RECORDING EARTHQUAKES IN THE OCEANS
FOR GLOBAL SEISMIC TOMOGRAPHY
BY FREELY-DRIFTING ROBOTS

Joel David Simon

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
GEOSCIENCES

Adviser: Frederik J. Simons

June 2020
Abstract

This dissertation describes the treatment of data recorded by a novel instrument named MERMAID, short for Mobile Earthquake Recording in Marine Areas by Independent Divers. MERMAID is a passively-drifting float that autonomously records and transmits seismograms of teleseismic earthquakes from the global oceans. The primary goal of the instrument is to fill a seismic data gap. The primary goal of this dissertation is to describe how MERMAID data are analyzed and interpreted to achieve this end. We begin by developing a new method to pick multiscale arrival times of seismic phases in its noisy seismograms and further to assign uncertainty estimates and confidence intervals to those times. Our method is first applied to data collected in the Indian Ocean and the Mediterranean Sea. It is able to identify multiple phases that traveled different routes through the Earth with high accuracy and low uncertainty. Next we describe an ongoing deployment of 50 MERMAIDS in the South Pacific. There we study the seismicity rates yet reported and project for the total volume of data we expect returned over the lifetime of the deployment. We again use our method of arrival-time and uncertainty estimation to add to our ever-growing catalog of MERMAID travel-time residuals, and compute a similar catalog using seismograms from traditional seismometers installed on islands proximal to our floats. Their agreement proves that MERMAID records tomographically-useful data. We use these residuals to provide the first analyses of travel-time anomalies for the new ray paths sampling the mantle under the South Pacific. Finally, we end with a description of the open-source software developed over the course of this research and note that it may be used to reproduce the results of this dissertation.
Acknowledgments

From the jump I have to thank my best friend and fiancée Katie. You moved across the country for me and I am so glad you did. There is not much to say other than I love you. And of course my family, literally for everything. Again, I do not know what to say other than thank you, and I love you Mom, Dad, and Audrey.

I thank my adviser, Frederik Simons. Getting to this point was not an easy process for me. I have realized many times over that often my brain works in different ways from the brilliant people I have been surrounded by while at Princeton. But Frederik challenged—really pushed me—and he taught me so much. I came to Princeton having never opened a Linux terminal and I leave here having produced a suite of software to manage and analyze data collected by robots drifting around the remote South Pacific. With his guidance I can honestly say that I have produced my best work possible. Further, the support and help I have had from my committee and other professors has been essential. Jessica Irving, Guust Nolet, Jeroen Tromp, and Allan Rubin, thank you. I am honored to have written papers with two of you.

To Gerard Roe, my undergraduate adviser at the University of Washington, thank you for being my first scientific mentor and encouraging me to pursue a graduate degree. Your belief in me was inspiring.

Room 308A has seen a lot of graduate students come and go since I took root there many years ago. I have been very lucky to be around such a supportive and bright group of people, and if I have been a fraction as helpful to them as they have been to me, I have done well. I especially want to thank Wenbo, who was a great help and an even better friend. And just down the hall sat James, another great friend. I look back fondly on the afternoon runs the three of us would take on the towpath near campus.

The entire Geosciences family has been so supportive and helpful but I have to give special thanks in particular to Dawn, Mary Rose, and Sheryl. I hope you three know how much I appreciate
you. And of course I want to thank everyone on the Coprolites, the Geosciences summer-league softball team. Those summertime games provided me the most joy during my time here. I have to thank in particular Kaz, Rachel, and Chris for leading us to back-to-back championships.

And my cohort. We came in hot as a huge group and have been slowing trickling out the Guyot doors over the last few years. You were all great to me. Dani, Akshay, and Scott—you three already know what you mean to me, thank you. We will meet back at the Ivy soon.

Finally, I would be remiss if I did not thank my third grade teacher Mrs. Muller at Duniway Elementary in Portland, Oregon. She instilled in me the insatiable drive to always ask “why.” In doing so I quickly realized that humans, quite generally, do not know many things. So here I am, having achieved the highest academic degree one can, and I can confidently report that humans still do not know many things. However, with this dissertation I am pleased to deepen our collective wisdom, and going forward I know that I will continue to do so because I forever remain discontented with not understanding.
To nurses, everywhere.
## Contents

- Abstract ................................................................. iii
- Acknowledgments ......................................................... iv
- List of Figures ........................................................... x
- List of Tables ............................................................. xii
- Preface ................................................................. xiii

1 Introduction ........................................................... 1
   1.1 The Seismic Data Gap .............................................. 1
   1.2 One Solution: MERMAID ......................................... 4
   1.3 A Roadmap to This Dissertation .................... 5

2 Multiscale Estimation of Event Arrival Times and Their Uncertainties in Hydroacoustic Records from Autonomous Oceanic Floats 7
   2.1 Abstract .......................................................... 7
   2.2 Introduction ..................................................... 8
   2.3 General Considerations ......................................... 9
      2.3.1 A simple model of noise and signal .................... 9
      2.3.2 Maximum-likelihood changepoint estimation .......... 12
      2.3.3 Expected behavior ....................................... 16
      2.3.4 The summed log-likelihood .............................. 18
      2.3.5 Take-home message 1 .................................... 21
   2.4 The Akaike Information Criterion (AIC) ................. 22
      2.4.1 AIC-based changepoint estimation ................. 23
      2.4.2 Signal-to-noise ratio ................................... 26
   2.5 Formalizing the AIC Timing Uncertainty ............. 28
      2.5.1 Method I: Monte-Carlo resimulation .................. 29
      2.5.2 Method II: using the shape of the AIC curve ....... 32
      2.5.3 Comparison of Methods I and II ..................... 38
      2.5.4 Take-home message 2 .................................... 39
   2.6 Multiscale Analysis Methodology ............................. 40
      2.6.1 The discrete wavelet transform ...................... 40
      2.6.2 Multiscale AIC-based changepoint estimation ....... 42
      2.6.3 Timescale to time mapping: forward transform ....... 43
      2.6.4 Multiscale changepoints to multiscale arrival times 44
      2.6.5 Take-home message 3 .................................... 44
   2.7 Multiscale Analysis in Practice ...................... 45
2.7.1 Multiscale AIC-based changepoint estimators .......................... 46
2.7.2 Timing uncertainty of changepoint estimators ......................... 46
2.7.3 Multiscale AIC-changepoint timing uncertainty .................... 47
2.7.4 Coherence across scales ........................................... 49
2.7.5 From synthetics to actual data ..................................... 51
2.8 The Data Set ................................................................... 51
  2.8.1 The instrument ......................................................... 52
  2.8.2 The seismograms ..................................................... 53
  2.8.3 The initial events catalog ......................................... 53
2.9 Application to Mermaid Seismograms ..................................... 56
  2.9.1 Multiscale AIC measurements and their uncertainty ............ 56
  2.9.2 Computing travel-time residuals ................................ 60
  2.9.3 Distribution of travel-time residuals ............................. 63
  2.9.4 The updated events catalog ..................................... 66
2.10 Conclusion .................................................................... 68
2.11 Data and Resources ........................................................ 69
2.12 Acknowledgments ........................................................... 69
2.13 Supplemental Material ....................................................... 70
  2.13.1 Description ............................................................ 70
  2.13.2 Timescale to time mapping: inverse transform .................. 70
  2.13.3 Handling edges ....................................................... 71
    2.13.3.1 Edge sensitivity in the timescale domain ................. 71
    2.13.3.2 Edge sensitivity in the time domain ...................... 72
  2.13.4 Non-Gaussian models ............................................. 73
  2.13.5 A final data example ............................................. 73

3 Recording Earthquakes for Tomographic Imaging of the Mantle Beneath the South Pacific by Autonomous MERMAID Floats 76
  3.1 Abstract ...................................................................... 76
  3.2 Introduction and Motivation ............................................ 77
  3.3 The MERMAID Instrument .......................................... 82
  3.4 The EarthScope-Oceans Consortium ............................... 85
  3.5 The MERMAID SPPIM Deployment ............................... 86
    3.5.1 MERMAID drifts with the currents .......................... 87
  3.6 Matching MERMAID Seismograms to Earthquakes ............... 90
    3.6.1 Automated preliminary matching ............................ 91
    3.6.2 Manual winnowing and sorting .............................. 94
  3.7 The MERMAID Seismic Catalog .................................... 96
    3.7.1 Catalog summary: in pictures ................................. 96
    3.7.2 Catalog summary: by the numbers .......................... 101
    3.7.3 Catalog completeness and statistics .......................... 104
    3.7.4 Estimating the final size of the MERMAID catalog ....... 107
  3.8 Estimating First-Arrival Times and Their Uncertainties .......... 112
    3.8.1 The arrival-time pick ........................................... 113
    3.8.2 The travel-time residual ....................................... 113
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The seismic data gap.</td>
<td>2</td>
</tr>
<tr>
<td>2.1</td>
<td>Summed log-likelihoods: changepoint correct.</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Summed log-likelihoods: changepoint incorrect.</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>Summed log-likelihoods: changepoint early, correct, and late.</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Summed log-likelihoods: at every changepoint, in theory and in practice.</td>
<td>20</td>
</tr>
<tr>
<td>2.5</td>
<td>Changepoint detection using an Akaike information criterion (AIC).</td>
<td>26</td>
</tr>
<tr>
<td>2.6</td>
<td>AIC weights: low- and high-SNR.</td>
<td>27</td>
</tr>
<tr>
<td>2.7</td>
<td>Uncertainty estimation Method I: Monte-Carlo resimulation.</td>
<td>30</td>
</tr>
<tr>
<td>2.8</td>
<td>Method I applied to low-SNR time series.</td>
<td>31</td>
</tr>
<tr>
<td>2.9</td>
<td>Uncertainty estimation Method II: using the shape of the AIC curve.</td>
<td>35</td>
</tr>
<tr>
<td>2.10</td>
<td>Method II applied to low-SNR time series, plotted as a function of $\alpha$.</td>
<td>36</td>
</tr>
<tr>
<td>2.11</td>
<td>Method II applied to low-SNR time series, plotted as a function of time.</td>
<td>37</td>
</tr>
<tr>
<td>2.12</td>
<td>Multiscale changepoint estimate errors at various SNRs.</td>
<td>48</td>
</tr>
<tr>
<td>2.13</td>
<td>Ray paths of earthquakes recorded in the Indian Ocean and Mediterranean Sea.</td>
<td>55</td>
</tr>
<tr>
<td>2.14</td>
<td>Multiscale arrival-time estimation: low-SNR $P$ wave.</td>
<td>57</td>
</tr>
<tr>
<td>2.15</td>
<td>Multiscale arrival-time estimation: high-SNR $P$ wave.</td>
<td>59</td>
</tr>
<tr>
<td>2.16</td>
<td>Multiscale arrival-time estimation: multiple core phases.</td>
<td>61</td>
</tr>
<tr>
<td>2.17</td>
<td>Multiscale arrival-time estimation: multiple $P$ phases.</td>
<td>62</td>
</tr>
<tr>
<td>2.18</td>
<td>High-quality travel-time residuals in the public GéoAzur catalog.</td>
<td>64</td>
</tr>
<tr>
<td>2.19</td>
<td>Our updated catalog with multiscale picks and their associated uncertainties.</td>
<td>67</td>
</tr>
<tr>
<td>2.20</td>
<td>Changepoint detection using non-Gaussian time series: $t$-distribution.</td>
<td>74</td>
</tr>
<tr>
<td>2.21</td>
<td>Changepoint detection using non-Gaussian time series: $F$-distribution.</td>
<td>74</td>
</tr>
<tr>
<td>2.22</td>
<td>Multiscale arrival-time estimation: core and reflected phases.</td>
<td>75</td>
</tr>
<tr>
<td>3.1</td>
<td>Closing the seismic data gap.</td>
<td>78</td>
</tr>
<tr>
<td>3.2</td>
<td>The MERMAID instrument.</td>
<td>82</td>
</tr>
<tr>
<td>3.3</td>
<td>The MERMAID dive cycle.</td>
<td>83</td>
</tr>
<tr>
<td>3.4</td>
<td>MERMAID drift trajectories from deployment through 2019.</td>
<td>88</td>
</tr>
<tr>
<td>3.5</td>
<td>Record sections for $M_w$ 5 through $M_w$ 8 earthquakes.</td>
<td>90</td>
</tr>
<tr>
<td>3.6</td>
<td>MERMAID seismogram annotated with preliminary earthquake matches.</td>
<td>93</td>
</tr>
<tr>
<td>3.7</td>
<td>Distributions of earthquake magnitudes, epicentral distances, and their SNRs.</td>
<td>97</td>
</tr>
<tr>
<td>3.8</td>
<td>Ray paths for shallow, intermediate, and deep-focus earthquakes.</td>
<td>99</td>
</tr>
<tr>
<td>3.9</td>
<td>The MERMAID seismic catalog recorded in the South Pacific.</td>
<td>105</td>
</tr>
<tr>
<td>3.10</td>
<td>Travel-time residuals of first-arriving $p$ and $P$ waves</td>
<td>117</td>
</tr>
<tr>
<td>3.11</td>
<td>MERMAID record sections showing the arrival of $S$ waves.</td>
<td>119</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>3.12</td>
<td>MERMAID seismogram showing the arrival of surface waves.</td>
<td></td>
</tr>
<tr>
<td>3.13</td>
<td>MERMAID seismogram showing the arrival of $T$ waves.</td>
<td></td>
</tr>
<tr>
<td>3.14</td>
<td>MERMAID record section showing the arrival of core phases.</td>
<td></td>
</tr>
<tr>
<td>3.15</td>
<td>Core-phase travel-time residuals computed against $PKPbc$.</td>
<td></td>
</tr>
<tr>
<td>3.16</td>
<td>MERMAID seismogram with unidentified local events not recorded anywhere else.</td>
<td></td>
</tr>
<tr>
<td>3.17</td>
<td>MERMAID record sections including seismograms from nearby island stations.</td>
<td></td>
</tr>
<tr>
<td>3.18</td>
<td>Residuals recorded by traditional, MERMAID, and Raspberry Shake instruments.</td>
<td></td>
</tr>
<tr>
<td>3.19</td>
<td>Smeared travel-time residuals recorded by MERMAID in the South Pacific.</td>
<td></td>
</tr>
<tr>
<td>3.20</td>
<td>Residuals for common events: those recorded by all three instrument classes.</td>
<td></td>
</tr>
<tr>
<td>3.21</td>
<td>Highest-SNR common-event seismograms: traditional seismometers.</td>
<td></td>
</tr>
<tr>
<td>3.22</td>
<td>Highest-SNR common-event seismograms: MERMAID.</td>
<td></td>
</tr>
<tr>
<td>3.23</td>
<td>Highest-SNR common-event seismograms: Raspberry Shake.</td>
<td></td>
</tr>
<tr>
<td>3.24</td>
<td>SACPZ validation via displacement waveform comparison.</td>
<td></td>
</tr>
</tbody>
</table>
List of Tables

2.1 The approximate frequency sensitivity of wavelets. 54
2.2 Phases with associated high-quality residuals in the public GéoAzur catalog. 65

3.1 The MERMAID seismic catalog recorded in the South Pacific. 101
3.2 Global $M_4$–$4.9$ earthquakes, missed or reported by our MERMAIDS. 108
3.3 Global $M_5$–$5.9$ earthquakes, missed or reported by our MERMAIDS. 109
3.4 Global $M_6$–$6.9$ earthquakes, reported or missed by our MERMAIDS. 109
3.5 Global $M_7$–$7.9$ earthquakes, reported or missed by our MERMAIDS. 110
3.6 Global $M_8$–$8.9$ earthquakes, reported or missed by our MERMAIDS. 110
3.7 Nearby island seismic stations: data available from IRIS. 128
3.8 Nearby island seismic stations: data provided by Dr. Olivier Hyvernaud. 128
Preface

I am the first author of all components of this dissertation.

I am the lead author of all research detailed herein, the entirety of which is original work. At the outset I wish to highlight three chapters and their ongoing contribution to scientific knowledge.

Chapter 2 was published in 2020 in the journal Bulletin of the Seismological Society of America. The authorship of that study is Joel D. Simon, Frederik J. Simons, and Guust Nolet. As published, it spans 28 journal pages plus an additional four-page supplement, and it includes a catalog of data for use by other scientists that stretches 8,686 lines.

Chapter 3 is ready for publication. The authorship of that study is Joel D. Simon, Frederik J. Simons, and Jessica C.E. Irving. As typeset for submission it stands at 29 journal pages, and it includes an additional 11 page supplement. Similar to Chapter 2 it will be published with multiple complementary data-products. In total, nearly 10,000 lines of data to foster scientific discovery were generated in the course of that study.

Finally, Chapter 4 summarizes the software I have developed during my doctoral research. Every bit of scientific programming I have completed while at Princeton is published online and is freely and publicly accessible at github.com/joelsimon/omnia. My software package is built from neatly organized, reusable functions that are documented, version-controlled, and tested on various Linux and Macintosh architectures. They have already proven their utility to the broader scientific community, and they are currently in use by other researchers around the world. Finally, every byte of code used to generate every number, figure, data-table entry, etc. in the articles just mentioned, and this dissertation as a whole, is available there, so that these documents shall remain reproducible in perpetuity.
Chapter 1

Introduction

1.1 The Seismic Data Gap

More than 70% of Earth’s surface is covered by water. Seismic data recorded in the global oceans are sparse in both spatial and temporal coverage, especially in the Southern Hemisphere. Figure 1.1 proves this point by mapping, in blue, the location of every seismic station retrievable from the Incorporated Research Institutions for Seismology (IRIS). While the map is indubitably incomplete, and the recorded presence of a station does not imply that the data are also available, it illustrates the sparsity of seismic sampling in the oceans, especially in the Southern Hemisphere. This dissertation is about closing that seismic data gap with the help of a relatively new instrument named MERMAID for Mobile Earthquake Recording in Marine Areas by Independent Divers. MERMAID is a passively-drifting, oceangoing float that autonomously returns triggered hydroacoustic time series of teleseismic earthquakes recorded from within the water column. I show in this dissertation that MERMAID records tomographic-quality seismic data. Before describing the current MERMAID float, let us briefly review some of the important milestones in the relatively short history of seismic instrumentation of the oceans (only about one hundred seismic records from the deep-ocean bottom existed by the 1960s, according to Bradner, 1964) to place MERMAID in its proper historical perspective (see also Simons et al., 2009).

Historically, seismic studies in and of the oceans have proven complex and costly. Early attempts to instrument the oceans for regional and global seismology came in the form of en-
1.1. The Seismic Data Gap

Figure 1.1: All seismic stations (46295 small blue triangles) ever reported to the Incorporated Research Institutions for Seismology (IRIS). There is an egregious lack of instrumentation in the oceans, especially in the Southern Hemisphere.

cased seismometers dropped in free fall onto the seafloor from a ship, with or without anchored tether, and with a variety of mechanisms for recovery and data retrieval (Ewing & Vine, 1938; Bradner, 1964; Whitmarsh, 1970). Progress toward true instrument autonomy came in the form of freely-drifting telemetered devices, either neutrally-buoyant mid-column floating versions of ocean-bottom sensors (e.g. Bradner et al., 1970), or sonobuoys, with a hydrophone loosely suspended from a surface buoy (e.g. Reid et al., 1973). Most of these experiments were short-lived due to power restrictions. Longer-lived moored sonobuoys (e.g. Kebe, 1981) and moored hydrophones (e.g. Fox et al., 1993) could provide continuous hydroacoustic data at the expense of requiring seafloor cables to power them, restricting their spatial range of coverage.

In the last three decades, ocean bottom seismometry with long-life robust, three-component broadband sensors has positively flourished (Zhao et al., 1997; Webb, 1998; Webb & Crawford, 2003; Suetsugu & Shiobara, 2014). Nevertheless, to this day such instruments remain physically large and expensive to install (Beauduin et al., 1996; Collins et al., 2001), requiring a specialized research vessel both for deployment and recovery (Stephen et al., 2003), as establishing semi-
1.1. The Seismic Data Gap

permanent installations (e.g. Duennebier et al., 2002; Romanowicz et al., 2006) worldwide remains a developing goal for the international community (Montagner et al., 1998; Romanowicz & Giardini, 2001; Favali & Beranzoli, 2006).

Two ambitious multi-station, multi-instrument cabled arrays have been rooted on the seafloor off the coast of Japan (Hirata et al., 2002; Shinohara et al., 2014) and in the Canadian Northeast Pacific (Barnes et al., 2013; Matabos et al., 2016) for the long-term monitoring of subduction zones. These installations provide high-quality data with low latency, but they require massive upfront costs, demand costly maintenance, are limited by cables and, being permanent, cannot be rapidly reassigned in the case of developing seismic crises (e.g. Duennebier et al., 1997).

The current fleet of recoverable ocean bottom seismometers (OBS) is autonomous but unable to transmit data while deployed, hence data acquisition and processing are separated by months or years, unless catastrophe precludes recovery (Tolstoy et al., 2006). More recently, wave-powered gliders which float at the surface and may be remotely controlled to remain in the vicinity of an ocean-bottom station have been used as a go-between to relay data from seafloor to shore via acoustic modem and satellite uplink (Berger et al., 2016). This coupling of technologies allows the delivery of seismic data from the seafloor in near real-time. While they have shown promise, such solutions remain fragile and costly to operate, and they have not yet enjoyed large-scale deployment. Other solutions to the logistical problem of data recovery are currently being tested. These include ocean-bottom systems that periodically release data pods from the seafloor, each with a self-contained telecommunications unit to relay data via satellite upon surfacing (Hammond et al., 2019). Finally, while the age where the cables themselves may act as seismic sensors appears to have arrived (e.g. Sladen et al., 2019; Williams et al., 2019), such technology is in its infancy.

Despite those advances in technology, no single seismic instrument has solved all the issues just presented: the ability to deliver high-quality data with autonomy, low cost, low latency, and nimbleness. Nor should we assume that any single instrument can be designed to optimize for all. Our instrument, MERMAID, fills a gap in instrumentation by providing low-cost hydroacoustic records suitable for global seismology (Simons et al., 2006b) from the oceans in near real-time (Hello
et al., 2011) without the requirement of a research vessel for deployment and, being unrecovered, negating the need for a recovery cruise.

While MERMAID’s hydroacoustic time series, collected by a single limited-bandwidth hydrophone floating at mid-column water depths, forever will remain less “complete” seismic data sets in comparison with a well-coupled three-component broadband ocean-bottom seismometer, its benefits are its lower manufacturing costs, its logistical simplicity, and its algorithmic flexibility (Sukhovich et al., 2011, 2014) in selecting promising seismic phases to report with each surfacing. Hence, MERMAID can be thought of as a 21st century sonobuoy without the previous century’s drawbacks.

1.2 One Solution: MERMAID

The purpose of the MERMAID float is to return seismic data of tomographic quality from the global oceans in near real-time. The instrument and its dive cycle were inspired by oceanic floats (Swallow, 1955; Rossby & Webb, 1970; Davis et al., 1992, 2001), which have become ubiquitous in the global oceans (see Gould, 2005, for historical perspective). The international Argo program has been continuously providing the scientific community with a wealth of temperature, salinity, and trajectory data over the last several decades (Lavender et al., 2000; Roemmich et al., 2009; Davis, 2005; Abraham et al., 2013). Along with the payload required for in situ observations and hydrographic profiling, a contemporary APEX Argo float is equipped with a hydraulic pump which modulates an expandable bladder that allows it to be neutrally buoyant at many mid-column depths, a Global Positioning System (GPS) for location tracking, and a satellite link for data transmission.

Argo floats collect and transmit data over repeated dive cycles. A typical cycle begins with the float deflating its bladder to achieve negative buoyancy so that it may sink to a predetermined parking depth (generally between 1000 m to 2000 m below the sea surface), at which point it passively drifts at depth for a set amount of time (usually around 10 days), before finally reinflating its bladder to slowly rise back to the surface. During this ascent it samples and processes a roughly
vertical column of water via a conductivity-temperature-depth (CTD) sensor. Once at the surface it acquires a GPS fix, transmits the new data via satellite, and repeats the process. Because they are autonomous and drift at the whim of ocean currents, Argo floats are practically guaranteed to sample the water column at a previously unsampled location every time they ascend. As of 11 April 2020 there were 4060 Argo floats actively reporting from within every ocean on Earth, and on average some 800 are being deployed yearly to maintain the fleet and, like MERMAID, they are not designed to be recovered.

The first-generation MERMAID float was a modified Sounding Oceanographic Lagrangian Observer (SOLO) float (Davis et al., 2001), fitted with a hydrophone and a custom algorithmic processing unit so that it returned seismologically viable hydroacoustic data recorded at its parking depth (Simons et al., 2006b, 2009). The second-generation MERMAID (Hello et al., 2011; Sukhovich et al., 2015) was a modified APEX float built by Teledyne Webb Research. The third-generation MERMAID is a redesign from the ground up by Yann Hello at GéoAzur and French engineering firm OSEAN SAS. It is an autonomous robotic float consisting of a High Tech HTI-96-MIN_HEX hydrophone, a Gardner DENVER pneumatic pump, a Garmin GPS 15 unit, a Motorola 9522 two-way Iridium communication module, Electrochem lithium batteries, and dedicated onboard detection and discrimination software (Sukhovich et al., 2011). Once deployed MERMAID sinks to a predetermined depth (usually 1500 m) and records the ambient acoustic wavefield while freely drifting with the mid-column currents. If triggered by seismic activity, or once a threshold time is reached, MERMAID surfaces, transmits the new data, downloads mission-command files via satellite, and repeats the process.

1.3 A Roadmap to This Dissertation

This dissertation analyzes data returned by the second- and third-generation MERMAID floats. Chapter 2 describes a method we developed to pick multiscale seismic phase arrival times with high precision in their noisy hydroacoustic records. Further, it details two methods by which we
1.3. A Roadmap to This Dissertation

assign uncertainty estimates and confidence intervals to those arrival-time picks. There we apply our method to data that was recorded by a handful of second-generation MERMAIDS while deployed both in the Indian Ocean and Mediterranean Sea. Finally, we compare our observed arrival times against theory to compute a catalog of multiscale travel-time residuals. This catalog is the foundation for future tomographic studies, and our uncertainty estimates will inform the data weights during the inversion.

In chapter 3 we describe an ongoing large-scale deployment of 50 third-generation MERMAIDS into the South Pacific as part of the South Pacific Plume Imaging and Modeling (SPPIM) project (see section 3.5). I oversaw the data management entailing retrieval, processing, analysis, interpretation, etc., of 16 of those floats, the others being operated by our international EarthScope-Oceans collaborators (see section 3.4). In that chapter we expand our focus to explore other facets of MERMAID’s utility. For example, we report seismicity rates recorded by MERMAID to answer questions regarding the speed at which MERMAID is closing the seismic data gap. There we also apply our arrival-time picker and uncertainty estimator of chapter 2 both to MERMAID’s hydroacoustic records and seismograms recorded by other seismic instruments on islands in the general neighborhood of the SPPIM deployment. We show that data recorded by MERMAID and traditional seismometers consistently agree, proving that MERMAID records high-quality, tomographically useful data. Similar to chapter 2, chapter 3 resulted in multiple large catalogs of travel-time residuals and their associated uncertainties, both for MERMAID and other seismic stations. These catalogs are ready to be slotted into tomographic inversions as-is because our uncertainties provide a relative measure of the quality of every residual, irrespective of instrument type.

Both chapter 2 and chapter 3 resulted in the development of many pieces of software to automate the processes described there, from matching MERMAID seismograms with earthquakes, to picking the arrival times of various seismic phases, and retrieving complementary data from nearby island seismic stations. Those codes are explained in chapter 4. They are all freely and publicly accessible at github.com/joelsimon/omnia and enjoy use by our EarthScope-Oceans collaborators around the world.
Chapter 2

Multiscale Estimation of Event Arrival Times and Their Uncertainties in Hydroacoustic Records from Autonomous Oceanic Floats

2.1 Abstract

We describe an algorithm to pick event onsets in noisy records, characterize their error distributions, and derive confidence intervals on their timing. Our method is based on an Akaike information criterion (AIC) that identifies the partition of a time series into a noise and a signal segment which maximizes the signal-to-noise ratio. The distinctive feature of our approach lies in the timing uncertainty analysis, and in its application in the time domain and in the wavelet timescale domain. Our novel data are records collected by freely floating Mobile Earthquake in Marine Areas by Independent Divers (MERMAID) instruments, mid-column hydrophones that report triggered segments of ocean-acoustic time series.


2.2 Introduction

We wish to pick multiscale seismic arrival times and estimate their uncertainties in noisy hydroacoustic records. Our approach to onset detection centers on the evaluation of an Akaike information criterion (AIC) function at multiple scales of a wavelet-decomposed time series. In the most general sense our procedure identifies the changepoint in the time series where the discrepancy between the segments to the left (the “noise”) and the right (the “signal”) is maximized. We present two changepoint estimates derived from the AIC function, two methods to estimate their uncertainties, and investigate their application in both the time and timescale domains.

Laying out preliminaries and developing a simple signal model in section 2.3, we define the Akaike information criterion in section 2.4. Next we formalize a general scheme for uncertainty-analysis in section 2.5. In section 2.6 we extend our procedures to the multiresolution domain via wavelet decomposition of the time series. There, we detail the mapping between the time and timescale domains and apply it in section 2.7 to restate our general problem of changepoint estimation in both domains. By section 2.8 we have exhausted the theory and introduce the MERMAID instrument and its data in preparation to apply our method in section 2.9. There we analyze 445 publicly available MERMAID seismograms, and use the algorithm we have developed to identify arrival times and estimate their uncertainties. We use a public catalog of associated earthquakes to compute their travel-time residuals with respect to the ak135 velocity model. We summarize the statistics of the multiscale residuals in section 2.9.3, and introduce our updated events catalog in section 2.9.4, with all identified seismic phases, travel-time residuals, and timing uncertainties in the data set. In the supplemental material of this chapter we conclude with a remark on the general utility of our changepoint and uncertainty estimation procedures beyond the scope of this study.

Further, the catalog available as supplemental data to this dissertation provides the foundation for future tomographic studies that will use MERMAID travel-time residuals. Our uncertainty estimates will serve as measures for data weighting in the inversion. The software developed for this work is freely and publicly available, and is ready for application to diverse problems in time-series analysis.
2.3 General Considerations

We begin by posing the task of identifying the signal from the noise as a problem of statistical inference via likelihood analysis.

2.3.1 A simple model of noise and signal

To start with a very simple description, we model the seismogram, \( x(k) \), where \( k = [1, \ldots, N] \), as the concatenation of two separate and distinct segments joined at sample index \( k_o \). We label the first segment noise, \( n(k) \), and the second segment signal, \( s(k) \):

\[
x(k) = \begin{cases} 
n(k), & \text{if } 1 \leq k \leq k_o, \\
s(k), & \text{if } k_o + 1 \leq k \leq N.
\end{cases}
\]  

A true seismogram, of course, contains some level of noise throughout the interval under consideration, although, as a starting point to describe our automatic arrival identifier, this simplified model suffices. Our goal is to find \( k_o \), the changepoint, which “best” separates the noise segment from the signal segment.

We further assume that both the noise and the signal segments are samples from Gaussian (normal) parent distributions with different population parameters. Individually, they contain independently and identically distributed (i.i.d.) data, and they are mutually uncorrelated. Aware of the oversimplification, we stick to it for the time being. We formulate the problem of finding \( k_o \) in the context of maximum-likelihood estimation (MLE), specifically finding the sample index at which the time series is split into two segments that are most likely individually i.i.d.

Labeling the individual sample indices in the time series as \( x_i \), with \( i = [1, \ldots, N] \), in our description the probability density of any such individual point is

\[
f(x_i; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ \frac{-(x_i - \mu)^2}{2\sigma^2} \right\},
\]  

(2.2)
2.3. General Considerations

whose expectation, $\mu$, and variance, $\sigma^2$ will be identified as $\mu_1$, $\mu_2$, and $\sigma_1^2$, $\sigma_2^2$, depending on whether $i \leq k_o$ or $i > k_o$, respectively. The population parameter sets are $\theta_1 = \{\mu_1, \sigma_1^2\}$ (for the noise) and $\theta_2 = \{\mu_2, \sigma_2^2\}$ (for the signal).

The likelihood of an initial portion of the time series $x_1, \ldots, x_k$, where $k = [1, \ldots, N]$, being derived from the noise parent distribution is

$$
L_1(\mu_1, \sigma_1^2; x_1, \ldots, x_k) = \prod_{i=1}^{k} f(x_i; \mu_1, \sigma_1^2)
$$

$$
= (2\pi \sigma_1^2)^{-\frac{k}{2}} \exp \left[ -\frac{1}{2\sigma_1^2} \sum_{i=1}^{k} (x_i - \mu_1)^2 \right],
$$

(2.3)

and the log-likelihood function is

$$
\ell_1(\theta_1; x_1, \ldots, x_k) = -\frac{1}{2} \left[ k \ln(2\pi) + k \ln(\sigma_1^2) + \frac{1}{\sigma_1^2} \sum_{i=1}^{k} (x_i - \mu_1)^2 \right].
$$

(2.4)

The maximum-likelihood estimate (MLE), $\hat{\theta}_1 = \{\hat{\mu}_1, \hat{\sigma}_1^2\}$, solves

$$
\frac{\partial \ell_1(\hat{\theta}_1; x_1, \ldots, x_k)}{\partial \theta_1} = 0,
$$

(2.5)

which takes the well-known form

$$
\hat{\mu}_1 = \frac{1}{k} \sum_{i=1}^{k} x_i,
$$

(2.6)

$$
\hat{\sigma}_1^2 = \frac{1}{k} \sum_{i=1}^{k} (x_i - \hat{\mu}_1)^2.
$$

(2.7)

Substituting the expressions (2.6) and (2.7) into equation (2.4) yields

$$
\ell_1(\hat{\theta}_1; x_1, \ldots, x_k) = -\frac{k}{2} \left[ \ln(2\pi) + \ln(\hat{\sigma}_1^2) + 1 \right].
$$

(2.8)
Similarly, the likelihood of the remaining portion of the time series \( x_{k+1}, \ldots, x_N \), where \( k = [1, \ldots, N] \), being drawn from the signal parent distribution is

\[
\mathcal{L}_2(\mu_2, \sigma_2^2; x_{k+1}, \ldots, x_N) = \prod_{i=k+1}^{N} f(x_i; \mu_2, \sigma_2^2) = (2\pi\sigma_2^2)^{-\frac{(N-k)}{2}} \exp \left[ -\frac{\sum_{i=k+1}^{N} (x_i - \mu_2)^2}{2\sigma_2^2} \right],
\]

and its log-likelihood function is

\[
\ell_2(\theta_2; x_{k+1}, \ldots, x_N) = -\frac{1}{2} \left[ (N - k) \ln(2\pi) + (N - k) \ln(\sigma_2^2) + \frac{1}{\sigma_2^2} \sum_{i=k+1}^{N} (x_i - \mu_2)^2 \right].
\]

The equivalents to equations (2.6) and (2.7) are the elements of \( \hat{\theta}_2 \):

\[
\hat{\mu}_2 = \frac{1}{(N - k)} \sum_{i=k+1}^{N} x_i,
\]

\[
\hat{\sigma}_2^2 = \frac{1}{(N - k)} \sum_{i=k+1}^{N} (x_i - \hat{\mu}_2)^2,
\]

and substitution of equations (2.11) and (2.12) into (2.10) yields

\[
\ell_2(\hat{\theta}_2; x_{k+1}, \ldots, x_N) = -\frac{(N - k)}{2} \left[ \ln(2\pi) + \ln(\hat{\sigma}_2^2) + 1 \right].
\]

The sum of the logarithmic likelihoods (2.4) and (2.10) is

\[
\ell(\Theta; x; k) = \ell_1(\theta_1; x_1, \ldots, x_k) + \ell_2(\theta_2; x_{k+1}, \ldots, x_N),
\]

denoting \( \Theta = \{\theta_1, \theta_2\} = \{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\} \). Likewise, we sum the evaluated log-likelihoods in
2.3. General Considerations

equations (2.8) and (2.13) to

\[
\ell(\hat{\Theta}; x; k) = \ell_1(\hat{\theta}_1; x_1, \ldots, x_k) + \ell_2(\hat{\theta}_2; x_{k+1}, \ldots, x_N) = -\frac{1}{2} \left[ k \ln(\hat{\sigma}_1^2) + (N - k) \ln(\hat{\sigma}_2^2) + C \right],
\]

(2.15)

where \(\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2\} = \{\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2\}\), and \(C = N[\ln(2\pi) + 1]\).

2.3.2 Maximum-likelihood changepoint estimation

In Figure 2.1 we inspect the behavior of the log-likelihoods (2.4), (2.10) and (2.14) evaluated at different parameter values \(\Theta = \{\theta_1, \theta_2\}\), for synthetically generated data sets \(x = \{x_1, \ldots, x_{k_0}; x_{k_0+1}, \ldots, x_N\}\), in other words, for \(x = \{n, s\}\) drawn from known Gaussian parent distributions \(N(\Theta_o)\) with specific true population parameters \(\Theta_o = \{\theta_{1_o}, \theta_{2_o}\}\). In this figure we hold \(k = k_o\) fixed and consider it to be known. Similarly, we fix the expectations to zero. Figure 2.1d contains one such realization, color-coded blue for the noise segment and red for the signal segment. In the same colors, Figure 2.1b,c show their log-likelihoods \(\ell_1(\theta_1; x_1, \ldots, x_{k_o})\) and \(\ell_2(\theta_2; x_{k_o+1}, \ldots, x_N)\), respectively, as a function of the normalized-variance parameters \(\sigma_1^2/\sigma_{1_o}^2\) and \(\sigma_2^2/\sigma_{2_o}^2\), i.e. normalized by the true variances. The maximum-likelihood estimates \(\hat{\sigma}_1^2\) and \(\hat{\sigma}_2^2\) and their log-likelihoods, \(\ell_1(\hat{\theta}_1; x_1, \ldots, x_{k_o})\) and \(\ell_2(\hat{\theta}_2; x_{k_o+1}, \ldots, x_N)\), respectively, are marked by black-filled circles, whose average over all the 25 trials in this example is rendered as a white-filled circle at the correct location on the abscissa axes, and offset from their average log-likelihood, marked by the horizontal line, by an an arbitrary amount for clarity. Whiskers extend two standard deviations in both directions on both axes. Figure 2.1a shows the summed log-likelihoods \(\ell(\Theta; x; k_o)\). We use normalized-variance-sum coordinates denoted by \(\sigma^2/\sigma_o^2\) such that the summed log-likelihood value at \(\sigma^2/\sigma_o^2 = 1\) in Figure 2.1a is the sum of the log-likelihood values of Figure 2.1b at \(\sigma_1^2/\sigma_{1_o}^2 = 1\) and Figure 2.1c at \(\sigma_2^2/\sigma_{2_o}^2 = 1\). Similar to Figure 2.1b and Figure 2.1c, the black-filled circles correspond to the sums of the maximum-likelihood estimates, the white-filled circle their average over the 25 trials in this experiment, and the horizontal black line the average
2.3. General Considerations

Figure 2.1: Maximum-likelihood estimation (MLE) of the sample variance of synthetic realizations of time series $x(k) = n(k) + s(k)$ that follow the simple model in equation (2.1), assuming uncorrelated Gaussian distributions with a different variance for each segment. Panel (a) contains 25 examples of the sum of the log-likelihoods $\ell = \ell_1 + \ell_2$ from equation (2.14), and panels (b) and (c) the log-likelihoods $\ell_1$ from equation (2.4), $\ell_2$ from equation (2.10), rendered in the color corresponding to the noise (blue) or signal segment (red) of the synthetic time series, one realization of which is displayed in panel (d). The true variances $\sigma_{1o}^2 = 1$ and $\sigma_{2o}^2 = 2$ serve as normalization on the abscissa axes. The population expectations are $\mu_{1o} = \mu_{2o} = 0$. The changepoint between the noise and signal segments occurs at $k_o = 500$, the half-way point in each of these length $N = 1000$ simulations (which turns the color for the summed log-likelihoods into magenta, the sum of an even mixture of blue and red). The black-filled circles mark the MLE of the variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, and their averages are shown by white-filled circles, offset vertically for clarity from their average log-likelihoods marked by the horizontal lines, with whiskers extending two standard deviations in both directions in both coordinates.
of those summed log-likelihoods. As expected, the maximum-likelihood variance estimates are on average very close to the truth; that is, given a range of parameter values to test (the normalized abscissa axes in Figure 2.1a–c), the MLE, $\hat{\Theta}$, of the population parameters, $\Theta$, converges to the truth $\Theta$.

In Figure 2.2 we investigate the effect on the likelihoods and their ratios of varying $k$, holding $k_0$ fixed as an unknown truth. We hold all the population parameters unchanged from the cases presented in Figure 2.1. What varies is the sample index where we split the time series into an assumed “noise” and an assumed “signal” segment. Figure 2.2e,f contain two realizations of our same process, but now we color blue the first segment, until the index $k$ marked by the dotted line, which we may consider to be “noise,” and we color red the second segment, which we may consider to be “signal.” Figure 2.2e is an example of where our identification is tardy, $k > k_0$, Figure 2.2f is an example of where our identification is early, $k < k_0$. As in Figure 2.1, the true noise segment contains sample indices $k = [1, \ldots, 500]$, and the signal $k = [501, \ldots, 1000]$. Corresponding to Figure 2.2e, and for 25 such identical experiments, we show the log-likelihoods of the misidentified “noise” segments as the blue curves in Figure 2.2b, and, corresponding to Figure 2.2f, we show the log-likelihoods of the misidentified “signal” segments as the red curves in Figure 2.2c. We are not showing the log-likelihoods of the corresponding red signal segments in Figure 2.2e, nor of the blue noise segments in Figure 2.2f, since those do not consist of wrongly mixed models, and their shapes are similar to those plotted in Figure 2.1c,b.

The black-filled circles on each log-likelihood curve in panels Figure 2.2a–d mark the MLEs of the variances and their corresponding log-likelihoods. The horizontal black line marks the average likelihood value of the MLEs of the variances for the 25 trials shown, while the white-filled circle, arbitrarily offset on the ordinate axis, marks the average MLE, with whiskers extending two standard deviations in both directions on both axes.

Figure 2.2a contains the summed log-likelihoods of the data segmentation of Figure 2.2e, combining the log-likelihoods shown in Figure 2.2b with those of the trailing segment (not shown for the reason stated above), and again rendered in the appropriately mixed colors. Figure 2.2d con-
2.3. General Considerations

Figure 2.2: Maximum-likelihood estimation (MLE) of the sample variance of synthetic time series that follow the same simple model as in Figure 2.1, but for the case of an unknown changepoint. Two realizations are shown, and the cases illustrated have $k_0 = 500$, marked by solid vertical lines, but pertain to a tardy ($k = 750$), in panel (e), and, in panel (f), an early ($k = 250$) assumed changepoint, marked by the dotted vertical lines. The log-likelihood that applies to the blue section (which includes the full noise segment and a portion of the signal) in panel (e) is shown in panel (b), and constitutes one term in the sum, the first term in equation (2.14), shown in panel (a). The log-likelihood of the corresponding red section (which only includes signal) of panel (e), the second term in the sum in equation (2.14), is not shown separately. Similarly, the log-likelihood of the red section (some of the noise segment and all of the signal) of panel (f) is shown in panel (c). Again, the log-likelihood of the corresponding blue section (containing only noise) of panel (f) is not shown. Panel (a) plots the summed log-likelihoods for the time series in panel (e), with the color revealing the relative amount of signal mixed in with “noise.” Panel (d) shows the equivalent summed log-likelihoods for the time series in panel (f), with the color betraying the relative amount of noise mixed in with the “signal.” As in Figure 2.1, MLEs are marked by black-filled circles and their average by white-filled circles with whiskers extending two standard deviations in both directions in both coordinates, offset vertically for clarity from the black horizontal lines that mark the average log-likelihoods of the MLEs.
tains the summed log-likelihoods of the data segmentation plotted in Figure 2.2f, combining the log-likelihoods of Figure 2.2c with those of the leading segment (not shown).

As in Figure 2.1, each log-likelihood curve is plotted on an abscissa axis that is normalized with respect to the true population parameters. For Figure 2.2a,d, as for Figure 2.1a, such an axis normalization amounts to summing along a 1:1 diagonal section through a two-dimensional likelihood surface parameterized in those normalized coordinates. Such a construct is theoretical: in section 2.3.3 we discuss its expected behavior, and in section 2.3.4 how to sum the likelihoods appropriately in a way that is diagnostic for model identification.

2.3.3 Expected behavior

When the estimated changepoint is late, \( k > k_0 \), the expectation of the sample variance of the first mixed segment of \( x \) (the blue “noise” portion in Figure 2.2e), is

\[
E\left[ \hat{\sigma}_1^2 \bigg| k > k_0 \right] = \frac{1}{k} \left[ k_0 \sigma_{1_0}^2 + (k - k_0) \sigma_{2_0}^2 \right]. \tag{2.16}
\]

The expectation of the sample variance of the remaining abbreviated signal segment is the trivial identity

\[
E[\hat{\sigma}_2^2 | k > k_0] = \sigma_{2_0}^2. \tag{2.17}
\]

When \( k > k_0 \) the expectation of the sample variance of the mixed “noise” segment, normalized in terms of the population variance of the true noise segment, is

\[
E\left[ \frac{\hat{\sigma}_1^2}{\sigma_{1_0}^2} \bigg| k > k_0 \right] = \frac{1}{k} \left[ k_0 + (k - k_0) \frac{\sigma_{2_0}^2}{\sigma_{1_0}^2} \right], \tag{2.18}
\]

which evaluates to 1.333 in the case shown in Figure 2.2b, where \( \sigma_{1_0}^2 = 1, \sigma_{2_0}^2 = 2 \), and \( k = 750 \). The mean value over 25 tests, 1.324, is marked by the white-filled circle in Figure 2.2b, implying suitable convergence. The corresponding log-likelihood at this incorrect candidate changepoint is given by equation (2.8), and is approximately equal to \(-1172\), which is close to the value.
marked by the horizontal line in Figure 2.2b. The expectation of the sample variance of the remaining abbreviated signal segment (the red signal portion in Figure 2.2e, whose likelihoods are not shown) is uninteresting because it simply equals the true population variance (in which $\sigma^2 = 2$ and $E[\hat{\sigma}^2 / \sigma^2_{2_o}]_{k>k_o} = 1$). Its corresponding log-likelihood is given by equation (2.13), and is approximately equal to $-441$.

Similarly, when the estimated changepoint is early, $k < k_o$, the expectation of the sample variance of the mixed second segment of $x$ (the red “signal” portion in Figure 2.2f), is

$$E[\hat{\sigma}^2_{2}]_{k<k_o} = \frac{1}{(N-k)} \left[ (k_o - k)\sigma^2_{1_o} + (N - k_o)\sigma^2_{2_o} \right].$$

(2.19)

The expectation of the sample variance of the shortened noise segment is again trivial,

$$E[\hat{\sigma}^2_{1}]_{k<k_o} = \sigma^2_{1_o}.$$  

(2.20)

When $k < k_o$ the expectation of the sample variance of the mixed “signal” segment, normalized in terms of the population variance of the true signal segment, is given by

$$E \left[ \frac{\hat{\sigma}^2_{2}}{\sigma^2_{2_o}} \right]_{k<k_o} = \frac{1}{(N-k)} \left[ (k_o - k)\frac{\sigma^2_{1_o}}{\sigma^2_{2_o}} + (N - k_o) \right],$$

(2.21)

which evaluates to $0.833$ in the case shown in Figure 2.2c, where $\sigma^2_{1_o} = 1$, $\sigma^2_{2_o} = 2$, and $k = 250$. The mean value over 25 tests, 0.826, is marked by the white-filled circle in Figure 2.2c, implying suitable convergence as in Figure 2.2b. The log-likelihood associated with this changepoint is given by equation (2.13), and is approximately equal to $-1256$. Like the abbreviated signal section of Figure 2.2e, the abbreviated noise segment of Figure 2.2f (in blue, whose likelihoods are not shown) contains no improperly mixed sample indices and thus the expectation of its sample variance equals the corresponding true population variance ($\sigma^2_{1_o} = 1$; $E[\hat{\sigma}^2_{1}/\sigma^2_{1_o}]_{k<k_o} = 1$). The corresponding log-likelihood is given by equation (2.8) and is approximately equal to $-355$.

In the jointly normalized abscissa coordinates of Figures 2.1a, 2.2a, and 2.2d, the location of
the maximizer along the diagonal of the summed log-likelihoods lies at a linear mixture of the sample variances of the misidentified segments proportional to the length of the segments over which they are assumed to apply:

\[
\hat{\sigma}_2^2 = \frac{1}{N} \left[ \frac{1}{\sigma_1^2} k \hat{\sigma}_1^2 + \frac{1}{\sigma_2^2} (N - k) \hat{\sigma}_2^2 \right]. \tag{2.22}
\]

Substituting equations (2.16) and (2.17) into the generalized summed log-likelihood expression of equation (2.22) yields

\[
\mathbb{E} \left[ \frac{\hat{\sigma}_2^2}{\sigma_2^2} \right]_{k > k_0} = 1 + \frac{1}{N} \left[ (k - k_0) \left( \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) \right]. \tag{2.23}
\]

Substituting equations (2.19) and (2.20) into the generalized summed log-likelihood expression of equation (2.22) yields

\[
\mathbb{E} \left[ \frac{\hat{\sigma}_2^2}{\sigma_2^2} \right]_{k < k_0} = 1 + \frac{1}{N} \left[ (k_0 - k) \left( \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) \right]. \tag{2.24}
\]

In Figure 2.2a, \( k = 750 \) and equation (2.22) evaluates to 1.250. In Figure 2.2d, \( k = 250 \) and equation (2.22) evaluates to 0.875. These are indeed the expected linear sums of the sample variances, normalized proportionally to the true population variances, for both Figure 2.2e,f, to which the experiments converged, as shown by the white-filled circles in Figure 2.2a,d.

### 2.3.4 The summed log-likelihood

The summed log-likelihoods shown in Figure 2.1d and Figure 2.2a,d are diagonal profiles through a two-dimensional surface in two ratio-variables (they are normalized variances). No estimation method is expected to hug this diagonal unless \( \sigma_2^2 / \sigma_1^2 \), the true signal-to-noise ratio, is unity. Hence, a general method would need to construct a two-dimensional summed log-likelihood surface for a suitably large cross-product space of ratios in either variable, for a given candidate changepoint \( k \), find and record the maximum of the summed log-likelihood, and then repeat the process at every new candidate changepoint. See Figure 2.3 for a graphical illustration at three different candidate changepoints.
2.3. General Considerations

Figure 2.3: Summed log-likelihood plots in the two variables of interest for the cases where the changepoint estimate is early, correct, and late, respectively. In each panel, the left and lower axes are quoted in terms of the variances of the assumed “signal” and “noise” segmentations, one or both of which is composed entirely of the true signal or noise segment, normalized by their true variances, respectively; and the right and upper axes are in terms of the variances of the same segmentations not normalized by their true values. The color maps are identical and the crosses mark the average MLEs over the 1000 tests considered here. Figure 2.3a (changepoint early), Figure 2.3b (changepoint correct) and Figure 2.3c (changepoint late) correspond to the segmentations of \( x \) analyzed in Figure 2.2d, Figure 2.1a, and Figure 2.2a, respectively. What differs here is that the estimate of the variances maximizes the entire summed log-likelihood surface, not the version restricted to the 1:1 diagonal shown in the earlier figures. This two-dimensional procedure at each changepoint model \( k \) correctly identifies the normalized variances of the two segmentations, converging to the true values of \((1.000, 0.8333)\) in (a), at \((1.000, 1.000)\) in (b), and at \((1.333, 1.000)\) in (c). Further, the summed log-likelihood value, considering all three tests shown, is maximized when the changepoint is exactly correct, as will be shown in Figure 2.4.

A procedure properly diagnostic of the true changepoint \( k_o \) would recover both it and the variances of the noise and signal segments, \( \sigma^2_{1o} \) and \( \sigma^2_{2o} \). A profile through the three-dimensional likelihood volume (in the parameters \( k, \sigma^2_1, \) and \( \sigma^2_2 \)) does indeed peak at the correct triplet of true values \( k_o, \sigma^2_{1o} \) and \( \sigma^2_{2o} \). This behavior is borne out by Figure 2.4, which portends the success of a method that simply takes \( \hat{\Theta} \), containing the sample means \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) and variances \( \hat{\sigma}^2_1 \) and \( \hat{\sigma}^2_2 \), of the segmentation of the time series \( x \), calculates the summed log-likelihood \( \ell(\hat{\Theta}; x; k) \) according to equation (2.15), and defines the maximum-likelihood estimate for the changepoint as the argument that maximizes \( \ell(\hat{\Theta}; x; k) \) over \( k \).

Figure 2.4 plots those evaluated summed log-likelihoods at the sample indices \( k = [1, \ldots, N] \),
2.3. General Considerations

Figure 2.4: Summed log-likelihoods (equation 2.15) plotted at every changepoint model, $k = [1, \ldots, N]$. The colored curve is the expected behavior, when $\hat{\Theta} = \Theta_o$, with the color-coding based on the amount of incorrect mixing between the (blue) noise and (red) signal sections. The three likelihood values marked by the filled circles correspond to, from left to right, the mixtures shown in Figures 2.2d, 2.1a, and 2.2a, respectively. The gray curves plot 25 evaluations of equation (2.15) for synthetically generated time series like those in Figures 2.1d and 2.2e and 2.2f, and where $\hat{\Theta}$ is being estimated from those data. The black curve is the average of the 25 tests shown in gray.

both in expectation (colored curve, substituting $\hat{\Theta} = \Theta_o$), and for 25 realizations of $x$ (gray curves, when in general, $\hat{\Theta} \neq \Theta_o$) and their mean (black curve). As in Figures 2.1d and 2.2e and 2.2f, and for all tests here, $N = 1000$, $k_o = 500$, and $\Theta_o = \{\mu_{1o}, \sigma_{1o}^2, \mu_{2o}, \sigma_{2o}^2\} = \{0, 1, 0, 2\}$. The colored curve in Figure 2.4 uses a color gradient as in Figures 2.1 and 2.2, revealing the amount of incorrect mixing between the noise and signal segments at any given changepoint. Red dominates the first half of the curve, implying that at early changepoints the “noise” model includes too much signal; the second half grades to blue implying the reverse. At every changepoint, the sample variances of the noise and mixed “signal” segments ($k < k_o$, equations 2.20 and 2.19), or the mixed “noise” and signal segments ($k > k_o$, equations 2.16 and 2.17), are computed and substi-
tuted into equation (2.15). The filled circles highlight three sample indices, \( k = \{250, 500, 750\} \), and their respective summed log-likelihoods, \( \ell \approx \{-1611, -1592, -1613\} \). These correspond exactly to the maximum summed log-likelihood values obtained at the MLEs of variances in Figures 2.3a (changepoint early), 2.3b (changepoint correct), 2.3c (changepoint late), and to the values approached by the (slightly differently) summed log-likelihoods of the similarly-mixed time series underlying the curves of Figures 2.2d, 2.1a, and 2.2a, and drawn there as horizontal lines.

It is clear that, on average, the likelihood associated with the correct changepoint model, \( k = k_0 \), is greater than any of the incorrect changepoint models, \( k \neq k_0 \), wherein either the “noise” or “signal” segment contains samples from two distinct processes, and thus is not i.i.d. The gray curves in Figure 2.4 plot the summed log-likelihood for 25 synthetically generated time series (examples of which were shown in Figures 2.1d and 2.2e and 2.2f). Here, then, the sample variances of the “noise” and “signal” segments are given by equations (2.7) and (2.12), respectively, calculated at every changepoint and substituted into equation (2.15). In the limit of many trials the average of these likelihood curves converges to the theoretical values of the colored curve. Figure 2.4 implies this to be the case.

### 2.3.5 Take-home message 1

From Figures 2.1 and 2.2 we learn that the expectation of the sample variance converges to the true population variance. Thus we can drop the requirement of a priori knowledge of the true population parameters and instead rely on the statistics of many trials to converge to the truth. Further, we find that the ratio of the sample variances of both examples in Figure 2.2 is smaller than that in Figure 2.1, and indeed note the correct changepoint lies at the sample index at which the ratio of the sample variances is largest. Figures 2.1a, 2.2a, and 2.2d also show us that for three different changepoint models, the summed log-likelihoods are informative, specifically telegraphing the “identicality” or “i.i.d.-ness” of the modeled “noise” and “signal” segments from which the data were most likely generated. The highest summed log-likelihood identifies the split in the time series where the “noise” and “signal” segmentations are simultaneously individually best fit
by a single-variance, i.i.d. process. The average summed log-likelihoods marked by horizontal lines in Figure 2.2a,d, which represent two incorrect changepoint models, are each lower than in Figure 2.1a, the correct changepoint model.

Figure 2.3 and Figure 2.4 cement our understanding in showing that the highest summed log-likelihood among all tested changepoint models \( k = [1, \ldots, N] \) corresponds to the case when the changepoint model is exactly correct, \( k = k_0 \). Therefore, a relative comparison of summed log-likelihoods for every changepoint model intuitively defines a scheme for seismic arrival identification whereby the highest likelihood after testing all possible models identifies the sample index at which the seismogram is most likely split into two distinct and individually i.i.d. processes: for example, ambient noise and a seismic arrival. Our experiments show the ability of summed log-likelihoods, evaluated at the maximum-likelihood variance estimate, to partition an incoming data stream into two segments that are most likely individually identifiable as being from distinct generating distributions. In the illustrations of our model, the two portions, separated at the changepoint, specifically differ only in their variance. In practice, we also estimate the mean. For every changepoint model we assume that the first segment is composed of noise and the remainder, signal. Signal-to-noise considerations, of course, are to follow.

In preparation for defining an algorithm to apply to real data we hereafter drop the quotes around “noise” and “signal,” by which we denoted the departure of a segment from a known truth, because real data have no “true” changepoint against which to test. For the remainder of this study we will refer to every segmentation at every model changepoint as either noise or signal, without quotes.

2.4 The Akaike Information Criterion (AIC)

In his seminal 1973 paper (reprinted as Akaike, 1998), Akaike links maximum-likelihood estimation via evaluated likelihoods to information theory. Briefly, Akaike shows that the dual problem of parameter estimation and model testing can be solved simultaneously. At its core, the crite-
2.4. The Akaike Information Criterion (AIC)

Criterion that bears his name seeks to identify a best-fitting model from a set of candidate models via minimization of a loss function. The loss function used is an estimate of the Kullback & Leibler (1951) divergence (K-L), which we understand as a measure of the information separation between two probability distributions, i.e., the “distance” between “responses” due to an evaluated model and the truth. Importantly, when real data are being considered, their true generating distributions, and thus their K-L divergences, can never be known and must therefore be estimated. Akaike’s contribution was the rigorous derivation of such an estimated loss function and the proof that their relative comparisons, after accounting for model complexity via a penalty term, is useful and appropriate for model discrimination.

For our purposes, as we illustrated in the 2.3.4 section, these are the ideas that allow us to make the logical leap from equation (2.4) to equation (2.8), and, similarly, from equation (2.10) to equation (2.13), and to using the likelihoods evaluated in equation (2.15) to discriminate between changepoint models and choose a “best” fit among them. In the examples shown in there we had access to the true generating distributions, but we showed that estimates suffice in the absence of such knowledge.

2.4.1 AIC-based changepoint estimation

Using the summed log-likelihood of equation (2.15) we write the Akaike information criterion (AIC) as

\[ \mathcal{A}(\hat{\Theta}; x; k) = -2 \ell(\hat{\Theta}; x; k) + \|\hat{\Theta}\|, \]  

(2.25)

whereby \(\|\hat{\Theta}\|\), the length of the vector of model parameters, accounts for the degrees of freedom in the system. The first term on the right side of equation (2.25) is a measure of the model fit. The second term is a bias-correction term which rewards model parsimony. Likelihoods of the exponential form in our model (e.g., equations 2.4 and 2.10) are distributed as \(\chi^2/2\) variates, which explains the factor of two in equation (2.25). In our case, \(\|\hat{\Theta}\| = 4\), but we may ignore it altogether because it does not depend on the changepoint model \(k\). The distributions that we consider remain
unchanged between possible segmentations, and they furthermore remain in the specific Gaussian form of equation (2.2). Morita & Hamaguchi (1984), Maeda (1985), and Sleeman & van Eck (1999) discuss a framework by which other noise and signal models, e.g. autoregressive ones, can be decorrelated into satisfying our model assumptions.

The use of the Akaike information criterion for seismological event detection is widespread (e.g., Maeda, 1985; Sleeman & van Eck, 1999; Leonard & Kennett, 1999; Zhang et al., 2003; Rastin et al., 2013; Zhang et al., 2017). At its essence the AIC approach is sensitive to changes in variance (second moments) between trial segments, which it is able to neatly separate with high temporal resolution. So-called short-term average over long-term average (STA/LTA) methods, based on comparing first-moment ratios (e.g., Allen, 1978), are most efficient at picking out segments, not points, of interest. Methods that utilize higher-order (e.g., squared-envelope, skewness, kurtosis) statistics (e.g. Baer & Kradolfer, 1987; Saragiotis et al., 2002; Baillard et al., 2014) base their estimates on exploiting the changing nature of the distribution over segments of fixed length, which imprints a certain time and frequency resolution to the event identification. However powerful and performant any of these alternative approaches, our method has no tunable parameters, it obtains excellent results for our data types, and it remains usefully insensitive to perturbations in our initial model assumptions, which, additionally, we remotivate in the supplemental material.

Substituting the evaluated summed log-likelihoods of equation (2.15) into equation (2.25), ignoring the constant of the former and bias-correction term of the latter, and simplifying the notation to make the dependence of the AIC value on the changepoint model \( k \) explicit, we write, for an input time series \( x \) of length \( N \),

\[
\mathcal{A}(k) = k \ln(\hat{\sigma}_1^2) + (N - k) \ln(\hat{\sigma}_2^2).
\]  

(2.26)

The discussion in the 2.3.4 section implies that a natural changepoint estimate is the sample index that maximizes equation (2.15), or indeed, minimizes the AIC in equation (2.26), over the set of
2.4. The Akaike Information Criterion (AIC)

changepoint models \( k = [1, \ldots, N] \). We call this estimator

\[
k_m = \text{arg min} [A(k)].
\]

(2.27)

Furthermore, we showed there that the absolute values of summed log-likelihoods are irrelevant, and instead it is their relative difference which bolsters the designation of a “best” changepoint model. Following the discussion of Burnham & Anderson (2002), we introduce the AIC difference, the distance along the ordinate axis between an AIC value obtained at a particular model \( k \) and the minimum AIC value of the ensemble,

\[
\Delta(k) = A(k) - A(k_m).
\]

(2.28)

A small AIC difference implies the model in question is relatively likely given the set of all models tested, and a large AIC difference provides reason to believe the opposite is true. We leverage the AIC differences as a tool for model discrimination, by interpreting the exponential form \( \exp[-\frac{1}{2} \Delta(k)] \) as a measure of the relative likelihood of the model \( k \) compared to the set. With Li et al. (2009), we define a second changepoint estimator, \( k_w \),

\[
k_w = \sum_{k=1}^{N} k w(k),
\]

(2.29)

using what are now commonly called Akaike weights,

\[
w(k) = \frac{\exp[-\frac{1}{2} \Delta(k)]}{\sum_{i=1}^{N} \exp[-\frac{1}{2} \Delta(i)]}.
\]

(2.30)

Figure 2.5 shows the difference between the estimators \( k_m \) of equation (2.27) and \( k_w \) of equation (2.29) for a synthetic time series where \( N = 1000 \), and \( k_o = 500 \), as derived from the AIC curve calculated using equation (2.26). In this example, \( k_m = 507 \), which is late compared to \( k_o \). In contrast, the second changepoint estimator is early, \( k_w = 495 \), preceding \( k_o \).
2.4. The Akaike Information Criterion (AIC)

Figure 2.5: A model time series, its associated AIC curve, and the changepoint estimators $k_m$ and $k_w$. (a) Synthetic time series, $x(k)$, as in Figures 2.1 and 2.2, $N = 1000$, with a true changepoint, $k_o$, at sample index 500, marked by a black vertical line. The noise is drawn from $N(\mu = 0, \sigma^2 = 1)$, and the signal from $N(\mu = 0, \sigma^2 = 2)$. (b) The AIC curve associated to the time series shown in (a), calculated from equation (2.26). Again, $k_o$ is marked in black, whereas the he estimator $k_m$ (equation 2.27) is marked in blue, and the estimator $k_w$ (equation 2.29) in red.

2.4.2 Signal-to-noise ratio

We estimate the signal-to-noise ratio (SNR) from the ratio of sample variances of the segments identified as signal and noise. For a particular changepoint index $k$, using equations (2.7) and (2.12), we define

$$\text{SNR} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}. \quad (2.31)$$

A time series characterized by a high-SNR will be revealed by a steep AIC curve that rapidly and almost surely monotonically decreases, and then rapidly and virtually monotonically increases, after reaching an easily identified, single, global minimum, and finally, flattening asymptotically. The Akaike weights (equation 2.30) are near zero everywhere, except within a small sample span about the true changepoint. In these cases $k_m$ and $k_w$ will generally coincide. Conversely, a low-
2.4. The Akaike Information Criterion (AIC)

Figure 2.6: High-SNR and low-SNR synthetic time series, their AIC curves, their Akaike weights, and the changepoint estimates. As before, \( N = 1000 \) and \( k_o = 500 \) for both examples shown here, and \( k_o \) is marked by a black vertical line. (a) High-SNR synthetic time series (SNR = 25) in gray, and a low-SNR synthetic time series (SNR = 2) in black. (b) Their associated AIC curves (equation 2.26), with the label on the left ordinate axis corresponding to the low-SNR example and the label on the right ordinate axis to the high-SNR example. (c) Akaike weights (equation 2.30) associated with both examples, again with a double ordinate axis, on a zoomed-in abscissa axis to show detail about the true changepoint.

SNR time series will have a flatter AIC curve with multiple local minima and no obvious global minimum, and the associated weights will be more broadly spread about the true changepoint. In those cases \( k_w \) and \( k_w^* \) may differ greatly.

Figure 2.6 shows this behavior. In Figure 2.6a, we plot two synthetic times series: a low-SNR = 2 example in black, and a high-SNR = 25 example in gray. Figure 2.6b plots their associated AIC functions (equation 2.26), and Figure 2.6c plots their Akaike weights (equation 2.30). As in all previous examples \( N = 1000 \) and \( k_o = 500 \), and \( k_o \) is marked with a black vertical line in Figure 2.6a–c. Both Figure 2.6b,c have a double ordinate axis: the left corresponds to the low-SNR time series and the right to the high-SNR time series. Note their order-of-magnitude difference in range.
2.5 Formalizing the AIC Timing Uncertainty

In Figure 2.6b, the low-SNR black AIC curve has many local minima surrounding the global minimum, and is much flatter compared to the gray high-SNR curve, which steeply and nearly monotonically decreases to an obviously global minimum, and then rapidly increases to an asymptote. The effect of these minima on the Akaike weights is apparent from Figure 2.6c, which is rescaled between sample indices 460 and 540 to show detail about the true changepoint. Here, the weights in the high-SNR case are narrowly distributed and nearly centered on \( k_o \), while the low-SNR case has a broad, multi-modal distribution. In the latter case, \( k_w \) is early at \( k_w = 491 \) and is marked by the red vertical lines, while \( k_m \) is late at \( k_m = 504 \) and is marked by the blue vertical lines. The two changepoint estimators for the high-SNR case coincide with \( k_o \).

2.5 Formalizing the AIC Timing Uncertainty

The novelty of our work lies in the development of a statistical framework to estimate the timing uncertainty of the changepoint estimates \( k_m \) and \( k_w \). We present two distinct methods. Both calculate error statistics using many realizations of AIC curves whose associated time series were generated by random sampling from distributions with known statistics, but they work in different coordinates. Method I collects the error along the sample index (or equivalently, time) axis of the AIC curve (a distance), while Method II performs hypothesis tests using the AIC values themselves (as percentages). We discuss the relative utility of both methods to inform the uncertainty estimation and the assignment of confidence intervals of the changepoint estimates \( k_m \) and \( k_w \).

A baseline scenario to compare the bias and variance of the changepoint estimators \( k_m \) and \( k_w \) involves testing their performance using a time series built of noise and signal segments randomly sampled from distributions with known statistical parameters \( \Theta_o \), concatenated at a known true changepoint \( k_o \). As before, and for all tests in this section, we generate low-SNR synthetic time series of length \( N = 1000 \), where sample indices \( k = [1, \ldots, 500] \) are drawn from \( \mathcal{N}(\mu = 0, \sigma^2 = 1) \), and samples \( k = [501, \ldots, 1000] \) from \( \mathcal{N}(\mu = 0, \sigma^2 = 2) \). Here, then, the SNR is 2, and the true changepoint that we attempt to locate is sample index 500, the last sample index of the noise segment.
2.5. Formalizing the AIC Timing Uncertainty

2.5.1 Method I: Monte-Carlo resimulation

The Method I (M1) test procedure directly measures the timing error, the discrepancy between a true changepoint, \( k_0 \), and an estimated changepoint, \( \hat{k} \), either \( k_m \) or \( k_w \), calculated for a synthetic time series. We define the timing error in terms of sample indices as the signed distance

\[
\hat{k} - k_0.
\]  

To evaluate the bias and variance of \( k_m \) and \( k_w \) we generate many random realizations of the low-SNR time series model (SNR = 2) described in the previous paragraph, calculate the corresponding AIC curve via equation (2.26), calculate changepoint estimates \( k_m \) and \( k_w \) with equations (2.27) and (2.29), respectively, and then tally the timing errors between these estimates and \( k_0 \) with equation (2.32). We summarize the error statistics after many realizations of this procedure.

Figure 2.7 displays one realization of the M1 testing scheme applied to the same AIC curve in Figure 2.5b, but shown with a zoomed-in abscissa axis. Here, \( k_0 \) is marked with a black-filled circle and intersecting vertical line, and \( k_m \) and \( k_w \) are marked with blue and red-filled circles, respectively. The distance in sample indices between each changepoint estimate and the truth is marked by a similarly color-coded horizontal bar of the appropriate length. In this example, the minimum-AIC estimator, \( k_m = 507 \), is late, and the weighted-average estimator, \( k_w = 495 \), is early.

Figure 2.8 summarizes the M1 error statistics after one million realizations of the procedure just described. The sample-distance errors are grouped into one-integer bins and plotted as histograms. The unfilled bars with black edges represent the distribution of the error of the estimator \( k_m \), while the gray bars represent the error of the \( k_w \) estimator. Overlain are two curves, blue for \( k_m \) and red for \( k_w \), respectively, which represent their corresponding best-fitting Gaussian probability distribution functions (pdfs) given the statistics quoted in the legend and elaborated upon below. The abscissa axis is limited to \( \pm 50 \) sample indices for display purposes, but mass extends beyond these limits for both histograms.
2.5. Formalizing the AIC Timing Uncertainty

Figure 2.7: One realization of the Method I (M1) testing procedure, tallying the distance in sample indices between the true changepoint and its estimates, applied to the same AIC curve in Figure 2.5b (SNR = 2), shown in detail about the true changepoint, $k_0$, marked by a black-filled circle. The estimator $k_m$ (equation 2.27) is marked by a blue-filled circle, and $k_w$ (equation 2.29) by a red-filled circle. The timing error is the difference between the estimate and the truth (equation 2.32), shown here as blue and red horizontal lines that connect $k_m$ and $k_w$ to the vertical line at $k_0$. In this example, $k_m$ is 7 sample indices late, while $k_w$ is 5 sample indices early.

Inset into the upper right corner of the Figure 2.8 is a quantile-quantile plot which displays the sample-distance error quantiles as a function of the quantiles of their best-fitting Gaussian pdfs, again color-coded blue for $k_m$ and red for $k_w$. A black line with a slope equal to 1 is also plotted for reference.

Invoking Figure 2.8 we submit that $k_w$ is a better estimator than $k_m$. Neither of them are Gaussian. Our test reveals that even in extremely low-SNR regimes $k_w$ is an unbiased estimator, whereas $k_m$ is biased, here with a mean error of 3.8 sample indices. We also find that $k_w$ has a narrower error distribution with a standard deviation of 21 sample indices compared with 25 sample indices for $k_m$.

The story grows somewhat more complex when the data are more fully inspected. For example,
2.5. Formalizing the AIC Timing Uncertainty

The $k_m$ errors have their mode at 0 sample indices and their median at 2 sample indices, while the $k_w$ errors have their mode at $-3$ samples and their median at $-1$. The peak at 0 sample indices in the $k_m$ error histogram shows that this estimator is about twice as likely as the $k_w$ estimator to be exactly correct, however the lower standard deviation of the $k_w$ error data shows that, on average, $k_w$ will be closer to the truth. Even though any given $k_w$ estimate is less likely than a $k_m$ estimate to be exactly correct, we still consider $k_w$ to be a better estimator than $k_m$ because it is unbiased, when $k_m$ is not, and lower-variance than $k_m$.

A positive bias of the minimum-AIC estimator was also documented by Leonard (2000). There, the author compares the slope of an autoregressive variant of our AIC curve immediately following

Figure 2.8: Histograms of estimation errors of the estimators $k_m$ (black-unfilled bars) and $k_w$ (gray-filled bars), after one million realizations of the Method I (M1) testing procedure, one of which is displayed in Figure 2.7. Overlaid on both histograms are their best-fitting Gaussian probability distribution functions, blue for $k_m$ and red for $k_w$, respectively, and whose means and standard deviations are quoted in the legend. Inset in the upper right is a quantile-quantile plot, again color-coded. In this low-SNR model ($\text{SNR} = 2$), the $k_w$ estimator is unbiased and has a lower standard deviation than the $k_m$ estimator, which has a positive bias revealing that, on average, it estimates the changepoint late.
Formalizing the AIC Timing Uncertainty

$k_m$ with analyst picks of the onset to derive an empirical relationship to adjust the typically tardy estimate backward in time.

2.5.2 Method II: using the shape of the AIC curve

The Method II (M2) test procedure comprises two hypothesis tests which employ a proxy for the sharpness of the AIC curve near $\hat{k}$. We know from numerical experiments that in high-SNR cases the AIC curve rapidly and monotonically descends to a single and obvious global minimum at the true changepoint, then rapidly and monotonically ascends to an asymptote. Conversely, in low-SNR cases the AIC curve is much flatter, with a lower overall range, and many local minima that amplify changepoint estimation uncertainty. The M2 test procedure proposed here seeks to quantitatively relate the shape of the AIC curve itself to a timing-confidence interval. Each test in the M2 suite asks a variation of the basic question: how steep is the AIC curve around the changepoint estimator $\hat{k}$, and how much greater is the AIC value at the true changepoint $k_o$ compared to the AIC value at its estimate? Like the M1 test procedure, statistics are generated via repetition of the M2 test over many randomly generated synthetic time series. Unlike M1, the confidence intervals returned are asymmetric, not measures of center and spread. Altogether, M2 produces probability curves that relate the shape of the AIC curve near $\hat{k}$ to a confidence interval on the estimate.

We define $\alpha$ as a percentage of the total range of the AIC curve (equation 2.26), and $\beta$ as the sum of the $\alpha$th fraction of the range and the AIC value of the changepoint estimator, $\hat{k}$, either $k_m$ or $k_w$.

$$\beta(\hat{k}, \alpha) = A(\hat{k}) + \frac{\alpha}{100} \left( \max \left[ A(k) \right] - \min \left[ A(k) \right] \right). \quad (2.33)$$

With this we define our first “unrestricted” $\beta$ hypothesis test,

$$H_0: A(k_o) > \beta(\hat{k}, \alpha), \quad (2.34)$$

$$H_1: A(k_o) \leq \beta(\hat{k}, \alpha). \quad (2.35)$$

In the unrestricted $\beta$-test the null hypothesis is rejected if the AIC value of the true changepoint is
equal to or less than the AIC value of the estimated changepoint plus a variably defined percentage of the total range of the AIC curve over the segment considered. We label the first and last sample indices whose AIC values are less than or equal to \( \beta \) as

\[
k' = \min \{ k : A(k) \leq \beta(\hat{k}, \alpha) \}, \quad (2.36)
\]
\[
k'' = \max \{ k : A(k) \leq \beta(\hat{k}, \alpha) \}. \quad (2.37)
\]

Lastly, we note a mapping exists that relates \( \alpha \) to the corresponding maximum sample span to which this test could apply, and call it the

unrestricted \( \beta \)-test sample span \( = 1 + (k'' - k') \). \quad (2.38)

Herein lies the origin of our use of the term “unrestricted,” because the elements in the set \( \{A(k) : k = k', \ldots, k''\} \) need not all be less than or equal to \( \beta \). As written in equations (2.34) and (2.35), the unrestricted \( \beta \)-test is applied in the ordinate direction: defined in terms of AIC values, not sample indices (potential changepoint models), at which those AIC values are obtained. The definitions of \( k' \) and \( k'' \) were introduced in order to formalize the \( \alpha \)-to-sample-span map which is pivotal to the utility of M2.

Conversely, in defining our second hypothesis test, the “restricted” \( \beta \)-test, we begin by highlighting the relevant sample span of that test, which we term the

restricted \( \beta \)-test sample span \( = 1 + (k^{\dagger \dagger} - k^{\dagger}) \), \quad (2.39)

where, for the relevant values of \( \alpha \),

\[
k^{\dagger} = \max \{ k < \hat{k} : A(k) > \beta(\hat{k}, \alpha) \} + 1, \quad (2.40)
\]
\[
k^{\dagger \dagger} = \min \{ k > \hat{k} : A(k) > \beta(\hat{k}, \alpha) \} - 1. \quad (2.41)
\]
If it is nonempty, the elements in the set \( \{ A(k) : k = k^\dagger, \ldots, k^{\dagger\dagger} \} \) are all equal to or less than than \( \beta \). A restricted \( \beta \)-test asks if \( k_0 \) lies within this contiguous set,

\[
H_0: k_0 \notin [k^\dagger, \ldots, k^{\dagger\dagger}], \tag{2.42}
\]

\[
H_1: k_0 \in [k^\dagger, \ldots, k^{\dagger\dagger}]. \tag{2.43}
\]

In graphical terms, the restricted \( \beta \)-test asks: (1) is \( A(k_0) \) below \( \beta \), and if so, (2) is \( k_0 \) in the same trough of the AIC curve that contains \( \hat{k} \)? Unlike an unrestricted \( \beta \)-test, where the null hypothesis may be rejected even if the AIC values between \( A(\hat{k}) \) and \( A(k_0) \) are greater than \( \beta \), a restricted \( \beta \)-test does not allow a local maximum, where the AIC curve rises above \( \beta \), to be crossed in search of \( A(k_0) \). In both unrestricted and restricted \( \beta \)-tests, as \( \alpha \) increases, the sample span (number of changepoint models) under consideration increases and the null hypothesis is rejected more frequently. Given these definitions we expect that restricted \( \beta \)-tests applied to \( k_m \) reject the null hypothesis at the lowest rate because they consider the smallest sample spans, and that unrestricted \( \beta \)-tests applied to \( k_w \) would reject the null hypothesis at the highest rate because they consider the largest sample spans.

The utility of the M2 testing procedure is in the generation of probability graphs that measure the shape of the AIC curve by relating \( \alpha \) to sample-span confidence intervals that include the truth some proportion of the time. As in M1, these curves are generated by applying M2 to many realizations of synthetic time series. With each realization, both an unrestricted and restricted \( \beta \)-test is performed, and their test results (null hypothesis rejection rates) and associated sample spans (equations 2.38 and 2.39) are recorded.

Figure 2.9 shows one such M2 test realization applied to the same section of the same AIC curve of Figure 2.7 (SNR = 2). The AIC value of the true changepoint, \( A(k_0) \), is marked by a black-filled circle bisected by a black horizontal line. This represents the value that \( \beta \) must exceed to reject the null hypothesis in an unrestricted test, and the value at or below which both \( A(\hat{k}) \) and \( A(k_0) \) must lie, in the same trough, to reject the null hypothesis in a restricted \( \beta \)-test. A blue-filled
2.5. Formalizing the AIC Timing Uncertainty

Figure 2.9: One realization of the Method II (M2) hypothesis test, applied to the same AIC curve shown in Figure 2.7. Again, $k_o$, $k_m$ and $k_w$ are marked by black, blue, and red filled circles, and $A(k_o)$ by a black horizontal line. Three values of $\beta$ (equation 2.33), corresponding to three percentages of the total range of the AIC curve, $\alpha$, of 0%, 3%, and 6%, are shown in blue and red as related to $k_m$ and $k_w$, respectively. The solid blue and red lines represent the sample spans to which a restricted $\beta$-test maps (equation 2.39), while the dashed lines represent the additive sample spans through which an unrestricted $\beta$-test maps (equation 2.38). For a specific $\alpha$, a restricted $\beta$-test (equations 2.42 and 2.43) asks if $k_o$ falls in the interval below the colored solid line, while an unrestricted $\beta$-test (equations 2.34 and 2.35) asks whether $k_o$ falls under either the colored solid or dashed lines. The null hypothesis in both cases remains that $k_o$ is found outside those intervals. In this example for both restricted and unrestricted $\beta$-tests the null hypothesis is only rejected for $k_m$ when $\alpha = 6\%$, though it is rejected for all $\alpha$ related to $k_w$.

circle marks $k_m$ and a red-filled circle marks $k_w$. Three $\beta$ values for both estimators are shown as horizontal lines, using the same color scheme, at $\alpha$ values equal to 0%, 3%, and 6%. Solid horizontal lines represent the sample spans to which the restricted $\beta$-tests map (equation 2.39), and dashed horizontal lines represent the sample spans to which an unrestricted $\beta$-test maps (equation 2.38) in addition to those already considered in a restricted $\beta$-test.

For $k_m$ in Figure 2.9, $\alpha = 0\%$, $\alpha = 3\%$, and $\alpha = 6\%$ correspond to sample spans equal to 1, 4, and 11, respectively, for a restricted $\beta$-test, and 1, 21, 24, respectively, for an unrestricted
2.5. Formalizing the AIC Timing Uncertainty

Figure 2.10: Probability of rejecting $H_0$ as a function of $\alpha$ for restricted and unrestricted $\beta$-tests for both changepoint estimators $k_m$ and $k_w$. These curves summarize the output of M2 after many test realizations, one example of which is displayed in Figure 2.9. Here, restricted (equations 2.42 and 2.43) and unrestricted $\beta$-tests (equation 2.34 and 2.35) were each applied 1000 times at every $\alpha$. As in Figure 2.9, the test curves associated with $k_m$ and $k_w$ are color-coded blue and red, respectively, and the solid and dashed lines represent the outcomes of restricted and unrestricted $\beta$-tests, respectively. Overall, restricted $\beta$-tests corresponding to $k_m$ generally map to the shortest sample spans (equation 2.39) and thus reject the null hypothesis at the lowest rate, while unrestricted $\beta$-tests corresponding to $k_w$ generally have the highest rejection rate at the expense of mapping to the longest sample spans (equation 2.39). Figure 2.11 replots these results as the probability of rejecting $H_0$ as a function of the average sample span of each $\beta$-test.

$\beta$-test. The null hypothesis is not rejected at $\alpha = 0\%$ or $\alpha = 3\%$, for either the restricted or unrestricted $\beta$-test, though it is rejected at $\alpha = 6\%$ in both tests.

In contrast, for $k_w$ in Figure 2.9, $\alpha = 0\%$, $\alpha = 3\%$, and $\alpha = 6\%$ correspond to sample spans equal to 13, 33, and 46, respectively, for a restricted $\beta$-test, and 25, 33, 46, respectively for an unrestricted $\beta$-test. The null hypothesis is rejected at all $\alpha$ values for both restricted and unrestricted $\beta$-tests.

Figure 2.10 displays the results of 1000 realizations of M2 applied to synthetic time series of the
2.5. Formalizing the AIC Timing Uncertainty

Figure 2.11: Probability of rejecting \( H_0 \) as a function of the average sample span considered under each \( \beta \)-test. These curves render the same simulation results as Figure 2.10 except they are now plotted in terms of the average sample span of each \( \beta \)-test. Vertical lines tied to the curves reaching 0.68 and 0.95 on the ordinate axis, from left to right, define the corresponding confidence intervals. The 68% confidence interval is bounded on the left at 22 sample indices by the restricted \( \beta \)-test curve corresponding to \( k_m \), and at 37 sample indices on the right by the unrestricted \( \beta \)-test curve corresponding to \( k_w \). The 95% confidence interval is bounded on the left at 62 samples by the unrestricted \( \beta \)-test curve corresponding to \( k_m \), and on the right at 69 sample indices by the unrestricted \( \beta \)-test curve corresponding to \( k_w \).

Figure 2.11: Probability of rejecting \( H_0 \) as a function of the average sample span considered under each \( \beta \)-test. These curves render the same simulation results as Figure 2.10 except they are now plotted in terms of the average sample span of each \( \beta \)-test. Vertical lines tied to the curves reaching 0.68 and 0.95 on the ordinate axis, from left to right, define the corresponding confidence intervals. The 68% confidence interval is bounded on the left at 22 sample indices by the restricted \( \beta \)-test curve corresponding to \( k_m \), and at 37 sample indices on the right by the unrestricted \( \beta \)-test curve corresponding to \( k_w \). The 95% confidence interval is bounded on the left at 62 samples by the unrestricted \( \beta \)-test curve corresponding to \( k_m \), and on the right at 69 sample indices by the unrestricted \( \beta \)-test curve corresponding to \( k_w \).

same model as in the previous section (SNR = 2), one example of which is shown in Figure 2.5a. Here the test proceeds for \( \alpha \) in the inclusive range \( \alpha = [0, \ldots, 100] \), tested in \( \alpha = 0.1 \) increments, for both restricted and unrestricted \( \beta \)-tests, for both estimators \( k_m \) and \( k_w \). The data are plotted in terms of the probability of rejecting \( H_0 \) as a function of \( \alpha \). As before, the probability curves are color-coded blue and red for changepoint estimators \( k_m \) and \( k_w \), respectively. Like in Figure 2.9, the solid curves represent the results of restricted \( \beta \)-tests, while the dashed curves show the results of unrestricted \( \beta \)-tests. As expected, Figure 2.10 shows that, in general, \( \beta \)-tests associated with \( k_m \) reject the null hypothesis at a lower rate than those associated with \( k_w \), and restricted \( \beta \)-tests reject the null hypothesis at a lower rate than their unrestricted relatives.
Figure 2.11 plots the same results as in Figure 2.10 except here the probability of rejecting $H_0$ is plotted as a function of the average sample span considered under each $\beta$-test. Again, curves corresponding to $k_m$ and $k_w$ are blue and red, respectively, and restricted $\beta$-tests curves are solid and unrestricted $\beta$-test curves dashed. Black horizontal dashed lines originating from the ordinate axis at 0.68 and 0.95 are plotted to guide the eye to their corresponding timing confidence intervals, respectively. These confidence intervals are marked by vertical solid lines that originate from the abscissa axis and intersect the horizontal dashed lines. From left, the first pair of vertical lines marks the total spread of the average of the sample spans considered under each $\beta$-test that included the truth 68% of the time. The lower bound of this confidence interval is defined by the restricted $k_m$ curve at 22 sample indices, and the upper bound is defined by the unrestricted $k_w$ curve at 37. The second pair of vertical lines marks the total spread of the average of the sample spans considered under each $\beta$-test that included the truth 95% of the time. The lower bound of this confidence interval is defined by the unrestricted $k_m$ curve at 62 sample indices, and the upper bound is defined by the unrestricted $k_w$ curve at 69 sample indices.

### 2.5.3 Comparison of Methods I and II

The $\beta$-test curves of Figure 2.11 connect the M2 hypothesis testing procedure to the M1 Monte-Carlo resimulation. In Figure 2.8 we find that the M1 one-standard-deviation error estimates of $k_w$ and $k_m$ are 21 and 25 sample indices, respectively. From Figure 2.11 we see that the M2 one-standard deviation confidence interval is between 22 and 37 sample indices, a similar though larger range than what was found in M1. However, there is a cluster of $\beta$-test curves near the lower end of the M2 one-standard-deviation confidence interval, implying the true one-standard-deviation confidence interval is nearer 22 sample indices than 37. The large spread in $\beta$-test sample spans between $k_m$ restricted tests and $k_w$ unrestricted tests is greatly diminished in higher-SNR cases, where the AIC curve is less flat and the four tests see roughly the same sample spans.

The lack of agreement between the M1 two-standard deviation error estimate between 42 and 50 sample indices (Figure 2.8) and the M2 two-standard deviation confidence interval at 62 and 69
2.5. Formalizing the AIC Timing Uncertainty

sample indices (Figure 2.11) is likely due to a few factors. Firstly, the asymmetric nature of the AIC curve may play a part. AIC curves tend to have a lower gradient immediately before the true changepoint compared to the gradient immediately after it (note this phenomenon in Figure 2.9). What this means is that the sample span considered under an unrestricted $\beta$-test to the left of an incorrect changepoint estimate may be quite large, even if the estimate differs from the truth by only a single sample index. In addition, M2 confidence intervals will always be at least as large, but likely larger, than the error estimates returned in M1. This is due to the fact that M1 simply tallies the error from an estimate to a truth with no overshoot, while any given $\beta$-test will more than likely include more sample indices than are required to find the truth. Stated another way: there is likely to be overshoot in the sample span of a $\beta$-test in the event that enough sample indices are considered to reject the null hypothesis. This overshoot can be reduced by refining the discretization of $\alpha$ at the expense of increase computational complexity.

The utility of M2 is the generation of curves that negate the need for resimulation of synthetic data as in M1. Assuming a time series of similar SNR has previously been processed through the M2 procedure, a researcher may immediately quote any arbitrary confidence interval given a probability range of interest simply by querying the appropriate $\beta$-test curve.

2.5.4 Take-home message 2

The contribution of this study is a scheme for the rapid estimation of timing-error confidence intervals by inspection of the shape of the AIC curve, without the need for resimulation using synthetic data. After the generation of a library of curves detailing the probability of rejecting $H_0$ as a function of the average sample span of each $\beta$-test, like the one shown in Figure 2.11, for seismograms with various SNRs, no new synthetic time series for testing need be generated. Instead, a researcher may simply ask: “what is the probability that the truth lies within a sample span equal to (an arbitrary number of sample indices) that also includes $\hat{k}$?” or alternatively, “what is the sample span that includes $\hat{k}$ and has a probability of (an arbitrary value) of including the truth?”
2.6 Multiscale Analysis Methodology

The waveforms of various arriving seismic phases have distinct frequency signatures, and that information can and needs to be explicitly taken into account during inversion (e.g., Luo & Schuster, 1991; Dahlen et al., 2000; Yuan & Simons, 2014). In what follows we prepare to apply the concepts of sections 2.4 and 2.5 to a very specific type of timescale analysis via the wavelet transform (Strang & Nguyen, 1997; Mallat, 1998).

Our philosophy is perhaps closest to that of Zhang et al. (2003), and our contribution can be understood as picking up where they left off. While Zhang et al. (2003) use an Akaike information criterion (AIC) on wavelet-coefficient time series, they only use those picks and their consistency across neighboring scales to identify waveform segments of interest, which they then analyze with the AIC on the original time series. In contrast, here, our goal is to use AIC and wavelet analysis to determine scale-dependent seismic arrival-time picks and their associated uncertainties.

2.6.1 The discrete wavelet transform

We transform the time series into the multiresolution wavelet domain. Various orthogonal or biorthogonal sets of self-similar scaled (dilated) and shifted (translated) basis functions (wavelets, $\psi(k)$, scaling functions, $\phi(k)$, and their duals) can be used to decompose our time series $x(k)$. The analysis yields a set $j = [1, \ldots, J]$ of scaling coefficients $a_j(l)$, containing approximations at a certain scale $j$ (where a high number denotes a coarse resolution of $x$), and wavelet coefficients, $d_j(l)$, which provide the details missing to proceed to higher-resolution (at a lower scale number). The index sets in the timescale domain are scale-dependent, $l = [1, \ldots, M(j)]$, hereinafter implied where not explicitly stated, and encode the translations of the basis functions, giving them two indices, $\psi_{j,l}(k)$ and $\phi_{j,l}(k)$. The identity

$$x(k) = \sum_{j=1}^{J} \sum_{l=1}^{M(j)} d_{j,l} \psi_{j,l}(k) + \sum_{l=1}^{M(J)} a_{J,l} \phi_{J,l}(k) \quad (2.44)$$
leads to two ways of implementing our detection algorithm: either in the timescale domain, i.e. directly on the coefficient sequences

\[ d_{j,l}, \quad (2.45) \]
\[ a_{j,l}, \quad (2.46) \]

or in the time domain using the timescale subspace projections,

\[ x_j(k) = \sum_{l=1}^{M(j)} d_{j,l} \psi_{j,l}(k), \quad (2.47) \]
\[ \bar{x}_j(k) = \sum_{l=1}^{M(j)} a_{j,l} \phi_{j,l}(k). \quad (2.48) \]

The nested structure implied by equations (2.44)–(2.48) explains why the detail coefficients \( d_{j,l} \) are colloquially referred to as “differences” because they hold the information lost in the successively coarsening approximations \( a_{j,l} \), or “averages” A similar procedure was followed by Anant & Dowla (1997), who called the sequences (2.47) and (2.48) “interpolated coefficients.”

The timescale coefficients are inner products of the input time series with the basis functions \( \psi \) and \( \phi \). We utilize wavelets with compact support that have nonzero, real values only over a finite interval. As the scale number increases, the support of the wavelet and scaling functions doubles. The detail coefficients derive from a high-pass or differencing filter, while the approximation coefficients are the results of low-pass or moving-average filters. Our wavelet transform is non-redundant: details and approximations at increasing scales are obtained by iterating on the low-pass branch of what amounts to a filter bank (Strang & Nguyen, 1997). Various fast algorithms for wavelet analysis and synthesis are in use: here, we use the lifting scheme of Sweldens (1996).

Many wavelet bases exist and their choice depends on multiple factors including the domain of application, computational complexity, symmetry and smoothness of the basis functions and the data themselves, as well as timescale tiling considerations. Compared to alternative decompositions as, for example, the discrete-time short-time Fourier transform, where the window length
must be specified a priori and thus the timing resolution for all frequencies is constant, wavelet methods subdivide the timescale domain into tiles of variable size, whose duration is inversely proportional to frequency (e.g., Chakraborty & Okaya, 1995; Tibuleac & Herrin, 1999). At the lowest scale numbers, \( j \), we thus experience excellent time resolution and are able to extract high-frequency information, while at the higher scale numbers the opposite is true. With the wavelet transform, the character of abrupt (e.g., the arrival of high frequency \( P \)-wave energy) or emergent (e.g., the slow onset of a \( T \)-wave ) signals is revealed upon inspection (e.g., Gendron et al., 2000; Simons et al., 2009; Sukhovich et al., 2011, 2014).

Let us return to the problem of the paper. The application of the wavelet transform to the time series of interest, \( x(k) \), yields two new sets of time series, sensitive to information at different scales, \( j \). One type is a coefficient series indexed in the timescale domain, i.e. the \( d_{j,t} \) and \( a_{j,t} \) of equation (2.44), the other is a regular time series containing the subspace projections, i.e. the \( x_j(k) \) and \( \pi_j(k) \) of equations (2.47) and (2.48).

### 2.6.2 Multiscale AIC-based changepoint estimation

Acting on \( x(k) \), equation (2.26) led to the estimates \( k_m \) and \( k_w \) in equations (2.27) and (2.29). We can now substitute any of the four coefficient or time series just derived into equation (2.26), to return eight new changepoint estimators. Forgoing \( k_m \), which we showed to be biased, we focus on the multiscale analogs to \( k_w \) to define the set

\[
\begin{align*}
    k_{w,j} & \quad \text{derived from the wavelet-space projection } x_j(k), \quad (2.49) \\
    \overline{k}_{w,j} & \quad \text{derived from the scaling-space projection } \pi_j(k), \quad (2.50) \\
    l_{w,j} & \quad \text{derived from the wavelet coefficient series } d_{j,t}, \quad (2.51) \\
    \overline{l}_{w,j} & \quad \text{derived from the scaling coefficient series } a_{j,t}. \quad (2.52)
\end{align*}
\]

The first two are available for immediate use in seismology. The latter two require a method to map a timescale domain coefficient index to the time-domain, in practice: the range of time-
2.6. Multiscale Analysis Methodology

domain indices that a timescale domain index is sensitive to under the forward and inverse wavelets transforms. The mapping under the forward transform is described next, and under the inverse transform in the supplemental material, allowing for the possibility that the forward and inverse transforms need to be considered separately, as is the case for any biorthogonal transform. Readers uninterested in the mechanics of the algorithmic implementation are invited to skip forward to section 2.7, where we illustrate that the method works, or to section 2.9, where we illustrate its application to real data.

2.6.3 Timescale to time mapping: forward transform

A particular timescale coefficient index \( l \) at a particular scale \( j \) may be associated to a set of time-domain sample indices, identifying those points in the time domain from which a particular coefficient in the timescale domain receives contributions under the forward wavelet transform, via the mapping

\[
\hat{F}_j(l) = \hat{k}_{j,l}^*.
\]  

The star reminds us that the output of \( \hat{F}_j(l) \) for a single timescale domain index \( l \) at a single scale \( j \) is a contiguous time-domain sample span and not a single time-domain index \( k \). Such a mapping exists for every index \( l \) at every scale \( j \), and is particular to the basis used. The support of a wavelet or scaling basis function used in the forward transform, at a certain scale and translation, is completely described by the corresponding set \( \hat{k}_{j,l}^* \).

We call out three time-domain sample indices in each set, \( \hat{k}_{j,l}^* \): the “left” edge, the “middle,” and the “right” boundary:

\[
\hat{k}_{j,l}^+ = \min(\hat{k}_{j,l}^*),
\]
\[
\hat{k}_{j,l}^- = \left\lfloor \frac{1}{2} (\hat{k}_{j,l}^* + \hat{k}_{j,l}^T) \right\rfloor,
\]
\[
\hat{k}_{j,l}^T = \max(\hat{k}_{j,l}^*),
\]  

43
where \([\lfloor\cdot\rfloor]\) in equation (2.55) signifies rounding to the nearest integer. Equations (2.54)–(2.56) allow us to tack a changepoint estimate determined in the timescale domain, and thus representing a time smear, to a single time-domain sample index.

The supplemental material contains the complementary map for the inverse wavelet transform, and details our treatment of edges in both transform directions.

2.6.4 *Multiscale changepoints to multiscale arrival times*

The time-domain changepoint estimate (equation 2.29) by our definition is the last sample of the noise segment, hence, the reported arrival time of a specific seismic phase will be at the sample index that immediately follows, \(k_{w} + 1\).

Similarly, the multiscale changepoint estimates of equations (2.49) and (2.50) effortlessly map into arrival times by adding one to their time-domain sample indices, \(k_{w,j} + 1\) and \(\overline{k}_{w,j} + 1\). We note that these may correspond to different seismic phases, or differently resolved specific arrivals.

For their part, the changepoint estimates (2.51) and (2.52) need to be mapped into the time-domain sample indices that capture arrival times in a way that is not similarly trivial. Indeed, in this case all of our operations (AIC picking, SNR estimation, uncertainty estimation) occur natively in the timescale domain, not after conversion to the time-domain. We leverage the mapping established in section 2.6.3 to label the time-domain sample spans \(\hat{k}_{j,l,w+1}^{*}\) and \(\overline{k}_{j,l,w+1}^{*}\), where we draw attention to the fact that the addition of one index occurs in the subscripted timescale domain. From there on, we can use equations (2.54)–(2.56) to report individual time domain sample indices.

The final conversion of course, of any and all of these estimates, will map time-domain sample indices into time by taking into account any initial offset and the sampling rate.

2.6.5 *Take-home message 3*

The careful mapping of time-domain sample indices of a time series of interest onto timescale domain coefficients, and vice versa, for orthogonal and biorthogonal wavelet and scaling functions,
under the wavelet transform and its inverse, allows us to carry out changepoint estimation via
the AIC in either domain, giving us three options: directly in the time-domain, directly on the
timescale domain coefficient sequences, or on partially reconstructed subspace projections in the
time domain, leading to true multiscale arrival-time detection.

2.7 Multiscale Analysis in Practice

As we did in section 2.5 with the time-domain changepoint estimation, here in this section we
test the performance of the multiscale method and its implementation in code. We stick with the
CDF(2,4) wavelet basis (Cohen et al., 1992) based on prior experience with actual data: their basis
functions are short and capture the $P$-wave onset with just a handful of diagnostic coefficients
(Simons et al., 2006a; Sukhovich et al., 2011; Yuan & Simons, 2014).

The changepoint estimators that we study in detail are: those derived by applying the AIC to the
time-domain wavelet-subspace projection of the time series $x_j(k)$ at a particular scale $j = [1, \ldots, J]$, namely

$$k_{wj} \quad \text{for} \quad j = [1, \ldots, J], \quad \text{from equation (2.49)}, \quad (2.57)$$

one derived from the complementary scaling-space projection $\pi, j(k)$ at the coarsest scale $J$, namely

$$\overline{k}_{wj} \quad \text{for} \quad j = J, \quad \text{from equation (2.50)}, \quad (2.58)$$

and lastly, the pairs of triplets that result from the application of the timescale to time-domain
mappings of equations (2.54)–(2.56) to the estimates of equations (2.51) and (2.52), and which we
write here as

$$\dot{k}_{\perp j,l}^{\perp} \quad \text{for} \quad j = [1, \ldots, J] \quad \text{and} \quad \dot{k}_{\perp J,lw}^{\perp}, \quad (2.59)$$

$$\dot{k}_{\perp j,l}^{\downarrow} \quad \text{for} \quad j = [1, \ldots, J] \quad \text{and} \quad \dot{k}_{\downarrow J,\perp}^{\downarrow}, \quad (2.60)$$

$$\dot{k}_{\perp j,l}^{\uparrow} \quad \text{for} \quad j = [1, \ldots, J] \quad \text{and} \quad \dot{k}_{\uparrow J,\perp}^{\uparrow}, \quad (2.61)$$
Each of these various scale-dependent estimates will be generically referred to as $\hat{k}_j$ if they serve to estimate a certain changepoint $k_o$ in a time series as modeled in section 2.3.

### 2.7.1 Multiscale AIC-based changepoint estimators

It is important to note that we are testing the performance of measurements made at a certain scale $j \neq 0$ that estimate changepoints $k_o$ that, under our model espoused in section 2.3.1, are changes at scale $j = 0$. This apples-to-oranges comparison is necessary because it might correspond to how our algorithm is used in practice. The comparison of how a particular $\hat{k}_j$ is successful at estimating a particular $k_o$, at the same scale $j$ is exactly what has been covered in section 2.5: after all, our model is that of a certain sequence that changes at a certain index, regardless of scale.

Should our model be that of a time series that had a changepoint at a particular scale, even multiple different ones, the multiresolution wavelet analysis would produce coefficient sequences which, thanks to mutual orthogonality, would be uncorrelated between different scales, and the AIC-based methodology would lead to the proper identification of the changepoint at the appropriate scale. An example would be where different seismic phases with intrinsically different time-frequency signatures (e.g., $P$ or $SS$) would be properly identified in time and scale (Tibuleac et al., 2003).

### 2.7.2 Timing uncertainty of changepoint estimators

The uncertainty estimates of section 2.5 are found in the domain in which the changepoint is estimated. For both time-domain and timescale domain cases the procedures of sections 2.5.1 and 2.5.2 are followed at each individual scale using the domain-specific indices of interest: $k$ in the time domain, and $l$ in the timescale domain.

For uncertainty estimation of multiscale changepoint estimates in the time domain, equations (2.32)–(2.43) are computed as written, acting on $k$, except they are repeated at every scale, $j$, and each can be considered reproduced here as such with that subscript.

For uncertainty estimation associated with scale-dependent changepoint estimates $\hat{l}$ made in the
timescale domain, those same equations are computed with \( l \) in place of \( k \), again at each scale \( j \). In that case, the M1 statistic (equation 2.32) becomes \( \hat{l} - l_0 \), and the index spans relevant to the two M2 hypothesis tests (equations 2.38 and 2.39) become the timescale domain coefficient index spans, \( 1 + [l', \ldots, l''] \), and \( 1 + [l', \ldots, l'''] \). Finally, multiplication of these statistics by the length of the appropriate span, either \( k^*_{j,lw} \) or \( \hat{k}^*_{j,lw} \), maps them back to the time domain.

In summary, at any given scale a unit error associated with either \( k_{lw} \) or \( k_{wJ} \) maps to one time domain sample index; whereas a unit error of \( l_w \) or \( l_{wJ} \) maps to the length of the time smear associated with the changepoint estimate itself, \( k^*_{j,lw} \) or \( \hat{k}^*_{j,lw} \). Therefore, uncertainty estimates associated with changepoint estimates derived in the timescale domain see their uncertainty dilate in concert with the wavelets and scaling functions themselves.

### 2.7.3 Multiscale AIC-changepoint timing uncertainty

Let us examine what happens when we follow the Monte Carlo resimulation method M1 detailed in section 2.5.1, using many realizations of synthetic time series of length \( N = 4000 \) with a known true changepoint, \( k_0 \), at time-domain sample index 2000. Each synthetic time series was generated by concatenation at \( k_0 \) of two random samples drawn from Gaussian distributions with zero expectation, whose variances differed by a prescribed factor. As with all previous synthetic time series, the first segment, time-domain sample indices \( k = [1, \ldots, k_0] \), was noise drawn from the standard normal distribution, \( \mathcal{N}(\mu = 0, \sigma^2 = 1) \), whereas the second segment, \( k = [k_0 + 1, \ldots, N] \), was signal drawn from a density \( \mathcal{N}(\mu = 0, \sigma^2 = \text{SNR}) \), where SNR was varied between 2 and 1024, increasing in powers of two. We decompose each time series to five wavelet scales.

We evaluate the performance of a generic single-scale changepoint estimate via the statistics of the signed distance \( \hat{k}_j - k_0 \), the time-domain sample-index error. Figure 2.12 shows the result of 1000 tests for each \( \hat{k}_j \) at every SNR considered. For ease of reporting we label results for \( k_{wj} \) and \( \overline{k}_{wj} \) and likewise for all the other pairs of changepoint estimates ultimately derived from wavelet-space or scaling-space representations, respectively, on a common axis, where we will draw a bar over the scale number to identify its scaling-space nature: results for, e.g. \( k_{w5} \) will be reported at
2.7. Multiscale Analysis in Practice

Figure 2.12: Multiscale changepoint estimation errors at various signal-to-noise ratios (SNR, equation 2.31) for timescale and time domain changepoint estimates at 5 wavelet scales, and one approximation scale, labeled $\tilde{s}$. The individual panels (a)–(j) each have the same abscissa axis (low- to high-scale, or equivalently, high- to low-resolution) from left to right, and ordinate axis ($-200$ to 200 time domain sample indices), and are arranged from high- to low-SNR tests (noted in the upper left corner of each panel). In each of (a)–(j) the filled markers (circles or diamonds), connected as a curve, represent the mean changepoint estimation error (via the M1 procedure, $\hat{k}_j - k_o$) for 1000 test realizations where $\hat{k}_j$ is one of: $k_{w_j}$ or $\overline{k}_{w_j}$ (purple), marked with filled diamonds; or $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$ (teal), $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$ (red), or $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$ (green), marked with filled circles. The ticks extend up and down from each mean estimate by once the standard deviation. Each test realization was performed on a synthetic time series of $N = 4000$ generated via the concatenation at $k_o = 2000$ of two random samples drawn from Gaussian distributions of zero expectation that differ in variance by the SNRs indicated in the inset boxes.

The points marked $\tilde{s}$, while results for, e.g., $\overline{k}_{w_j}$ will be drawn at $\tilde{s}$ on the abscissa axis.

In each of Figure 2.12a–j the colored curves connect the observed average error at every scale, marked as a filled diamond or circle, with the ticks extending vertically in both directions from the marker indicating the width of the distribution by mirroring their standard deviations. The curves are color-coded to differentiate the changepoint estimator: those derived in the time-domain, $k_{w_j}$ and $\overline{k}_{w_j}$, whose average errors are marked with purple diamonds; and those derived in the timescale domain and mapped back to the time domain, whose average errors are marked with filled circles, either teal for $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$, red for $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$, or green for $\hat{k}_{j,l,w}$ or $\hat{k}_{j,l,w}$. The axes for
all of the panels are equal, with the ordinate axis representing the changepoint estimation error and bounding the span of $-200$ and $200$ time domain sample indices, while the abscissa axis denotes the scale at which the M1 test was performed, with the overline notation for the scaling-space as discussed above.

The estimator pairs $(\hat{k}_{j,l}^l, \hat{k}_{j,l}^r)$ and $(\overline{k}_{w_j}, \overline{k}_{w_j})$, perform the best, the former having acceptable bias even at high-scales and low SNR, and the latter being better than the others at all but the lowest-SNRs. Regardless of SNR all changepoint estimate pairs $(k_{w_j}, \overline{k}_{w_j})$ display either large biases and/or large standard deviations at the highest scale, $j = 5$. Of course the “edge” changepoint estimators, $(\hat{k}_{j,l}^l, \hat{k}_{j,l}^r)$ and $(\overline{k}_{j,l}^l, \overline{k}_{j,l}^r)$, perform most poorly, showing growing negative (early) and positive (late) biases regardless of SNR, as the scales increase. They represent the two end-member cases for estimates made in the timescale domain and mapped back to the time domain. As mentioned, the compromise between the two, $(\hat{k}_{j,l}^l, \hat{k}_{j,l}^r)$, outperforms either of its two siblings, especially at low-SNRs, though it does display bias at high-SNRs.

### 2.7.4 Coherence across scales

A takeaway from Figure 2.12 is that all of the changepoint estimators $\hat{k}_j$ perform well at high resolution, when the scale is low, $j = 1$ or $j = 2$, regardless of SNR. Indeed, even at the lowest SNR tested, SNR $= 2$, all such estimators were on average within 5 time-domain sample indices of the truth. The short time-domain support at the lowest scales lessens the mapped time smears of the basis functions, $\hat{k}_{j,l}^*$. A shorter basis function always results in more precise changepoint estimates as long as the time window is sufficiently long to capture it.

As the scales increase timing resolution degrades because the basis functions dilate to capture longer-period features. The shape of the curves in Figure 2.12 reveals the twofold dilation of the basis functions at every scale. In particular for the CDF(2,4) basis that we used, the primal wavelets and scaling functions have support of 3 and 9 time-domain sample indices, respectively. At the highest scale, after five successive dilations, the effective support balloons to 153 for $\psi_{5,l}$ and 249 for $\phi_{5,l}$. In the frequency domain, a simple rule-of-thumb holds that the first scale of a
wavelet-decomposed time series is approximately sensitive to the frequency band spanning $\frac{1}{2}$ to $\frac{1}{4}$ the sampling rate of the input time series. Each subsequent scale further halves the sensitivity frequency band of the previous, leading to a recursive relation approximating the frequency sensitivity of a wavelet at scale $j$ as

$$\sim \left[ \frac{f_s}{2^j} \leftrightarrow \frac{f_s}{2^{(j+1)}} \right], \quad (2.62)$$

where $f_s$ is the sampling rate of the input time series.

The behaviors in this section illustrate the coherence and the range over which it persists, between changepoint estimates made at different scales, that can be expected, and indeed, is often implicitly assumed or enforced, with the notion of there being a single onset happening at the raw sampling rate. In what we propose our picks are single-scale at multiple scales, and no interscale coherence is required for their validation.

Figure 2.12 may be used to intuit the accuracy of a specific changepoint method, but should not be construed to represent the precision associated with each. It is presented partly to show that in extremely low-SNR cases, e.g., $\text{SNR} = 2$, if accuracy is paramount to uncertainty estimation, one may wish to derive changepoint estimates in the timescale domain and tack their time domain time smears to a single sample index with the pairs $(\hat{k}|_{j,w}, \hat{k}|_{J,w})$, instead of picking them in the time domain with the pairs $(k_{w,j}, \overline{k}_{w,j})$. However, while an estimator like $k|_{j,w}$ may be more accurate than $k_{w,j}$ at low-SNRs, at high scales its precision degrades.

For this reason we adopt as our preferred arrival-time estimates $k_{w,j}$ and $\overline{k}_{w,j}$, found at the subspace projections $x_j$ and $\overline{x}_j$, rather than the estimates $\hat{k}|_{j,w}$ or $\hat{k}|_{J,w}$ (or any individual sample within that span) found using the raw detail and approximation coefficients. In section 2.9 we apply our multiscale-AIC method in both domains and the differences in timing-uncertainty estimates will become more clear.

Uncertainty estimation in practice, with real data when $k_\circ$ is not known, is discussed next, with particular emphasis given to the problem of uncertainty estimation associated with changepoint estimates derived in the timescale domain. In that case, the assumed “truth,” $l_\circ$, must exist distinctly
2.8. The Data Set

at every scale \( j \) in the timescale domain, which differs from the error measurements of this section, which occurred after mapping \( l_w \) and \( \overline{l_w} \) back to the time domain and assigning to each of them a single time-domain sample index against which signed distances could be measured.

2.7.5 From synthetics to actual data

Until now we have discussed changepoint estimation methods and timing uncertainty appraisal on synthetically modeled data, where the true changepoint \( k_o \) or \( k_o_j \) is known. With actual data, which we will discuss in section 2.9, we replace \( k_o_j \) with the changepoint estimate made on the real data at the relevant scale: one of either \((l_w, \overline{l_w})\) or \((k_w_j, \overline{k_w_j})\) depending on the domain in which it is determined.

Synthetic seismograms generated for uncertainty estimation via the M1 resimulation and M2 curvature analysis methods of section 2.5 are constructed in the same domain that the changepoint estimate was made, at every scale. At each scale, many randomized synthetic seismograms are constructed by concatenation at \((l_w, \overline{l_w})\) or \((k_w_j, \overline{k_w_j})\) of two samples drawn from two different and distinct Gaussian distributions: \( N(\hat{\mu}_1, \hat{\sigma}_1^2) \) and \( N(\hat{\mu}_2, \hat{\sigma}_2^2) \), the parameters of which are computed from the noise and signal segments, respectively, of the seismogram at that scale. The means and variances \( \hat{\mu} \) and \( \hat{\sigma}^2 \) pertain to the coefficients \((d_{j,l}, a_{l,l})\) in the timescale domain, or to the projections \((x_j, \pi_j)\) in the time domain, as segmented into noise and signal by the changepoint estimators.

2.8 The Data Set

The parameters of many of our examples shown thus far to illustrate our procedures and methodology have been appropriate for a very specific data set collected by a very specific type of seismological instrument. In this section we turn to describing MERMAID (an acronym for Mobile Earthquake Recording in Marine Areas by Independent Divers) before applying our techniques to seismograms collected at sea in section 2.9.
MERMAID is an autonomous ocean-going diver with a hydrophone that continuously records and processes the ambient acoustic wavefield at mid-column depths. Its primary goal is to monitor worldwide earthquake activity, and specifically, to provide arrival times of teleseismic waves suitable for global seismic tomography (Simons et al., 2006b; Hello et al., 2011). A subset of the data collected, all of the data that we analyze in this study, are publicly available in Seismic Analysis Code and miniSEED formats (see section 2.11).

2.8.1 The instrument

While at depth, the acoustic data: a hydroacoustic time series, hereafter the “seismogram” (Joubert et al., 2015), are filtered between \([0.10 – 10] \) Hz and digitized in real time at a sampling rate of 40 Hz. The digitized data are then immediately processed by an STA/LTA algorithm (Allen, 1978) to identify segments of interest (Simons et al., 2009) where a possible signal rises above the level of the noise. When a predetermined STA/LTA trigger threshold is exceeded, MERMAID passes a windowed segment containing pre- and post-trigger data into the wavelet-based (Simons et al., 2006a; Sukhovich et al., 2011) detection algorithm of Sukhovich et al. (2014). This discrimination procedure inspects the energy partitioning between different wavelet scales and assigns a criterion value to the hypothesis that the waveform includes a \(P\)-wave arrival generated by a teleseismic earthquake, and not some other energy generated by a non-seismic source (e.g., ship propellers or ocean storms, which distribute energy differently over various scales). The on-board algorithm decomposes the 40 Hz data to 6 wavelet scales via the same lifting algorithm (Sweldens, 1996) and CDF(2,4) wavelet basis (Cohen et al., 1992) as in this study. If the signal is deemed to be a teleseismic arrival, a 200–250-second-long seismogram, containing the STA/LTA trigger at about 100 s, is returned to shore via the Iridium satellite constellation at MERMAID’s next surfacing. Because MERMAID freely drifts with the ocean currents its location at the time of recording the seismogram must be interpolated from multiple Global Positioning System locations fixed at the surface. The exact details of this procedure are described by Joubert et al. (2016).
2.8.2 The seismograms

By default MERMAID transmits the raw detail and approximation coefficient series at scales 2 – 6, omitting scale 1 to save data-transfer cost, and because scale 1 is not particularly useful for teleseismic P-wave analysis (at $f_s = 40$ Hz it covers the frequency band spanning roughly $[10 – 20]$ Hz). The seismogram is then reconstructed onshore via the inverse wavelet transform at an effective sampling rate of $f_s = 20$ Hz. In this study we decompose these 20 Hz seismograms to 5 scales so as to analyze the data at the same resolution as the on-board detection algorithm. In some cases, only 3 out of 6 wavelet coefficient sets are being returned, which leaves us with $f_s = 5$ Hz seismograms.

Table 2.1 lists the approximate frequency bands to which each wavelet is sensitive at the five scales in our numbering scheme, for seismograms sampled at 20 Hz and 5 Hz. Later, we compute arrival-time estimates and their uncertainties, and travel time-residuals considering the entire public MERMAID catalog. So as to compare arrival-time picks made at the same resolutions, scale 1 for 5 Hz seismograms will be analyzed with scale 3 for 20 Hz seismograms, scale 2 for 5 Hz seismograms will be analyzed with scale 4 for 20 Hz seismograms, and so on.

2.8.3 The initial events catalog

Prior work (Joubert et al., 2016; Sukhovich et al., 2015; Nolet et al., 2019) has resulted in a catalog of “identified” events (and a complement with “unidentified” seismograms). The classification indicates whether the seismograms were matched to known seismic events, i.e., earthquakes, as determined by querying published seismic catalogs available at the time. In this study we focus our attention on 445 identified MERMAID seismograms. These seismograms represent global earthquakes recorded at disparate times between December 2012 and February 2018 and various (time-variable) locations both in the Indian Ocean and Mediterranean Sea.

We maintained the event identifications but updated their details to the most up-to-date information available from the Incorporated Research Institutions for Seismology (IRIS). Since the original event identification consisted of an epicentral location, origin time, and magnitude, we
Table 2.1: The approximate frequency sensitivity of wavelets.

<table>
<thead>
<tr>
<th>Scale</th>
<th>$f_s = 20\text{ Hz}$</th>
<th>$f_s = 5\text{ Hz}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>[5.00 – 10.0] Hz</td>
<td>[1.25 – 2.50] Hz</td>
</tr>
<tr>
<td>2:</td>
<td>[2.50 – 5.00] Hz</td>
<td>[0.62 – 1.25] Hz</td>
</tr>
<tr>
<td>3:</td>
<td>[1.25 – 2.50] Hz</td>
<td>[0.31 – 0.62] Hz</td>
</tr>
<tr>
<td>3*</td>
<td>—</td>
<td>[0.1* – 0.31] Hz</td>
</tr>
<tr>
<td>4:</td>
<td>[0.62 – 1.25] Hz</td>
<td>—</td>
</tr>
<tr>
<td>5:</td>
<td>[0.31 – 0.62] Hz</td>
<td>—</td>
</tr>
<tr>
<td>5*</td>
<td>[0.1* – 0.31] Hz</td>
<td>—</td>
</tr>
</tbody>
</table>

The approximate frequency bands (equation 2.62) to which individual wavelets are sensitive to, based on the sampling rate ($f_s$) in Hz of the MERMAID seismogram. *The approximation, noted here with an overline at scale 3 or 5 depending on the sampling rate of the seismogram, in theory approximates a low-pass filter (with sensitivity down to 0 Hz), but in practice only has sensitivity to 0.1 Hz because MERMAID data are filtered on-board between [0.10 – 10] Hz before digitization (see section 2.8.1).

maintained the match by querying the latest catalogs within a buffer of 30 s and 1° in time and location, and of no more than one magnitude unit less than the original. In the few rare instances where our match was non-unique using these criteria we selected the one provided by the International Seismological Centre (2016) online bulletin. Only a handful of cases required manual intervention to complete our revised event catalog.

Figure 2.13 displays the updated locations of our events (red stars) and the interpolated positions of the MERMAID instruments that recorded them (yellow triangles). Great circles connect event-station pairs to give a sense of the geographic areas sampled most by our data set. Note that although only 13 individual MERMAID instruments are present, there are 445 unique receiver locations plotted, illustrating their passive drift with the ocean currents.

We computed arrival times of various seismic phases in the ak135 velocity model of Kennett et al. (1995). We first computed the theoretical travel times with MatTaup, specifically using the taupTime method. These codes are dated November 2002 but were distributed without a version number. They are based on TauP, described by Crotwell et al. (1999). The arrival time
2.8. The Data Set

Figure 2.13: Global map of all events and stations used in this study. Great-circle paths connect the event locations (red stars) to the MERMAID positions (yellow triangles) at the time the seismogram was recorded.

was obtained as the updated event time plus the theoretical travel time, minus the time at the first sample of the seismogram.

Inspired by our experience with our own data sets (e.g., Simons et al., 2009; Sukhovich et al., 2015; Nolet et al., 2019), and by work conducted by other researchers on hydroacoustic time series elsewhere (e.g., Bohnenstiehl et al., 2002; Smith et al., 2002; Dziak et al., 2004; McGuire et al., 2012), we consider the following phases likely to be present in MERMAID seismograms: $p$, $P$, $pP$, $PP$, $Pn$, $Pg$, $PcP$, $Pdiff$, $PKP$, $PKiKP$, $PKIKP$, $s$, $S$, $Sn$, $Sg$ (using the phase-naming convention of Crotwell et al., 1999). We allowed the computation of arrival times of extraordinary phases (e.g., $SKiKP$ for seismograms m31.20140910T053727.sac and m33.20150916T142424.sac) only if they were listed in the `events.txt` file distributed with the public MERMAID catalog. Such instances were rare, and we reserve the discussion on possible phase ambiguity for section 2.9.2.
2.9 Application to Mermaid Seismograms

We apply our AIC-based multiscale methods to the MERMAID data set described. We detrended
the seismograms and trimmed the last sample index to render all of them of even-length. We used
the CDF(2,4) wavelet transform to five scales for the $f_s = 20$ Hz seismograms and to three scales
for the $f_s = 5$ Hz set (see Table 2.1).

2.9.1 Multiscale AIC measurements and their uncertainty

Figures 2.14 and 2.15 provide the first illustrations of the determination of an independent set
of multiscale arrival-time estimates and their uncertainties from which to compute residuals with
the theoretical arrival times in the catalog. For these two example earthquakes in the Bali Sea
and the Tyrrhenian Sea, which yielded a low-SNR and a high-SNR seismogram, respectively, the
top set of panels, labeled (a), illustrate changepoint estimates made in the time domain (on the
subspace projections), whereas the bottom set, labeled (b), illustrate changepoint estimates made
in the timescale domain (on the coefficient series) as explained in section 2.6.2.

In both figures, for both sets (a) and (b), the topmost panel plots the MERMAID seismogram, $x$,
in blue, normalized between $-1$ and 1. Earthquake magnitude, epicentral distance (in degrees) to
the recording MERMAID, and event depth are listed in the legends. For the set (a), the panels below
the raw seismogram are the subspace projections after wavelet decomposition $(x_j, x_J)$, in gray,
normalized between $-1$ and 1. For the set (b), the panels below the first one show the absolute
values of the detail and approximation coefficients, $(d_{j,l}, a_{J,l})$, in gray, smeared over the time-spans
to which they are sensitive, $\hat{k}_{j,l}^*$, and normalized between 0 and 1. The abscissa axes, in seconds
offset from the start of the seismogram, ($t = 0$ at $k = 1$), are unchanged between subplots. The
subspace projection series in (a) are at least as truncated as the raw coefficient series in (b), but
usually more so, due to the increased influence of the edges in the time domain compared to the
timescale domain, as explained in the supplemental material.

Overlaid in black are the AIC functions, normalized per panel, as time-domain traces for the
Figure 2.14: (caption next page)
2.9. Application to Mermaid Seismograms

Figure 2.14: (previous page) Multiscale arrival-time estimation of a seismogram from the MERMAID data set in the low-SNR regime: a seismogram detected in the Mauritius-Réunion region corresponding to an earthquake in the Bali Sea. In both (a) and (b) the top panel plots the same raw seismogram (normalized between $-1$ and 1) in blue with the same theoretical arrival time of the $P$ wave marked with a black vertical line. Inset is the magnitude, great-circle distance between the epicenter and MERMAID, and the depth of the corresponding earthquake. The panels below the first one show, for the set (a), subspace projections at varying wavelet scales in gray (normalized between $-1$ and 1), with their AIC curves overlain in black. In the set (b), below the seismogram, the panels show the absolute values of the wavelet and scaling coefficient time series (normalized between 0 and 1), and their AIC function, rendered as the corresponding time-smears. Event detections were made at scales 1 through 3, in the time domain in set (a) where they are marked in purple, and in the timescale domain in set (b) using teal, red, and green to mark the beginning, middle, and end of the arrival time smears. The two-standard-deviations of the error distributions, obtained by Monte-Carlo resimulation, are listed to the right of the ordinate axis. Scales 4 and 5 each had SNRs less than or equal to 1 and thus no arrival time is reported.

The vertical lines mark the changepoint estimates: purple for the subspace-projection estimates in (a), and teal, red, and green for the coefficient series estimates in (b). The latter triplet of colors marks the beginning, middle, and end of the time-smears, ($\hat{t}_{j,l,w}^*, \hat{k}_{j,l,w}^*$), of the estimated arrival-time detail or approximation coefficients, as detailed in 2.6.3. These are the same colors used in Figure 2.12, but here we mark the estimated arrival times, and there we analyze the statistics of the estimated changepoints (the mapping between the two is discussed in section 2.6.4). The dilation of the wavelet with increasing scales is easily seen when plotted in this manner, especially in (b) of Figures 2.15–2.17, where arrival time estimates were made at every scale (all SNR $> 1$), and the time-smear of the basis function is seen to lengthen at every scale to over 12 s at scale 5.

The corresponding SNR at each scale is labeled in the lower-left corner inset of each panel, and when it is smaller than or equal to one no arrival-time estimate is reported. The two-standard deviation of the error distribution (after 1000 test realizations at each scale) is to the right of the ordinate axis, in s, as determined using the M1 test of section 2.5.1.

Figures 2.16 and 2.17 contain additional examples, presented in the same layout and with legends and labeling as in Figures 2.14 and 2.15. Figure 2.16 successfully separates core phases $PKIKP$ and $PKP$ and Figure 2.17 separates $P$ from $PP$. One more example is given in the supplemental material.
2.9. Application to Mermaid Seismograms

Figure 2.15: (caption next page)
Figure 2.15: (previous page) Multiscale arrival-time estimation of a seismogram from the MERMAID data set in the high-SNR regime: a seismogram detected near the Balearic Islands, Spain, corresponding to an earthquake in the Tyrrhenian Sea. Arrangement and labeling are as in Figure 2.14. Here we see clear \( P \)-wave detection coherently across all scales. The increase in low-frequency energy around 200 s is the arrival of an \( S \)-wave. This arrival is especially apparent in at scale 5, where the AIC function dips to nearly the same low value (high-likelihood of a second changepoint) as obtained at the \( P \)-wave pick, hinting at the prospect of an extended utility of our method by recursive implementation; i.e., reapplication of our AIC-based arrival-time detector on a shortened \((x_j, \pi_j)\) or \((d_{j,l}, a_{j,l})\) series that begins immediately after the initial \( P \)-wave detection. Also note here that while the ak135 travel-time residuals in (a) and (b) are similar at every scale, their estimated uncertainties are not, with those made in the timescale domain (b) much greater than those made in the time domain (a). This is due to the dilation of the wavelets themselves at increasing scales, which manifests as stretched time-smears, and which is well-illustrated in (b) by the increased separation between the teal and green vertical bars that mark the beginning and end, respectively, of the timescale coefficient smears of the estimated arrivals.

### 2.9.2 Computing travel-time residuals

We make the assumption that the theoretical phase arrival nearest in time to our AIC-based arrival-time estimate corresponds to the true seismic phase identified by our method, which gives us a scale-dependent travel-time residual against the ak135 velocity model. In Figures 2.14–2.17, the travel-time residuals and their associated phases are quoted to the right of the ordinate axes, at each scale. There will be ambiguities: for example, in Figure 2.17a at wavelet scale 5, where the arrival time is tagged as a \( pP \) wave instead of the preceding \( P \) wave, which is arguably more likely, and which one might disambiguate by taking amplitude information, or other attributes, into account. Similarly, the arrivals in scales 1 through 4 in Figure 2.17b might derive from either of two \( P \) waves, from the same earthquake, with predicted arrival times very near each other.

As explained in section 2.3.5 the MLE procedure which underpins our AIC-based method selects the index in the time series where the ratio of variances of the signal and noise segmentations is largest; it maximizes the SNR after testing all possible combinations. Without iteration, our method only allows for the identification of a single seismic phase at each scale. This is generally appropriate in our case, given that a MERMAID seismogram most typically includes a single \( P \)-wave arrival, as in Figure 2.14, or a single picked \( P \)-wave arrival, as in Figure 2.15. Further-
2.9. Application to Mermaid Seismograms

Figure 2.16: A mid-SNR seismogram detected near the south coast of France corresponding to an earthquake in the Kermadec Islands, New Zealand. This is the first illustration of our procedure resulting in multiphase detection of two distinct core-phases at separate scales: PKP-wave detections at scales 2 through 5 (at the resolution of the details) and PKIKP-wave detection at scale 5 (at the resolution of the approximation). There is an arrival pick at scale 1, though its SNR is low and thus its uncertainty is high. Such a pick is not considered high-quality and is not included in Figure 2.18 or Table 2.2.

61
Figure 2.17: A low-SNR seismogram detected over Broken Ridge corresponding to an earthquake in the southern Indian Ocean. Here we see the matching of different phases at scale 5 between arrival time picks made in the two domains: (a) a $pP$ wave in time, and (b), a $P$ wave in timescale. Furthermore, in (b) there are two distinct $P$ waves with ak135 arrivals very near in time to each other, one or the other of which minimizes the travel-time residual at one of the scales 1 through 5. Unresolved ambiguities of this sort are discussed in section 2.9.2.
more, the frequency bands of our wavelet decomposition are sufficiently narrow to successfully partition the arrival of distinctly pickable seismic phases over separate scales, as exemplified by Figures 2.16 and 2.17.

Picking up additional arrivals, should they contain substantial energy within the same scale, can be accomplished, if further decomposition to higher scales is not option, by iterating our method on the identified signal segment. A good candidate is shown in Figure 2.15, where the energy increase near 200 s is an S-wave arrival drowned out at wavelet scale 5, and in the shadow of the coda of the preceding P-wave arrival at approximation scale 5. Our algorithm could be run recursively at those scales, on a truncated time window beginning just after the P-wave arrival so as to bracket only the segment containing the S-wave arrival.

2.9.3 Distribution of travel-time residuals

We apply the procedure illustrated in Figures 2.14–2.17 to the complete data set of 445 identified MERMAID seismograms. We work in the time domain using the subspace projections \((x_j, \bar{x}_j)\) per the timing considerations discussed in section 2.7. For every seismogram at each scale we compute: (1) an AIC-based arrival time estimate; (2) the uncertainty associated with that arrival time using 1000 realizations of method M1; (3) the travel-time residual by considering all phases that might arrive in the time window of the seismogram, retaining the minimum travel-time residual and its associated phase.

Figure 2.18 shows the multiscale distribution of travel-time residuals in our data set. We only show those that we deem to be of “high-quality,” falling within 6 s of a theoretical phase arrival, and whose two-standard deviation uncertainties per the M1 method are smaller than 1 s. These quality criteria resolve some of the possible issues with phase ambiguity, guarantee sufficient SNR to reduce the likelihood of falsely triggering on spurious energies not related to a phase arrival, and, lastly, they produce histograms that are in line with expectations for mantle P waves (e.g., Simmons et al., 2012) without generating long tails beyond the 6 s cutoff.

Figure 2.18a–f plot histograms of travel-time residuals at scales 1 through 5 for 20 Hz seis-
2.9. Application to Mermaid Seismograms

Figure 2.18: High-quality travel-time residuals in the public MERMAID catalog. Figure 2.18a–f each are a histogram of residuals from high to low resolution. Each histogram displays only travel-time residuals whose absolute values are less than or equal to 6 s and whose two-standard-deviation error by the M1 method of section 2.5.1 are smaller than 1 s. In each panel, the mean and standard deviation of the residuals at that resolution are listed inside, the total number of residuals and their percentage relative to all arrival-time estimates (SNR \( > 1 \)) at that resolution are listed above, and the scales considered and their approximate frequency sensitivities (2.1) at that resolution are listed below.

While we are confident that, taken as a whole, the residuals presented in this study and summarized here faithfully record the signal of the Earth, we are aware that false triggers and mismatched phases may exist in our catalog, as is true for all automated arrival-time identification and phase-picking methods. Future work will necessarily include waveform modeling to better tack our AIC-based arrival times to their associated phases in MERMAID seismograms.

Of the 445 seismograms considered, 339 were sampled at 20 Hz and 106 at 5 Hz. The limits of the ordinate axes in Figure 2.18a–f are adjusted to reflect the fact that roughly 25% more data are available for consideration in Figure 2.18c–f than in Figure 2.18a–b. The total numbers of arrival-time estimates (SNR \( > 1 \)) and residuals computed at each resolution of Figure 2.18a–f are 238, 296, 410, 410, 401, and 384, with the numbers of high-quality residuals actually plotted displayed.
2.9. Application to Mermaid Seismograms

Table 2.2: Phases in Figure 2.18.

<table>
<thead>
<tr>
<th>Phase</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$:</td>
<td>44</td>
<td>106</td>
<td>134</td>
<td>129</td>
<td>100</td>
<td>110</td>
<td>623</td>
</tr>
<tr>
<td>$pP$:</td>
<td>6</td>
<td>26</td>
<td>48</td>
<td>46</td>
<td>43</td>
<td>33</td>
<td>202</td>
</tr>
<tr>
<td>$PcP$:</td>
<td>8</td>
<td>14</td>
<td>25</td>
<td>22</td>
<td>27</td>
<td>29</td>
<td>125</td>
</tr>
<tr>
<td>$PKIKP$:</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>33</td>
<td>96</td>
</tr>
<tr>
<td>$PKP$:</td>
<td>1</td>
<td>10</td>
<td>15</td>
<td>16</td>
<td>13</td>
<td>6</td>
<td>61</td>
</tr>
<tr>
<td>$PKiKP$:</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>$PP$:</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>$Pn$:</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>$p$:</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>$Pdiff$:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$S$:</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>$Sn$:</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>$SKiKP$:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td>64</td>
<td>179</td>
<td>255</td>
<td>246</td>
<td>225</td>
<td>251</td>
<td></td>
</tr>
</tbody>
</table>

Identified phases associated with the high-quality travel-time residuals of Figure 2.18. The columns correspond to the histograms in Figure 2.18a–f, listed from fine to coarse resolution. The last column sums the number of high-quality identifications across all scales for the phase specified. The bottom row totals the number of high-quality travel-time residuals collected at each scale and is the same number listed above each histogram in Figure 2.18. We did not actually observe an arrival associated with the phase $SKiKP$ (nor $PKiKP$, see the example in the supplemental material) but kept the entry as an example of the issue of phase ambiguity associated with our phase-matching scheme. See the discussion in sections 2.8.3 and 2.9.2.

above each plot. The bracketed numbers are the percentages of high-quality residuals contained in the histogram, relative to the total available at each resolution. We immediately see that the lowest percentage of high-quality residuals is computed at scale 1 for 20 Hz data. Otherwise, the percentage of high-quality residuals hovers near or just above 60% at all other resolutions. Computing these travel-time residuals is the first step in the inverse problem of tomography, and their bias alerts us to velocity perturbations encountered along the ray paths. From the discussion in section 2.5.1 we know that our AIC-based changepoint estimator is unbiased. Hence, these residuals are not an artifact of our arrival-time estimation procedure, but rather a direct measure of velocity anomalies in the Earth.
2.9. Application to Mermaid Seismograms

2.9.4 *The updated events catalog*

We have compiled an updated events catalog of all of the seismograms in our data set, without applying any quality criteria. Figure 2.19 shows two entries of the updated catalog available in the supplemental material. The upper half of each block (ending at the line beginning “Updated”) lists general event information, including the date that the IRIS database was last queried for metadata related to that event, and the “Initial” (which contained no magnitude type) and “Updated” event parameters as returned on the date of the last query. The initial and updated event information are often (but not always) the same, as shown in the examples here.

The lower half of each block in Figure 2.19 lists phase-arrival information starting with the name of the phase initially associated with this seismogram as reported in the `events.txt` file distributed by GéoAzur, but reported there without an arrival time. The lines that follow list the multiscale arrival picks, travel-time residuals, and uncertainties found in this study. The first alerts the reader to the number of scales used in the wavelet decomposition: always either 5 if the sampling rate was 20 Hz or 3 if the sampling rate was 5 Hz. Next, a header-line describes the forthcoming columns: the phase name associated with this travel-time residual; the observed arrival time in s into the seismogram, computed by our changepoint estimation procedure; the time residual in s between our arrival-time pick and the theoretical arrival time of the listed phase; the SNR of this arrival; and the mean and two-standard-deviation of the estimated error distribution in s, per the M1 procedure of section 2.5.1 after 1000 realizations. Each line corresponds to one scale starting at scale 1 and ending at the scale corresponding to the resolution of the approximation, either scale $\frac{5}{3}$ or scale $\frac{3}{5}$ as shown here.

For the event described in the first block in Figure 2.19, only the travel-time residual and two-standard deviation of the uncertainty estimate made at scale 1 are smaller than 6 s and 1 s, respectively. Hence that is the only travel-time residual included in Figure 2.18, with its associated phase, $P$, listed in Table 2.2. The estimates made at all other resolutions are characterized by a large travel-time residual and/or a high two-standard-deviation uncertainty, and thus are not plotted in Figure 2.18 or listed in Table 2.2. Similarly, none of the estimated arrival times or their
Figure 2.19: Two event blocks in the catalog that accompanies this study in the supplemental material. The upper portion of each block (ending with “Updated”) lists event metadata, including the originally-reported event information recorded in the SAC header, and the updated event information per the last query to IRIS, noted as the “Last updated” date. The lower portion of each block lists phase information, beginning with the phase associated with this seismogram as initially reported without an arrival time. Next we list, from the lowest to the highest scale, the phase name; arrival time; travel-time residual; SNR; and the mean and two-standard-deviation of the error estimate derived from 1000 realizations of the M1 method, identified by our arrival-time estimation procedure. Considering both events shown, only the estimated travel-time residual (−2.32 s) and phase (P) associated with the arrival-time estimate made at scale 1 for the top event is included in Figure 2.18 and listed in Table 2.2. All other travel-time residuals and/or the two-standard-deviation errors of the uncertainty estimates are too great to be considered of “high quality.”
associated phases in the second event block in Figure 2.19 are plotted in Figure 2.18 or listed in Table 2.2, because the two-standard-deviation uncertainty at scales 1 through 3 are too large to be considered likely arrivals, and the travel-time residual at scale $3$ is too great to truly match the assumed $P$-wave arrival listed there.

2.10. Conclusion

We have developed a method to detect arrival times of seismic energy at multiple scales in noisy seismograms, and estimate their uncertainties. Our procedure centers on the wavelet-multiscale application of an AIC-based changepoint-detection scheme, which we have applied to the problem of computing travel-time residuals in low-SNR hydroacoustic records. Our uncertainty estimation procedure provides a quantitative metric for weighting residuals during tomographic inversions.

We have defined two changepoint estimators useful for arrival-time identification: the minimum, $k_m$, and the weighted average $k_w$, respectively, of the AIC function. We investigated two methods to compute the uncertainty associated with these estimates, one using brute-force Monte Carlo resimulation, and another through the analysis of the shape of the AIC curve itself. The former gives an estimated error distribution while the latter assigns confidence intervals to the estimates. We showed that the weighted-average $k_w$ estimator is unbiased, unlike the oft-used $k_m$ estimator. For this reason we suggest the adoption of the former. We discussed the nuances of applying our procedure in the time and timescale domains and recommend the former to ensure the lowest uncertainties.

We apply our preferred method to 445 MERMAID seismograms to identify seismic phase arrivals and compute their travel-time residuals. Our multiscale AIC-based method is able to identify seismic phases with low uncertainty. The majority of the events in our MERMAID catalog corresponds to $P$ waves, but we are also able to identify a few $S$-waves as well as core phases.

We discussed the multiscale distributions of high-quality travel-time residuals in the MERMAID data set. In the supplemental material we provide an updated events catalog that details those
2.11 Data and Resources

We use MERMAID seismograms with identified events available from geoazur.unice.fr/ftp/mermaid/. These data were last accessed March 2019. We rely on irisFetch.m version 2.0.10, available from IRIS, to query seismic catalogs available through the International Federation of Digital Seismograph Networks (FDSN). Manual event matching, where required (as described in section 2.8.3), was performed using the International Seismological Centre (2016) online bulletin, last accessed March 2019. We use MatTaup, written in MATLAB by Qin Li while at the University of Washington in 2002, to compute theoretical travel times in the ak135 velocity model of Kennett et al. (1995). We maintain all of those codes, with minor modifications, at github.com/joelsimon/omnia, which furthermore contains all of our software developed for this study. The supplemental material contains our catalog of MERMAID arrival times and their uncertainties, examples of our method applied to non-Gaussian synthetic and more seismic data, and details on our treatment of the edges of the seismogram during the forward and inverse wavelet transform.

2.12 Acknowledgments

Part of this work was supported by the National Science Foundation (DGE-1656466 to JDS and EAR-1550732, EAR-1736046, and OCE-1917058 to FJS). MERMAID data were obtained with financing from the European Research Council (Advanced Grant 226837 to GN). Constructive comments by Eric Chael, the Associate Editor of the journal Bulletin of the Seismological Society of America, and an anonymous reviewer, were greatly appreciated.
2.13 Supplemental Material

2.13.1 Description

This supplement contains details concerning the algorithmic implementation of our method, as well as more examples of it applied to synthetic and real data. Specifically it contains, in order: a mapping similar to that described in section 2.6.3, but described here for the inverse transform; the treatment of the edges by our multiscale method; our method applied to non-Gaussian synthetic time series; and our method applied to real data containing a supposed PKiKP-wave arrival. The latter is provided both to display an interesting seismogram as well as to illustrate the issue of phase ambiguity with our automated method.

Finally, as supplementary data to this study we also provide a separate plain text catalog of MERMAID arrival times and their uncertainties, the format of which is detailed in section 2.9.4.

2.13.2 Timescale to time mapping: inverse transform

In preparing to describe our scheme to handle the edges of the time series we first define the inverse map which complements that described in section 2.6.3. The mapping between a timescale domain coefficient index \( l \) at scale \( j \), and the time-domain sample-span to which \( l \) is sensitive under the inverse wavelet transform, is

\[
\hat{F}_j(l) = \hat{k}_{j,l}^*. 
\]

As in the case of \( \hat{F}_j \), the output of \( \hat{F}_j \) is a time-domain sample span. Similarly to \( \hat{k}_{j,l}^* \), the support of a wavelet or scaling basis function of the inverse wavelet transform is completely described by \( \hat{k}_{j,l}^* \).

Similar to equations (2.54) and (2.56) we define the left and right boundaries by, respectively,

\[
\hat{k}_{j,l}^+ = \min(\hat{k}_{j,l}^*), 
\]

\[
\hat{k}_{j,l}^- = \max(\hat{k}_{j,l}^*). 
\]
2.13. Supplemental Material

2.13.3 Handling edges

In the main text we ignored the edges of any times series considered. Rather than building them into the construction of the transform itself, we opt to remove spurious wavelet values, those influenced by the edges, before changepoint estimation.

2.13.3.1 Edge sensitivity in the timescale domain

At every scale, the last, and first, timescale domain coefficient indices, which sense the left, and right, edges of $x$ during forward wavelet transformation are

\[
\hat{l}_L = \max\{l : 1 \in \hat{k}_L^j\}, \quad (2.66)
\]
\[
\hat{l}_R = \min\{l : N \in \hat{k}_R^j\}, \quad (2.67)
\]

respectively. Upon mapping to the time domain the first and last sample indices which are not influenced by the edges after forward wavelet transformation are

\[
\hat{k}_L = \min(\hat{k}_{lL}^j, \hat{k}_{L+1}^j), \quad (2.68)
\]
\[
\hat{k}_R = \max(\hat{k}_{R}^j, \hat{k}_{R-1}^j), \quad (2.69)
\]

respectively. In the forward wavelet transformation every timescale domain coefficient is independent and unaware of the edge unless it receives contributions directly from the edge.

At every scale, before applying the multiscale analog of equation (2.26) directly on the timescale domain coefficient series $d_{j,l}$ and $a_{j,l}$, en route to finding the estimators $l_w$ and $\overline{l}_w$ of equations (2.51) and (2.52), we remove from consideration the edge-sensitive timescale domain coefficients.
2.13.3.2 Edge sensitivity in the time domain

Similarly, at every scale, the last and first timescale coefficient indices which sense the left and right edges, respectively, of $x$ during inverse wavelet transformation are

\[
\hat{l}_L = \max\{l : 1 \in \hat{k}_{j_{L}}^*\}, \quad (2.70)
\]
\[
\hat{l}_R = \min\{l : N \in \hat{k}_{j_{R}}^*\}. \quad (2.71)
\]

During reconstruction via the inverse wavelet transformation, temporally overlapping wavelets potentially propagate edge-effects deeper into the partially-reconstructed time series, $x_j$ or $\pi_j$, than in the case presented in section 2.13.3.1. Therefore, during reconstruction we cannot simply augment $l$ by $\pm 1$ to locate the time-domain sample indices which are surely not influenced by the edge, as we did in equations (2.68) and (2.69). Indeed, this overlap implies that

\[
\max(\hat{k}_{j_{L}}^\perp) \geq \min(\hat{k}_{j_{L+1}}^\perp), \quad (2.72)
\]
\[
\max(\hat{k}_{j_{L}}^\perp) \geq \min(\hat{k}_{j_{L+1}}^\perp), \quad (2.73)
\]
\[
\min(\hat{k}_{j_{R}}^\top) \leq \max(\hat{k}_{j_{R-1}}^\top), \quad (2.74)
\]
\[
\min(\hat{k}_{j_{R}}^\top) \leq \max(\hat{k}_{j_{R-1}}^\top). \quad (2.75)
\]

The left-hand sides in equations (2.72)–(2.75) are time-domain sample indices that are assured to be free of edge influence: the first and last being given respectively by equations (2.72) and (2.74) after forward wavelet transformation, and equations (2.73) and (2.75) after inverse wavelet transformation. To be conservative we keep the larger and smaller of these as the first and last time domain sample indices, respectively, which are not influenced by the edges of $x$, defining

\[
\hat{k}_L = \max\{\hat{k}_{j_{L}}^\perp, \hat{k}_{j_{R}}^\perp\} + 1, \quad (2.76)
\]
\[
\hat{k}_R = \min\{\hat{k}_{j_{R}}^\top, \hat{k}_{j_{R}}^\top\} - 1. \quad (2.77)
\]
At every scale, before applying the multiscale analog of equation (2.26) on the time-domain subspace projection series $x_j$ and $\pi_j$, en route to finding the estimators $k_{w_j}$ and $\bar{k}_{w_j}$ of equations (2.49) and (2.50), we remove the edge-sensitive time-domain sample indices from consideration.

2.13.4 Non-Gaussian models

We have intimated that our AIC-based event detection method will enjoy broad application beyond the scope of this study. Indeed, our multiscale changepoint estimation procedure is completely agnostic of seismology. The general form of equation (2.26) in terms of natural logarithms multiplied by the sample variances of the segments implies that it can be readily applied to any time series that can be modeled as concatenated samples from distributions in the larger exponential family. We forgo showing such examples here.

Future applications that may opt for entirely different synthetic model formulations may simply rederive the appropriate AIC function and deploy it as part of our workflow. On the other hand, our AIC formulation quite simply compares ratios of variances and thus it is likely to remain useful even when applied to time series whose generating distributions are not in the exponential form. Figures 2.20 and 2.21 recreate Figure 2.5, assuming that $x$ is drawn from two concatenated Student $t$-distributions, or from two $F$-distributions, respectively, with different variances. Our changepoint detection method remains on target, although the longer tails of these generating distributions will lead to broader error distributions than is the case of the assumed Gaussian models of this study.

2.13.5 A final data example

Presented in the same layout as Figures 2.14–2.17, Figure 2.22 discriminates the (reported) inner-core reflection $PKiKP$ wave from the reflected mantle phase $PP$. Note that the former is more likely a $PKIKP$ wave that bottoms just inside the inner core, as this phase also appears around 114° in the ak135 velocity model for an earthquake at 607 km depth.
2.13. Supplemental Material

Figure 2.20: A recreation of Figure 2.5 using a non-Gaussian generating distribution. Here \( x \) is a sample of length \( N = 1000 \) drawn from a Student \( t \)-distribution with 10 degrees of freedom. Sample indices \( k = [501, \ldots, 1000] \) are multiplied by \( \sqrt{2} \), therefore generating a time series with \( \text{SNR} \approx 2 \) and a true changepoint at sample index 500. In this example \( k_m = 499 \) and \( k_w = 507 \).

Figure 2.21: A recreation of Figure 2.20, this time assuming that the time series under consideration \( x \) is drawn from the \( F \)-distribution with parameters \((10, 10)\). Like Figure 2.20, sample indices \( k = [501, \ldots, 1000] \) are multiplied by \( \sqrt{2} \) to yield \( \text{SNR} \approx 2 \) and \( k_o = 500 \). In this example the estimated changepoints are \( k_m = 493 \) and \( k_w = 495 \), illustrating again that our procedure remains broadly valid, while uncertainties and confidence intervals will require suitable adaptation.
Figure 2.22: A mid-SNR seismogram detected in the southern Indian Ocean corresponding to an earthquake in the Sea of Okhotsk. Here our procedure picks a reported $PKiKP$ wave (but more likely a $PKIKP$ wave; see section 2.9.2) at scales 1 and 2 and a $PP$ wave at scale 3, in both domains. Note that the sampling rate of this seismogram is only 5 Hz, and therefore was decomposed only to three scales (see Table 2.1).
Chapter 3

Recording Earthquakes for Tomographic Imaging of the Mantle Beneath the South Pacific by Autonomous MERMAID Floats

3.1 Abstract

We present the first 16 months of data returned from a mobile array of 16 freely-floating diving instruments, named MERMAID for Mobile Earthquake Recording in Marine Areas by Independent Divers, launched in French Polynesia in late 2018. Our 16 are a subset of the in total 50 MERMAIDS deployed over a number of subsequent cruises in this vast and understudied oceanic province as part of the collaborative South Pacific Plume Imaging and Modeling (SPPIM) project under the aegis of the international EarthScope-Oceans consortium. Our common objective is the hydroacoustic recording, from within the oceanic water column, of the seismic wavefield generated by earthquakes worldwide, and the nearly real-time transmission to satellite of these data, collected directly above and on the periphery of the South Pacific Superswell. This region, characterized by anomalously elevated oceanic crust and myriad seamounts, is believed to be the surface expression of a deeply-rooted mantle plume. Tomographically imaging Earth’s mantle under the South Pacific with data from equipment of this novel kind requires a careful examination of the earthquake-to-MERMAID travel-time residuals of the high-frequency P-wave detections within the windows selected for reporting by the discrimination algorithms on board. We discuss a work-

*This chapter will be submitted to a journal with the authors Simon, J.D., Simons, F.J., & Irving, J.C.E.
flow suitable for a fast-growing mobile sensor database to pick the relevant arrivals, match them to known earthquakes in the global earthquake catalogs, calculate their travel-time residuals with respect to a global seismic reference model, characterize their quality, and estimate their uncertainty. We detail seismicity rates as recorded by MERMAID in its first 16 months, break these statistics down by magnitude to quantify the completeness of our catalog, and discuss magnitude-versus-distance relations of detectability for our network. The projected lifespan of an individual MERMAID is five years, allowing us to estimate the final size of the data set that will be available for future study. To prove their utility for seismic tomography we compare the MERMAID data quality against traditional land seismometers and their low-cost Raspberry Shake counterparts, using waveforms recovered from instrumented island stations in the geographic neighborhood of our floats. Finally, we provide the first analyses of travel-time anomalies for the new ray paths sampling the mantle under the South Pacific over the first 16 months of operation of our array.

3.2 Introduction and Motivation

More than 70% of Earth’s surface is covered by water. Seismic data recorded in the global oceans are sparse in both spatial and temporal coverage, especially in the Southern Hemisphere. Figure 3.1 proves this point by mapping, in blue, the location of every seismic station retrievable from the Incorporated Research Institutions for Seismology (IRIS). While the map is indubitably incomplete, and the recorded presence of a station does not imply that the data are also available, it illustrates the sparsity of seismic sampling in the oceans, especially in the Southern Hemisphere.

Historically, seismic studies in and of the oceans have proven complex and costly. What follows is a brief recapitulation of the relatively short history of the field (only about one hundred seismic records from the deep-ocean bottom existed by the 1960s, according to Bradner, 1964) to place MERMAID in its proper historical perspective (see also Simons et al., 2009).

Early attempts to instrument the oceans for regional and global seismology came in the form of encased seismometers dropped in free fall onto the seafloor from a ship, with or without an-
Figure 3.1: All seismic stations (46295 small blue triangles) ever reported to the Incorporated Research Institutions for Seismology (IRIS), and the locations (orange and gray large triangles) of the Mobile Earthquake Recording in Marine Areas by Independent Divers (MERMAID) floats in the South Pacific, at the time of their deployment. The 16 MERMAID floats maintained by Princeton whose data are discussed in this study are highlighted in orange. The black rectangle is the boundary of the region searched for nearby island stations, the details of which are discussed in section 3.10.
3.2. Introduction and Motivation

chored tether, and with a variety of mechanisms for recovery and data retrieval (Ewing & Vine, 1938; Bradner, 1964; Whitmarsh, 1970). Progress toward true instrument autonomy came in the form of freely-drifting telemetered devices, either neutrally-buoyant mid-column floating versions of ocean-bottom sensors (e.g. Bradner et al., 1970), or sonobuoys, with a hydrophone loosely suspended from a surface buoy (e.g. Reid et al., 1973). Most of these experiments were short-lived due to power restrictions. Longer-lived moored sonobuoys (e.g. Kebe, 1981) and moored hydrophones (e.g. Fox et al., 1993) could provide continuous hydroacoustic data at the expense of requiring seafloor cables to power them, restricting their spatial range of coverage.

In the last three decades, ocean bottom seismometry with long-life robust, three-component broadband sensors has positively flourished (Zhao et al., 1997; Webb, 1998; Webb & Crawford, 2003; Suetsugu & Shiobara, 2014). Nevertheless, to this day such instruments remain physically large and expensive to install (Beauduin et al., 1996; Collins et al., 2001), requiring a specialized research vessel both for deployment and recovery (Stephen et al., 2003), as establishing semi-permanent installations (e.g. Duennebier et al., 2002; Romanowicz et al., 2006) worldwide remains a developing goal for the international community (Montagner et al., 1998; Romanowicz & Giardini, 2001; Favali & Beranzoli, 2006).

Two ambitious multi-station, multi-instrument cabled arrays have been rooted on the seafloor off the coast of Japan (Hirata et al., 2002; Shinohara et al., 2014) and in the Canadian Northeast Pacific (Barnes et al., 2013; Matabos et al., 2016) for the long-term monitoring of subduction zones. These installations provide high-quality data with low latency, but they require massive upfront costs, demand costly maintenance, are limited by cables and, being permanent, cannot be rapidly reinstalled or reassigned in the case of developing seismic crises (e.g. Duennebier et al., 1997).

The current fleet of recoverable ocean bottom seismometers (OBS) is autonomous but unable to transmit data while deployed, hence data acquisition and processing are separated by months or years, unless catastrophe precludes recovery (Tolstoy et al., 2006). More recently, wave-powered gliders which float at the surface and may be remotely controlled to remain in the vicinity of an
3.2. Introduction and Motivation

Ocean-bottom stations have been used as a go-between to relay data from seafloor to shore via acoustic modem and satellite uplink (Berger et al., 2016). This coupling of technologies allows the delivery of seismic data from the seafloor in near real-time. While they have shown promise, such solutions remain fragile and costly to operate, and they have not yet enjoyed large-scale deployment. Other solutions to the logistical problem of data recovery are currently being tested. These include ocean-bottom systems that periodically release data pods from the seafloor, each with a self-contained telecommunications unit to relay data via satellite upon surfacing (Hammond et al., 2019). Finally, while the age where the cables themselves may act as seismic sensors appears to have arrived (e.g. Sladen et al., 2019; Williams et al., 2019), such technology is in its infancy.

Despite those advances in technology, no single seismic instrument has solved all the issues just presented: the ability to deliver high-quality data with autonomy, low cost, low latency, and nimbleness. Nor should we assume that any single instrument can be designed to optimize for all. Our instrument, MERMAID, fills a gap in instrumentation by providing low-cost hydroacoustic records suitable for global seismology (Simons et al., 2006b) from the oceans in near real-time (Hello et al., 2011) without the requirement of a research vessel for deployment and, being unrecovered, negating the need for a recovery cruise.

While MERMAID’s hydroacoustic time series, collected by a single limited-bandwidth hydrophone floating at mid-column water depths, forever will remain less “complete” seismic data sets in comparison with a well-coupled three-component broadband ocean-bottom seismometer, its benefits are its lower manufacturing costs, its logistical simplicity, its algorithmic flexibility (Sukhovich et al., 2011, 2014) in selecting promising seismic phases to report with each surfacing—and its longevity, currently projected to be about five years (∼250 dive cycles) on a single battery charge. Hence, MERMAID can be thought of as a 21st century sonobuoy without the previous century’s drawbacks. Fulfilling the promise of the first-generation MERMAID instrument (Simons et al., 2009) and substantiating the record accumulated by MERMAIDs of the second generation (Sukhovich et al., 2015; Nolet et al., 2019), the over 1300 records presented here, collected by the current third generation of instruments, constructed by OSEAN SAS of Le Pradet, France,
3.2. Introduction and Motivation

are closing the seismic data gap in the world’s oceans.

Studying the interior of the Earth using seismic tomography (Nolet, 2008; Romanowicz, 2008; Rawlinson et al., 2010), primarily of $P$ delay times, was, and remains to date, MERMAID’s primary strength and objective. Joubert et al. (2016) and Nolet et al. (2019) have shown that the accuracy of MERMAID’s position underwater, interpolated from multiple surfacings, and the accuracy with which the arrival time of seismic $P$ phases can be determined from the sometimes noisy acoustic records, are of sufficiently high quality to constrain velocities for tomographic inversion. Simon et al. (2020) presented a new algorithm for the multiscale estimation of event arrival times and their precision, which closes the loop from detection and discrimination of $P$ waves in the ocean, to the accurate determination of their travel times, to the assessment of their uncertainties.

In this paper we bank on all of these developments and present the first 16 months of data returned by the 16 MERMAIDs owned and operated by Princeton University that were deployed in French Polynesia in late Summer 2018. We compare their waveforms with traces available from 20 seismic island stations in the same region, and with records from a set of five comparatively less expensive, but increasingly more abundant, Raspberry Shake (Anthony et al., 2019) instruments.

We study the statistics of our growing catalog of seismic data, a lasting product of this study, to comment on its completeness, and to estimate the total number of tomographic-quality records that can be expected to be returned per MERMAID over its projected five-year lifetime. We compute MERMAID travel-time residuals against the one-dimensional (1-D) ak135 velocity model (Kennett et al., 1995), corrected for bathymetry and MERMAID’s cruising depth. We estimate their uncertainties, compute signal-to-noise ratios, and compare these statistics with a complementary data set derived from traditional seismometers and Raspberry Shake stations installed on ocean islands. These travel-time residuals will be the inputs for future tomographic studies, with uncertainties to serve as weights in the inversion.

For a taste for the likely signals from the Earth’s mantle that will emerge from our data collection we further correct the residuals for the three-dimensional (3-D) elliptical $P$-wave speed model LLNL-G3Dv3 of Simmons et al. (2012), and project them onto their 1-D ray paths to reveal average velocity perturbations with respect to this Earth model that tomography will further image.
3.3. The MERMAID Instrument

The purpose of the MERMAID float is to return seismic data of tomographic quality from the global oceans in near real-time. The instrument (Figure 3.2) and its dive cycle (Figure 3.3) were inspired by oceanic floats (Swallow, 1955; Rossby & Webb, 1970; Davis et al., 1992, 2001), which have become ubiquitous in the global oceans (see Gould, 2005, for historical perspective). The international Argo program has been continuously providing the scientific community with a wealth of temperature, salinity, and trajectory data over the last several decades (Lavender et al., 2000; Roemmich et al., 2009; Davis, 2005; Abraham et al., 2013). Along with the payload required for in situ observations and hydrographic profiling, a contemporary APEX Argo float is equipped with a hydraulic pump which modulates an expandable bladder that allows it to be neutrally buoyant at many mid-column depths, a Global Positioning System (GPS) for location tracking, and a satellite link for data transmission.

Argo floats collect and transmit data over repeated dive cycles. A typical cycle begins with the float deflating its bladder to achieve negative buoyancy so that it may sink to a predetermined parking depth (generally between 1000 m to 2000 m below the sea surface), at which point it passively drifts at depth for a set amount of time (usually around 10 days), before finally reinflating...
3.3. *The MERMAID Instrument*

**Figure 3.3:** The first five dive cycles completed by MERMAID P012 after its deployment on the 10 August 2018. The parking depth for MERMAIDs discussed in this study was 1500 m, though, like nearly all mission parameters, this may be adjusted via the two-way Iridium satellite link. Typical (again, adjustable) descent speeds are on the order of -2.8±1.2 cm/s (-100.3±44.8 m/h), translating into 15.5±5.3 h to sink from the surface to the parking depth. Typical ascent speeds are on the order of 8.0±0.2 cm/s (289.8±8.4 m/h), meaning it takes MERMAID 5.1±0.2 h to ascend from depth to the surface. MERMAID’s onboard detection algorithm prompts immediate surfacing when it records a signal it considers with a high likelihood to be a teleseismic P wave, which explains the abbreviated second and third dives as compared to the first in this figure. This rapid triggering and transmission allowed us to receive record sections like those in Figure 3.5 within hours of large earthquakes. Depth data as determined by the MERMAID interpolation algorithm written by Sébastien Bonnieux.

its bladder to slowly rise back to the surface. During this ascent it samples and processes a roughly vertical column of water via a conductivity-temperature-depth (CTD) sensor. Once at the surface it acquires a GPS fix, transmits the new data via satellite, and repeats the process. Because they are autonomous and drift at the whim of ocean currents, Argo floats are practically guaranteed to sample the water column at a previously unsampled location every time they ascend. As of 11 April 2020 there were 4060 Argo floats actively reporting from within every ocean on Earth, and on average some 800 are being deployed yearly to maintain the fleet and, like MERMAID, they are not designed to be recovered.

The first-generation MERMAID float was a modified Sounding Oceanographic Lagrangian Observer (SOLO) float (Davis et al., 2001), fitted with a hydrophone and a custom algorithmic processing unit so that it returned seismologically viable hydroacoustic data recorded at its parking depth (Simons et al., 2006b, 2009). The second-generation MERMAID (Hello et al., 2011; Sukhovich et al., 2015) was a modified APEX float built by Teledyne Webb Research. The third-
3.3. The MERMAID Instrument

generation MERMAID is a redesign from the ground up by Yann Hello at GéoAzur and French engineering firm OSEAN SAS. It is an autonomous robotic float consisting of a High Tech HTI-96-MIN_HEX hydrophone, a Gardner DENVER pneumatic pump, a Garmin GPS 15 unit, a Motorola 9522 two-way Iridium communication module, Electrochem lithium batteries, and dedicated on-board detection and discrimination software (Sukhovich et al., 2011). Once deployed MERMAID sinks to a predetermined depth (usually 1500 m) and records the ambient acoustic wavefield while freely drifting with the mid-column currents. If triggered by seismic activity, or once a threshold time is reached, MERMAID surfaces, transmits the new data, downloads mission-command files via satellite, and repeats the process. Figure 3.3 shows the first five dive cycles completed by MERMAID P012 after its deployment on the 10 August 2018, and Figure 3.4 shows the drift trajectories of all 16 MERMAIDS discussed in this study.

The current onboard algorithm used to monitor and process the ambient acoustic wavefield (Sukhovich et al., 2011, 2014) was designed specifically to trigger on tomographic-quality teleseismic P-wave arrivals sensitive to mantle structure. Once parked at depth the hydrophone is switched on and data acquisition starts. The hydroacoustic data are processed in real-time by a short-term average over long-term average (STA/LTA) algorithm (Allen, 1978), and written to a Secure Digital (SD) card, which retains those data for one year. If the adjustable STA/LTA threshold is exceeded, a windowed section of those data are further interrogated via wavelet decomposition (Simons et al., 2006a), and its energy distribution across six wavelet scales is compared with statistical models of various signals known to exist in the oceans (many of which are not generated by seismic events).

A quality criterion is assigned to the signal that encodes the probability that the record under inspection includes a P-wave arrival, and if it is high enough, MERMAID is automatically triggered to cease data acquisition, regardless of how long it has remained at depth, and to surface and transmit that record immediately. While at the surface it will also offload any other records on the buffer that were marked for transmission, but whose criterion values were not high enough to warrant immediate surfacing. Currently as its default, and for all records discussed in this
3.4. The EarthScope-Oceans Consortium

study, MERMAID transmits the Cohen-Daubechies-Feauveau (2,4) wavelet and scaling coefficients (Cohen et al., 1992) from scales two through six of a time series originally sampled at 40 Hz and filtered between 0.1 Hz and 10 Hz before digitization. This means that, after reconstruction via inverse wavelet transformation, the MERMAID records presented here are hydroacoustic (pressure) time series of seismic conversions, sampled at 20 Hz.

MERMAID delivers seismic data from the oceans in near real time, with immediate surfacing and data transmission within hours of the largest events. MERMAIDS are individually programmable and mission parameters such as parking depth, maximum time to remain there, criterion thresholding values to trigger surfacing, and so on, all may be monitored and adjusted thanks to two-way Iridium communication. While the ability exists to request data from the MERMAID buffer for up to one year prior (which we have done with success), we found the default trigger algorithm to perform exceptionally well, and in this study we will restrict our discussion to only those triggered records which MERMAID sent us on its own accord. Indeed the default onboard algorithm was left untouched for the entirety of the deployment for all 16 floats discussed here.

3.4 The EarthScope-Oceans Consortium

The EarthScope-Oceans consortium was founded in 2016, and now counts members from the US (Princeton University, among whom the authors, IRIS Seattle, DBV Technologies North Kingstown), Japan (Kobe University, JAMSTEC, ERI), France (Géoazur Sophia Antipolis, EOST Strasbourg, IFREMER Plouzané, OSEAN Le Pradet), South Korea (KIGAM Daejeon), New Zealand (GNS Science, Lower Hutt), the UK (University of Oxford), and China (SUSTech, Shenzhen).

EarthScope-Oceans represents a multidisciplinary group of geoscientists who are coordinating efforts to create a global network of sensors to monitor the Earth system from within the oceanic environment. It intends to shepherd national projects into the international forum where globally relevant, applicable, and mutually agreed-upon decisions can be made on technological aspects of instrument development, science objectives and priorities on different time scales, data
management, dissemination, and archiving, and education and outreach efforts; much like IRIS (iris.edu) or ORFEUS (orfeus-eu.org) are doing for the land-based seismological communities today.

EarthScope-Oceans is partnered with the Joint IOC-World Meteorological Organization Technical Commission for Oceanography and Marine Meteorology and abides by the UNESCO agreements on global ocean observation systems, which spell out end-of-life provisions for MERMAID.

The Federation of Digital Seismic Networks (FDSN) has granted MERMAID data its own seismic network code (see fdsn.org/networks/detail/MH/ for detailed information). MERMAID floats generate location data, instrumental meta-data, and acoustic waveforms. All data recorded during the lifetime of MERMAID floats will be openly accessible from the IRIS Data Management Center (DMC) as soon as technically feasible and with a maximum two-year delay from collection.

### 3.5 The MERMAID SPPIM Deployment

The 16 Princeton-operated third-generation MERMAIDS whose data are the subject of this study are just one component of the South Pacific Plume Imaging and Modeling (SPPIM) project, an array of 50 MERMAIDS deployed into the South Pacific to study the underlying mantle composition and temperature with seismic tomography. Drifting united under the EarthScope-Oceans banner, these MERMAIDS are supported and maintained by our global consortium (earthscopeoceans.org).

A 24-hr trial run completed 12 April 2018 was led by Kobe University’s Hiroko Sugioka and JAMSTEC’s Masayuki Obayashi from the R/V Fukae Maru. During this test deployment MERMAID N03 recorded the magnitude $m_b$ 4.9±0.042 earthquake originating at 59.4±5.8 km depth, 56±6.6 km east of Ishinomaki, Japan (according to earthquake.usgs.gov), some 824 km distance from the instrument, which floated 500 m below the surface at the time.

The complete SPPIM array, shown as gray and orange upside-down triangles in Figure 3.1, was
deployed over several cruises led by Yann Hello (see Figure 3.2), research engineer at IRD/Géoazur, chief designer of MERMAID in its current third-generation, as currently commercially available from OSEAN SAS of Le Pradet, France.

On the first leg (Nouméa, New Caledonia to Mata-Utu, Wallis & Futuna, 21–28 June 2018), Yann Hello deployed two GéoAzur units from IRD/Genavir vessel R/V Alis. On the second leg (Mata-Utu to Papeete, Tahiti, French Polynesia, 3–13 August 2018), Hello deployed five Princeton units from the R/V Alis. On the third leg, Frederik Simons deployed 11 Princeton units from the R/V Alis, which departed Papeete on the 28 August 2018, returning to the same port on the 16 September of the same year (doi:10.17600/18000519). During this leg, the RV Alis completed a nearly circular trajectory, which can be traced in Figure 3.4 by connecting the deployment locations (dark blue) of MERMAID P013, P016, P017, etc., and continuing clockwise back to Tahiti (note that no instruments named P014 or P015 were ever deployed). It was during this cruise that the deployment of the 16 MERMAIDS operated by Princeton was completed.

Five Japanese units were launched from the R/V Mirai by Masayuki Obayashi, sailing from Shimizu, Japan to Valparaiso, Chile between 11 December 2018 and 24 January 2019.

The fourth leg (Papeete–Nouméa, 4–29 August 2019) was led by Hello, Obayashi, Zhen Guo and Yong Yu (SUSTech) from the R/V L’Atalante (doi:10.17600/18000882). This cruise saw the completion of the array with the deployment 23 SUSTech MERMAIDS and an additional four from Kobe University.

### 3.5.1 MERMAID drifts with the currents

It bears repeating that MERMAIDS drift with the ocean currents—they do not land on the seafloor like traditional ocean-bottom seismometers (although a “Lander” version is currently under development). Figure 3.4 shows the drift trajectories of all 16 Princeton-operated floats discussed in this study. Every dot represents one GPS fix taken by MERMAID while at the surface, color-coded to show the amount of time elapsed since its deployment (with dark blue representing the launch day, and dark red the last GPS fix of 2019). Each MERMAID trajectory is labeled by the corresponding
3.5. The MERMAID SPPIM Deployment

Figure 3.4: (caption next page)
3.5. The MERMAID SPPIM Deployment

Figure 3.4: (previous page) (a) MERMAID trajectories with time, and the locations of stationary “nearby” island stations. This map is a zoom-in of the rectangle drawn in Figure 3.1, itself representing bounding-box with edges framed roughly 2° beyond the extent (in all four cardinal directions) of the complete 50-mermaid SPPIM project. Here, only the drift-trajectories of the 16 Princeton-operated MERMAIDs that contributed data to this study are shown. The trajectories are color-coded by the time elapsed since deployment, where dark blue represents the location at the time of deployment, and dark red represents the last GPS fix of 2019. Therefore, these trajectories are a map of ocean currents at 1500 m, the parking-depth of MERMAID. The locations of nearby island seismic installations (Table 3.7 and 3.8) are marked by upside-down triangles, except in the case of a single, large right-side up triangle representing the collection of stations on Tahiti, French Polynesia. (b) Bathymetry and topography from the GEBCO 2019 model (Weatherall et al., 2015; GEBCO Bathymetric Compilation Group, 2019).

serial number of the float, excluding the “P0” prefix. By connecting these dots we therefore have an approximate (Davis, 2005) map of the ocean currents at 1500 m depth, the depth at which MERMAID spends the most time. See Nolet et al. (2019) for drift statistics broken down into surface and abyssal components.

Also labeled on this map are the locations of other seismic sensors against which MERMAID data is compared later in this study. Those station locations are marked by upside-down triangles, with the notable exception being the collection of stations on Tahiti, French Polynesia, marked by a larger right-side up triangle, to represent the many stations installed there and listed in the legend in the upper-right corner. For the sake of spacing in Figure 3.4, those station names have had their International Federation of Digital Seismograph Networks (FDSN) and/or network abbreviations removed, although those data are listed in Tables 3.7 and 3.8.

For added geologic and geodynamic context Figure 3.4 shows an elevation map of the same region. We immediately see myriad islands, seamounts (Wessel et al., 2010), hotspot tracks (Wessel & Kroenke, 1997), and, in lighter greens, large swaths of anomalously elevated oceanic crust known as the South Pacific Superswell (McNutt & Fischer, 1987; McNutt & Judge, 1990).
3.6. Matching MERMAID Seismograms to Earthquakes

Figure 3.5: Examples of MERMAID record sections. In every panel, different colors correspond to individual MERMAIDS reporting the relevant records, which are shown with unit scaling, filtered between 1–5 Hz. For the events identified in the titles theoretical travel-time curves in the ak135 velocity model are overlain as shown in the legend.

3.6 Matching MERMAID Seismograms to Earthquakes

Figure 3.5 is our first example of seismograms recorded by MERMAID as part of the SPPIM project. There we show record sections corresponding to four earthquakes, one each within the magnitude ranges (a) $M_5$–$5.9$; (b) $M_6$–$6.9$, (c) $M_7$–$7.9$, and (d) $M_8$–$8.9$. The seismograms in (a)–(d) are individually color-coded to distinguish the records reported by each MERMAID, whose instrument number is displayed before or after the each trace (excluding the “P0” prefix, and noting that the
3.6. Matching MERMAID Seismograms to Earthquakes

color assignments differ between panels). The seismograms are uncorrected pressure records in their native units of digital counts. By default, MERMAID sends 200 to 250-s-long seismograms, and the seismograms plotted here were demeaned, detrended, and tapered with a symmetric cosine (Hanning) taper, before being band-pass filtered between 1 and 5 Hz using a one pass, four pole Butterworth filter. Each trace is normalized for plotting purposes, resulting in arbitrary amplitudes within, and between, the panels Figure 3.5a–d. The black solid and/or dashed lines correspond to the theoretical arrival times of the phase(s) quoted in the legend, as computed in the velocity model ak135 of Kennett et al. (1995). Each of the four events shown in Figure 3.5 corresponds to the earthquake for which the highest number of records were reported within the respective unit-magnitude ranges shown in Figure 3.9, where they are bolded.

The MERMAID onboard detection algorithm sends time series determined via probabilistic wavelet-subspace analysis to be likely teleseismic $P$-wave arrivals. The algorithm does not provide arrival-time picks beyond the precision afforded by the underlying STA/LTA detection algorithm, nor is it privy to recent global seismicity. Therefore, to produce record sections like those of Figure 3.5 we must first determine if the seismograms sent by MERMAID match any events in the global catalog of recent events.

3.6.1 Automated preliminary matching

Upon receipt of a fresh seismogram transmitted by MERMAID we immediately wish to determine whether or not the signals it contains correspond to known seismic events. To that end we have developed a complete workflow in MATLAB to match untagged, raw seismograms to global seismic catalogs with minimal user intervention. This first step discussed next—the algorithmic querying of global catalogs, the tagging of likely events, the annotating of seismograms with their theoretical phase arrival times, and the multiscale detection of phases against which residuals are displayed—occurs automatically and without user intervention after a MERMAID transmits a new seismogram.

The preliminary matching process begins with the querying of global seismic catalogs with irisFetch.m, a software packaged and distributed by IRIS, for seismic events that occurred
3.6. Matching MERMAID Seismograms to Earthquakes

in the hour preceding the seismogram. Next, travel times are computed for seismic body waves that are likely to be present in the record using \texttt{taupTime.m} for the ak135 velocity model (see section 3.13 for more detail). Each event with one or more phase-arrivals in the time window of the seismogram is deemed a preliminary match, and all such events are sorted by magnitude (generally the single-greatest factor determining its identification) and saved together as individual structures (a convenient MATLAB data type) in a binary “unreviewed” file.

Preliminary matching generates two Portable Document Format (PDF) plots of the raw seismogram on which the theoretical phase-arrival times of possible events are marked, and with panels showing wavelet-subspace projections of the seismogram at five different scales overlain with AIC-based arrival-time picks, following the method of Simon et al. (2020). The first PDF, an example of which is Figure 3.6, displays the complete seismogram, and the second (not shown here) is truncated to show detail about a 100 s window centered on the first arrival of the event with the largest magnitude among all potential matches. Usually that is the true match, and thus the seismogram in the top panel of Figure 3.6 is annotated using those metadata, its first-arriving phase highlighted a solid red line to set it apart from all other possible phases in the time window of interest rendered in dashed black lines. All named phases are labeled in the top panel, with subscripts identifying the rank of the associated event in the magnitude-sorted preliminary match list. Hence, in Figure 3.6, $p_1$ is the model arrival time of a $p$ wave generated by the first preliminary event match, and, $S_4$ is the theoretical arrival time of an $S$ wave from the fourth potential event match.

These preliminary matches are automatically generated and the algorithm only requires a SAC file (Helffrich et al., 2013) as input; i.e., the only relevant information ingested by the algorithm in this preliminary-matching stage is a (mobile) receiver location and a time window (both, of course, being contained in the SAC file itself). Hence, our procedure is not specific to MERMAID data, and we may reasonably assume that it has application beyond the scope of this study, e.g., for single-station or array deployments of traditional broadband land instruments, perhaps in the context of Nuclear Testban Treaty verification, Raspberry Shakes (Calais et al., 2019), or various other forms of crowd-sourced “citizen” seismology, e.g., recorded by mobile phones (Kong et al., 2016).
3.6. Matching MERMAID Seismograms to Earthquakes

Figure 3.6: MERMAID seismogram after preliminary matching, as displayed for the researcher during manual event verification. The blue trace in the first panel is the raw seismogram, while the gray traces in the underlying panels are the wavelet-subspace projections at five scales, each overlain by the associated AIC curve (black) and AIC-based arrival time pick (purple) at that scale. The top panel is annotated with the theoretical arrival times of various phases from five distinct earthquakes, as noted in the subscript, computed in the ak135 velocity model, and marked in time by vertical lines. These represent all the phases which have theoretical arrival times within the time window of the seismogram, associated with known global seismic events in the catalogs queried from IRIS. The time of the first-arriving phase associated with the largest earthquake in the set \((p_1)\) is marked by a solid red vertical line. Its theoretical arrival-time agrees well with the AIC-based arrival time pick (which is agnostic of seismology) at the first three scales. The agreement of these two distinct estimated arrival times, each calculated in very different ways, lends itself to the confident assignment of “identified” to this seismogram. During manual review, this PDF (and secondary, a zoomed-in version) is displayed to the researcher, along with the event metadata for all potentially-matching events, and they are lead through a series of intuitive prompts in MATLAB for easy matching and sorting.
3.6. Matching MERMAID Seismograms to Earthquakes

or other low-cost accelerometers (Cochran et al., 2009), and for classroom seismic installations (Balfour et al., 2014), where an experienced researcher may not available to guide the matching process. Further, while the code ships with default parameters optimized for MERMAID data, these are easily tunable to specific seismic applications.

3.6.2 Manual winnowing and sorting

The second step of the matching procedure involves manual review of the preliminary matches. The review process is simple and intuitive and results in the seismograms being sorted into two classes: identified and unidentified. Those in the former class will have been assessed to contain energies consistent with phase arrivals corresponding to known earthquakes in global seismic catalogs, both by visual inspection and by considering their travel time residuals with respect to the AIC picks. For every SAC file reviewed, the two PDFs generated in the first step are automatically opened for inspection, and the interactive program guides the user through a series of prompts to determine if the event can be identified, and if so, which event(s) and phase(s) should be saved to a winnowed MATLAB binary file.

The process begins with a helpful printout of metadata on all potential events, with specific focus given to the largest event in the list and its corresponding residuals. At all times the user has quick access to all events, and their corresponding residuals, thanks to their MATLAB structures being loaded automatically with each seismogram under review.

We refer again to Figure 3.6, whose top panel plots the raw seismogram in blue. The arrival times marked on that top panel are the ak135 predictions. The panels below the first plot the subspace projection of the seismogram at five wavelet scales in gray, the amplitude of which corresponds to the left ordinate axis. Overlain in black in each panel is the associated AIC curve used to generate the arrival-time pick at that scale, the amplitude of which corresponds to the right ordinate axis. This curve is essentially an inverted likelihood-curve: where it is low, an arrival is likely. The specific AIC arrival-time pick is marked at each scale by a purple vertical line. Quoted in the legend are their corresponding signal-to-noise ratios, defined to be the ratio of the maximum-likelihood
estimates of the variances of the signal and noise segments,

\[
\text{SNR} = \frac{\hat{\sigma}_{\text{signal}}^2}{\hat{\sigma}_{\text{noise}}^2},
\]  

(3.1)

the “signal” being the segment after the AIC pick, and the “noise” the segment preceding it. Later in section 3.8 we redefine the noise and signal segments en route to identifying with high-precision the arrival times of first-arriving \( P \) waves in a single frequency band. There, the seismograms being analyzed have already been positively matched to an identified event, which differs from the procedure here, where we wish to inspect the full bandwidth of each seismogram via a multiscale decomposition.

It is important to note that the AIC-based picks are agnostic of seismology. It is their agreement, or lack thereof, with the theoretical arrival times of the phases from the match list that inform the decision to designate a seismogram as “(un)identified.” In the case presented in Figure 3.6, the purple AIC picks at subspace projection scales one \((x_1)\) through three \((x_3)\) agree well with the theoretical arrival time of the first-arriving \( p \) wave computed in ak135. The AIC picks at the other scales are either low-SNR, or very near the edge, the latter a common problem with AIC pickers that maximize the SNR by picking near an edge to obtain an extremely abbreviated (and thus low-variance) noise or signal segment. In these cases, visual inspection of those picks at the higher scales (lower frequencies) leave no question that those picks may be disregarded. This example is pretty clear: the AIC picks at high-frequency agree well with the theoretical arrival time of the \( p \) wave, and thus this seismogram would be counted among the “identified” category. This sorting is accomplished via simple prompts that guide the user through the winnowing process, which ultimately results in the seismogram being classified as “identified” or “unidentified,” and the relevant event data, and their associated residuals and uncertainties (the latter is briefly discussed in section 3.8.4, and at length by Simon et al., 2020), being saved to a MATLAB binary “reviewed” file.

Ultimately the decision to mark a seismogram as “identified” or “unidentified” comes down
to experience processing MERMAID seismograms like the one presented in Figure 3.6. The hope, however, is that the workflow developed here is simple enough for new researchers with some experience processing seismic data to quickly grasp and apply to their own untagged data with minimal training. Indeed, our workflow is already being successfully applied to the 23 SUSTech instruments included in the SPPIM deployment; albeit applied to the same type of data in this case, but importantly, matched by a different researcher.

### 3.7 The MERMAID Seismic Catalog

The process just outlined enabled the matching of seismograms to earthquakes in our growing MERMAID seismic catalog, the inspection of which is the focus of this section. We first take a broad look at the catalog itself, before drilling down to the statistics of the rate of return of identified events for individual MERMAIDs. We will also discuss the completeness of our catalog as compared to other global seismic catalogs available over the period of interest, considering the entire activity time of each of the 16 floats through the end of 2019. In this section, when we refer to the “catalog” we specifically mean the seismic catalog of recorded earthquakes, not the catalog of travel-time residuals, which we discuss later. The purpose of this section is to answer questions relevant to any new seismic instrument, such as: “How many earthquakes does MERMAID record per year, and what are the distributions of their magnitudes, epicentral distances, and locations?”; “What do the recorded magnitude-distance relations tell us about detectability thresholds?”; and “What is the probability that any single earthquake will be recorded by any single MERMAID, and how many earthquakes is each projected to record in its lifetime?”

#### 3.7.1 Catalog summary: in pictures

Figures 3.7 and 3.8 summarize the MERMAID catalog to date. Figure 3.7 plots histograms of earthquake magnitudes and distances, and combines those data with the SNR of the first-arrival in the first global earthquake-detectability diagram for third-generation MERMAID. Figure 3.8 plots
3.7. The MERMAID Seismic Catalog

Figure 3.7: (caption next page)
Figure 3.7: (previous page) Distributions of earthquake magnitudes, epicentral distances, and signal-to-noise ratios (SNR) considering the entire data set across all 16 Princeton MERMAIDS. In total, 668 MERMAID seismograms were identified to contain at least one phase arrival associated with one of 284 unique earthquakes. (a) The distribution of earthquake magnitudes has its minimum at $M_{4.2}$, its maximum at $M_{8.2}$, and its mean at $M_{6.1}$. (b) The distribution of earthquake epicentral distances has a roughly uniform distribution for epicentral distances out to around 90$^\circ$, except for an obvious peak around 10$^\circ$, which largely corresponds to light ($M_{4-4.9}$) and frequent earthquakes near Fiji that are sampled by the most proximal floats, mainly MERMAID P008. (c) A scatter plot of the data in (a) and (b), where the marker sizes represent the SNRs of individual arrivals. As expected, the highest-SNR records are associated with the largest and/or nearest earthquakes. Further, we find that for this data set a lower-detection threshold hovers just above $M_{6}$ near 160$^\circ$, as proven by MERMAID's identification of core phases. Note that the linear features in (c), for example, the horizontal string of points around $M_{7.5}$ that extends from roughly 20–100$^\circ$, are due to the fact that often more than one MERMAID identifies the same earthquake, leading to multiple detections of the same event at various epicentral distances. Figure 3.9 replots these data on a time axis such that the number of individual MERMAIDS detecting any single event is made more clear.

the ray paths of those earthquakes, connecting them to the locations of the MERMAIDS recording them, binned by event depth. In all, 668 MERMAID seismograms were identified as containing at least one phase arrival associated with one of 284 unique earthquakes.

Figure 3.7a shows that MERMAID sampled a fairly large range of earthquake magnitudes, recording quiet regional and local events. The smallest earthquake, a $m_b$ 4.2 at 97.8 km depth in the Tonga Islands, was recorded by P008 at an epicentral distance of 2.2$^\circ$. The largest event, a $M_w$ 8.2 at 600 km depth in the Fiji Islands region was recorded by five out of our 16 MERMAIDS. The specific number of MERMAIDS which reported specific earthquakes, broken down by magnitude unit, is discussed later in section 3.7.2.

Figure 3.7b is a histogram of those same earthquakes but now binned in terms of their epicentral distances. Again, we see fairly consistent sampling at a variety of epicentral distances, implying MERMAID samples tomographically useful data at the global scale, including phases which have transited the core of the Earth. Those arrivals are discussed later in section 3.9.4.

Finally, Figure 3.7c plots the SNR of the first-arriving phase, represented by the size of the marker, as a function of magnitude and epicentral distance. The SNRs plotted here differ slightly from those quoted in the legends of the subspace projections of Figure 3.6 (they are not multiscale),

98
3.7. The MERMAID Seismic Catalog

Figure 3.8: (caption next page)
3.7. *The MERMAID Seismic Catalog*

Figure 3.8: (previous page) Global source-receiver ray paths considering the entire data set across all 16 Princeton MERMAIDs, separated by event depth (from top to bottom): shallow-, intermediate-, and deep-focus. In each panel, the great-circle path (black curves) connects the earthquake location (red asterisks) with the interpolated location of MERMAID at the time of recording (yellow upside-down triangles). Each map is centered on Tahiti, French Polynesia, the approximate center of the SPPIM deployment. The geographic distributions of earthquake locations plotted here shows that MERMAID preferentially records subduction-zone earthquakes in the Pacific Rim.

and their derivation is discussed in detail in section 3.8, but the basic idea that it expresses the ratio of the variances of the seismogram after and before the AIC-based arrival time pick still rings true. Figure 3.7c is the first detectability curve available for the third-generation MERMAID—to be compared to the first-generation results shown by Simons et al. (2009) (their Figure 8). Here we see trends common to all seismic instruments: small events are preferentially recorded at short epicentral distances, before geometrical spreading and attenuation can sap them of their energy, while larger events (greater than $M_6$, in the case of MERMAID) may be recorded globally.

Figure 3.8 places the earthquake data of Figure 3.7 into their spatial context by plotting the ray paths between the earthquake and MERMAID locations at the time of recording (see section 3.5.1). The ray paths are binned by event depth from top to bottom as shallow-focus (a; less than 70 km) intermediate-focus (b; between 70 km and 300 km), and deep-focus (b; greater than 300 km) earthquakes. Listed above each map in Figure 3.8 is the total number of unique events recorded within those depth ranges. We find that MERMAID records shallow events most often, with 263 unique reports, though the counts at the other depths are overall similar, proving that MERMAID recorded earthquakes originating at depths ranging from the shallow crust to deep within subducting slabs. The shallowest earthquake in the catalog had its hypocenter at 2 km, under Northern Alaska, and the deepest ruptured at a depth of 652 km under the Fiji Islands region. Figure 3.8 also shows that MERMAID primarily recorded subduction-zone earthquakes occurring along the Pacific Rim, the so-called “Ring of Fire,” the nearly continuous chain of volcanoes fed by subducting oceanic crust that encircles the Pacific Ocean from New Zealand to Chile (Rinard Hinga, 2015). This is unsurprising given the location of SPPIM, roughly in the center of the Ring of Fire, and the fact
### 3.7. The MERMAID Seismic Catalog

Table 3.1: The seismic catalog of the 16 Princeton MERMAIDS, complete to end 2019.

<table>
<thead>
<tr>
<th>MERMAID</th>
<th>Deployment</th>
<th># Wks.</th>
<th># Seis.</th>
<th>% ID</th>
<th># Seis.</th>
<th># ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>P008:</td>
<td>05-Aug-2018</td>
<td>73.3</td>
<td>251</td>
<td>184</td>
<td>73.3%</td>
<td>179</td>
</tr>
<tr>
<td>P009:</td>
<td>06-Aug-2018</td>
<td>73.2</td>
<td>130</td>
<td>94</td>
<td>72.3%</td>
<td>93</td>
</tr>
<tr>
<td>P010:</td>
<td>07-Aug-2018</td>
<td>73.1</td>
<td>125</td>
<td>85</td>
<td>68.0%</td>
<td>89</td>
</tr>
<tr>
<td>P011:</td>
<td>09-Aug-2018</td>
<td>72.8</td>
<td>73</td>
<td>50</td>
<td>68.5%</td>
<td>52</td>
</tr>
<tr>
<td>P012:</td>
<td>10-Aug-2018</td>
<td>72.6</td>
<td>245</td>
<td>45</td>
<td>18.4%</td>
<td>176</td>
</tr>
<tr>
<td>P013:</td>
<td>31-Aug-2018</td>
<td>69.6</td>
<td>215</td>
<td>28</td>
<td>13.0%</td>
<td>161</td>
</tr>
<tr>
<td>P016:</td>
<td>03-Sep-2018</td>
<td>69.2</td>
<td>48</td>
<td>26</td>
<td>54.2%</td>
<td>36</td>
</tr>
<tr>
<td>P017:</td>
<td>04-Sep-2018</td>
<td>69.1</td>
<td>33</td>
<td>22</td>
<td>66.7%</td>
<td>25</td>
</tr>
<tr>
<td>P018:</td>
<td>05-Sep-2018</td>
<td>68.9</td>
<td>22</td>
<td>19</td>
<td>86.4%</td>
<td>17</td>
</tr>
<tr>
<td>P019:</td>
<td>06-Sep-2018</td>
<td>68.7</td>
<td>20</td>
<td>19</td>
<td>95.0%</td>
<td>15</td>
</tr>
<tr>
<td>P020:</td>
<td>08-Sep-2018</td>
<td>68.5</td>
<td>89</td>
<td>13</td>
<td>14.6%</td>
<td>68</td>
</tr>
<tr>
<td>P021:</td>
<td>09-Sep-2018</td>
<td>68.3</td>
<td>15</td>
<td>15</td>
<td>100.0%</td>
<td>11</td>
</tr>
<tr>
<td>P022:</td>
<td>10-Sep-2018</td>
<td>68.2</td>
<td>12</td>
<td>12</td>
<td>100.0%</td>
<td>9</td>
</tr>
<tr>
<td>P023:</td>
<td>13-Sep-2018</td>
<td>67.8</td>
<td>33</td>
<td>26</td>
<td>78.8%</td>
<td>25</td>
</tr>
<tr>
<td>P024:</td>
<td>13-Sep-2018</td>
<td>67.8</td>
<td>18</td>
<td>18</td>
<td>100.0%</td>
<td>14</td>
</tr>
<tr>
<td>P025:</td>
<td>14-Sep-2018</td>
<td>67.6</td>
<td>16</td>
<td>12</td>
<td>75.0%</td>
<td>12</td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td>1118.9</td>
<td>1345</td>
<td>668</td>
<td>49.7%</td>
<td>983</td>
</tr>
<tr>
<td>Mean:</td>
<td>29-Aug-2018</td>
<td>69.9</td>
<td>84</td>
<td>42</td>
<td>49.7%</td>
<td>61</td>
</tr>
</tbody>
</table>

that roughly 90% of annual global seismicity occurs in this most active of regions.

Figure 3.7 and Figure 3.8 plot compiled data considering all 16 floats in the Princeton-operated fleet. As combined, these numbers mask the variability in the rate of seismicity recorded by individual floats. In what follows we parse the catalog by specific float numbers to capture the idiosyncrasies of each.

### 3.7.2 Catalog summary: by the numbers

Table 3.1 is a breakdown of the rate of return of seismograms per MERMAID. The first column lists the MERMAID numbers, the second their deployment dates, and the third the total duration, in weeks, over which each MERMAID was active. The fourth and fifth columns list the total number of seismograms returned, and the subset of those identified, respectively, and the sixth column quotes the percentage of the latter. The seventh column lists the average number of seismograms returned per full year of activity, and the eighth column lists the same statistic pertaining to the identified
seismograms only. The penultimate row totals columns three through eight, while the ultimate row lists their averages. Columns four, five, seven, and eight (corresponding to a specific number of MERMAID seismograms) are rounded to the nearest integer. Their sums and means in the final two rows are also rounded, but only after the summing and taking the mean of the data, hence their values (for example in the penultimate row of seventh column) may differ slightly from the actual sum over that column.

Let us first take a bird’s-eye view of the data presented in Table 3.1 before teasing apart the statistics of the rate of return of individual MERMAIDS. From the penultimate row of Table 3.1 we see that our 16 MERMAIDS enjoyed a total of 1118.9 weeks (21.44 years) of deployment, over which time they autonomously recorded and transmitted 1345 seismograms from the South Pacific. Of those, 668 were positively matched to global catalogs (“identified”) available at the time using the methodology and software described in section 3.6. This means that roughly half the seismograms over this time period were identified. The others represent myriad diverse signals corresponding to small and/or local events that were missed by the global seismic networks (i.e., not recorded by any other seismic station on Earth, see section 3.9.5), oceanic $T$ waves from unidentified sources, as well as a substantial number of instrument glitches, which almost exclusively affected MERMAIDS P012, P013, and P025.

To belabor the first point: the MERMAID catalog contains many seismograms which are “unidentified” by the standards upheld here, but which do in fact record earthquakes that which otherwise went undetected by the global seismic network: not every “unidentified” event is just noise (more on this later). By summing column seven, the number of MERMAID seismograms divided by number of years that MERMAID was deployed, we find that we maintained a return rate of 983 seismograms per year of deployment. Lastly, by applying the historical percentage of identifications as a ratio of total seismograms from column six, we find that our 16 MERMAIDS averaged 486 identifications per year.

These data are further distilled in the last row of Table 3.1, where we list the rate of the return of an “average” MERMAID in our fleet. There we quote the arithmetic means of the columns, i.e.,
not weighted by the length of time that any individual MERMAID was deployed. Ergo, the final value in this row is the number of identified seismograms one may expect to receive from any MERMAID in any given year. Of course, our sample size of 16 is small, and limited in time and space (on average each MERMAID was deployed for around 1.3 years in a very specific part of the world), but this number is the first step towards defining the expected long-term output of a single, “average” MERMAID. The final value in this row is perhaps most relevant for future MERMAID deployments: we find that on average each instrument returned 30 identified seismograms per year. With a projected lifespan of five years according to the manufacturer, we thus expect a return of approximately 150 identified seismograms over the lifetime of each MERMAID.

However, and perhaps not unexpectedly, the values in the final column which contribute to this mean of are broadly distributed, ranging from a maximum of 131 returned by P008, to a minimum of 9 returned by both P022 and P025. This spread is not due to implicit differences in the floats—indeed they all are identical both in manufacturing and software, and the programmable parameters (e.g., parking depth, detection criteria thresholds etc.) were left unchanged for the duration of the deployments. Rather, this variance is most likely due to the geographic distribution of MERMAIDS. MERMAID P008, the busiest of the group, returned so many identifiable seismograms because it cruised the oceanic region between Fiji and Samoa, near enough to the former (and drifting closer—see Figure 3.4) to record many seismograms matched to light and moderate earthquakes whose energy never reached the more distant MERMAIDS in the open ocean (see Figure 3.4). Not shown is the distribution of identifications per MERMAID for light earthquakes ($M_4$-4.9) because in total 85 seismograms were matched to 79 unique earthquakes, fully 70 of which P008 recorded.

A secondary factor controlling the rate of return of individual MERMAIDS, apart simply from simple source-receiver distance considerations, must also be the oceanic and bathymetric settings around and below the floats themselves. The SNRs of signals received by MERMAID are affected by a number of factors beyond those of common terrestrial stations; of course, like an ordinary terrestrial station the noise is time-variable, but perhaps more importantly, the impedance along the ray path between a repeating earthquake and MERMAID is also time-variable, in contrast to
3.7. The MERMAID Seismic Catalog

As MERMAID drifts it may find itself over oceanic regions with varying sedimentary cover, attenuating or amplifying incident P-wave energy, resulting in weaker or stronger acoustic conversion in the water column (Ewing et al., 1957; Stephen, 1988). Multiple additional factors such as the water depth underlying the float (Lewis & Dorman, 1998; Weatherall et al., 2015), nearby seamounts and other kinds of rough bottom topography (Dougherty & Stephen, 1991), the width and depth of the Sound Fixing and Ranging (SOFAR) channel (Munk, 1974) over the ∼20° and across the seasons covered, and other, understudied and as yet unknown factors may also all play a role in the conversion of energy (Tolstoy & Ewing, 1950; Okal, 2008) and in determining the local ambient noise field (Gualtieri et al., 2019).

As no modeling of the acoustic conversion under the floats was performed in this study, indeed even the bathymetry is not well constrained near many of our floats, nor were wave or storm records correlated with our seismic data (arguably the main driver of time-variable background noise levels, see, e.g. Webb & Cox, 1986; Babcock et al., 1994; Gualtieri et al., 2013; Farra et al., 2016), we cannot yet separate the various factors that contributed most to the large variance in the rate of return after correcting for distance and magnitude considerations. It is an interesting question, though beyond the scope of this study, to probe if the MERMAIDs which sent the least data spent the most time in the noisiest areas stalled over areas of the seafloor with inefficient seismic-acoustic coupling, were muted by some other unidentified disturbance, or some combination of all of these factors.

3.7.3 Catalog completeness and statistics

We now move to comparing our seismic catalog with other global catalogs available at the time. How “complete” is our catalog compared to those others? Conversely, how many global earthquakes did MERMAID miss? No catalog can include all earthquakes of all magnitudes, globally (Kagan, 2003), and in our own section 3.9 we show an example of an earthquake in the MERMAID catalog that was not found in any other global catalog, but for the purpose of this section we do take the number of events recorded in global seismic catalogs to be the true population size against
Figure 3.9: Identified seismic events recorded by MERMAID between 5 August 2018 and 31 December 2019. In each of the panels (a), (c), (e), and (g), a single event is bolded in black—those correspond to the earthquakes which resulted in the most identifications in that magnitude range, shown in Figure 3.5a–d, respectively.

which we will derive completeness statistics.

Figure 3.9 plots the MERMAID seismic catalog, both by the rate of return considering the entire fleet, and in sum considering each float individually. It further breaks these numbers down by magnitude, ranging from $M_5$ in Figure 3.9a–b through $M_8$ in Figure 3.9g–h. The stem plots in the left column (a,c,e,g) show the number of MERMAIDs reporting each positively identified earthquake as a function of time, beginning from the first deployment of P008 on 5 August 2018 through the end of 2019. The histograms in the right column (b,d,f,h) aggregate these data over time, but separate them by float, to identify which floats reported the most earthquakes within a specific magnitude range.
To get at the question of completeness of our catalog versus other global seismic catalogs available at the time, in the stem plots we also represent missed events, not reported by any MERMAID, as crosses placed below the zero line for clarity. For example, Figure 3.9e, corresponding to all $M7$ earthquakes that occurred globally while MERMAID was deployed, shows that six earthquakes went unreported by the entire Princeton fleet. Conversely, Figure 3.9e shows that no events were missed in the magnitude range $M8+$. So many $M5$–$5.9$ events went undetected that rather than plotting each of them in Figure 3.9a, the mean miss-rate (around 4 events per day) is reported below the zero line. Note the different scaling of the ordinate axes in Figure 3.9a, which highlights the fact that the rate of return for earthquakes in the magnitude range $M5$–$5.9$ was lower than for the other magnitudes shown. Listed above each stem plot is the total number of unique global events in that magnitude range over the time period considered, and in parenthesis the number, also as a fraction in per cent, which were positively identified by at least a single MERMAID. Finally, in each of the stem plots one event is highlighted in black. These are the events reported by the largest number of MERMAIDS within each magnitude range and previously rendered in the record sections of Figure 3.5.

The histograms in the right column of Figure 3.9 parse the cumulative return of each individual float. Figures 3.9b ($M5$–$5.9$) and 3.9d ($M6$–$6.9$) visualize the observation of Table 3.1 that P008 outpaced all the other floats in terms of reporting identifiable earthquakes, which we attribute to its geographic proximity to Fiji and the Tonga, as mentioned in section 3.7.1. Listed above the histograms is the total number of identified events reported by any MERMAID in the fleet, and in parentheses the average over all 16 instruments.

The complementary distribution for $M4$–$4.9$ earthquakes is not shown. Most of the faint events in that category were missed, but we do note that, of the 85 total events reported (of which 79 were unique identifications), fully 70 were reported by P008.

We now summarize the statistics presented in Figure 3.9 for individual magnitudes from $M4$ (not shown) through $M8$, for the time period from 5 August 2018 through the end of 2019. Note that the following statistics, e.g., the average number of identified events per MERMAID, are not
rounded as they were in Figure 3.9.

In the magnitude range $M4–4.9$ there were 14535 unique global events, of which 79 (0.5%) were positively identified by at least one of our 16 MERMAIDS. If an event was identified, on average 1.1 MERMAIDS reported that unique event. In total, 85 event identifications were reported by 16 MERMAIDS, meaning that on average each MERMAID reported 5.3 unique events.

For the magnitude range $M5–5.9$, there were 2245 unique global events, of which 108 (4.8%) positively identified by at least one MERMAID. If an event was identified, there were on average 1.6 MERMAIDS reporting that unique event. In total, 173 event identifications were reported by 16 MERMAIDS, on average each MERMAID reported 10.8 unique events.

The magnitude range $M6–6.9$ comprised 198 unique global events, 81 (40.9%) of which were positively identified by at least one MERMAID. If an event was identified, on average 3.4 MERMAIDS reported it. In total, 276 event identifications were reported, for an average of 17.2 unique events reported per MERMAID.

Magnitude range $M7–7.9$ counted 20 unique events, of which 14 (70.0%) were positively identified by at least one of our instruments. If an event was identified, there were on average 8.4 MERMAID reports of it. In total, 117 event identifications were reported, on average 7.3 unique events per MERMAID.

Finally, in magnitude range $M8–8.9$, there were precisely two events, both positively identified by at least one MERMAID in our fleet, with an average 8.5 MERMAIDS reporting each. In total, 17 event identifications were reported by 16 MERMAIDS, meaning that on average each MERMAID reported 1.1 unique events.

There were no magnitude 9 events during our study period.

### 3.7.4 Estimating the final size of the MERMAID catalog

With Table 3.1 we found that, between deployment in Fall 2018 and the end of 2019, each MERMAID in our fleet returned an average of 42 identified seismograms, or, normalizing for the amount of time each was deployed, about 30 per year. With Figure 3.9 we saw how those 42 identifications
Table 3.2: Global M4–4.9 earthquakes, missed or reported by our MERMAIDs.

<table>
<thead>
<tr>
<th>MER.</th>
<th># EQ</th>
<th># ID</th>
<th>% ID</th>
<th># ID</th>
<th># ID</th>
<th># ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>P008</td>
<td>14535</td>
<td>70</td>
<td>0.5%</td>
<td>50</td>
<td>249</td>
<td>59</td>
</tr>
<tr>
<td>P009</td>
<td>14496</td>
<td>9</td>
<td>0.1%</td>
<td>6</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>P010</td>
<td>14475</td>
<td>6</td>
<td>0.0%</td>
<td>4</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>P011</td>
<td>14420</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P012</td>
<td>14379</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P013</td>
<td>13559</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P016</td>
<td>13489</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P017</td>
<td>13464</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P018</td>
<td>13430</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P019</td>
<td>13371</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P020</td>
<td>13315</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P021</td>
<td>13256</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P022</td>
<td>13223</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P023</td>
<td>13151</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P024</td>
<td>13146</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P025</td>
<td>13115</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>218824</td>
<td>85</td>
<td>0.0%</td>
<td>60</td>
<td>302</td>
<td>71</td>
</tr>
<tr>
<td>Mean:</td>
<td>13677</td>
<td>5</td>
<td>0.0%</td>
<td>4</td>
<td>19</td>
<td>4</td>
</tr>
</tbody>
</table>

were distributed across M5+ earthquakes. In this section we extrapolate those historical data to estimate the final size of the complete MERMAID catalog. We will use a projected five-year lifespan of MERMAID, as quoted by the manufacturer, to make these estimates. Like the numbers in Table 3.1, the numbers quoted here are rounded to the nearest integer (excepting percentages). As such, some values quoted there, which themselves are the sums of the analogous statistics quoted here, may differ slightly.

Tables 3.2–3.6 break down the rate of return of identified events per magnitude M4 through M8 for each float, and also use these numbers to project how many identified seismograms within those magnitude ranges each float is likely to return in its lifetime. As in Table 3.1, the first column in Table 3.2–3.6 lists the MERMAID serial number. The second quotes the total number of earthquakes that occurred over the complete deployment of that specific float. For example, the value in this column in the first row of Tables 3.2–3.6, corresponding to P008, is the same number quoted above the stem plots in Figure 3.9. This number represents the maximum number
3.7. The MERMAID Seismic Catalog

Table 3.3: Global M5–5.9 earthquakes, missed or reported by our MERMAIDS.

<table>
<thead>
<tr>
<th>MER. #</th>
<th># EQ</th>
<th>% ID</th>
<th># ID</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>P008:</td>
<td>2245</td>
<td>64</td>
<td>2.9%</td>
<td>46</td>
<td>228</td>
<td>46</td>
<td>230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P009:</td>
<td>2242</td>
<td>34</td>
<td>1.5%</td>
<td>24</td>
<td>121</td>
<td>25</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P010:</td>
<td>2240</td>
<td>34</td>
<td>1.5%</td>
<td>24</td>
<td>121</td>
<td>25</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P011:</td>
<td>2230</td>
<td>14</td>
<td>0.6%</td>
<td>10</td>
<td>50</td>
<td>10</td>
<td>51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P012:</td>
<td>2222</td>
<td>9</td>
<td>0.4%</td>
<td>6</td>
<td>32</td>
<td>7</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P013:</td>
<td>2096</td>
<td>4</td>
<td>0.2%</td>
<td>3</td>
<td>15</td>
<td>3</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P016:</td>
<td>2085</td>
<td>3</td>
<td>0.1%</td>
<td>2</td>
<td>11</td>
<td>2</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P017:</td>
<td>2081</td>
<td>1</td>
<td>0.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P018:</td>
<td>2070</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P019:</td>
<td>2051</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P020:</td>
<td>2039</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P021:</td>
<td>2036</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P022:</td>
<td>2032</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P023:</td>
<td>2019</td>
<td>9</td>
<td>0.4%</td>
<td>7</td>
<td>35</td>
<td>7</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P024:</td>
<td>2019</td>
<td>1</td>
<td>0.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P025:</td>
<td>2013</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>33720</td>
<td>173</td>
<td>0.5%</td>
<td>124</td>
<td>621</td>
<td>126</td>
<td>630</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean:</td>
<td>2108</td>
<td>11</td>
<td>0.5%</td>
<td>8</td>
<td>39</td>
<td>8</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: Global M6–6.9 earthquakes, reported or missed by our MERMAIDS.

<table>
<thead>
<tr>
<th>MER. #</th>
<th># EQ</th>
<th>% ID</th>
<th># ID</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
<th>E[# ID]</th>
<th>yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>P008:</td>
<td>198</td>
<td>42</td>
<td>21.2%</td>
<td>30</td>
<td>149</td>
<td>26</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P009:</td>
<td>198</td>
<td>42</td>
<td>21.2%</td>
<td>30</td>
<td>150</td>
<td>26</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P010:</td>
<td>198</td>
<td>34</td>
<td>17.2%</td>
<td>24</td>
<td>121</td>
<td>21</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P011:</td>
<td>198</td>
<td>24</td>
<td>12.1%</td>
<td>17</td>
<td>86</td>
<td>15</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P012:</td>
<td>198</td>
<td>22</td>
<td>11.1%</td>
<td>16</td>
<td>79</td>
<td>14</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P013:</td>
<td>181</td>
<td>14</td>
<td>7.7%</td>
<td>10</td>
<td>52</td>
<td>9</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P016:</td>
<td>181</td>
<td>15</td>
<td>8.3%</td>
<td>11</td>
<td>57</td>
<td>10</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P017:</td>
<td>181</td>
<td>13</td>
<td>7.2%</td>
<td>10</td>
<td>49</td>
<td>9</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P018:</td>
<td>181</td>
<td>13</td>
<td>7.2%</td>
<td>10</td>
<td>49</td>
<td>9</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P019:</td>
<td>180</td>
<td>12</td>
<td>6.7%</td>
<td>9</td>
<td>46</td>
<td>8</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P020:</td>
<td>178</td>
<td>7</td>
<td>3.9%</td>
<td>5</td>
<td>27</td>
<td>5</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P021:</td>
<td>178</td>
<td>8</td>
<td>4.5%</td>
<td>6</td>
<td>31</td>
<td>5</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P022:</td>
<td>176</td>
<td>5</td>
<td>2.8%</td>
<td>4</td>
<td>19</td>
<td>3</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P023:</td>
<td>175</td>
<td>9</td>
<td>5.1%</td>
<td>7</td>
<td>35</td>
<td>6</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P024:</td>
<td>175</td>
<td>11</td>
<td>6.3%</td>
<td>8</td>
<td>42</td>
<td>8</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P025:</td>
<td>175</td>
<td>5</td>
<td>2.9%</td>
<td>4</td>
<td>19</td>
<td>3</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>2951</td>
<td>276</td>
<td>9.4%</td>
<td>202</td>
<td>1011</td>
<td>177</td>
<td>885</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean:</td>
<td>184</td>
<td>17</td>
<td>9.4%</td>
<td>13</td>
<td>63</td>
<td>11</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 3.7. The MERMAID Seismic Catalog

Table 3.5: Global $M7$–$7.9$ earthquakes, reported or missed by our MERMAIDs.

<table>
<thead>
<tr>
<th>MER.</th>
<th># EQ</th>
<th># ID</th>
<th>% ID</th>
<th># ID</th>
<th>$E[# ID]$</th>
<th>$E[# ID]$</th>
<th>$E[# ID]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P008</td>
<td>20</td>
<td>7</td>
<td>35.0%</td>
<td>5</td>
<td>25</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>P009</td>
<td>20</td>
<td>8</td>
<td>40.0%</td>
<td>6</td>
<td>29</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>P010</td>
<td>20</td>
<td>9</td>
<td>45.0%</td>
<td>6</td>
<td>32</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>P011</td>
<td>20</td>
<td>10</td>
<td>50.0%</td>
<td>7</td>
<td>36</td>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>P012</td>
<td>20</td>
<td>12</td>
<td>60.0%</td>
<td>9</td>
<td>43</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>P013</td>
<td>17</td>
<td>10</td>
<td>58.8%</td>
<td>7</td>
<td>37</td>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td>P016</td>
<td>17</td>
<td>7</td>
<td>41.2%</td>
<td>5</td>
<td>26</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>P017</td>
<td>17</td>
<td>7</td>
<td>41.2%</td>
<td>5</td>
<td>26</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>P018</td>
<td>17</td>
<td>5</td>
<td>29.4%</td>
<td>4</td>
<td>19</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>P019</td>
<td>16</td>
<td>6</td>
<td>37.5%</td>
<td>5</td>
<td>23</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>P020</td>
<td>16</td>
<td>5</td>
<td>31.2%</td>
<td>4</td>
<td>19</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>P021</td>
<td>16</td>
<td>6</td>
<td>37.5%</td>
<td>5</td>
<td>23</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>P022</td>
<td>16</td>
<td>6</td>
<td>37.5%</td>
<td>5</td>
<td>23</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>P023</td>
<td>16</td>
<td>7</td>
<td>43.8%</td>
<td>5</td>
<td>27</td>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>P024</td>
<td>16</td>
<td>5</td>
<td>31.2%</td>
<td>4</td>
<td>19</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>P025</td>
<td>16</td>
<td>7</td>
<td>43.8%</td>
<td>5</td>
<td>27</td>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td>280</td>
<td>117</td>
<td>41.8%</td>
<td>87</td>
<td>435</td>
<td>88</td>
<td>438</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>18</td>
<td>41.8%</td>
<td>5</td>
<td>27</td>
<td>5</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 3.6: Global $M8$–$8.9$ earthquakes, reported or missed by our MERMAIDs.

<table>
<thead>
<tr>
<th>MER.</th>
<th># EQ</th>
<th># ID</th>
<th>% ID</th>
<th># ID</th>
<th>$E[# ID]$</th>
<th>$E[# ID]$</th>
<th>$E[# ID]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P008</td>
<td>2</td>
<td>1</td>
<td>50.0%</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>P009</td>
<td>2</td>
<td>1</td>
<td>50.0%</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>P010</td>
<td>2</td>
<td>2</td>
<td>100.0%</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P011</td>
<td>2</td>
<td>2</td>
<td>100.0%</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P012</td>
<td>2</td>
<td>2</td>
<td>100.0%</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P013</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P016</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P017</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P018</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P019</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P020</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P021</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P022</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P023</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P024</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>P025</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>17</td>
<td>81.0%</td>
<td>13</td>
<td>63</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>1</td>
<td>81.0%</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
of earthquakes that each float could have individually identified during its deployment. The third and fourth columns list the number, and percentage, of those events that were identified. The fifth is analogous to the final column of Table 3.1, except here it is further parsed by magnitude, while there it was the sum across all magnitudes.

The sixth column of Tables 3.2–3.6 lists our first estimates of the expected total number of identified seismograms that any individual MERMAID may return over its projected five-year lifespan. It is simply the historical yearly rate of return of identified seismograms (the previous column), multiplied by five. For light and moderate earthquakes, especially, this method of estimation is likely sound because there are so many earthquakes within those magnitude ranges annually, that the year-to-year variance in the earthquake sample size (each of which MERMAID either does or does not identify) is relatively small. Conversely, one could imagine a case where the historical rate we derived for, e.g., great earthquakes, was sampled during an anomalous year, and was thus a poor estimator of the true annual population. In that case, projections based on those values could greatly skew our estimates.

To combat the potential issue of anomalous sample sizes skewing the projections of Tables 3.2–3.6 we pulled a data set of all events cataloged by IRIS from 1985 through to the end 2014. We choose to base our updated annual seismicity rates on those dates because: 30 years of data would surely provide a large enough sample size within each magnitude unit to converge to the true population values; 2014 was far enough in the past to ensure that the ISC catalog (International Seismological Centre, 2016) had been reviewed and published (it generally lags behind the PDE, the nearly instantaneous, but not necessarily most accurate, source of earthquake data from IRIS, by a few years); while 1985 was still recent enough to ensure that a robust and relatively modern seismic network was installed globally, which all but guaranteed that the resultant catalogs would be relatively complete. We found that over that 30-year span there were a total of 365378 $M_4$, 48511 $M_5$, 3650 $M_6$, 396 $M_7$, and 28 $M_8$ earthquakes, resulting in an average of 12179 $M_4$, 1617 $M_5$, 122 $M_6$, 13 $M_7$, and 1 $M_8$ earthquakes per year. We use the latter numbers to compute the values in the final two columns of Tables 3.2–3.6, where the overline, $\bar{y}$, denotes an “average”
3.8. Estimating First-Arrival Times and Their Uncertainties

year. In column seven we multiplied these average seismicity rates by the percentage of the total that each float identified (column four) to compute a second estimate of the expected total number of identified seismograms any individual MERMAID may return in a year. In column eight, we again multiplied this number by five to project the final number of earthquakes each float may be expected to identify in its lifetime.

The final two rows of Tables 3.2–3.6 summarize the data in much the same way as Table 3.1—the penultimate row tallies the totals of the columns, and the ultimate row reports their means. Like there, the final number in the final row carries perhaps the most meaning: it is our best guess of the total number of identified earthquakes that any given MERMAID will report within a specific magnitude range over its projected lifetime. We find these numbers to be 22 $M_4$, 39 $M_5$, 55 $M_6$, 27 $M_7$, and 4 $M_8$ earthquakes, or just under 150 earthquakes in total. For our fleet of 16, this equates to nearly 2400 identified earthquakes. However, as we have seen, the variance in the rate of return among the floats is large, and some, for example P008 with its 184 identified events, have already surpassed their expected lifetime-total return.

Ultimately, we return the conclusion of section 3.7.2, that it is likely the geographic location, the frequency and severity of nearby storms, and perhaps to a lesser extent the geologic setting around where MERMAID drifts, that most drives the rate of return of identified seismograms by any individual float. This point is made well in Table 3.2 columns two and three, where we see that P008 was privy to only 115 more $M_4$ earthquakes than P011, but the former identified 70, and the latter identified none.

3.8 Estimating First-Arrival Times and Their Uncertainties

Having exhausted our study of the ability of MERMAID to detect, or not, global earthquakes, we now move to discussing the seismograms themselves. This is the main thrust of this study; the high-precision picking of first-arriving $P$ or $p$ waves, the estimation of uncertainty about those times, and what their residuals against various velocity-model predictions may tell us about mantle structure.
We shared a preview of our preferred method of phase-picking in Figure 3.6 of section 3.6.1. Here we elaborate slightly on our procedure, discussing specifically how we applied it in this study to accurately identify first-arrival times and quantify their uncertainties.

### 3.8.1 The arrival-time pick

We developed an automated AIC-based arrival time estimation scheme (Simon et al., 2020) to rapidly and accurately pick seismic phase arrival times. Our procedure relies on computing the likelihood that a time-pick best partitions the seismogram into two distinct segments: noise and signal. We do so by maximizing the signal-to-noise SNR over all possible splits, from start to end. Simon et al. (2020) also include two methods to assess the uncertainty of our arrival-time estimate. We use the first of those methods, Monte-Carlo resimulation and re-picking, to estimate the arrival-time uncertainties reported in this study.

We used the same picking procedure for every seismogram analyzed in this study regardless of whether it was recorded by MERMAID, a “traditional” seismometer, or a Raspberry Shake, with which we will be comparing and validating the MERMAID results. First, a 60 s segment of the demeaned and detrended seismogram, centered on the theoretical phase arrival time, was isolated for inspection. Then it was multiplied by a symmetric window, flat in its 30 s interior and with a 15 s cosine taper at either end. Next, the tapered seismogram was band-pass filtered between 1 and 5 Hz by a one-pass, four-pole Butterworth filter. Finally, our AIC-based picking scheme was run on the 30 s interior time window.

### 3.8.2 The travel-time residual

We define the travel-time residual to be the time difference between our time pick, $t_{AIC}$, and the theoretical arrival time computed in the 1-D spherical ak135 velocity model of Kennett et al. (1995).

For all records of traditional sensors and Raspberry Shake stations on nearby islands that we
will be discussing, the travel-time residual is simply

\[ t_{\text{res}} = t_{\text{AIC}} - t_{\text{ak135}}. \]  
(3.2)

Computing MERMAID travel-time residuals in the same model requires adjusting for bathymetry and MERMAID cruising depth,

\[ t^*_{\text{res}} = t_{\text{AIC}} - t^*_{\text{ak135}}, \]  
(3.3)

as explained in section 3.8.3.

We perform this simple 1-D comparison first to show that the distributions of residuals from more traditional seismic instruments and MERMAID agree well, proving that MERMAID is returning tomographically useful data. Later we recompute residuals for MERMAID using the fully-3-D, elliptical LLNL-3DGv3 velocity model of Simmons et al. (2012), defining

\[ t^\oplus_{\text{res}} = t_{\text{AIC}} - t_{\text{LLNL}} \]  
(3.4)

to interrogate the geographic distribution of velocity perturbations in Earth’s mantle as recorded by MERMAID.

### 3.8.3 Adjusting for bathymetry and MERMAID cruising depth

Equation 3.3 required adjusting the ak135 travel-time for bathymetry, the water layer, and a submerged receiver. There, \( t^*_{\text{ak135}} \) is the theoretical arrival time computed in the adjusted ak135 velocity model,

\[ t^*_{\text{ak135}} = t_{\text{ak135}} + t_{\text{adj}}, \]  
(3.5)

where \( t_{\text{adj}} \) is the difference between the travel time in the standard and the adjusted models. Because the theoretical ray paths are identical in both models until reaching the seafloor, \( t_{\text{adj}} \) equals the difference between the travel time of the converted phase from the seafloor to MERMAID, and in a
3.8. Estimating First-Arrival Times and Their Uncertainties

rock layer equal in thickness to the local water depth,

\[ t_{\text{adj}} = \frac{z_w - z_{\text{MER}}}{v_w \cos \theta_w} - \frac{z_w}{v_r \cos \theta_r}. \]  \hfill (3.6)

In this convention \( z \) is depth in m positive down below the surface, \( v \) is the acoustic velocity in m/s, \( \theta \) is the angle of incidence in degrees, and subscripts ‘w’ and ‘r’ denote those values in water and rock, respectively. Local bathymetry at the location \( (z_w) \) of the recording MERMAID is interpolated from GEBCO 2014 (Weatherall et al., 2015), and MERMAID depth at the time of trigger \( (z_{\text{MER}}) \) was measured via its onboard pressure sensor and written to the header of the SAC file. The standard default dive depth is 1500 m. We assume an acoustic velocity of 1500 m/s for the water layer and use 5800 m/s for rock, in keeping with the upper layer in ak135. The incidence angle of the converted phase in the water column is given by Snell’s law (Nolet, 2008),

\[ \theta_w = \sin^{-1} \left( \frac{v_w \sin \theta_r}{v_r} \right). \]  \hfill (3.7)

Equation 3.6 yields an adjustment of +0.98 s for a \( P \) wave incident at 0° on the seafloor of a 4000 m deep ocean, and recorded by MERMAID at a cruising depth of 1500 m, in other words, for an “average” ocean depth and an “average” MERMAID cruising depth.

Considering the MERMAID’s design goal of reporting teleseismic waveforms that bottomed in the lower mantle and are thus incident at small angles on the seafloor, a good rule of thumb holds that 1 s should be added to arrival times computed in the ak135 velocity model (or, equivalently, 1 s should be \textit{removed} from MERMAID travel-time residuals computed against ak135 as in eq. (3.2)). The values reported by Simon et al. (2020) for the second-generation MERMAID data did not account for bathymetry or cruising depth, and hence this rule is to be applied to the residuals reported there.
3.8.4 The uncertainty on the residual

Our AIC-based picking procedure simultaneously provides uncertainty estimates associated with each arrival time. Method 1 of Simon et al. (2020), used here, leverages the statistics of the seismogram to generate synthetic sequences from which timing-error distributions are generated via Monte-Carlo resimulation. Every such “seismogram” is simply modeled as a noise segment preceding a signal segment, individually generated by an uncorrelated Gaussian distribution and concatenated at the presumed arrival time. The means and variances of the two segments are estimated from the data themselves as part of the AIC picking procedure. In practice, zero-mean noise and zero-mean signal sequences result in synthetics whose two segments differ only in variance, and which match the SNR and the picked “changepoint” of the seismogram after which they were modeled. A new AIC arrival-time was picked on each synthetic, and the signed distance between it and the AIC pick on the real seismogram (the assumed truth) was tallied over 1000 simulations to generate the error distribution. For this study we use twice the standard deviation of this distribution, $2\text{SD}_{\text{err}}$, as our standard measure of timing uncertainty, quoted in seconds.

Figure 3.10 shows 12 MERMAID seismograms and the arrival-time picks and uncertainty estimates following the procedures just described. The rows are ordered from low- to high-uncertainty, with the first three seismograms (Figure 3.10a–c) representing the lowest-uncertainty records in the MERMAID catalog, and the final three (Figure 3.10j–l) representing the last three seismograms with picking uncertainties equal to or less than 0.15 s. The middle rows, Figure 3.10d–f and Figure 3.10g–i, show the seismograms for which the corresponding uncertainties straddle the 33rd and 66th percentiles between these two uncertainty bounds, respectively. Each panel of Figure 3.10 plots 30 s of one MERMAID seismogram in blue, centered on the theoretical arrival time of the first-arriving $P$ or $p$ wave in an adjusted ak135 velocity model (dashed black vertical line, eq 3.5). This is the complete segment, after tapering and filtering, which was considered for the AIC arrival-time pick (solid red vertical line, with its estimated uncertainty shown as dashed red vertical lines at $\pm$two-standard deviations along the time axis), for which the corresponding adjusted residual is quoted above each panel (eq. 3.3). Moving clockwise from the upper left corner,
3.8. Estimating First-Arrival Times and Their Uncertainties

Figure 3.10: (caption next page)

[20180819T042909-09,5HTA4C26.MER.DET.WLT5.sac]

(a) $t_0^\text{obs} = -1.36$ s [max. 1.55 s later]

SNR = 5.2e+05
2 SD$_\text{SNR} < 1/f_s$

Time relative to $p$ phase (a)

[20180819T000202-09,5HTA4C26.MER.DET.WLT5.sac]

(b) $t_0^\text{obs} = -3.99$ s [max. 1.45 s later]

SNR = 6.2e+05
2 SD$_\text{SNR} < 1/f_s$

Time relative to $p$ phase (a)

[20180819T020101-09,5HTA4C72.MER.DET.WLT5.sac]

(c) $t_0^\text{obs} = -3.30$ s [max. 1.50 s later]

SNR = 2.9e+04
2 SD$_\text{SNR} < 1/f_s$

Time relative to $p$ phase (a)

[20180506T212523-08,5CD0F92D.MER.DET.WLT5.sac]

(d) $t_0^\text{obs} = 1.97$ s [max. 0.50 s later]

SNR = 1.3e+02
2 SD$_\text{SNR} = 0.06$ s

Time relative to $P$ phase (a)

[20180506T212523-08,5CD0F92D.MER.DET.WLT5.sac]

(e) $t_0^\text{obs} = 2.15$ s [max. 0.40 s later]

SNR = 1.3e+02
2 SD$_\text{SNR} = 0.06$ s

Time relative to $P$ phase (a)

[20180506T212523-08,5CD0F92D.MER.DET.WLT5.sac]

(f) $t_0^\text{obs} = 2.47$ s [max. 0.90 s later]

SNR = 8.0e+01
2 SD$_\text{SNR} = 0.06$ s

Time relative to $P$ phase (a)

[20180201T162007-24,5C54DF84.MER.DET.WLT5.sac]

(g) $t_0^\text{obs} = -0.95$ s [max. 1.25 s later]

SNR = 3.8e+01
2 SD$_\text{SNR} = 0.10$ s

Time relative to $P$ phase (a)

[20180201T162007-24,5C54DF84.MER.DET.WLT5.sac]

(h) $t_0^\text{obs} = 2.13$ s [max. 1.60 s later]

SNR = 3.2e+01
2 SD$_\text{SNR} = 0.10$ s

Time relative to $P$ phase (a)

[20180222T102627-23,5C707946.MER.DET.WLT5.sac]

(i) $t_0^\text{obs} = 1.49$ s [max. 0.15 s later]

SNR = 3.9e+01
2 SD$_\text{SNR} = 0.10$ s

Time relative to $P$ phase (a)

[20180222T102627-23,5C707946.MER.DET.WLT5.sac]

(j) $t_0^\text{obs} = 2.84$ s [max. 0.55 s later]

SNR = 1.8e+01
2 SD$_\text{SNR} = 0.15$ s

Time relative to $P$ phase (a)

[20191109T112553-09,5DCA3596.MER.DET.WLT5.sac]

(k) $t_0^\text{obs} = -1.62$ s [max. 1.60 s later]

SNR = 1.6e+01
2 SD$_\text{SNR} = 0.15$ s

Time relative to $P$ phase (a)

[20191120T083504-09,5D544AS2.MER.DET.WLT5.sac]

(l) $t_0^\text{obs} = 0.44$ s [max. 1.75 s later]

SNR = 1.5e+01
2 SD$_\text{SNR} = 0.15$ s

Time relative to $P$ phase (a)

[20191204T201757-19,5DF1DDE4.MER.DET.WLT5.sac]
3.9 Beyond P Waves: S, T, Surface Waves, and Core Phases

Figure 3.10: (previous page) MERMAID seismograms showing detail in a 30 s window centered on the theoretical first arrival in the adjusted ak135 model (eq. 3.5). In each panel the blue trace is the seismogram filtered between 1 and 5 Hz and plotted in counts, the dashed vertical line at 0 s is the theoretical arrival time of the first-arriving p or P wave, the solid red vertical line is our automatic AIC-based arrival time pick ($t_{AIC}$) with its estimated uncertainty shown as dashed vertical lines extending ±two-standard deviations in both directions along the time axis. Listed above each panel is the adjusted travel-time residual (eq. 3.3) and the delay between our pick and the time at the maximum (or minimum) amplitude of the signal, the latter being listed in brackets on the ordinate axis. Inset in each panel clockwise from top left is the earthquake magnitude, the earthquake depth and its distance, the two-standard deviation error estimate of our pick using Method 1 of Simon et al. (2020), and the estimated SNR of the seismogram.

The seismograms in Figure 3.10 represent the complete set which contributed residuals data to Figure 3.19c–d, the uncertainty threshold there being 0.15 s. This quality criterion was decided upon after inspecting all seismograms in the MERMAID catalog and finding that the trained eye began to distrust the picks with larger uncertainties. For reference, the highest uncertainty among all first-arriving P waves in the catalog is 1.1 s. However, it is important to remember that the uncertainties quoted here correspond to a pick made in a single frequency band in which seismic energy for any given event may be missing and/or is emergent. The picking procedure detailed by Simon et al. (2020) and shown in Figure 3.6 inspects the full bandwidth of the MERMAID seismogram via multiscale wavelet decomposition, and treats every wavelet scale (in practice, scales are roughly analogous to frequency bands) separately for arrival-time estimation, the consideration of all being necessary to the informed matching of seismograms to events.

3.9 Beyond P Waves: S, T, Surface Waves, and Core Phases

Before returning to the main thrust of this study, the high-precision picking of P waves in the MERMAID catalog for future tomographic inversions, we will take a slight detour to review some of the other, unexpected signals that MERMAID has recorded. As mentioned in section 3.3, the MERMAID
3.9. Beyond P Waves: S, T, Surface Waves, and Core Phases

algorithm was written with the express purpose to trigger on tomographic-quality mantle P waves. However, MERMAID has, on occasion, sent us other phases which extend its utility beyond P-wave tomography, including S waves, surface waves, T waves, and even core phases.

3.9.1 S waves

Figure 3.11 is a record section for a large ($M_w$ 6.7) and deep (564.1 km) earthquake in the Fiji Islands region. The layout is similar to that of the record sections plotted in Figure 3.5, except here the seismograms are filtered between a much lower frequency band of 0.1 to 0.2 Hz. Within this band we clearly see the arrival of both $P$ and $S$ waves, the theoretical arrival times of which are marked by black and gray curves, respectively. These are the first published examples of $S$ waves recorded by MERMAID. Because shear waves cannot travel through water, what is actually being recorded here are the signals of vertically-polarized ($S_v$) waves that underwent seismic-
Figure 3.12: Unfiltered MERMAID seismogram showing the arrival of a surface wave from a nearby shallow event in the Tonga Islands.

Acoustic conversion at the seafloor. Indeed, the secondary phases in Figure 3.11 traversed the mantle as S waves, and after conversion at the seafloor, as acoustic waves through the water column to MERMAID. Note that the seismograms in the record section of Figure 3.11 were filtered at a much lower frequency band than those of Figure 3.5, and they were not tapered to show the secondary-arrivals near the edges of the seismograms.

3.9.2 Surface Waves

Figure 3.12 shows a clear surface-wave detection in an unfiltered MERMAID seismogram. This is the first example ever published of a surface wave recorded by MERMAID. Its dominant period is around 10 s, the longest-period resolvable by MERMAID. We know it must be a Rayleigh wave as opposed to a Love wave, because the latter produces (theoretically) no vertical displacement on the (assumed perfectly flat) seafloor, which is required to generate the hydroacoustic pressure signals that MERMAID records. The corresponding event was a large (MW 6.0), shallow (10 km), and proximal (4.1°) earthquake in the Tonga Islands region. Plotted as vertical lines from left to right are the theoretical arrival times as computed in ak135 of the first-arriving P and S waves, as black solid and dashed lines, respectively, and as a red solid line, the theoretical arrival time of a surface wave with a velocity of 3.5 km/s.
3.9. Beyond P Waves: S, T, Surface Waves, and Core Phases

Figure 3.13: High-frequency MERMAID seismogram showing the arrival of a T wave likely from a local shallow earthquake in the Fiji Islands.

We see that the largest-amplitude signal corresponds to the (relatively) low-frequency phase which arrives around the theoretical arrival time of the surface wave. We do not see a large arrival associated with the earlier S wave, as we do in Figure 3.11, at least in the unfiltered data.

Surface-wave amplitudes decay slower than do body wave amplitudes, with the former going as as $1/\sqrt{r}$, and the latter as $1/r$, where $r$ is the path length. Given that the amplitude of the surface wave is so large in this example, we would expect to find other large-amplitude surface waves in the MERMAID data set, if we looked in the right time and place. In this example we were lucky to record this surface wave, because, as explained, the algorithm triggered on the P wave arrival around 100 seconds, not on the surface-wave arrival around 70 seconds later. In the future we will request segments of interest for large and shallow events like this using the two-way Iridium communication built into every MERMAID.

3.9.3 T waves

Figure 3.13 shows an example of a T-wave arrival in a high-frequency MERMAID seismogram. Here, the data are filtered between 5 and 10 Hz, and the theoretical arrival time of the first-arriving P wave is marked around 100 seconds with a black vertical line. The corresponding event was
3.9. Beyond P Waves: S, T, Surface Waves, and Core Phases

a local (2.3°) and shallow (10 km) $m_b$ 5.2 earthquake in the Fiji Islands. We speculate that the secondary arrival that has its maximum amplitude around 150 seconds later is a $T$ wave, and mark in red the theoretical arrival time of phase with a with a velocity of 1.5 km/s. Various observations make us confident this is in fact a $T$ wave: (1) it is of very high frequency; (2) it is of large amplitude compared to the $P$ wave; (3) its arrival is emergent and decays nearly symmetrically. The last point proves it is not a body wave because those, especially at these short distances, arrive impulsively.

3.9.4 Core phases

Figure 3.14 is a record section corresponding to a $M_w$ 6.8 earthquake under the Ionian Sea. It shows the arrivals of phases that transited the core of the Earth, and then were recorded by four MERMAIDS in the South Pacific. Unlike the unambiguous phase arrivals of the cases presented in Figure 3.5, there all direct $p$ or $P$ waves, here there exist four potential phase arrivals in the time window of each seismogram. Also unlike Figure 3.5, here we color the individual MERMAID traces black, and instead color-code the four theoretical-travel time curves, black for $PKPbc$, green for $PKPa$, magenta for $PKiKP$, and blue for $PKiKP$. The number of the recording MERMAID is marked outside the right ordinate axis. Like in Figure 3.5, these travel-time curves are computed in the ak135 velocity model, and they are not adjusted for bathymetry or MERMAID cruising depth. Therefore, the rule of thumb described in section 3.8.3, of adding approximately 1 s to those travel-time curves, applies.

We hypothesized that in cases like Figure 3.14, where multiple core-phase arrivals coexist in each seismogram, the dominant phase actually being detected was $PKPbc$. This is based on a few key observations. First, we can immediately reject the possibility that the inner-core reflection $PKiKP$ was detected, because that phase is rarely detected by a single station (Ohtaki & Kaneshima, 2015), instead requiring multi-station array methods to boost the signal to detectable levels. Second, it was predicted that arrivals associated with either of the two $PKP$ branches would have higher SNRs than those associated with $PKiKP$, because the former transited the extremely
low-attenuating outer core (often approximated as having a bulk quality factor $Q_k \approx \infty$), while the latter dove into the comparatively highly-attenuating inner core (Romanowicz & Mitchell, 2015). Last, beyond the caustic around 145°, $PKP_{bc}$ arrives before $PKP_{ab}$. We therefore expect that the latter-arriving phase would be drowned out by the persistent reverberations in the water column that dominate MERMAID seismograms for many tens of seconds after the first arrival. Such behavior is generally observed in Figure 3.14, leave for MERMAID P009, which appears to record both outer-core phases.

Figure 3.15 tests our hypothesis that we are observing $PKP_{bc}$ arrivals. It plots observed travel-time residuals, and theoretical differential travel-time residuals, against $PKP_{bc}$ computed in an adjusted ak135 velocity model, as a function of epicentral distance. The observed residuals, plotted as black diamonds, were computed in exactly the same manner as the first-arrival residuals of section 3.8, except here the window used for our AIC pick was centered on the adjusted theoretical arrival time of $PKP_{bc}$, and the bandwidth of the filter was narrowed to consider only those frequencies between 1 and 2 Hz. A black line at 0 s is drawn to show the time around which we would expect these residuals to cluster, assuming our hypothesis that we are detecting $PKP_{bc}$ arrivals is correct. The theoretical differential residuals computed against $PKP_{bc}$, each individually adjusted for bathymetry and MERMAID cruising depth, are color-coded as in Figure 3.14. They are marked as open circles at epicentral distances where MERMAID actually recorded seismograms, and they are connected as dotted lines to guide the eye. Note that these curves, which would be smooth for a single earthquake at one depth, exhibit notches because four earthquakes at different depths contributed residual data to this figure.

We submit as evidence Figure 3.14 and Figure 3.15 to prove our hypothesis that in potentially ambiguous cases when MERMAID may be recording both inner- and outer-core phases, it is most likely recording the latter. Figure 3.14 shows that the actual arrivals are much too delayed compared to the $PKIKP$ theoretical arrival time (in blue). And Figure 3.15 shows that when all travel time adjustments are made, and bathymetry and MERMAID cruising depth are accounted for, the residuals observed by MERMAID fall nearest to the expected arrival time of $PKP_{bc}$. Every $PKP_{bc}$
3.9. Beyond P Waves: S, T, Surface Waves, and Core Phases

Figure 3.14: MERMAID record section for a single event in the Ionian Sea, displaying PKPbc outer-core phase arrivals.

Figure 3.15: Adjusted travel-time residuals (diamonds; eq. 3.3), and adjusted theoretical differential travel-time residuals (colored curves), computed against PKPbc, considering all outer-core phases recorded by MERMAID.
3.9. Beyond P Waves: S, T, Surface Waves, and Core Phases

Figure 3.16: MERMAID seismogram displaying what are likely two separate unidentified events, not recorded anywhere else.

residual in the catalog displays a positive bias, meaning it is delayed with respect to the theoretical arrival time. Their mean residual is 2.82 s.

We end here with a note that MERMAID records \textit{PKIKP} phases as well. Indeed, the MERMAID catalog also includes two tomographically-useful \textit{PKIKP} arrivals, not shown in Figure 3.14 or Figure 3.15, which were recorded at epicentral distances where there existed no ambiguity in what phase was being detected.

3.9.5 Unidentified (local) events

We end this tour of the data MERMAID returns beyond teleseismic \textit{P} waves with an example of unidentified events in Figure 3.16. In this example we see two distinct arrivals: the first is around 80 seconds, and the second larger arrival is just under one minute later. Neither of these two arrivals match with any theoretical phase-arrival times associated with any known events in the global seismic catalogs. Our interpretation is that these two signals are both \textit{p-} (or, less likely, \textit{P-}) wave arrivals from two distinct, very nearby events.

There are a few observations that support this hypothesis. First, both signals are impulsive, especially the secondary signal, proving it cannot be a \textit{T} wave associated with the earlier signal.
Second, it is also unlikely that the second, larger arrival is an $S$ wave because: (1) it is of very high frequency, which is not expected of the shear conversion, and (2) the $S$–$P$ delay time would imply an epicentral distance of approximately 500 km (both arguments can be made against a surface wave as well, the epicentral distance computed in that case being even larger). The last argument, especially, makes it very unlikely that Figure 3.16 is showing an $S$ wave because any event large enough to generate high-frequency, high-amplitude $S$ waves that propagated $\sim5^\circ$ would have assuredly been recorded and located by other seismometers in the region. Additionally, no other MERMAID recorded these events, further proving their extremely local nature. Note as well that both arrivals are still very distinct at frequencies up to 10 Hz, not shown here, but which is uncommon for identified MERMAID signals, except in the case of local events like in Figure 3.13. Thus we are left to conclude that Figure 3.16 includes two distinct arrivals, from two very proximal earthquakes.

3.10 Nearby Island Seismic Stations

To this point we have discussed the MERMAID instrument, its seismograms, our procedure to match those to known global earthquakes, our methods to pick various arrivals with high precision and to estimate their uncertainties, and we have jaunted through a tour of the various signals present in the MERMAID data set. For the remainder of the study we will remain solely focused on the travel-time residuals of first-arriving $p$ and $P$ waves. In this section we aim to prove the tomographic utility of MERMAID residuals by comparing their statistics against measurements made for the same events at island stations located in the oceanographic neighborhood of the slowly dispersing SPPIM array.

We compare our uncalibrated MERMAID hydroacoustic pressure records (the MERMAID “seismograms”) with velocity seismograms from land-based seismic sensors. We then compare the MERMAID catalog of travel-time residuals, uncertainties, and SNR estimates, with a similar catalog that we construct for island seismic stations in the vicinity of MERMAID. Figure 3.1 shows the bounding box of MERMAID’s oceanic neighborhood, drawn with an approximately $2^\circ$ buffer.
3.10. Nearby Island Seismic Stations

around the maximum extent of the SPPIM array, as deployed. At 32 million km$^2$, it spans a large portion of the South Pacific and some of the North Pacific, nearly 6.5% of Earth’s surface, or roughly double the area of Russia.

3.10.1 Data retrieval

We queried IRIS for terrestrial seismometers in this neighborhood with data publicly available after July 2018. This returned 19 stations: fourteen “traditional” seismic sensors from GEOSCOPE (G), the Australian National Seismograph Network (AU), the Red Sismológica Nacional (C1), and the Global Seismograph Networks IRIS/IDA (II) and IRIS/USGS (IU); and five low-cost Raspberry Shake instruments (AM). Table 3.7 lists these stations and their locations. They amount to one for every 2.3 million km$^2$, an area larger than Saudi Arabia, and very inhomogeneously clustered on islands. Additionally we obtained data from six short-period seismometers in the Réseau Sismique Polynésien (RSP) maintained by the Centre Polynésien de Prévention des Tsunamis in Papeete, Tahiti, French Polynesia. Those stations and their locations are listed in Table 3.8.

To construct the complementary data set we retrieved every available seismic trace from these nearby stations, beginning five minutes before the arrival of the first ak135 phase for all 284 identified events in our MERMAID catalog.

For each station listed in Table 3.7 we requested traces from every location, for all M* (mid period; sampling rate between 1–10 Hz), B* (broadband; 10–80 Hz), H* (high broadband; 80–250), S* (short period; 10–80 Hz), and E* (extremely short period; 10–80 Hz) vertical channels. No data from mid-period instruments were returned, and all Raspberry Shake stations were short-period or extremely short-period instruments. This yielded 7305 traces. Of those, 6885 were from the “traditional” sensors, representing data recorded during all 284 earthquakes in the MERMAID catalog, and 420 were from Raspberry Shake instruments, accounting for data recorded during a subset of only 164 of those same earthquakes. The latter instrument class had less data available because, unlike the “traditional” stations that were all installed before MERMAID P008 was deployed, not all Raspberry Shake stations in Table 3.7 were installed before the deployment of the SPPIM array.
3.10. Nearby Island Seismic Stations

Table 3.7: “Nearby” stations with data available from Incorporated Research Institutions for Seismology (IRIS).

<table>
<thead>
<tr>
<th>FDSN code</th>
<th>Station</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>FUTU</td>
<td>-14.307673</td>
<td>-178.121070</td>
</tr>
<tr>
<td>G</td>
<td>PPTF</td>
<td>-17.589600</td>
<td>-149.565250</td>
</tr>
<tr>
<td>G</td>
<td>TAOE</td>
<td>-8.854920</td>
<td>-140.147750</td>
</tr>
<tr>
<td>AU</td>
<td>NIUE</td>
<td>-19.076300</td>
<td>-169.927200</td>
</tr>
<tr>
<td>Cl</td>
<td>VA02</td>
<td>-27.160299</td>
<td>-109.434502</td>
</tr>
<tr>
<td>II</td>
<td>MSVF</td>
<td>-17.744801</td>
<td>178.052795</td>
</tr>
<tr>
<td>II</td>
<td>RPN</td>
<td>-27.126699</td>
<td>-109.334396</td>
</tr>
<tr>
<td>IU</td>
<td>AFI</td>
<td>-13.909300</td>
<td>-171.777298</td>
</tr>
<tr>
<td>IU</td>
<td>FUNA</td>
<td>-8.525900</td>
<td>179.196594</td>
</tr>
<tr>
<td>IU</td>
<td>KNTN</td>
<td>-2.774400</td>
<td>171.718597</td>
</tr>
<tr>
<td>IU</td>
<td>PTCN</td>
<td>-20.071301</td>
<td>-130.095306</td>
</tr>
<tr>
<td>IU</td>
<td>RAO</td>
<td>-29.245001</td>
<td>-177.929001</td>
</tr>
<tr>
<td>IU</td>
<td>RAR</td>
<td>-21.212500</td>
<td>-159.773300</td>
</tr>
<tr>
<td>IU</td>
<td>XMAS</td>
<td>2.044800</td>
<td>-157.445694</td>
</tr>
<tr>
<td>AM</td>
<td>R028A</td>
<td>-17.693694</td>
<td>-149.574629</td>
</tr>
<tr>
<td>AM</td>
<td>R06CD</td>
<td>-17.567568</td>
<td>-149.570661</td>
</tr>
<tr>
<td>AM</td>
<td>R0EF4</td>
<td>-17.720721</td>
<td>-149.297984</td>
</tr>
<tr>
<td>AM</td>
<td>*RC78F</td>
<td>-17.531532</td>
<td>-145.509009</td>
</tr>
<tr>
<td>AM</td>
<td>RF737</td>
<td>-17.531532</td>
<td>-149.474642</td>
</tr>
</tbody>
</table>

Table 3.8: “Nearby” stations from the Réseau Sismique Polynésien (RSP) network, whose data were made available to us by Dr. Olivier Hyvernaud.

<table>
<thead>
<tr>
<th>Network</th>
<th>Station</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSP</td>
<td>PAE</td>
<td>-17.661050</td>
<td>-149.579700</td>
</tr>
<tr>
<td>RSP</td>
<td>TVO</td>
<td>-17.782500</td>
<td>-149.251667</td>
</tr>
<tr>
<td>RSP</td>
<td>PMOR</td>
<td>-15.002222</td>
<td>-147.894167</td>
</tr>
<tr>
<td>RSP</td>
<td>VAH</td>
<td>-15.236589</td>
<td>-147.628472</td>
</tr>
<tr>
<td>RSP</td>
<td>TBI</td>
<td>-23.348889</td>
<td>-149.460833</td>
</tr>
<tr>
<td>RSP</td>
<td>RKT</td>
<td>-23.124790</td>
<td>-134.972000</td>
</tr>
</tbody>
</table>

128
3.10. Nearby Island Seismic Stations

From the short-period instruments at the stations in Table 3.8 we obtained 1534 traces, corresponding to all of the 284 MERMAID events. These data are not publicly distributed or long-term archived—we thank Dr. Olivier Hyvernaud for sharing them.

3.10.2 Data processing

Each trace had its mean and trend removed, and was tapered at both ends with a symmetric cosine taper of 5% the length of the trace (per SAC defaults). The instrument responses on record in the pole-zero (SACPZ) files were removed using SAC, deconvolving the traces from digital counts to velocity records. Each trace was high-pass filtered above 0.1 Hz and low-passed below 10 Hz to maintain the bandwidth between 0.1–10 Hz. These frequencies were chosen to correspond as closely as possible to the sensitivity band of a MERMAID instrument, whose pressure records are filtered between those bounds before digitization, and whose instrument gain is reported by the manufacturer to be flat within that bandwidth.

SACPZ and/or StationXML files with response data were readily available for the stations in Table 3.7. SACPZ files were not available for the stations in Table 3.8. The Supplemental Material contains the necessary details and the software to perform instrument correction, which will be of use to others.

Figure 3.17 replots Figure 3.5 to include the velocity seismograms from nearby stations, normalized for easy viewing. Station names are labeled inside the right ordinate axis. As in Figure 3.5, the MERMAID traces are color-coded for easy differentiation, while those from nearby stations are left gray. For clarity, overlapping traces (e.g., corresponding to stations on Tahiti, and those with multiple channels) were removed. To mimic a MERMAID seismogram we trimmed all seismograms to a length of 250 s, with the theoretical first-arrival time at 100 s (the approximate time of the STA/LTA trigger in MERMAID seismograms), and tapered them with the same Hanning window that was applied the seismograms in Figure 3.5. When required, they were decimated from their native sampling frequency to 20 Hz (or 25 Hz) before band-pass filtering to mirror the sampling frequency of MERMAID. In this way, we may qualitatively compare the waveforms and their
3.10. Nearby Island Seismic Stations

Figure 3.17: Record sections, as in Figure 3.5, but now also including, in gray, data from nearby terrestrial, traditional and Raspberry Shake, stations. The gray records were trimmed to a length of 250 s and then processed in exactly the same manner as the MERMAID data so that SNR comparisons of the three instrument platforms may be made easily by eye. We see that MERMAID SNRs for the events shown compare favorably to the island stations. This comparison is formalized in Figure 3.18d–f for high-quality residuals culled from all three instrument platforms.

SNRs between the instrument types, as more formally discussed next.

Figure 3.17 serves merely as a visual aid to appreciate the types of signals that MERMAID records compared with other stations, given the same earthquake. We do not use the gray waveforms as shown to make first-arrival picks. Rather, for every first-arrival time reported in this study, regardless of instrument, we make the arrival-time picks on segments like those in Figure 3.10, not like those shown in Figure 3.5 or Figure 3.17. Hence, regardless of instrument type, each trace
was processed as described in section 3.8. For the island stations, the only difference was that, if required, they were decimated to 20 Hz or 25 Hz to match the sampling frequency of MERMAID, and no bathymetric (or elevation) time corrections were applied. Seismograms were rejected if they were less than 200 s long due to missing data, if they had any missing data within the taper window described in section 3.8.1, or if the theoretical first-arrival time was near enough to an edge to result in the deconvolution taper used to remove the instrument response overlapping with the taper used for arrival-time picking.

### 3.10.3 MERMAID residuals versus nearby island stations

Figure 3.18 shows the distributions of travel-time residuals (top row), their SNRs (middle row), and their two-standard deviation uncertainty estimates (bottom row), for “traditional” seismometers, MERMAID, and Raspberry Shake stations, left, middle, and center, respectively. We consider this a substantiation that MERMAID records tomographically-useful $p$- and $P$-wave arrivals.

Starting with the top row of Figure 3.18 we plot the travel-time residuals for all first-arriving $p$ and $P$ waves whose residuals fell within 10 s, either before or after, their theoretical arrival time as computed in ak135. For MERMAID (Figure 3.18b), we use the adjusted travel times explained in section 3.8.3, and label them appropriately as $t^*_\text{res}$, while for the “traditional” (Figure 3.18a), and Raspberry Shake stations (Figure 3.18c), we plot the unadjusted residuals. The only quality-control applied to the MERMAID residuals was the rejection of those that exceeded (positively or negatively) the 10 s cutoff. For the residuals in Figures 3.18a and 3.18c we applied the additional quality criteria that their SNRs (middle row), and the two-standard deviations of their uncertainty estimates (bottom row), must be equal to or greater than, or equal to and less than, the corresponding values in the MERMAID first-arrival data set (middle column), respectively. This was done to mimic what the author’s eye had already rejected. Indeed, during manual review, often the author would reject, for various reasons, a phase-arrival pick, $t_{\text{AIC}}$, which aligned nicely with a theoretical arrival time. This is because our AIC picker will always report an arrival as long as its SNR is greater than one. Therefore, often by algorithmic necessity the picker will trigger on something
3.10. Nearby Island Seismic Stations

Figure 3.18: MERMAID travel-time residuals (top row), their SNRs (middle row), and their estimated uncertainties (bottom row) compared to “traditional” seismometers and Raspberry Shake stations installed on islands in the general neighborhood of the SPPIM deployment. MERMAID data (middle column; blue) most closely resemble those of the “traditional” stations (left column; green), and display much higher mean and maximum SNRs, and much lower mean estimated uncertainties, than its Raspberry Shake neighbors (right column; raspberry).

that is extremely low-SNR (just above one), which by coincidence also aligns nicely with some theoretical first-arrival, but which the human eye would readily reject. Every single MERMAID seismogram discussed here was reviewed by a human and a phase-arrival was verified to exist, but not every seismogram from the nearby stations was reviewed (rather, each was simply run though the same phase-picking and processing algorithm). Therefore, it was determined that the minimum SNR and maximum two-standard deviation of the uncertainty estimates of the verified data set could serve to approximately winnow the nearby data to the standards by which the author’s eye accepted or rejected a pick. This is clearly an imperfect process, and in doing it we are not
implying that noise levels across all three instrument classes are equal (they are not, see the middle row). The number of traces that passed this winnowing process and contributed data to the histograms of Figure 3.18a–c is quoted above each panel.

In total, we see that the distribution of first-arrival MERMAID residuals in Figure 3.18b agrees well with the complementary distribution from “traditional” seismometers in Figure 3.18a, and to a lesser extent the same distribution computed for the Raspberry Shake stations in Figure 3.18c. All are positively biased, meaning that, on average, the arrival time of the first-arriving $p$ or $P$ wave was late compared to the ak135 reference model. The means and standard deviations of the distributions are quoted inside each panel, and the former is marked by a vertical dashed line in each. We find that MERMAID reports a mean residual that falls between the other two instrument classes. Similarly, the standard deviation of the MERMAID residuals is less than the same statistic computed for the other two instrument classes. These findings bolster our claim that MERMAID reports data useful for seismic tomography.

The middle row of Figure 3.18 displays histograms of the SNRs of the first-arrival residuals. Quoted inside each panel is the minimum, median, and maximum SNR of the data set, and their means are marked by a dashed vertical line. Recall that the minimum SNR is the same for all three histograms because that is the minimum SNR in the MERMAID data set, and it was used as a quality threshold for the others. Despite our best attempts to winnow the data from nearby stations, we believe we are still seeing the result of human intervention to a greater extent in the shape of the MERMAID histogram when compared to the others. We note that the MERMAID SNR histogram has its mode nearer its mean than in either of the other two cases, and it ramps up to its maximum value as opposed to starting at or near it, and then decreases by some exponential curve. Likely, this is the result of the author rejecting those lower SNR picks in the MERMAID data set via manual review, and not performing the same manual intervention for the other data.

Like the residual data, we see that, very generally, MERMAID SNRs fall between those computed from “traditional” stations and Raspberry Shake instruments. The median SNRs of the MERMAID and “traditional” data sets are identical, and both are higher than the same statistic for the
Raspberry Shake data set. Interestingly, the maximum SNR among the three differs greatly, with that of the “traditional” stations being much greater than MERMAID, which is much greater than Raspberry Shake. Note that the abscissa axes in Figure 3.18e–f are in base-ten logarithmic scales, and mass extends beyond the limits plotted. To give a complete picture of the data set we compiled, here we plot all available data that passed our winnowing procedure from each instrument class. In Figure 3.20 of the Supplementary Material we recreate Figure 3.18 considering only the earthquakes for which all three instrument classes had at least one station returning data—i.e., for only those events common to the three catalogs (meaning they occurred after the installation of Raspberry Shake stations, the most data-limited instrument class in our study). That figure presents a proper apples-to-apples comparison of the SNRs returned by the three instrument classes, which we now quote. Considering the subset of events common to all catalogs, at their best: “traditional” seismometers recorded the first-arriving phase with an SNR of $2.6 \times 10^7$; MERMAID with an SNR of $6.3 \times 10^4$; and Raspberry Shake with and SNR of $4.6 \times 10^2$. In our definition of the SNR (eq. 3.1), for the same amplitude signal, that equates to a 26 dB reduction in the noise level of a “traditional” station as compared to MERMAID, and a 21 db reduction in the noise level of MERMAID compared with Raspberry Shake. One caveat to this, however, is that we have seen that MERMAID seismograms can contain high-amplitude reverberations for many tens of seconds after the initial arrival that can artificially inflate the SNR of those seismogram. Proper noise-spectra comparisons are the target of future work, but for now these approximations suffice.

For those who are interested in a more detailed comparison of the waveforms than can be gleaned from Figure 3.17, we have included in the Supplementary Material Figure 3.21, Figure 3.22, and Figure 3.23, which each plot the 12 highest-SNR seismograms from the three instrument classes, presented in the same format as Figure 3.10, considering only the data in the catalog common to all. Thus the first panel (a) in each plots the highest-SNR seismogram just discussed for noise comparison. There it is immediately obvious that the noise levels for both “traditional” stations and MERMAID stations are much lower than Raspberry Shake, with MERMAID being more like the former than the latter.
3.10. Nearby Island Seismic Stations

The final row of Figure 3.18 shows histograms of the corresponding two-standard deviation uncertainty estimates, $2\text{SD}_{\text{err}}$, discussed in section 3.8.4, associated with each first-arrival residual. These are the values quoted in the lower-right legends in each panel of Figure 3.10. As in the SNR histograms of the middle row, the minimum, median, and maximum values of the uncertainty estimates are labeled inside each panel, and their means are marked by dashed vertical lines. Similarly, like in the middle row, mass extends beyond the limits of these histograms. We see that the uncertainties associated with “traditional” stations in Figure 3.17g display a satisfying exponential decay, with their mode nearest the lower end of the uncertainty scale. The uncertainties associated with MERMAID residuals in Figure 3.17h display a softer exponential decay, and they do enjoy the lowest uncertainties at nearly the same frequency as the “traditional” stations. The distribution of Raspberry Shake uncertainties in Figure 3.17i is quite different from either of the other two instrument classes. It does not display the obvious peakedness at the lower-end, and it is approximately uniform across the full range from low to high uncertainty. Further, while the lowest and highest uncertainties corresponding to each instrument class are roughly the same, the median values are quite different, with that associated with Raspberry Shake thrice that of the others.

One caveat concerning our method of uncertainty estimation developed in Simon et al. (2020) is that for comparisons to be usefully made across various instrument types, as we have done here, the data must all be downsampled to match the sampling frequency of the lowest-sampled instrument. Our method relies on estimating uncertainties in terms of samples, which are converted to time via multiplication with the sampling interval. Therefore, given the same estimated sample uncertainty, the timing-uncertainty associated with a 100 Hz Raspberry Shake seismogram that has not been downsampled would be reported as being much lower than that associated with a 20 Hz MERMAID seismogram. In that sense, rather than considering the timing-uncertainty estimates output by our method as absolute times, they may better serve as relative metrics for comparisons between and across data sets. Practically speaking, they map an SNR to a time, which “the eye” may or may not agree with, but they nonetheless provide a convenient means to sort and winnow data, as we will do next.
3.11 Toward South Pacific P-Wave Tomography

Having proven the quality of MERMAID residuals compared to the best data available from permanent island stations in the area, we finally move to placing them in their geographic context to explore the velocity perturbations that they reveal of the mantle under the South Pacific.

Figure 3.19 plots the highest-quality first-arrival $p$- and $P$-wave travel-time residuals of our MERMAID data set. The residuals are color-coded blue for fast (the first-arrival is early compared with theory) and red for slow (the first-arrival is late compared to theory), and they are smeared along their ray paths from source to receiver. We plot them against three velocity models: ak135 at the top (a; eq. 3.2); ak135 adjusted for bathymetry and MERMAID cruising depth in the middle (c; eq. 3.3); and the fully-elliptical 3-D model LLNL-G3Dv3 at the bottom (e; eq. 3.4). In all three cases the initial residuals were computed in the adjusted ak135 model, $t^\star_{ak135}$, as is shown in Figure 3.10, and then each was individually readjusted using the relative travel-time difference between that model and ak135 or LLNL-G3Dv3 to generate Figure 3.19a and and Figure 3.19e, respectively. This means that the residuals shown here were not re-picked using slightly adjusted windows based on the updated model, which is acceptable because the maximum absolute 3-D--1-D travel-time difference for all residuals plotted in Figure 3.19 is 4.66 s, thus still well within a 30 s window centered on the theoretical first-arrival, as in Figure 3.10.

We only include the highest-quality residuals—those about which we are most confident—in Figure 3.19. To ensure that our AIC-based picker triggered on a legitimate phase-arrival, and not on some other spurious energy, we rejected any residuals which exceeded positive (or negative) 10 s. More importantly, we also rejected any residuals whose two-standard deviation uncertainty estimates were greater than 0.15 s, the limit beyond which the author’s eye began to distrust the picks and/or it was felt that the uncertainties were underestimated. For reference, Figure 3.10 displays the full range the data plotted in Figure 3.19c and Figure 3.19d: from the four lowest-uncertainty residuals (top row), through the 33rd and 66th percentiles of uncertainty (second and third row, respectively), to the four highest-uncertainty residuals which passed muster in the bottom row of Figure 3.10.
3.11. Toward South Pacific P-Wave Tomography

Figure 3.19: Smeared travel-time residuals computed against ak135 (a; eq. 3.2), ak135 corrected for bathymetry and MERMAID cruising depth (c; eq. 3.3), and LLNL-G3Dv3 (e; eq 3.4), and their distributions of those residuals in (b), (d), and (f), respectively. Here we only show the highest-quality residuals in the MERMAID data set: those with maximum two-standard deviation estimated uncertainties less than 0.15 s (the final row of Figure 3.10 shows the three “worst” seismograms that made the cut). The colorbar is in units of absolute time, and its color is the residual between our pick and the theoretical arrival time of the reference model (blue is fast, red is slow). The final map (e) and its corresponding histogram (f) includes 3-D mantle and ellipticity corrections absent in the two prior sets, and thus the residuals shown there are the truest picture of the velocity perturbations yet recorded by MERMAID.
In total, 500 residuals passed these quality thresholds for the unadjusted 1-D model, 503 for the adjusted 1-D model, and 502 were retained in the 3-D case. The distribution of those residuals, corresponding to each model, are displayed in the histograms in the right column of Figure 3.19. The mean, standard deviation, and skewness of the histogram is listed inside each panel, and the former is marked by a dashed vertical line. Two numbers are bracketed in the upper right corner of each histogram. The first quotes the number residuals plotted in the histogram (mass may extend beyond the limits of the abscissa axis), and the second reports the total number of residuals which passed quality-thresholding and are plotted as smeared residuals on the respective map to the left of each histogram. The statistics quoted for each histogram were computed using the latter set.

Starting with the smeared residuals in Figure 3.19a, and their corresponding distribution in Figure 3.19b, we generally see large positive anomalies (red; tardy) associated with equatorial ray paths, and lower-amplitude negative anomalies (blue; early) associated with more polar ray paths. There are two explanations for this: (1) these data are not corrected for bathymetry or MERMAID cruising depth, which we have shown to add 1 s to the ak135 residuals in normal circumstances; and (2) the ak135 model is spherical and thus does not account for ellipticity, resulting in larger delays for the (longer in the real Earth) equatorial ray paths. The first point results in an overall mean-shift of around 1 s for all residuals in the histogram in Figure 3.19b, and the second point adds an additional error whose geographic distribution is governed by seismological back azimuth.

The residuals in the middle row of Figure 3.19 have been adjusted for bathymetry and MERMAID cruising depth, though they remain in the spherical ak135 velocity model. As such, the mean-shift in Figure 3.19b has been reduced by over 1 s in Figure 3.19d, but the signal of Earth’s ellipticity remains visible in Figure 3.19c. In fact, that signal is now more pronounced in the North Pacific, those data generally displaying negative, or at most weakly-positive, residuals before applying the adjustment. This image proves why it will be absolutely vital in the ultimate tomographic inversion to use a fully-3-D reference velocity model to compute the travel-time residuals.

Finally, the residuals presented in the bottom row of Figure 3.19 are the closest to the real signal of velocity perturbations within the Earth’s mantle yet presented. They are computed against the
3.12. Conclusion

fully-3-D and elliptical velocity model LLNL-G3Dv3. Immediately we see that we have finally removed the signal of Earth’s ellipticity; the ray paths through the North Pacific no longer display generally large negative anomalies, and residuals smeared along equatorial ray paths see their generally large positive residuals lowered slightly. The map is still very red, however, implying that the majority of the 3-D residuals recorded by MERMAID displays positive anomalies for all back azimuths. Figure 3.19f proves this to be the case, showing that, on average, the residuals recorded by MERMAID in the South Pacific are around 1 s late compared to LLNL-G3Dv3. This means that, more often than not, MERMAID recorded seismic waves that traversed slow regions of the mantle. Interestingly, the distribution of these residuals in Figure 3.19f is actually of higher-mean than the analogous distribution for the adjusted 1-D-model in Figure 3.19d. It is also satisfying to note that adjustment to the 3-D model lowered their standard deviation to the smallest value among all three models (albeit marginally). Further, like the other histograms in Figure 3.19, this one displays positive skewness, but most interestingly, it shows the largest positive skewness among all three. Indeed, 339 of the 502 residuals plotted are positive. Further, we recorded no large-negative residuals—the lowest (earliest arrival compared to theory) 3-D residual in Figure 3.19e and Figure 3.19f is \(-4.24\) s. Conversely, the positive residuals display an exponential decay that continues to, and extends beyond, the abscissa axis limit of eight seconds. The maximum 3-D residual the plotted in Figure 3.19e is 9.49 s, corresponding to a \(M_W\) 5.3 earthquake at 10 km depth the Samoa Islands region. It was recorded by MERMAID P012, which was 12.74° to the east, sampling an extremely low-wavespeed mantle along the way.

3.12 Conclusion

We described a new seismic instrument, the third-generation Mobile Earthquake Recording in Marine Areas by Independent Divers (MERMAID), which records earthquakes and transmits their seismograms in nearly real-time from the global oceans. They dive to 1500 m depth below the sea surface and passively drift with the mid-column currents while monitoring the ambient acoustic
3.12. Conclusion

wavefield, surfacing only to relay seismic data, their location, and to download new command files. We discussed the South Pacific Plume Imaging and Modeling (SPPIM) project, which has launched an array of 50 MERMAIDs in the South Pacific Ocean, deployed and maintained by a global consortium, EarthScope-Oceans. The array was completed in August 2019, and as of this writing 48 MERMAIDs are reporting data (see earthscopeoceans.org), and will be for many years to come. We highlighted the time-variable nature of the locations of the subset of 16 Princeton-operated MERMAIDs, from their deployment in August 2018 through the end of 2019, whose data were the focus of this study.

We proposed a workflow to quickly process the continuous data stream of incoming untagged seismograms, and to match them with earthquakes in the global catalog. We reported on the quality and size of the resultant MERMAID seismic catalog, a data product of this study, built up over 16 months of deployment: which earthquakes MERMAID recorded, and which it missed. We found that our MERMAIDs averaged around 30 event detections per year, equating to an expected 150 over their projected five-year lifespan, though we found these numbers to be highly variable between different MERMAIDs, largely due to their proximity to areas of different earthquake rates. We discussed the statistics of completeness for our MERMAID seismic catalog and parsed its numbers by magnitude to reveal the types of earthquakes to which MERMAID proved itself most sensitive. We found that for “typical” global earthquakes, an “average” MERMAID had around a 0.5% chance of recording a M5, a 9% chance of a recording a M6, a 42% chance for an M7, and an 81% chance of recording a M8 earthquake.

We summarized a procedure to pick, with high precision, the arrival times of phases in MERMAID seismograms, and to estimate their uncertainties. We commented on the discovery of phases other than mantle $P$ waves (on which the MERMAID algorithm was designed to trigger), via our phase-picking method. Among these many signals we highlighted $S$ waves, surface waves, $T$ waves, core phases, and unidentified earthquakes, the latter being small and local earthquakes, likely recorded by no other instrument on Earth.

We compared our catalog of first-arrival residuals, another data product of this study, against
3.13. **Data and Resources**

A similarly-derived catalog computed using all seismic instruments in the general vicinity of the SPPIM deployment. In all, we collected nearly 9000 seismograms from 25 island stations, corresponding to the 284 unique earthquakes that MERMAID recorded. We compared the distributions of first-arrival travel-time residuals, SNRs, and their uncertainties between “traditional” seismic stations, MERMAID, and Raspberry Shake instruments, and found the MERMAID data had more in common with the former than the latter, proving that MERMAID is indeed in the process of recording tomographically useful data.

Finally, we winnowed our set of first-arrival $p$- and $P$-wave travel time residuals down to the highest-quality subset—a little over 500 picks—which we compared against the fully-3-D and elliptical model LLNL-G3Dv3. We found that, on average, phase arrivals at MERMAID were delayed approximately 1 s, revealing that the novel ray paths sampled in this study navigated slow regions of the Earth’s mantle. We displayed these residuals smeared along their ray paths to gain a geographic sense for the signature of those velocity anomalies under the South Pacific. These residuals, their weights being dictated by the associated uncertainties computed here, will form the basis of future tomographic inversions to probe the structure beneath the South Pacific Superswell.

### 3.13 Data and Resources

We rely on irisFetch.m version 2.0.10, available from IRIS, to query seismic catalogs available through the International Federation of Digital Seismograph Networks (FDSN). We use MatTaup, written in MATLAB by Qin Li while at the University of Washington in 2002, to compute theoretical travel times in the ak135 velocity model of Kennett et al. (1995). We maintain all of those codes, with minor modifications, at github.com/joelsimon/omnia, which furthermore contains all of our software developed for this study.
3.14 Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656466. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are grateful to Dr. Olivier Hyvernaud from the Centre Polynésien de Prévention des Tsunamis in Papeete, Tahiti, French Polynesia, for providing seismic data from the Réseau Sismique Polynésien array, as well as guidance on how to remove their instrument response. Constructive comments by Dr. Guust Nolet were greatly appreciated.

3.15 Supplemental Material

3.15.1 Comparing the highest-SNR seismograms across the instrument classes

Figure 3.20 remakes Figure 3.18 considering only the subset of data representing events in the MERMAID catalog that occurred while at least one “traditional” and one Raspberry Shake instrument were installed—i.e., the catalog of events common to all instrument classes. This comparison is made here because many of the largest events present in the data plotted in Figure 3.10 and Figure 3.18 are not in the Raspberry Shake catalog because those stations had not yet been installed.

Figure 3.21, Figure 3.22, and Figure 3.23 each plot the 12 highest-SNR signals in this catalog of common events for “traditional” island stations, MERMAID, and Raspberry Shake island stations, respectively. It is readily apparent that Raspberry Shake instruments are generally noisier than either of the other two instrument classes because the variance of the gray noise segment that precedes the signal is often visible, whereas for the other two instrument classes this not the case (the noise looks flat at this range of ordinate values). Also, the uncertainties associated with Raspberry Shake seismograms are generally higher than those of the other two instrument classes.
3.15. Supplemental Material

Figure 3.20: Figure 3.18 remade considering only the subset of events for which data existed for at least one station within each instrument class.

3.15.2 Writing SAC pole-zero files for Réseau Sismique Polynésien stations

3.15.2.1 The SACPZ file

Seismic Analysis Code (SAC) pole-zero (SACPZ) files specify the frequency response of a digital seismic instrument. They describe how a seismometer converts ground motion to digital counts. Therefore, the output in digital counts of a seismometer may be considered to be a record of true ground motion multiplied by the response of the instrument in the frequency domain. With a SACPZ file, one may recover an accurate record of ground motion via deconvolution (division in the frequency domain) of the seismogram with the frequency response. This process is commonly referred to as “removing the instrument response,” and it can be accomplished in the SAC program with the TRANSFER command.
Figure 3.21: The 12 highest-SNR signals recorded by "traditional" island stations considering the catalog of events common to all three instrument classes, presented in the same format as Figure 3.10.
Figure 3.22: The 12 highest-SNR signals recorded by MERMAID considering the catalog of events common to all three instrument classes, presented in the same format as Figure 3.10.
3.15 Supplemental Material

Figure 3.23: The 12 highest-SNR signals recorded by Raspberry Shake island stations considering the catalog of events common to all three instrument classes, presented in the same format as Figure 3.10.
Ignoring optional comments, SACPZ files contain three main components: poles, zeros, and a constant. The first two are the complex roots (poles: denominator; zeros: numerator) of the transfer function of the analog component of the instrument, and the last is a constant factor that describes the gain of the entire system. By analog we mean the physical seismic instrument—for example, an inertial mass held in place by a varying electric current, such that the input is ground motion (e.g., m/s) and the output is voltage (V). Following the analog stage, seismometers pass their data through multiple stages of digitization where the voltage is converted to digital counts. Ignoring any frequency effects during digitization (which SACPZ files do not include), the poles and zeros are enough to describe the phase response of the system—i.e., with no constant, the ungained seismogram after deconvolution will have the proper shape but incorrect amplitude. Note that phase shifts acquired during the digital stages are usually negligible and can be ignored. This fact is noted (in bold) on page 152 of the Standards for the Exchange of Earthquake Data (SEED) Reference Manual Version 2.4 (2012, fdsn.org/pdf/SEEDManual_V2.4.pdf), page 409 of the Seismic Analysis Code Users Manual Version 101.6a (2014, ds.iris.edu/files/sac-manual/sac_manual.pdf), and has been independently verified by the authors by comparing waveforms deconvolved with SACPZ and RESP files (the latter of which take into account all digitization stages). We include this comment to make the point that our method of removing the instrument response using the humble SACPZ file (and not other file standards like RESP or StationXML, which encode information concerning the full cascade of digital filters) is sufficient to recover an accurate record of ground motion for the RSP instruments used in this study.

A SACPZ file may be specified in terms displacement, velocity, or acceleration, with the values varying in each case for the same seismometer. Very commonly a seismometer will physically measure ground velocity, in which case the poles and zeros will likely be reported for the velocity transfer function of the analog stage, and the gain constant (also called the “sensitivity”) will be given in units like counts/(m/s). SACPZ files were not available for the six stations used in this study from the Réseau Sismique Polynésien (RSP) network. However, we were provided the poles
and zeros, and a gain constant at a specific frequency, which is enough to write our own SACPZ file. It is important to note that a gain constant at a single frequency is not the same thing as the constant of a SACPZ file. To be unambiguous we will hereafter refer to the former as the sensitivity. Indeed the sensitivity describes a gain factor at a single frequency, while the SACPZ constant describes the gain factor at all frequencies. Using the notation of the SEED manual, and the pole-zero representation of the transfer function, the frequency response at any stage of the system is (eq. 4 pg. 158),

$$G(f) = S_d R(f),$$

(3.8)

where $R(f)$ is some function of frequency and the $S_d$ is the sensitivity. For the analog stage,

$$R(f) = A_0 H_p(s),$$

(3.9)

where $A_0$ is a normalization factor at frequency $f_s$ in Hz (note that the normalization factor may be derived at a frequency, $f_n$, different from $f_s$, but this is goes against the SEED convention (pg. 157), and is not done here), and $H_p(s)$ is the transfer function at $s = 2\pi i f$ rad/s. Note as well that it is assumed here the poles and zeros of $H_p(s)$ are in terms of rad/s (SEED type “A”), and not in Hz (SEED type “B”), as is the convention used in the manual, and which has been our experience when retrieving RESP and StationXML files from both the IRIS and the International Federation of Digital Seismogram Networks (FDSN) Web Services. At all stages $R(f)$ is defined such that its modulus is unity at the specified frequency of the sensitivity, $f = f_s$,

$$|R(f_s)| = 1,$$

(3.10)

leading to the relationship at the analog stage,

$$A_0 = 1/H_p(s_s).$$

(3.11)

Therefore, ignoring frequency effects beyond the analog stage, and defining $S_D$ to be the multi-
plicative combination of sensitivities at all stages, the complete frequency response of the entire system at any frequency in Hz is

\[ G(f) = S_D A_0 H_p(s) \]  \hspace{1cm} (3.12)

\[ = C H_p(s), \]  \hspace{1cm} (3.13)

where \( C \) is the constant included in the SACPZ file.

3.15.2.2 The SACPZ constant

With the delivery of seismic data from the Réseau Sismique Polynésien (RSP) network by Dr. Olivier Hyvernaud, a geophysicist at Laboratoire de Géophysique Centre Polynésien de Prévention des Tsunamis (CPPT), we were also provided poles, zeros, and a sensitivity corresponding to each station. Equal for all six stations were their two zeros \((0;0)(0;0)\) and two poles \((-4.44;-4.44)(-4.44;4.44)\). The sensitivity of stations PAE and TVO was given as 0.5236 (nm/s)/LSB at 1 Hz, and for stations PMOR, VAH, TBI, and RKT as 0.212 (nm/s)/LSB at 1 Hz. Here, LSB means least significant bit, and in this case refers to digital counts. Therefore, for all six stations, the poles, zeros, and sensitivity frequency of \( f_s = 1 \) Hz are identical, but the sensitivities differ.

Before computing the constant of equation 3.13 for each station we must transform the given data slightly. First, the sensitivities were given in terms of velocity per counts, whereas the convention used in the SEED manual and the IRIS and FDSN Web Services specifies those data in terms of counts per unit of ground motion. Therefore, the sensitivities were inverted to convert them to counts/(nm/s). Next, they were converted from nm to m by multiplication (the ground-motion unit is now in the denominator of the sensitivity) with \( 10^9 \) to conform to the SEED convention that the transfer function be given in SI units (pg. 12).

Finally, we converted the pole and zero data from velocity (describing the transformation from digital counts to m/s) to displacement (counts to m). This was done to conform to the SAC standard that a TRANSFER to NONE (deconvolution in the SAC program) results in a displacement seis-
mogram (otherwise, if left as is, TRANSFER to NONE would produce a velocity seismogram—
when using SACPZ files, the SAC TRANSFER function does not automatically convert ground
motion units to displacement, if required, like it does with RESP files (SAC Manual, pg. 406)).
To that end, the sensitivities were multiplied by $2\pi f_s$ (recalling that the sensitivities are true at a
specific frequency in Hz, but were computed from the transfer function in rad/s), and a zero was
added to the set of poles and zeros (resulting from the integration of the complex transfer func-
tion). Note that SAC does not use SI units, but rather assumes (and populates the relevant header
variables accordingly) that a TRANSFER to NONE results in a displacement seismogram in units
of nm/s. However, we chose to prioritize SEED standards over SAC standards (which, again, is
true of the IRIS and FDSN Web Services in our experience), and thus we were careful to apply the
proper multiplication factor of $10^9$ in the SAC program after deconvolution such that the resultant
waveforms were in nm (or nm/s for a TRANSFER to VEL, as was done with all data in the main
text from nearby island stations), to properly match the units written to the SAC header variable
“IDEP.”

We therefore report the following displacement SACPZ files in SI units for RSP stations PAE
and TVO,

ZEROS 3
+0.000000e+00 +0.000000e+00
+0.000000e+00 +0.000000e+00
+0.000000e+00 +0.000000e+00
POLES 2
-4.440000e+00 -4.440000e+00
-4.440000e+00 +4.440000e+00
CONSTANT 2.699191e+09
and for stations PMOR, VAH, TBI, RKT,

Zeros 3
+0.000000e+00 +0.000000e+00
+0.000000e+00 +0.000000e+00
+0.000000e+00 +0.000000e+00

Poles 2
-4.440000e+00 -4.440000e+00
-4.440000e+00 +4.440000e+00

Constant 6.666493e+09

We include with this study software to compute the SACPZ constant, $C$, as well as the normalization factor, $A_0$. The functions relevant to this section are printed at the end of this supplement and are accessible at github.com/joelsimon/omnia/. Included there as well is transfunc.m, a function which may be of use to those interested in the conversion between SACPZ, RESP, and StationXML files, as well as the transformation of them between velocity and displacement responses.

3.15.2.3 Verification

Figure 3.24 proves that the displacement SACPZ files we wrote for RSP stations is correct. It compares the unfiltered (apart from those corner frequencies specified during deconvolution, see section 3.10.2) seismograms, plotted in displacement (nm), corresponding to a great earthquake in the Fiji Islands region that was recorded by three nearby stations. The traces are each aligned on the theoretical arrival time of the first-arriving $P$ wave computed in the ak135 velocity model. Two of the seismograms were recorded by stations in the RSP network (PAE and PMOR, in purple and red seismograms), each serving as the archetypal station for their respective group’s SACPZ file written by the authors, and the other by station G.PPTF (in gray), for which the displacement SACPZ file was available from IRIS. The distance each RSP station was from G.PPTF is listed
Figure 3.24: Unfiltered seismograms from RSP.PAE (purple), RSP.PMOR (red), G.PPTF (gray) of a nearby great earthquake. The SACPZ files corresponding to the two RSP stations were written by the authors, and that corresponding to G.PPTF was provided by IRIS. The similarity of the waveforms, both in phase and amplitude, proves that our SACPZ files are correct.

inside the axis (8.1 km for PAE and 338.5 km for PMOR), and they are near enough to one another that we would expect the ground motion at the three stations to be very similar, given the magnitude of earthquake. We see that the waveforms agree very well, both in amplitude and phase, both before and after the first-arrival, but especially for the first fifteen seconds after the first arrival. Therefore, we conclude that the two SACPZ files we wrote in section 3.15.2.2 corresponding to six stations in the RSP network are correct.

We include as further verification an example in the header of sacpzconstant.m, software that we wrote and which is explained in the next section, which rederives the values printed on an IRIS help page that explains how to convert a velocity RESP file to a displacement SACPZ file (ds.iris.edu/ds/support/faq/24/what-are-the-fields-in-a-resp-file/). That example shows that our SACPZ constant agrees with the one provided by IRIS to within 0.003%, well within acceptable error.

3.15.2.4 Software

Finally we reproduce below the two functions written in MATLAB used to derive the RSP SACPZ files of section 3.15.2.2. First we print sacpzconstant.m, a function that is not specific to
RSP data, and which will compute a displacement SACPZ constant, $C$, and normalization factor, $A_0$, assuming the poles, zeros, and sensitivity are provided in the format explained in section 3.15.2.2. Next we print the function, `cppsacpzconstant.m`, which is specific to the RSP SACPZ files derived in section 3.15.2.2. Within that code we explain how to convert transfer function data in non-SI velocity units to SI displacement units so that they may be processed by `sacpzconstant.m`.

Note that in both cases the codes are short and the comments are long, reflecting our experience that the interplay between these variables and files is often confusing, and the fact that these codes will live a life separate of this supplement. Also, we have done our best to compile this information from various sources, but errors may have been made, and we will correct them to the best of our ability as they arise. As such, we leave these codes printed here for the sake of posterity, and note that the most up-to-date versions may be found at [github.com/joelsimon/omnia/](https://github.com/joelsimon/omnia/).
function [CONSTANT, A0] = sacpzconstant(SD, fs, P, Z)

% [CONSTANT, A0] = SACPZCONSTANT(SD, fs, P, Z)

% SACPZCONSTANT returns the CONSTANT and A0 normalization factor for a
% displacement SACPZ file.

% Input:
% SD  Gain constant or sensitivity, true at fs [counts/m]
% fs  Frequency at which the gain constant is true [Hz]
% P   Complex poles of the transfer function [rad/s]
% Z   Complex zeros of transfer function [rad/s]

% Output:
% CONSTANT  SACPZ file CONSTANT (TRANSFER to NONE = meters)
% A0   A0 normalization factor at fs

% In the parlance of the SEED Manual v2.4, SD (uppercase "D;" my notation)
% corresponds to the combined sensitivity considering all stages, and Sd
% (lowercase "d"; their notation) is the sensitivity at a single stage, e.g.,
% the analog stage. In the example on pg. 166, the first stage is the
% seismometer [V/(m/s²)], and the second stage is the digitizer [counts/V].
% These two sensitivities are multiplied in stage 0 to compute the total
% sensitivity (SD) of the system (we are ignoring other digital stages (3+; FIR
% filters etc.), which also contribute to the stage 0 sensitivity, but
% negligibly). Ultimately, the combined sensitivity, SD, of the system is
% quoted in units like [counts/(m/s²)], though here it must be input in terms
% of counts/m.

% Therefore, using eq. (6) on pg. 159, ignoring frequency effects after the
% analog stage, and substituting the total sensitivity at SD for the
% analog-stage sensitivity, Sd, "...at any frequency f (in Hz) the response is,"
%
% \[ G(f) = SD \cdot A0 \cdot Hp(s) \]
%       = CONSTANT \cdot Hp(s), \ (author's interpretation)

% where (pg. 158), "...Hp(s) represents the transfer function ratio of
% polynomials specified by their roots," the roots being the poles and zeros
% when \( s = 2\pi i f \) rad/s \( f \) in Hz).

% While I have never seen it explicitly stated in either the SEED or SAC manuals
% that the SACPZ CONSTANT = SD \cdot A0, I have:
3.15. Supplemental Material

% (1) Seen that statement in ObsPy (1.2.0) source code --
% https://docs.obspy.org/_modules/obspy/io/sac/sacpz.html
% (2) Seen that statement on the IRIS' help pages --
% https://ds.iris.edu/ds/support/faq/24/what-are-the-fields-in-a-resp-file/
% (3) Verified that relation through pers. comm. with Olivier Hyvernaud
% (4) Concluded it must be so given the definition of the G(f)

% SACPZCONSTANT assumes the input poles and zeros correspond to a a "Transfer
% function type: A", i.e., "Laplace transform analog response, in rad/sec"
% (pg. 53). Further, the transfer function must be described by its roots
% (poles, 'P', and zeros, 'Z'), not the the coefficients of its numerators and
% denominators. This is the SEED "preferred" standard (pg. 159), and the only
% way I have ever seen these data represented (they are called "Pole-Zero
% files," not "transfer-function coefficient files" for a reason. Finally, 'SD'
% is sensitivity at 'fs' Hz, and must be given in counts/m to conform to the
% SEED standard of SI units (meters, not nanometers), and the SAC standard that
% a TRANSFER to NONE results in a displacement seismogram. Note these conflict:
% in SAC a TRANSFER to NONE assumes displacement units of nanometers. Here we
% prioritize SEED standards.

% All this is to say that if you get any random RESP or StationXML file, those
% values are more than likely already in the UNITS required as input here to
% result in a DISPLACEMENT poles and zeros file, however, their VALUES may need
% to be adjusted to move from, e.g., velocity to displacement (add one zero;
% multiply SD by 2pi*fs).

% Ex:( velocity RESP file to displacement SACPZ file, following the example of: )
% ( https://ds.iris.edu/ds/support/faq/24/what-are-the-fields-in-a-resp-file/ )
% SD_vel = 9.630000E+08; % counts/(m/s)
% fs = 0.02; % Hz
% P_vel = [-0.0123+0.0123i, -0.0123-0.0123i, -39.1800+49.1200i, -39.1800-49.1200i];
% Z_vel = [0, 0];
% % Convert SD from vel. to disp, by multiplication with 2pi*fs (SD computed in rad/s)
% SD_disp = SD_vel *2*pi*fs; % counts/m
% % Convert PZ from vel. to disp. by addition of one zero (poles unchanged)
% Z_disp = [Z_vel 0];
% P_disp = P_vel;
% [CONSTANT, A0] = SACPZCONSTANT(SD_disp, fs, P_disp, Z_disp);
% fprintf('Displacement SACPZ CONSTANT and A0 give by IRIS: 3.802483e+12, 31421.7\n')
% fprintf('Displacement SACPZ CONSTANT and A0 compute here: %.6e, %.1f\n', CONSTANT, A0)
% fprintf('CONSTANT differs by %.4e\%', abs((CONSTANT-3.802483e+12)/3.802483e+12)*100, '\%')
% For myriad verifications,
% see also: transfunc.m
%
% Author: Joel D. Simon
% Contact: jdsimon@princeton.edu | joeldsimon@gmail.com
% Last modified: 07-Apr-2020, Version 9.3.0.948333 (R2017b) Update 9 on MACI64

% Convert transfer function from pole-zero representation to
% numerator-denominator coefficient sets.
[b, a] = zp2tf(Z(:,), P(:,), 1);

% Convert sensitivity frequency from Hz to rad/s.
fs = 2*pi*fs;

% Compute the complex frequency response of the transfer function, which
% requires at least two frequencies as input. Don’t multiply ‘w’ by complex ‘i’
% (or ‘j’) because freqs.m does that internally with the frequency vector.
w = [fs-pi, fs, fs+pi];
Hp = freqs(b, a, w);

% Return the frequency response evaluated at the sensitivity frequency.
Hp_fs = Hp(2);

% The normalization factor is the modulus of the transfer function evaluated at
% the sensitivity frequency, and the SACPZ CONSTANT is that factor multiplied by
% the sensitivity.
A0 = 1/abs(Hp_fs);
CONSTANT = A0*SD;
function [CONSTANT1, CONSTANT2] = cpptsacpzconstant

% [CONSTANT1, CONSTANT2] = CPPTSACPZCONSTANT
%
% Returns the displacement SACPZ CONSTANT for six stations in the Reseau
% Sismique Polynesien (RSP) network.
%
% Input:
% --none--
%
% Output:
% CONSTANT1  Displacement SACPZ CONSTANT corresponding to PAE, TVO
% CONSTANT2  Displacement SACPZ CONSTANT corresponding to PMOR, VAH, TBI, RKT
%
% See also: sacpzconstant.m
%
% Author: Joel D. Simon
% Contact: jdsimon@princeton.edu | joeldsimon@gmail.com
% Last modified: 07-Apr-2020, Version 9.3.0.948333 (R2017b) Update 9 on MACI64

%% Relevant bits of original email.
%%______________________________________________________________________________________%%
% SUBJECT: Polynesian seismic data
% SENT: Fri 24-Jan-2020
%
% Hi Joel,
%
% Sensitivity :
%
% 0.5236 nm/s/LSB at 1 Hz for PAE, TVO
% 0.212 nm/s/LSB at 1 Hz for PMOR, VAH, TBI, RKT
%
% Response for PAE, TVO, PMOR, VAH, TBI, RKT (high pass filter at 1 Hz, order 2) :
%
% 2 zeroes : (0;0) (0;0)
%
% 2 poles : (-4.44;-4.44)(-4.44;4.44)
%
% Regards,
%
% Olivier

%%______________________________________________________________________________________%%
%% Frequency of sensitivity.

% The frequency of sensitivity is always quoted in Hz.
fs_Hz = 1;
fs_rad = 2*pi*fs_Hz;

%% Poles and zeros (the same for all stations).

% They were provided to me in terms of velocity.
Z_vel = [0+0i ... 
        0+0i];
P_vel = [-4.44-4.44i ... 
        -4.44+4.44i];

% Add one zero to convert to displacement. The poles are unchanged.
Z_disp = [Z_vel ... 
        0+0i];
P_disp = P_vel;

%% Sensitivities (Sd in SEED parlance) -- NB, Sd is the sensitivity at a single 
%% stage (e.g., analog; V/(m/s)); I use SD to represent the combined 
%% sensitivities after all stages (digital; counts/(m/s)).

% NB, the equality of (nm/s/LSB)^-1 = counts/(nm/s) was verified by Olivier 
% Hyvernaud 22-Feb-2020, pers. comm.

% We must convert SD from <physical_unit>/count --> count/<physical_unit> (the 
% convention in the SEED manual, and how SAC/PZ, RESP, and StationXML files are 
% delivered from IRIS). Next we must convert SD velocity to displacement, such 
% that a TRANSFER to NONE in SAC produces a displacement seismogram. Finally, in 
% keeping with SEED standard of SI units, we must convert SD from 
% counts/nanometer to counts/meter.

% I will refer to stations PAE, TVO as group 1 and all others as group 2. These 
% are the sensitivities as provided in nm/s/LSB.
SD1_vel = 0.5236;
SD2_vel = 0.212;

% Convert from nm/s/LSB to counts/(m/s).
SD1_vel = SD1_vel^-1 * 1e9;
SD2_vel = SD2_vel^-1 * 1e9;
% Convert from velocity sensitivities to displacement sensitivities. This
% requires multiplying by the frequency of sensitivity in rad/s because the
% sensitivity (while quoted in Hz) was computed in rad/s.
SD1_disp = SD1_vel*fs_rad;
SD2_disp = SD2_vel*fs_rad;

% Finally, compute the constants.
CONSTANT1 = sacpzconstant(SD1_disp, fs_Hz, P_disp, Z_disp);
CONSTANT2 = sacpzconstant(SD2_disp, fs_Hz, P_disp, Z_disp);
Chapter 4

A Guide to OMNIA

4.1 Software Description

OMNIA, Latin for “everything,” is an open-source repository, wholly written by me, containing every bit of software developed over the course of my doctoral research. It is freely and publicly accessible at github.com/joelsimon/omnia, and it contains everything required to reproduce the science described in this dissertation. The purpose of this chapter is to briefly review the components of OMNIA, and to highlight some of its most useful functions. Note that required dependencies, all of which are maintained separately from my own work, and most of which are contained in the conspicuously-named directory NOTMYCODE, are ignored here for obvious reasons.

By necessity, some of the hundreds of functions I have published online were written explicitly to process the noisy and complex hydroacoustic time series returned by Mobile Earthquake Recording in Marine Area by Independent Divers (MERMAID) floats. However, and whenever possible, most were written very generally so as to be applicable to a broad range of problems in time-series analysis.

The OMNIA repository is organized into various directories which roughly separate its contents by use and purpose. However, because I prioritize shorter, easier to manage functions that perform one task over bulky, single-use scripts, there is a lot of inter-directory dependency worked into my software. The following subsections are named for each specific directory in OMNIA.
4.1 Software Description

4.1.1 BSSA2020

BSSA2020 contains all the software required to remake every Figure in Simon et al. (2020) and Chapter 2. A similar package containing the code developed for Chapter 3 will be released concurrently with the publishing of that chapter.

4.1.2 CHANGETPOINTS

CHANGETPOINTS contains three pieces of software that are essentially the crux of this entire document: cpest.m estimates the changepoint of a data vector; cpci.m estimates an uncertainty and a confidence interval associated with that changepoint; and changepoint.m combines these two processes into the wavelet-based multiscale estimation procedure of Simon et al. (2020) and Chapter 2. These codes are perfectly general and are thus applicable to any problem in time-series analysis. Indeed, the process of changepoint detection—identifying the location in a time series before and after which its statistical properties are “most different”—is not specific to seismology or MERMAID. As combined with the software of section 4.1.3, changepoint.m becomes cpsac2evt.m, which automates the process of matching untagged MERMAID seismograms (which contain no corresponding earthquake metadata) and computing their associated multiscale arrival-time estimates. The latter function was used to match every seismogram described in Chapter 3, and it is described in more detail in section 3.6. To this day, it remains the first step in the workflow to process every MERMAID seismogram.

4.1.3 EARTHQUAKES

EARTHQUAKES contains codes to collect, process, and analyze earthquake data. The function sac2evt.m automatically matches Seismic Analysis Code (SAC) files, a common seismic data format, to earthquakes in global seismic catalogs. After matching to an earthquake (or multiple), tres.m will compute the associated set of multiscale travel-time residuals using the software described in section 4.1.2. An example of these residuals is plotted in Figure 2.18. Furthermore,
4.1. Software Description

**EARTHQUAKES** contains various functions to track changes to those metadata and apply those updates to seismic catalogs generated locally, like the one described in section 3.7. This is necessary because global earthquake catalogs against which our **MERMAID** catalog is referenced (and its travel time residuals are computed) are constantly being modified. A main function here is `updateid.m`, which automates the process.

### 4.1.4 EXFILES

**EXFILES** contains the files necessary to run examples I provide in the documentation that accompanies every piece of code I write. Documentation is a tedious business but it is necessary, and I make a point to provide ample in-code documentation so that individual pieces may stand alone.

### 4.1.5 MERMAID

**MERMAID** contains the primary codes to intake and process the continuous flow of incoming **MERMAID** seismograms. The script `MERMAIDfetch` handles the complete workflow of: downloading binary data from the **MERMAID** server and converting them to SAC files using software written by Sébastien Bonnieux; generating a list of preliminary earthquake matches for review, and annotating the seismograms with the relevant phases (see section 3.6); and automatically picking their multiscale phase-arrival times, and writing them and their uncertainties to the disk using software described in section 4.1.2. Once I manually review seismograms, a simple process which is heavily automated by `reviewevt.m` in the EARTHQUAKES directory, the software goes on to further analyze the data. For example: `firstarrival.m` computes the travel-time residual of the first-arriving seismic phase in a seismogram, as described in section 3.8.2; `bathtime.m` computes an adjustment (eq. 3.6) to that residual due to bathymetry and a submerged receiver, as described in section 3.8.3; and `recordsection.m` plots record sections like those shown in Figure 3.5 and Figure 3.17. Despite these codes residing in the **MERMAID** directory, they remain applicable beyond **MERMAID** (a stated goal of my software), and they were successfully used in section 3.10 to process data from other instruments.
4.1. Software Description

4.1.5.1 CPPTSTATIONS

The CPPTSTATIONS subdirectory contains software specific to the retrieval and analysis of data from the Réseau Sismique Polynésien (RSP) network. Data from the RSP network was provided to me by Olivier Hyvernaud, a geophysicist at Laboratoire de Géophysique Centre Polynésien de Prévention des Tsunamis (CPPT). In particular, these are the stations listed in Table 3.8, and some of the stations included in Figure 3.4. Data from those stations was compared with MERMAID data in section 3.10. A few functions of note are: requestcppttraces.m, which automatically generates requests for seismic data, based on the MERMAID seismic catalog (see section 3.7), to be sent to Olivier; cpptsacpzconstant.m, which computes a vital constant used in the removal of the instrument response, and which is detailed in section 3.15.2.1; and rmcpptresp.m which uses that derived constant to remove the instrument response from the raw seismic data.

4.1.5.2 GEOAZUR

The GEOAZUR subdirectory contains software specific to the analysis of MERMAID data from GéoAzur, publicly available at geoazur.unice.fr/ftp/mermaid/. That website contains the complete set of MERMAID seismograms analyzed in Chapter 2. This subdirectory contains the software required to write the catalog that was generated there and which is included as supplemental data with this dissertation.

4.1.5.3 NEARBYSTATIONS

The NEARBYSTATIONS subdirectory contains software specific to the retrieval and analysis of data from stations that are in the general vicinity of the current South Pacific Plume Imaging and Modeling MERMAID deployment (see section 3.5). In particular, these are the stations listed in Table 3.7, and some of the stations included in Figure 3.4. This subdirectory is similar to CPPTSTATIONS except that these data are available online and thus it contains functions like fetchnearbytraces.m to retrieve seismograms from these stations when MERMAID identifies a new earthquake, and fetchnearbypz.m to concurrently update any instrument-response
metadata associated with those stations. Finally, similar to CPPTSTATIONS, this subdirectory also contains \texttt{rmnearbyresp.m} to remove the instrument response from retrieved seismograms.

### 4.1.6 NORMLY

NORMLY contains software to compute log-likelihood values of various input time series assuming those inputs are composed of Gaussian (normal) distributions, either with prescribed or unknown parameters. It is an entire software suite to test how likelihood values vary when the parameters being tested start to stray from the truth. These codes are the basis for our understanding and derivation of the Akaike information criterion (AIC; reprinted as Akaike, 1998) arrival-time estimator used throughout this dissertation. A notable function here is \texttt{cpsumly.m}, which computes a changepoint (seismic-phase arrival time) by maximizing the summed log-likelihood of a time series that is modeled as two segments concatenated at the changepoint. This is the essential theory underlying our application of an AIC in equation 2.26.

### 4.1.7 OMNEALIUD

OMNEALIUD translates from Latin to “everything else,” and this directory stores just that—functions that do not fall into any specific category, but rather are generally useful for a variety of day-to-day problems. One basic but robust function that is called in nearly every code in OMNIA that deals with timing (which is basically all of them) is \texttt{timewindow.m}, which artfully isolates a segment of interest in a time series so that it may be more closely inspected. Also, this subdirectory stores codes that I wrote to facilitate the interfacing of MATLAB with \texttt{git}, especially as it pertains to the former modifying files that are tracked by the latter (see \texttt{gitrmfile.m} as an example). These codes run in the background of the entire MERMAID workflow to avoid common pitfalls associated with, for example, improper modification or deletion of files tracked by \texttt{git}.
4.2. Software Summary

4.1.8 PlotBits

PlotBits contains software specific to the generation of beautiful plots in MATLAB. My favorite is latimes.m, which converts all fonts in a figure window to Times and updates the interrupter to \LaTeX so that an extended set of symbols may be included in the output graphics. This function was used for all figures in this dissertation that were generated in MATLAB.

4.1.9 Wavelets

Wavelets contains software for the multiscale analysis of data. These codes underpin the execution of changepoint.m, described in section 4.1.2. Some notable functions include wtspy.m and iwtspy.m, which compute the time-smear associated with the forward and inverse wavelet transforms (eq. 2.53 and eq. 2.63, respectively), and wtrmedge.m, which handles the removal of edge effects during those transforms (explained in detail in section 2.13.3).

4.2 Software Summary

I have built a complete software workflow to handle the retrieval and processing of seismic data returned by MERMAID. The process is highly automated and it seamlessly interacts with git such that all data are continuously backed up and their modifications and deletions respect git best practices. My suite of software is tested and documented, and is currently being used to process data from the 16 MERMAIDS that I oversee, and an additional 24 operated by our EarthScope-Oceans collaborators at SUSTech. Most everything I write is general and thus in OMNIA I have published many functions which are useful beyond the scope of this dissertation and even my scientific field.
Chapter 5

Conclusion

5.1 My Contribution

We described a relatively new seismic instrument, MERMAID (Mobile Earthquake Recording in Marine Areas by Independent Divers), which records earthquakes from within the oceans, and proved that seismic data it returns are of tomographic quality. We developed tools to analyze their data and to pick the arrival times of various seismic phases in their hydroacoustic records. New methods of uncertainty estimation were developed and applied throughout this dissertation so that the quality of each record could be quantitatively compared to the set. Our research resulted in the construction of multiple data sets that will form the backbone of future seismic studies utilizing MERMAID data. The uncertainties reported therein will be used in future tomographic inversions to weight the data. These data sets, and the software used to compute them, are published with our research.

5.2 Looking Forward: A Data-Rich Future

In chapter 3 we discussed the ongoing South Pacific Plume Imaging and Modeling (SPPIM) project, an array of 50 MERMAIDS deployed between 2018 and 2019, 16 of which are maintained by Princeton. In total the array is a collaborative effort overseen by the international consortium, EarthScope-Oceans. In that chapter we showed that over the last 16 months, each MERMAID in
our subset returned an average of 30 tomographic-quality earthquake detections. With a projected lifespan of five years according to the manufacturer that equates to 150 high-quality seismograms delivered by each MERMAID over the course of its deployment. If the array continues to thrive the result will be a data set of 7500 earthquake detections, sampling novel ray paths through the mantle under the South Pacific.

The numbers just quoted, and by and large this dissertation, refer only to those seismograms which were matched to known earthquakes in global seismic catalogs. They ignore the hundreds of seismograms MERMAID returns that record small and local earthquakes that go unnoticed by every other seismic instrument on Earth. There is a wealth of research yet to be conducted—locating, characterizing, and adding them to our incomplete global seismic catalogs. Indeed, the original goal of MERMAID was to fill the seismic data gap in the world’s oceans by recording teleseisms from cataloged earthquakes, but I have shown that it will serve a dual role of expanding the global seismic catalogs themselves. And further, there are many other interesting scientific questions outside the field of seismology that may be investigated with its data. For example, the drift of the instrument itself may be used to generate maps of mid-column currents, and its hydroacoustic data may prove useful to investigate and track ocean storms and other weather-related phenomena.

I am confident that the analysis techniques I have developed and coded, and the data sets I have compiled, will bear fruit for many years to come. In the near term, I will continue to process and analyze the MERMAID data to bolster the already substantial data set just presented. I have also started working with Incorporated Research Institutions for Seismology (IRIS) to have these data archived in their data center. This is a proximal goal of mine, so that other scientists may begin to analyze this rich data set that has, up to now, had only a few sets of eyes on it. Looking forward I will step into a larger role at EarthScope-Oceans to facilitate the management and rapid analysis of the continuous data stream in a more centralized manner. MERMAID data is unlike others that currently reside at IRIS, and I will work with both parties to formalize the data products and standards so that they may be of the greatest utility to the most scientists. In this role I will continue to develop techniques to analyze these data, and apply my expertise to inform their interpretation.
References


References


References


References


