
ub<-aB[195]
lb<-aB[5]
int.a<-c(lb,ub)
print(int.a)

m<-c(1:6)
class<-c(1:200)
intup<-c(1:length(m))
intdown<-c(1:length(m))

for (i in m){
    class<-dia[i]
    class2<-sort(class[[1]])
intup[i]<-class2[195]
intdown[i]<-class2[5]

intervals<-data.frame(rbind(intup, intdown))
names(intervals)<-nameSp

#Results
sum(ctB)
summary(testing$Species)
print(int.a)
print(intervals)
print(ctB)
RANDOM FOREST MODEL

# Bootstrap
n<-c(1:200)
resB<-c(1:200)
er<-c(1:200)
acc<-c(1:200)
dia<-data.frame(matrix(nrow=length(n),ncol=6))
names(dia)=nameSp

for(j in n) {
  #apply the function
  splits <- splitdf(DATA, seed=593+j)
lapply(splits,nrow)
  training <- splits$trainset
  testing <- splits$testset
  model.rfB <- randomForest (Species~.
    , data = training
    , mtry = 2
    , ntree = nt
    , do.trace = FALSE
    , proximity = TRUE
    , importance = TRUE
    , keep.forest = TRUE
    )

  pred.rfB <- predict(
    model.rfB,
    newdata=testing,
    importance=TRUE,
    proximity=TRUE)
  # Assess the accuracy of the prediction
  # percent correct for each category of Species
  ctB <- table(testing$Species, pred.rfB$predicted)
  diag(prop.table(ctB, 1))
  # total percent correct
  resB[j]<-sum(diag(prop.table(ctB)))
}
print(resB)

er[j]<-model.rfB$err.rate[nt, 1]
acc[j]<-1-model.rfB$err.rate[nt, 1]
dia[j,]<-diag(prop.table(ctB, 1))
}

aB<-sort(resB)
ub<-aB[195]
lb<-aB[5]
int.a<-c(lb, ub)
print(int.a)
aOOB<-sort(acc)
ubOOB<-aOOB[195]
lbOOB<-aOOB[5]
int.aOOB<-c(lbOOB, ubOOB)
print(int.aOOB)

m<-c(1:6)
class<-c(1:200)
intup<-c(1:length(m))
intdown<-c(1:length(m))

for (i in m){
class<-dia[i]
class2<-sort(class[[1]])
intup[i]<-class2[195]
intdown[i]<-class2[5]
}

intervals<-data.frame(rbind(intup, intdown))
names(intervals)<-nameSp

#Results
sum(ctB)
summary(testing$Species)
print(int.a)
print(intervals)
print(ctB)
SUPPORT VECTOR MACHINE MODEL

#parameters of the model
c0<-1000  #model cost
cr<-0  #cross validation value
ga<-0.45  #gamma

real.val<-DATA$Species

#bootstrap
n<-c(1:200)
a<-c(1:200)
b<-c(1:200)

nameSp<-c("CorAlg","EncrCoral","MassCoral","Microbialite","Sand","Silt")
dia<-data.frame(matrix(nrow=length(n),ncol=6))
names(dia)=nameSp

for(j in n) {#B<-DATA[sample(1:nrow(DATA),replace=TRUE),]
  splits <- splitdf(DATA, seed=593+j)
lapply(splits,nrow)
  training <- splits$trainset
  testing <- splits$testset

  #multiclass
  model.svm.boo<-svm(Species~.,data=training,type="C-classification",cost=c0,gamma=ga,kernel="radial",cross=cr)
pred.mctt<-predict(model.svm.boo,newdata=testing)

  #confusion matrix
  con.pairB<-table(testing$Species,pred.mctt)
b[j]<-sum(diag(prop.table(con.pairB)))
dia[j,]<-diag(prop.table(con.pairB, 1))
}

print(b)

b<-sort(b)
ubb<-b[195]
lbb<-b[5]
int.b<-c(lbb,ubb)
print(int.b)

m<-c(1:6)
class <- c(1:200)
intup <- c(1:length(m))
intdown <- c(1:length(m))

for (i in m){
    class <- dia[i]
    class2 <- sort(class[[1]])
    intup[i] <- class2[195]
    intdown[i] <- class2[5]
}

intervals <- data.frame(rbind(intup, intdown))
names(intervals) <- 'nameSp'

# Results
sum(con.pairB)
summary(testing$Species)
print(int.b)
print(intervals)
print(con.pairB)
% 1D diffusion model for chloride concentration in deep clays
% Measured FF and porosity
% Programmed for chloride concentration
% code units: m and yr

clear all; close all; clc;

%Loading files (example)
load Chl_U1371.txt %measured chloride concentration
File_Chl=Chl_U1371;
load U1371_FF.txt; %formation factor
coreFF=U1371_FF;
load U1371_Por.txt; %porosity
corepor=U1371_Por;
load sealevel4.txt %sea level height
sl=sealevel4;
load sealevel5.txt %sea level height
sl2=sealevel5;
%Chloride concentration measurements
m_Chl=(35/1000).*File_Chl(:,2);
m_z=File_Chl(:,1);
%Parameters
D=1e-9; %molecular diffusion coefficient (m^2/s)
d=0.5; %distance between nodes of depth
%Steps
dt=5e7; %per year 6e12; %time step 1e8s=3.1710 yr
pres_time=1700; %present time value
change=365*24*60*60; %change from yrs to seconds
%Formation Factor measurements
nzz=coreFF(:,1); %nodes for depth (adimensional)
zz=coreFF(:,2); %depths (m)
F=coreFF(:,3); %ff (adimensional)
numzz=nzz(end); %number of measurements/nodes in depth
zzmax=zz(end); %maximum depth
%Loading the porosity data
nzz_p=corepor(:,1); %nodes for depth for porosity(adimensional)
zz_p=corepor(:,2); %depths for porosity (m)
P=corepor(:,3); %Porosity (adimensional)
numzz_p=nzz_p(end); %number of measurements/nodes in depth of porosity
zzmax_p=zz_p(end); %maximum depth for porosity

%max depth model
zmax_ex=max(zzmax,zzmax_p); %extended depth (m)

%Salinity values for boundary condition over time
yy=sl(:,2)*change; %Initially, time yr now changed into seconds
xx=sl(:,3); %Cl values
yr_max=yy(end); %Max time value
yr_min=yy(1);
y=[yr_min:dt:yr_max]; %Time vector for Cl boundary condition
chl=interp1(yy,xx,y); %Cl boundary condition for top

%Salinity values for boundary condition over time
yy2=sl2(:,2)*change; %Initially, time yr now changed into seconds
%(;1) negative years values
xx2=sl2(:,3); %Cl values
yr2_max=yy2(end); %Max time value
yr2_min=yy2(1);
y2=[yr2_min:dt:yr2_max]; %Time vector for Cl boundary condition
chl2=interp1(yy2,xx2,y2); %Cl boundary condition for top

%values of Chloride concentration for the BC
BC=mean(chl);

%time nodes
nt=length(chl); % number of time steps

%Smoothing the data fastsmooth (Y,w,type,ends)
%w=5, wide of the smooth; type=3, pseudo-Gaussian
%ends=1, the ends are smoothed with progressively smaller smooths the closer to the end.
sFF=fastsmooth(F,5,3,1); %smooth formation factor
sP=fastsmooth(P,5,3,1); %smooth porosity

%Interpolation of the data formation factor
z=[0:d:zzmax]; %array of depth for sampling each d separation
FF=interp1(zz,sFF,z); %FF interpolated
FF(1)=FF(2);

%Interpolation of the data porosity
phi=interp1(zz_p,sP,z); \%Porosity interpolated
phi(1:2)=phi(3);

\%Adkins & Schrag modeled tortuosity
mFF=(1-2.*log(phi))./phi; \%modeled formation factor

\% Phosphate matrix
Chl=BC*ones(numz,nt); \%matrix of zeros (nodes in z, time nodes)

curve=zeros(numz);
curve(20:40)=0.55;
BCd=19.90;

\% %Initial conditions
Chl(:,1)=(BC); \%Initial values for chlorinity for time step one

\%change initial condition in the sediment column
\%Chl(1:90,:)=BCd; \%step function initial condition
\%figure(6)
\%plot(Chl(:,15500))

IF=1./FF; \%inverse of the formation factors (adimensional)
mIF=1./mFF; \%inverse of the modeled formation factors (adimensional)

\%Loop to calculate the chlorinity at different time steps
lambda=(D*dt); \%lambda=1000;

rp=3:nnz;
rc=2:nnz-1;
rm=1:nnz-2;
B=zeros(nnz,1)';
oldB=zeros(nnz,1)';

for t=1:nt-1; \%loop over time

\% time derivative
B(rc)=((lambda./(phi(rc).*FF(rc).*d^2))).*(Chl(rp,t)'-2*Chl(rc,t)'+Chl(rm,t)')+
(lambda./(phi(rc).*d^2)).*(Chl(rp,t)'-Chl(rm,t)').*(IF(rp)-IF(rc))/2;

a(rc) = ((lambda./(phi(rc).*FF(rc).*d^2)))
- (lambda./(phi(rc).*d^2)).*(IF(rp)-IF(rc))/2;
b(rc) = -2*((lambda./(phi(rc).*FF(rc).*d^2)));
c(rc) = ((lambda./(phi(rc).*FF(rc).*d^2)))
+ (lambda./(phi(rc).*d^2)).*(IF(rp)-IF(rc))/2;

end
% B(rc)=TDMAsolver();
% time integration
if (t==1)
    Chl(rc,t+1) = Chl(rc,t) + B(rc)';
else
    Chl(rc,t+1)=Chl(rc,t)+1.5*B(rc)'-0.5*oldB(rc)';
end
oldB=B;
% boundary conditions
Chl(1,t+1)=chl(t+1);  %Top boundary condition
Chl(end,t+1)=(BC);  %Boundary condition infinite bottom
Chl(end,t+1)=chl2(t+1);  %BC as top BC with phase displacement 10000
Chl(end,t+1)=Chl(end-1,t+1);  %No-flux bottom BC
end

%Line graph chloride concentration changes over time
figure(1)
for tt=[1:500:nt-2]
    plot(Chl(:,tt),z)
    set(gca,'ydir','reverse')
    hold on
end
ylabel('Depth (m)')
xlabel ('Chloride concentration (g/kg)')
title('Chloride concentration diffusion for different times')

%Color graph with depth and time
figure(2)
time_axis=[-yr_max:dt:yr_min]/(1000*change);
imagesc(time_axis,z,Chl);
colorbar
title('Chloride concentration diffusion with time and depth')
ylabel('Depth (m)')
xlabel('Time (kyr)')

%Line graph last chloride concentration value
figure(3)
plot(Chl(1:end,end),z)
set(gca,'ydir','reverse')
hold on
plot(m_Chl,m_z,'*r')
ylabel('Depth (m)')
xlabel ('Chloride concentration (g/kg)')
title('Chloride concentration modeled and measured values')

%modeled formation factor
mB=zeros(numz,1)';
oldmB=zeros(numz,1)';
mChl=BC*ones(numz,nt); %matrix of zeros (nodes in z, time nodes)
mChl(:,1)=(BC); %Initial values for clorinity for time step one

for t=1:nt-1; %loop over time

% time derivative
mB(rc)=((lambda./(phi(rc).*mFF(rc)*d^2))).*(mChl(rp,t)'-2*mChl(rc,t)'+mChl(rm,t)')+...
       (lambda./(phi(rc)*d^2)).*(mChl(rp,t)'-mChl(rm,t)').*((mIF(rp)-mIF(rc))/2;
ma(rc) = ((lambda./(phi(rc).*mFF(rc)*d^2)))
       - (lambda./(phi(rc)*d^2)).*(mIF(rp)-mIF(rc))/2;
mb(rc) = -2*((lambda./(phi(rc).*mFF(rc)*d^2)));
mc(rc) = ((lambda./(phi(rc).*mFF(rc)*d^2)))
       + (lambda./(phi(rc)*d^2)).*(mIF(rp)-mIF(rc))/2;

% time integration
if (t==1)
   mChl(rc,t+1) = mChl(rc,t) + mB(rc)';
else
   mChl(rc,t+1) = mChl(rc,t)+1.5*mB(rc)'-0.5*oldmB(rc)';
end
oldmB=mB;
% boundary conditions
mChl(1,t+1)=chl(t+1); %Top boundary condition
mChl(end,t+1)=(BC); %Boundary condition infinite bottom
mChl(end,t+1)=chl(t+1); %BC as top BC
mChl(end,t+1)=mChl(end-1,t+1); %No-flux bottom BC
end

%Line graph last chloride concentration value
figure(4)
plot(Chl(1:end,end),z)
set(gca,'ydir','reverse')
hold on
plot(mChl(1:end,end),z,'g')
hold on
plot(m_Chl,m_z,*r*)
ylabel('Depth (m)')
xlabel ('Chloride concentration (g/kg)')
title('Chloride concentration modeled and measured values')
legend([Cl] modeled FF,[Cl] modeled Adkins & Schrag,measured [Cl],Location,'SouthEast')
# APPENDIX VI

## WOCE LINES AND LOCATIONS USED FOR MODERN SALINITY

(Chapter 4)

Table 1: WOCE lines and locations used for modern salinity.

<table>
<thead>
<tr>
<th>Site</th>
<th>Lat.</th>
<th>Long.</th>
<th>WOCE line</th>
<th>Lat.</th>
<th>Long.</th>
<th>Diff. long. (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQP10</td>
<td>20.683333</td>
<td>-143.3565</td>
<td>P16 150°W</td>
<td>~21</td>
<td>-150</td>
<td>~7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) North</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQP11</td>
<td>30.355333</td>
<td>-157.8705</td>
<td>P16 150°W</td>
<td>~30</td>
<td>-150</td>
<td>~7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) North</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1365</td>
<td>-23.850822</td>
<td>-165.644373</td>
<td>P15 165°W</td>
<td>~24</td>
<td>-165</td>
<td>~0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) South</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1370</td>
<td>-41.852148</td>
<td>-153.106332</td>
<td>P16 150°W</td>
<td>~42</td>
<td>-150</td>
<td>~3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) South</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1371</td>
<td>-45.964153</td>
<td>-163.184188</td>
<td>P15 165°W</td>
<td>~46</td>
<td>-165</td>
<td>~3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) South</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1225</td>
<td>2.770782</td>
<td>-110.571482</td>
<td>P18 105°W</td>
<td>~3</td>
<td>-105</td>
<td>~6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(500:1) North</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>