Statistical Modelling and Analysis of Pacific Sea Surface Temperatures

Georgina Davies

University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/theses1

Recommended Citation

https://ro.uow.edu.au/theses1/418

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
Statistical Modelling and Analysis of Pacific Sea Surface Temperatures

Georgina Davies BA BSc(Hons)

This thesis is presented as required for the conferral of the degree:

Master of Philosophy

Supervisor:
Noel Cressie

Co-supervisor:
Pavel N. Krivitsky

The University of Wollongong
School of Mathematics and Applied Statistics

November 2018
Declaration

I, Georgina Davies BA BSc(Hons), declare that this thesis is submitted in fulfilment of the requirements for the conferral of the degree Master of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Georgina Davies BA BSc(Hons)

November 26, 2018
Sea surface temperature (SST) in the Pacific Ocean is a key component of many global climate models and of the El Niño Southern Oscillation (ENSO) phenomenon. We analyse SST for the period November 1981 – December 2014. To study the temporal variability of the ENSO phenomenon, we have selected a subregion of the tropical Pacific Ocean, namely the Niño 3.4 region, as it is thought to be the area where SST anomalies indicate most clearly ENSO’s influence on the global atmosphere. SST anomalies, obtained by subtracting the appropriate monthly averages from the data, are the focus of the majority of previous analyses of the Pacific and other oceans’ SSTs. Preliminary data analysis showed that not only Niño 3.4 spatial means but also Niño 3.4 spatial variances varied with month of the year. In this thesis, we conduct an analysis of the raw SST data and introduce diagnostic plots (here, plots of variability versus central tendency). These plots show strong negative dependence between the spatial standard deviation and the spatial mean. Outliers are present, so we use robust regression to obtain intercept and slope estimates for the twelve individual months and for all-months-combined. Based on this mean-standard deviation relationship, we define a variance-stabilising transformation. The transformation we derive is logarithmic, monotonic, nonlinear, and it respects the variability seen in SSTs from month to-month during the year. On the raw SST and transformed scales, we describe the Niño 3.4 SST time series with statistical models that are linear, heteroskedastic, and dynamical. We also derive a back-transform to take our forecasts on the transformed scale back to degrees Celsius. We compare the two forecasting methods via in-sample forecasting the data the model was trained on, November 1981 – December 2014, and then out-of-sample forecasting from January 2015 – December 2017. Our results indicate that the forecasts on the transformed scale perform better when predicting up to and into boreal spring, while the forecasts on the original scale perform better when predicting across and from boreal spring into summer. We also provide visualisations of the forecast error bias and variance which can be used to better identify and understand the (boreal) spring barrier.
Acknowledgments

I would like to thank all the people who assisted me with this project and acknowledge the following important people who have supported me throughout this work.

To my primary supervisor, Distinguished Professor Noel Cressie thank you for your expertise, guidance, and persistence throughout this project.

My co-supervisor, Dr Pavel Krivitsky thank you for your insight and careful consideration when reviewing my work.

Finally to my parents and my partner, thank you for always giving me your love and support.
# Contents

1 Introduction ........................................... 1

1.1 ENSO Indices ........................................ 3

1.1.1 Southern Oscillation Index. ...................... 4

1.1.2 Oceanic Niño Index. ............................. 4

1.1.3 Coupled ENSO Index. ............................. 5

1.1.4 Multivariate ENSO Index. ....................... 6

1.1.5 Other Indices. ................................. 7

1.2 Dataset ............................................ 9

1.2.1 Spatial Region ................................ 10

1.2.2 Temporal Period ............................... 11

1.2.3 Other Oceanic and Atmospheric Variables ..... 11

1.3 Summary ........................................... 11

2 Some Exploratory Spatio-Temporal Data Analysis .... 15

2.1 Visualisations ................................... 15

2.2 Empirical Orthogonal Functions .................. 17

2.3 Principal Oscillation Patterns .................... 18

2.4 Canonical Correlation Analysis ................... 19

2.5 Neural Networks ................................ 20

2.6 Dimension Reduction ............................. 20
3 Mean-Standard Deviation Dependence

3.1 Spatial Mean ................................................................. 22
3.2 Spatial Variance and Spatial Standard Deviation ...................... 25
3.3 Diagnostic plots of Mean-Standard Deviation Dependence .......... 25

4 Robust Linear Regression

4.1 Ordinary Least Squares for Simple Linear Regression ............... 30
   4.1.1 Gauss-Markov Theorem ............................................. 31
   4.1.2 Alternative Derivation of the OLS Regression Estimates ....... 32
   4.1.3 Statistical Properties of OLS Estimates ....................... 34
   4.1.4 Example ............................................................... 37
4.2 Robust Regression Methods ............................................. 41
4.3 Theil-Sen Method ....................................................... 41
   4.3.1 Statistical Properties of Theil-Sen Estimates .................. 42
4.4 Weighted Median ........................................................ 44
4.5 Weighted-Theil-Sen Method ............................................ 45
   4.5.1 Statistical Properties of Weighted Theil-Sen (WTS) Estimates .. 46

5 Further Exploration of the Mean-Standard Deviation Dependence 50

5.1 Ordinary Least Squares (OLS) Estimation in the Diagnostic Plots .. 50
5.2 Weighted Theil-Sen (WTS) Estimation in the Diagnostic Plots ....... 52

6 Bootstrap for Regression

6.1 Standard Bootstrap Algorithm ........................................ 56
6.2 Alternative Bootstrap Algorithm ..................................... 57
6.3 Bootstrap Results ..................................................... 59
7 Variance Stabilising Transform 64

7.1 Dependence of the Spatial Standard Deviation on the Spatial Mean 64

7.2 Variance-Stabilising Transform ............................................. 65

7.2.1 All-months-combined ......................................................... 68

7.2.2 ENSO Spring Barrier .......................................................... 73

7.2.3 Individual Months .............................................................. 74

7.2.4 Chosen Transform .............................................................. 78

8 Forecast Model Development for SSTs 85

8.1 Autoregressive Processes ....................................................... 85

8.1.1 Autocorrelation and Partial Autocorrelation Functions ............. 87

8.1.2 Akaike Information Criterion ............................................... 88

8.1.3 Autoregressive Processes in SST modelling ............................ 89

8.2 Fitting Autoregressive Processes to Time Series of Tropical Pacific
SSTs ....................................................................................... 90

8.2.1 Original SST Data: Possible Outliers ....................................... 91

8.2.2 Transformed SST Data: Possible Outliers ................................. 93

8.2.3 Combined Outliers ............................................................... 95

8.2.4 Original SST Data: AR Fitting ............................................... 96

8.2.5 Transformed SST Data: AR Fitting ......................................... 100

8.3 Forecasting Equations ............................................................. 104

8.3.1 Forecasting Equations: Original SST ..................................... 104

8.3.2 Forecasting Equations: Transformed SST ............................... 105
8.3.3  Parametric Bootstrap ........................................ 107
8.4   AR(1) Simulation Study ............................................ 110
  8.4.1  Theoretical Calculation of $k_1$ and $k_2$ .............. 112
  8.4.2  Simulation Study Parametric Bootstrap for $k_1$ and $k_2$ . 118
  8.4.3  Comparison of Theoretical and Parametric Bootstrap $k_1$ and $k_2$ values ........................................ 120
  8.4.4  Forecasting using Theoretical and Parametric Bootstrap $k_1$ and $k_2$ values ........................................ 123

9  Forecasting SSTs ....................................................... 127
  9.1   In-Sample Forecasting ........................................... 127
    9.1.1  In-Sample Forecasting: Original SST Data ............ 127
    9.1.2  In-Sample Forecasting: Transformed SST Data ........ 134
    9.1.3  Comparison of In-sample Forecasts ....................... 141
  9.2   Out-of-Sample Forecasting ....................................... 160
    9.2.1  Out-of-Sample Forecasts: Original SST Data ........ 160
    9.2.2  Out-of-Sample Forecasts: Transformed Data .......... 161
    9.2.3  Comparison of Out-of-Sample Forecasts ................. 164

10 Discussion and Conclusions ........................................ 169

Bibliography .............................................................. 172
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>ACF</td>
<td>Autocorrelation Function</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autoregressive Moving Average</td>
</tr>
<tr>
<td>CCA</td>
<td>Canonical Correlation Analysis</td>
</tr>
<tr>
<td>CEI</td>
<td>Coupled ENSO Index</td>
</tr>
<tr>
<td>EBLUP</td>
<td>Empirical Best Linear Unbiased Prediction</td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory Data Analysis</td>
</tr>
<tr>
<td>ENSO</td>
<td>El Niño Southern Oscillation</td>
</tr>
<tr>
<td>EOF</td>
<td>Empirical Orthogonal Function</td>
</tr>
<tr>
<td>EOT</td>
<td>Empirical Orthogonal Teleconnections</td>
</tr>
<tr>
<td>est.</td>
<td>estimate</td>
</tr>
<tr>
<td>iid</td>
<td>independent and identically distributed</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile Range</td>
</tr>
<tr>
<td>KPCA</td>
<td>Kernel Principal Component Analysis</td>
</tr>
<tr>
<td>LHS</td>
<td>Left hand side</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>MEI</td>
<td>Multivariate ENSO Index</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimate</td>
</tr>
<tr>
<td>MSPE</td>
<td>Mean Squared Prediction Error</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
</tr>
<tr>
<td>NOAA</td>
<td>National Oceanic and Atmospheric Administration</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
</tr>
<tr>
<td>ONI</td>
<td>Oceanic Niño Index</td>
</tr>
<tr>
<td>PACF</td>
<td>Partial Autocorrelation Function</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDA</td>
<td>Pacific Decadal Oscillation</td>
</tr>
<tr>
<td>POP</td>
<td>Principal Oscillation Pattern</td>
</tr>
<tr>
<td>RHS</td>
<td>Right hand side</td>
</tr>
<tr>
<td>RS</td>
<td>Relative Skill</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>SLP</td>
<td>Sea Level Pressure</td>
</tr>
<tr>
<td>SOI</td>
<td>Southern Oscillation Index</td>
</tr>
<tr>
<td>SST</td>
<td>Sea Surface Temperature</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>THL</td>
<td>Theoretical</td>
</tr>
<tr>
<td>TS</td>
<td>Theil-Sen</td>
</tr>
<tr>
<td>WTS</td>
<td>Weighted Theil-Sen</td>
</tr>
</tbody>
</table>
List of Figures

1.1 SST data and anomalies January 1983 ................................. 2
1.2 Sea-level pressure Darwin and Tahiti January 2010–July 2015 .... 5
1.3 Niño 3.4 region ........................................................... 6
1.4 El Niño and La Niña events November 1981–December 2014 ... 6
1.5 The SOI, NOAA ONI, and MEI: November 1981–December 2014 .. 8
1.6 The global SST dataset for November 1981 ......................... 10
1.7 SST data and anomalies: September 1997 – February 1998 ....... 13
1.8 SST data and anomalies: March – August 1988 .................... 14

2.1 Hovmöller diagram of SST in the Niño 3.4 region: November 1981–December 1989 ......................................................... 16
2.2 Time series of the SST data in the Niño 3.4 region: November 1981–December 1989 ......................................................... 17

3.1 Spatial mean of the SST dataset in the Niño 3.4 region ............ 23
3.2 Spatial mean of the SST dataset in the Niño 3.4 region grouped by ENSO phase ......................................................... 24
3.3 Spatial variance of the SST dataset in the Niño 3.4 region ....... 26
3.4 Spatial variance of the SST dataset in the Niño 3.4 region grouped by ENSO phase ......................................................... 26
3.5 Spatial standard deviation versus spatial mean in the Niño 3.4 region 27

3.6 Spatial standard deviation versus spatial mean in the Niño 3.4 region by season ........................................... 28

3.7 Spatial standard deviation versus spatial mean in the Niño 3.4 region by month ................................................ 29

4.1 *Iris virginica* dataset \( \{ (x_i, y_i) : i = 1, \ldots, n \} \) with OLS fitted line superimposed; \( n = 50 \) ........................................ 37

4.2 *Iris virginica* dataset with two OLS fitted lines superimposed .......... 38

4.3 Altered *Iris virginica* dataset .................................................. 39

4.4 Altered *Iris virginica* dataset with OLS fitted line, and original OLS line .......................................................... 40

4.5 Altered *Iris virginica* dataset with the Theil-Sen and OLS fitted lines 44

4.6 Altered *Iris virginica* dataset with the weighted Theil-Sen, Theil-Sen, and OLS lines ............................................. 48

4.7 Altered *Iris virginica* dataset with the weighted Theil-Sen fitted line and OLS line fitted to the unaltered data ...................... 49

5.1 Spatial standard deviation versus spatial mean in the Niño 3.4 region for all months with OLS-fitted lines ....................... 51

5.2 Spatial standard deviation versus spatial mean in the Niño 3.4 region for each month with OLS-fitted lines .......................... 52

5.3 Spatial standard deviation versus spatial mean in the Niño 3.4 region for all months with WTS-fitted lines .......................... 54

5.4 Spatial standard deviation versus spatial mean in the Niño 3.4 region for each month with WTS-fitted lines .......................... 55

6.1 Histogram of the 1000 bootstrap values for OLS-fitted slopes from regressing \( S_z(t) \) on \( \overline{z}(t) \), in the Niño 3.4 region by month .......... 59
6.2 Histogram of the 1000 bootstrap values for OLS-fitted slopes from regressing $S_z(t)$ on $\tau(t)$, in the Niño 3.4 region for all months.

6.3 Histogram of the 1000 bootstrap values for WTS-fitted slopes from regressing $S_z(t)$ on $\tau(t)$, in the Niño 3.4 region by month.

6.4 Histogram of the 1000 bootstrap values for WTS-fitted slopes from regressing $S_z(t)$ on $\tau(t)$, in the Niño 3.4 region for all months.

7.1 Monthly slope estimates, upper and lower limits from point-wise 95% bootstrap percentile confidence intervals: (a) OLS-fitted estimates, (b) WTS-fitted estimates.

7.2 Monthly OLS- and WTS-fitted slope estimates.

7.3 Monthly intercept estimates, upper and lower limits from point-wise 95% bootstrap percentile confidence intervals: (a) OLS-fitted estimates, (b) WTS-fitted estimates.

7.4 Monthly OLS- and WTS-fitted intercept estimates.

7.5 Spatial standard deviation versus the spatial mean of the transformed data $\{u(s,t)\}$, in the Niño 3.4 region with the WTS-fitted line.

7.6 Monthly spatial standard deviation versus the spatial mean of the transformed data $\{u(s,t)\}$, in the Niño 3.4 region with the WTS-fitted line.

7.7 Monthly WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data $\{u(s,t)\}$.

7.8 Monthly WTS-estimated slope and intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data when two different transforms are used (one for March, April, May and June, and one for the other months).

7.9 Spatial standard deviation versus the spatial mean of the transformed data $\{v(s,t)\}$ in the Niño 3.4 region, for each month with the WTS-fitted lines.
7.10 Monthly WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{v(s, t)\} \) .................................................. 76

7.11 Monthly WTS-estimated intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{v(s, t)\} \) .................................................. 77

7.12 Spatial standard deviation versus the spatial mean of the transformed data \( \{v(s, t)\} \) in the Niño 3.4 region, for each month with the WTS-fitted lines .............................................................. 79

7.13 Monthly WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{v(s, t)\} \) .................................................. 80

7.14 Monthly WTS-estimated intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{v(s, t)\} \) .................................................. 81

7.15 Comparison of the raw SST data, \( z(s, t) \), to the variance-stabilised transformed data, \( v(s, t) \), for January 1983 in the Niño 3.4 region ........ 82

7.16 Comparison of the raw SST data, \( z(s, t) \), to the variance-stabilised transformed data, \( v(s, t) \), for July 1988 in the Niño 3.4 region ........ 83

7.17 Comparison of the raw SST data, \( z(s, t) \), to the variance-stabilised transformed data, \( v(s, t) \), for all \( t \in D_t \) in the Niño 3.4 region ........ 84

8.1 Centred time series, \( \{\tilde{z}_c(t)\} \) ................................................................. 91

8.2 Histogram of the time series, \( \{\tilde{z}_c(t)\} \) ................................................................. 92

8.3 Centred time series, \( \{\tilde{v}_c(t)\} \) ................................................................. 94

8.4 Histogram of the time series, \( \{\tilde{v}_c(t)\} \) ................................................................. 95

8.5 PACF plot for \( \{\tilde{z}_c(t)\} \) ................................................................. 96

8.6 AIC plot for \( \{\tilde{z}_c(t)\} \) ................................................................. 96
8.7 Residuals $\{z_c(t)_{\text{res}}\}$ ................................. 97
8.8 Histogram of the residuals, $\{z_c(t)_{\text{res}}\}$ ................................. 98
8.9 ACF of the residuals, $\{z_c(t)_{\text{res}}\}$ ................................. 98
8.10 PACF of the residuals, $\{z_c(t)_{\text{res}}\}$ ................................. 99
8.11 PACF plot for $\{\bar{v}_c(t)\}$ ................................. 100
8.12 AIC plot for $\{\bar{v}_c(t)\}$ ................................. 100
8.13 Residuals, $\{\bar{v}_c(t)_{\text{res}}\}$ ................................. 102
8.14 Histogram of the residuals, $\{\bar{v}_c(t)_{\text{res}}\}$ ................................. 102
8.15 ACF of the residuals, $\{\bar{v}_c(t)_{\text{res}}\}$ ................................. 103
8.16 PACF of the residuals, $\{\bar{v}_c(t)_{\text{res}}\}$ ................................. 103
8.17 WTS fitted lines for February and October, $\tau = 1, 3, 7$ ................................. 109
8.18 Histogram of $\tilde{z}$ minus $\hat{z}$ ................................. 115
8.19 Histogram of $\tilde{z}$ minus $\hat{z}$ where $\tau = 1, 3, 7$ ................................. 116
8.20 Quantiles of $\tilde{z}$ minus $\hat{z}$ where $\tau = 1, 3, 7$ ................................. 117
8.21 Comparison of OLS, WTS and THL lines for February and October
where $\tau = 1, 3, 7$ ................................. 119
8.22 Comparison of the OLS, WTS and THL values for $k_1$ ................................. 121
8.23 Comparison of the OLS, WTS and THL values for $k_2$ ................................. 122
8.24 Boxplots comparing of $\tilde{z}_{\text{OLS}}(t + \tau) - \bar{z}(t + \tau)$, $\tilde{z}_{\text{WTS}}(t + \tau) - \bar{z}(t + \tau)$,
and $\tilde{z}_{\text{THL}}(t + \tau) - \bar{z}(t + \tau)$ ................................. 124
8.25 Comparison of MSPE as defined in Equation 8.67 for $\tilde{z}_{\text{OLS}}(t + \tau)$,
$\tilde{z}_{\text{WTS}}(t + \tau)$, and $\tilde{z}_{\text{THL}}(t + \tau)$ ................................. 125
8.26 Comparison of relative skill for OLS/THL, WTS/THL, and OLS/WTS126
9.22 In-sample forecast-error variance relative skill . . . . . . . . . . . . . 158

9.23 Out-of-sample forecast errors \( \{ e(\hat{z}(t + \tau)) \} \) . . . . . . . . . . . . . . 162

9.24 Out-of-sample forecast errors \( \{ e(\tilde{z}(t + \tau)) \} \) . . . . . . . . . . . . . . 163

9.25 Comparison of out-of-sample bias . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 165

9.26 Comparison of out-of-sample MSPE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 167

9.27 Raster plot of out-of-sample MSPE relative skill . . . . . . . . . . . . . . . . . . . . . . . . 168
List of Tables


5.1 Slope estimates from linear regression of the spatial standard deviation on the spatial mean in the Niño 3.4 region ................. 53

5.2 Slope estimates from WTS regression of the spatial standard deviation on the spatial mean in the Niño 3.4 region ................ 55

6.1 Bootstrap measures of uncertainty for the OLS-fitted slopes from regressing $S_z(t)$ on $\bar{z}(t)$, in the Niño 3.4 region ............... 62

6.2 Bootstrap measures of uncertainty for the WTS-fitted slopes from regressing $S_z(t)$ on $\bar{z}(t)$, in the Niño 3.4 region ............... 63

7.1 Monthly WTS slope and intercept estimates, for a regression of spatial standard deviation on spatial mean in the Niño 3.4 region .......... 69

7.2 WTS model-based slope estimate and point-wise 95% bootstrap percentile confidence interval by month and for all-months-combined, for spatial standard deviation versus the spatial mean of the transformed data, $\{u(s,t)\}$, in the Niño 3.4 region............................. 70

7.3 WTS model-based slope and intercept estimates and point-wise 95% bootstrap percentile confidence interval of the slope for each month for spatial standard deviation versus the spatial mean of the transformed data, $\{v(s,t)\}$, in the Niño 3.4 region.............. 78

8.1 The values $\{\bar{v}^M\}$ and $\{I_v^M\}$ in Equation 8.19 for each month $M \in \{\text{Jan}, \text{Feb}, \ldots, \text{Dec}\}$.................. 94
8.2 The observations identified as outliers in the time series, $\{z_c(t)\}$ and $\{\pi_c(t)\}$.

8.3 The WTS fitted value of $k_1$

8.4 The WTS fitted value of $k_2$
Chapter 1

Introduction

Sea surface temperature (SST) is an important component of many global climate models. The thermal inertia of the oceanic surface layer means the air-sea interaction that occurs at the surface of the ocean is important to global temperature models and prediction. Monthly SST datasets are a combination of satellite, ship, and buoy observations. Typically, these are interpolated to produce a cohesive dataset, such as can be found at NOAA’s (National Oceanic and Atmospheric Administration) website [9]. SST anomalies, rather than raw SST data, are often analysed; SST anomalies are determined by calculating the long-term (usually 30-year) temporal average for each month and then subtracting the appropriate monthly average from each data point. For an illustration, the SST data and the calculated anomalies for January 1983 in the tropical Pacific Ocean are shown in Figure 1.1, where the climatology base period is 1971–2000.

In the tropics, SST influences the atmosphere through diabatic heating. Diabatic heating occurs where the change in temperature of air is not related to its vertical displacement, but is rather related to such effects as evaporative cooling or warming from the surface of the earth. Conversely, the temperature of the sea surface is influenced by the motion of the atmosphere through the surface winds. Tropical SST anomalies have their greatest effect in the western Pacific Ocean where the SST is normally higher than the global average SST. [83]

SST in the tropical Pacific Ocean is a key component of the El Niño Southern Oscillation (ENSO) phenomenon [16, 25]. The term “El Niño” was originally used to refer to the upwelling of warm water in the Pacific Ocean off the South American coast [83]. However, the term is now used to describe a broader range of interconnected oceanic effects. ENSO describes the distribution of warmer-than-average
Figure 1.1: Comparison of the SST data to the SST anomaly data for January 1983. The climatology base period for the anomalies is 1971–2000. Units on the colour scale are °C.

waters in the tropical Pacific Ocean, the associated atmospheric variations, and the resulting weather conditions. ENSO has two canonical states or regimes, an El Niño event (warmer eastern tropical Pacific) and a La Niña event (cooler eastern tropical Pacific). There are also periods where the ocean is in a transition phase between these two states; this is referred to as the neutral phase.

**El Niño event.** During an El Niño event, the trade winds are weaker. This allows the warm water in the western tropical Pacific Ocean to spread eastwards into the central and eastern Pacific Ocean [32, 105]. There is an increase in dry conditions (reduced precipitation) in Australia, particularly in the east and across Indonesia, New Guinea, Micronesia, Fiji, New Caledonia, and Hawaii [119, 132]. Concordantly, there is higher precipitation in the central Pacific Ocean, western and southern United States, Central America, and South America [124, 131, 132]. El Niño is also associated with anticyclonic anomalous flow in the upper troposphere on either side of the equator [19].

**La Niña event.** During a La Niña event, the trade winds are stronger, the thermocline across the Pacific Ocean gets steeper, and the western tropical Pacific Ocean has warmer-than-normal sea surface temperatures [105]. There is increased rainfall
in Australia, and La Niña events have been correlated with an increased number of tropical cyclones during cyclone season \[36, 133\]. In the south western United States, a La Niña event is typically associated with unusually dry conditions \[43\].

**Neutral phase.** In the neutral phase, there is low surface pressure in the western tropical Pacific over the warm ocean around Indonesia, high surface pressure in the eastern tropical Pacific, and the trade winds (the prevailing pattern of easterly tropical surface winds) blow across the tropical Pacific ocean \[83\].

El Niño and La Niña have serious implications for coastal communities because the frequency and intensity of tropical cyclones (also known as typhoons and hurricanes) is related to the ENSO state. Pielke and Landsea \[117\] showed that the mean and median value (in US dollars) of US Atlantic hurricane damage was significantly different between El Niño and La Niña years, and that more damage occurred in La Niña years (between 1925–1997). The negative correlation between the number of Atlantic hurricanes (and tropical storms) and El Niño events has also been shown by Gray (1900–1982) \[73\] and Tang and Neelin (1979–2003) \[150\]. El Niño has also been correlated with increases in tropical cyclone activity in the eastern part of the northwest tropical Pacific Ocean (1948–1982) \[35\], in the vicinity of Hawaii (1949–1995) \[42\], and in Australia (1950–1975) \[113\]. A relationship has also been found between where tropical cyclones form in the South Pacific and the Southern Oscillation Index (SOI), which is an indicator of ENSO \[125\].

Although ENSO events are characterised in the equatorial Pacific Ocean, the global atmosphere and oceans are highly interconnected and, thus, climatic effects in regions of the world outside the Pacific can be correlated with ENSO phases. For example, the ENSO phenomenon teleconnects with precipitation during the monsoon season in India, the rainy season in southeastern Africa \[132, 133\], and global precipitation \[49, 48, 111\]. Consequently, El Niño and La Niña events have worldwide implications, and their accurate forecasting is an extremely important problem that has still not been solved satisfactorily.

### 1.1 ENSO Indices

There have been a number of spatial analyses of the SST field, and these are summarised in Chapter 2. By averaging over spatial subdomains, several indices comparing atmospheric and oceanic measurements have been developed. The temporal behaviour of these indices can be used to study the dynamics of ENSO.
1.1.1 Southern Oscillation Index.

The Southern Oscillation Index (SOI) is calculated by taking the difference between the atmospheric sea-level pressure (SLP) at Darwin and that at Tahiti. The Southern Oscillation is an inter-annual and large-scale (global) fluctuation reflecting a shift of atmospheric mass between the Indonesian equatorial low pressure cell (represented by the Darwin measurement) and the South Pacific subtropical anticyclone (represented by the Tahiti measurement) [38]. Figure 1.2 compares the SLP at the two sites, January 2010–July 2015. Other methods of calculating an SOI have been proposed by Walker and Bliss [26], Troup [158], Chen [38], and Trenberth [153, 154]. Typically, these methods calculate monthly anomalies of smoothed data (using a window of 3–5 months) and then normalise the indices to have zero mean and unit variance [134]. Trenberth [153] proposes normalising the Tahiti and Darwin anomalies separately, so that each site is represented equally. As a consequence, the index values no longer have unit variance. Troup [158] subtracts the Darwin measurement from the Tahiti measurement before calculating anomalies and normalising the indices. The Australian Bureau of Meteorology uses the Troup SOI, and sustained negative (positive) values indicate an El Niño (La Niña) event [2].

1.1.2 Oceanic Niño Index.

Indices based on smoothed (i.e., running mean) monthly SST anomalies in the Niño 3.4 region (defined below) are each referred to as being an Oceanic Niño Index (ONI). The mean (averaged spatially) SST of other Niño regions are sometimes used as an ENSO index, however the Niño 3.4 region has emerged as the most widely used, for example, by NOAA [174]. The Niño 3.4 region was introduced in 1996, and it combines part of each of the Niño 3 region (5S–5N, and 90W–150W) and the Niño 4 region (5S–5N, and 160E–150W). Barnston et al. [21] defined the region as the latitude range, 5S to 5N, and the longitude range, 120W to 170W; see Figure 1.3. A Niño 1+2 region (0–10S, and 80–90W) has also been defined [152]. Trenberth [156] suggested using five-month running means of the Niño 3.4 anomalies, which we refer to as the Trenberth ONI. He defined El Niño (La Niña) events as follows: If the five-month running means of Niño 3.4 anomalies exceed 0.4 (−0.4) degrees Celsius (°C) for six consecutive months or more, then an El Niño (La Niña) event is said to have occurred. NOAA [3] suggest using three-month, instead of five-month, running means which we refer to as the NOAA ONI. NOAA defines El Niño (La Niña) events as follows: an El Niño (La Niña) event is said to have occurred if
Figure 1.2: Comparison of the sea-level pressure (mb) measured at Darwin [4] and Tahiti [6]: January 2010–July 2015.

the three-month running means of Niño 3.4 anomalies exceeds +0.5 (−0.5) degrees Celsius (°C) for 5 consecutive months or more. Table 1.1 lists the El Niño and La Niña events, based on the NOAA ONI, that have occurred during the period, November 1981–December 2014. These events are also shown in Figure 1.4.

1.1.3 Coupled ENSO Index.

The Coupled ENSO Index (CEI) proposed by Gergis and Fowler [68] is a composite of the ONI and SOI. They also devised a classification scheme based on these two indices to identify synchronous atmospheric and oceanic anomalies, to generate a time series of El Niño, La Niña, and neutral months. At least six consecutive El Niño months or La Niña months were required for identification of an ENSO event, with the following caveat: This six-month period could contain a maximum of two neutral months. The authors also performed sensitivity analyses for the previously published thresholds on each index used to define an El Niño event and determined that an SOI threshold of ±0.2mb gave suitable agreement with the NOAA ONI definition of El Niño and La Niña events. [68]
Figure 1.3: The Niño 3.4 region (5S–5N, and 120W–170W) is indicated by the black rectangle. The left pastel-blue region indicates the Niño 4 region (5S–5N, and 160E–150W), and right bright-blue region indicates the Niño 3 region (5S–5N, and 90W–150W).

Figure 1.4: El Niño and La Niña events that have occurred, November 1981–December 2014, based on the NOAA ONI [3] and event definitions, for the Niño 3.4 region.

### 1.1.4 Multivariate ENSO Index.

The Multivariate ENSO Index (MEI), introduced by Wolter and Timlin [173], combines six observed surface fields in the tropical Pacific: SLP; zonal and meridional components of the surface wind; SST; surface air temperature; and total cloudiness. This index is the first unrotated principal component of the multivariate observations [104, 173, 174]. If at least one bi-monthly period (e.g., December 1990/January 1991 or September/October 2000) reaches or exceeds the 20-th percentile threshold (moderate intensity) of the distribution, then Wolter defines it to be an El Niño (upper tail) or La Niña (lower tail) event [175]. Mazzarella et al. [104] used spectral analysis to identify a 60-month cycle in the MEI (1950–2008) and used this to
Table 1.1: El Niño and La Niña events that have occurred November 1981–December 2014, based on the NOAA ONI for the Niño 3.4 region.

<table>
<thead>
<tr>
<th>El Niño</th>
<th>La Niña</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 2009–April 2010</td>
<td></td>
</tr>
</tbody>
</table>

forecast an El Niño event in 2013 and a La Niña event in 2016. (Both the 2013 El Niño and 2016 La Niña forecasts proved to be incorrect.)

1.1.5 Other Indices.

There are a number of other ENSO indices that have been developed. Some examples include the Cold Tongue SST index, which summarizes equatorial SST anomalies from the dateline to the South American coast; the Bivariate ENSO Time Series, which combines SOI and Niño 3.4 indices similar to the CEI; the Trans-Niño Index, which calculates the difference between normalised SST anomalies from the Niño 1+2 and the Niño 4 regions; and nonlinear dynamical heating, which calculates the heat budget of the ocean surface layer using SST and surface-wind velocities [18, 157, 146, 176, 177]. The SOI, the NOAA ONI, and the MEI are compared visually in Figure 1.5. As in Davies and Cressie [51], in this thesis I shall use the NOAA ONI, based on the Niño 3.4 region.
Figure 1.5: The SOI [1], the NOAA ONI [3], and the MEI [5]: November 1981–December 2014.
1.2 Dataset

The dataset we shall analyse is a subset of the global monthly SST from the Climate Modeling Branch of NOAA using the Reynolds and Smith optimum-interpolation-version-2 algorithm [129, 9]. The optimum interpolation analysis is produced weekly using \textit{in situ} (buoys and ships) and bias-corrected satellite observations, combined with SST estimations based on sea-ice cover.

The buoy observations are from both moored and drifting buoys, and they are considered to be more accurate than ship observations [129]. Prior to 1998, the \textit{in situ} observations were obtained from the Comprehensive Ocean-Atmosphere Data Set; these observations are now obtained in real time from the World Meteorological Organisation’s Global Telecommunication System. Satellite measurements of SSTs have been obtained from geostationary and polar low-earth-orbiting platforms, using the advanced very high resolution radiometer and advanced microwave scanning radiometer instruments. The satellite observations measure approximately the top millimetre of the ocean, while the ship and buoy measurements measure the top few metres. [50, 54, 127, 128]

There are a number of differences between the \textit{in situ} and satellite observations, and each source has individual biases. The ship observations are skewed because they are more commonly collected from major shipping lanes, and ships avoid areas of ice. Also, the ship intake temperatures are believed to be higher than the ocean they were sampled from. Reynolds [127] found that the satellite observations are on average approximately 0.5 degrees Celsius (°C) less than \textit{in situ} observations. There are typically more daytime satellite observations than night-time, and the daytime and night-time observations use different channels on the instruments, which also potentially introduce biases. The satellite observations are affected by aerosols in the atmosphere; for example, the eruptions of El Chichón, in March–April 1982 resulted in negative biases of over 2 degrees Celsius (°C) for the satellite observations. Also, the satellite observations are affected if the cloud-detection algorithms fail, because clouds are colder than the sea surface. [127, 128, 129]

The dataset we analyse is referred to as “optimum interpolation version 2” because in November 2001 the fields were recalculated from November 1981 onwards. Optimum interpolation is the method by which the irregularly spaced observations in time and space are combined to form a cohesive dataset. The improved algorithm incorporates a new sea-ice algorithm (which primarily affects observations in higher latitudes) and reduces the biases in the satellite observations. The SST value for sea ice is set to
−1.8 degrees Celsius (°C), as that is the freezing point of seawater with a salinity of 34 psu (practical salinity unit). For frozen freshwater in the Great Lakes in North America, the value is set to 0 degrees Celsius (°C) [129].

To obtain the monthly SST fields, the weekly fields are linearly interpolated to daily fields, and then the daily fields are averaged for the month. The data are defined on a 1 × 1 degree latitude-longitude grid and are in units of degrees Celsius (°C). An example of global SST for November 1981 is given in Figure 1.6.

![Figure 1.6: The global SST dataset for November 1981.](image)

### 1.2.1 Spatial Region

To study the temporal variability of the ENSO phenomenon, we have selected a subregion of the tropical Pacific Ocean, namely the Niño 3.4 region. Recall that the Niño 3.4 region contains part of both the Niño 3 and Niño 4 regions and is defined by the latitude range, 5S–5N, and the longitude range, 120W–170W; see Figure 1.3 [156]. We chose the Niño 3.4 region for our analysis as it is widely used in tropical Pacific SST studies, and it is thought to be the area where SST anomalies indicate most clearly ENSO’s influence on the global atmosphere [33].

In what follows, the $i$-th grid cell in the Niño 3.4 region is referenced by its centroid $s_i$. These latitude-longitude coordinates are one degree apart and take half-degree values. The spatial region of interest (Niño 3.4 region) is made up of $10 \times 50 = 500$ ocean pixels, namely $D_s = \{s_1, \ldots, s_{500}\}$. 


1.2.2 Temporal Period

As in Davies and Cressie [51], we analyse the SST dataset for the period November 1981 – December 2014. Thus, the temporal period of interest is $D_t = \{1, \ldots, 398\}$, where each $t$ corresponds to a time period of one month. For some of our analyses, we subset $D_t$ into $D_{t,\text{Jan}}, D_{t,\text{Feb}}, \ldots, D_{t,\text{Dec}}$, corresponding to the months January, February, . . . , December, respectively; in this case $t = 1$ corresponds to 1981, $t = 2$ corresponds to 1982, and so forth. Examples of the evolution of an El Niño event and a La Niña event from this temporal period are given in Figures 1.7 and 1.8, respectively.

1.2.3 Other Oceanic and Atmospheric Variables

SST models can often be improved by the addition of other oceanic and atmospheric measurements. A number of authors include SLP (e.g., [22, 52, 55, 72, 79, 80, 107, 167, 175]) or components of the surface winds (e.g., [24, 55, 71, 72, 97, 107]). Other variables that have been considered in analyses with SST data include total solar irradiance [122], atmospheric aerosols [110], and cloudiness [104, 173].

1.3 Summary

Understanding the variability of SST in the tropical Pacific Ocean in different ways can help climate scientists better predict El Niño and La Niña events and potentially give more insight into recent changes in climate. This thesis adds to this understanding through a new set of diagnostic plots, where a relationship between the spatial mean and the spatial standard deviation is demonstrated on a monthly scale. The dynamical aspects of SST in the tropical Pacific Ocean are summarised in new ways by separating out the original plot into twelve plots, corresponding to the twelve months of the year.

Chapter 2 discusses exploratory spatio-temporal data analysis and includes selected visualisations of our dataset. In Chapter 3, we explore the mean-standard deviation dependence of tropical Pacific SST data. Most analyses and models found in the literature work directly with the SST anomalies. The mean-standard deviation relationship we found indicates that there is structure in the raw data that is missed by previous analyses. Our exploration of the mean-standard deviation relationship
identifies a need for robust regression methods, which are summarised in Chapter 4. The dependence between the spatial mean and the spatial standard deviation of tropical Pacific SST data is explored further in Chapter 5 using ordinary-least-squares regression and robust regression. These results make distributional assumptions to obtain $p$-values and confidence intervals which we avoid in Chapter 6 by introducing the non-parametric bootstrap.

The aforementioned diagnostic plots and the relationship between the spatial mean and the spatial standard deviation were published in Davies and Cressie (2016) in the journal, *Advances in Statistical Climatology, Meteorology and Oceanography* [51]. That article also describes a methodology to derive a variance-stabilising transformation of the Niño 3.4 SST data, which is given in detail in Chapter 7 of this thesis.

The article then went on to compare forecasting on the transformed scale to forecasting using the original data, and this thesis also presents a comparison of two such forecasts. In Chapter 8, we define SST time series on the transformed and original scales and fit forecasting models to each time series. We also derive the back-transform to take our forecasts on the transformed scale back to degrees Celsius before using the models for forecasting in Chapter 9. There are a few differences from the article with what is presented in this thesis; these differences are explained towards the end of Section 9.1.3. The thesis also analyses forecasting performance in the context of the (boreal) spring barrier and introduces raster plots that can be used to gain an improved understanding of the influence of the spring barrier on the (relative) monthly forecasts. Our major findings and conclusions are discussed in Chapter 10.
Figure 1.7: Evolution of an El Niño event during September 1997 – February 1998, showing the warmer water off the South American coast. Left panel: SST data. Right panel: SST anomaly data (base period is 1971–2000).
Figure 1.8: Evolution of a La Niña event during March – August 1988, showing the cooler water in the central Pacific Ocean. Left panel: SST data. Right panel: SST anomaly data (base period is 1971–2000).
Chapter 2

Some Exploratory Spatio-Temporal Data Analysis

In this thesis, we shall model and analyse tropical Pacific SST data. Although most researchers consider seasonality of SST as well known and move quickly to analysing anomalies, we start here with the original SST observations. Ultimately, we wish to better understand the variability of SST in the tropical Pacific Ocean and build a prototype process model that could be incorporated into a full hierarchical system. The dynamical aspects of SST are arguably the least well understood, so we eventually focus on the temporal variability of SST averaged over the Niño 3.4 region. Leading up to that, this section considers exploratory-data-analysis (EDA) approaches to understanding the full spatio-temporal variability.

Visualisations are useful for exploring SST data and its relationships with other atmospheric or oceanic measurements. Section 1.1 summarised important research on SST data to develop indices to classify the SST data into ENSO phases. Other research into SST includes statistical and dynamical models of the spatial and spatio-temporal behaviour of the data, which are summarised in this section.

2.1 Visualisations

Numerous visualisation methods for spatio-temporal data exist. Marginal or conditional image plots and Hovmöller diagrams [85], are used widely to illustrate the variability of SST data; for examples of these see [30, 50, 63, 93, 99, 120, 122].

15
Hovmöller diagrams are best suited to representing the variable of interest in a narrow band of one spatial dimension through time. Typically, for Pacific SST, they are applied to the average over a representative sub-region such as the Niño 3, 3.4, or 4 regions. Alternatively, in a band along the equator, an average is taken over latitude resulting in a one-dimensional spatial process indexed by longitude. Suppose the equatorial band is defined by the Niño 3.4 region; Figure 2.1 shows a Hovmöller diagram of the SST data between November 1981 and December 1989. The El Niño events from April 1982–June 1983 and September 1986–February 1988, and La Niña events from October 1984–June 1985 and May 1988–May 1989 can be seen, although it should be remembered that such events are typically defined by the SST anomalies. [3]
Time series plots are informative for a single spatial location, the average over a
region (e.g., Figure 3.1) or each individual spatial location in a given spatial region
(e.g., Figure 2.2) [45]. Alternatively, spatial maps visualise all spatial locations at
given time points (e.g., Figures 1.7 and 1.8) [45].

Figure 2.2: Time series of the SST data in all of the 1x1 degree cells in the Niño 3.4

2.2 Empirical Orthogonal Functions

Spatial analysis of SST data often includes principal components analysis (PCA) or,
as it is often called in spatial statistics, empirical orthogonal function (EOF) analysis.
EOF analysis typically involves an eigendecomposition of the spatial covariance (or
correlation) matrix of the observations. EOFs are the most “efficient” possible
representation of the variability of the data [52]. They are used in SST analyses to
capture the spatial structure of the data and to understand the spatial complexity
of the variability of the SST data. The more EOFs required to capture a given
percentage of the variability in the SST data, the more spatially complex is the
variability [52]. The leading EOF of monthly SST anomalies is often called the ENSO
mode or El Niño mode [54, 87]. Deser et al. [54] found a very strong correlation
(namely 0.93) between the monthly time series obtained by projecting SSTs onto
the leading global EOF and the monthly time series of the Oceanic Niño Index.

Examples of the use of EOFs in the analysis and modelling of SST data include
the following works: Berliner et al. [24] incorporated EOF analysis into their hier-
archical Bayesian model to predict Pacific SST, Enfield and Mayer [62] used EOFs
when they determined that tropical Atlantic SST variability is correlated with Pacific ENSO variability, and Chand et al. [36] used EOFs with hierarchical clustering to determine the impact that different ENSO regimes have on tropical-cyclone genesis in the southwest tropical Pacific Ocean. More recently, Wills et al. [172] used EOF analysis in their research into the variability in Pacific SST data, where they concluded that the PDO (pacific decadal oscillation) and ENSO are separate processes, operating on different timescales. It should be noted that as EOF modes are calculated empirically, they do not necessarily correspond directly to the dynamical modes of the system [54].

A number of extensions of EOF analysis have also been developed for and used in the spatial analysis of SST data. These include rotated EOFs, empirical orthogonal teleconnections (EOTs), extended EOFs, complex EOFs, and kernel principal components analysis (KPCA). Rotated EOF analysis addresses some of the limitations of classical EOF analysis, however the resulting patterns can be overly localised in space [54]. The most popular method of rotated EOFs uses Varimax rotation, introduced by Kaiser in 1958 [91]. Examples of the application of rotated EOFs to tropical Pacific SST data include: [39, 92, 100, 107]. Extended EOFs, as described by Weare and Nasstrom [168], use a lagged covariance matrix, and the authors demonstrate the application of the method to tropical Pacific SST data. Tan [147] gives a detailed description of KPCA and its application to tropical Pacific SST data. In KPCA, the observed data are transformed using a kernel function, and then PCA is calculated on the transformed data. EOTs were proposed by van den Dool et al. in 2000 for spatio-temporal analysis [161]. Teleconnections are correlations between meteorological phenomena separated by large distances. EOTs are only orthogonal in either space or time and are found iteratively. This method differs from traditional EOFs, as EOFs are bi-orthogonal (orthogonal in both space and time) and calculated simultaneously. EOTs are arguably more easily interpreted as they can be easily linked to observed physical phenomena [161]. Further, EOTs were used to correct biases in the satellite data for the Reynolds and Smith optimum-interpolation -version-2-algorithm dataset used in our analyses [130].

2.3 Principal Oscillation Patterns

Principal oscillation pattern (POP) analysis is a multivariate technique used often in climate science to identify oscillation patterns and to explore the spatio-temporal behaviour of a complex system. POPs are the normal modes (eigenvectors) of a
linear system whose system matrix is estimated from the observed data by fitting a multivariate first-order Markov process. Thus, if the system has strong nonlinear behaviour, then POP analysis may not be suitable. However, if much of the variability of a nonlinear system is in fact controlled by linear dynamics, then this method may in a particular application be successful. POP analysis can also be considered as a trivial special case of principal interaction pattern analysis. [67, 78, 97, 163]

The leading POPs correspond to the most unstable modes of the linear system and, similar to EOF analysis, the leading oscillatory mode obtained from POP analysis of SST data has been linked to ENSO [67]. Gehne et al. [67] applied POP analysis to SST data from the Pacific and Indian Oceans. Pauluhn [115] used POP analysis to investigate circulation patterns in SST of the North Atlantic Ocean. POP analysis can also be used for prediction: Penland and Magorian [116] used POP to predict SST in the Niño 3 region. Some authors use POP analysis in conjunction with EOF analysis; the data are projected onto some of the EOFs and then POP analysis is performed on the resulting transformed observations (e.g., [67, 97]). Like EOF analysis, POP analysis has also been extended, and these extensions include cyclostationary POP analysis [27] and complex POP analysis [31]. In the paper by von Storch et al. [163], complex POP analysis was carried out on SST data and compared to complex EOF analysis.

2.4 Canonical Correlation Analysis

Canonical correlation analysis (CCA), introduced by Hotelling in 1936 [84], measures the linear relationship between two multi-dimensional datasets. It can be considered a generalisation of linear and multiple regression. CCA uses eigendecomposition (or singular value decomposition) to define a coordinate system (linear combination of the variables) that optimally describes the cross-covariance between two datasets. Cross-covariance measures the covariance of one spatial dataset with another spatial dataset, typically averaged over time. CCA can be used to explore the spatial and temporal behaviour of a single dataset. To do this, a spatially or temporally lagged dataset is often considered to be the second variable in the bivariate dataset to which CCA is applied. Examples of the use of CCA in SST analysis include Barnston and Ropelewski [22], who used CCA of SLP and SST to forecast SST and to predict ENSO episodes. Graham et al. [72] predicted equatorial Pacific SST from trade winds and SLP using CCA and extended EOF analysis. Tang et al. [148] found no significant difference in the ability of linear regression and CCA, versus nonlinear
neural networks, to predict equatorial Pacific SST in each of the Niño regions. They
used SST and lagged (3, 6, and 9 months) SLP observations as predictors and
included extended EOF analysis as a pre-processing, dimension-reduction step in
their analysis [148]. A note of caution is that CCA may find highly correlated pairs
of canonical variables that are scientifically uninteresting or not easily interpreted.
Also, CCA requires a large number of observations (certainly more than the number
of variables). [20, 22, 41, 72, 148]

2.5 Neural Networks

Neural network (NN) modelling is a machine-learning technique inspired by how
the human brain processes information. The input variables are combined and then
propagated through one or more hidden layers of neurons before output data or
predictands are produced. The number of hidden layers and number of neurons in
each layer are determined through the modelling process. More neurons and layers
makes the model able to capture more complex behaviour, however this results in
more parameters that have to be estimated. [11, 149].

Neural networks can extract linear and nonlinear patterns of behaviour from data. A
number of authors have used NNs to forecast SST anomalies in the tropical Pacific.
Aguilar-Martinez and Hsieh [11] use a Bayesian NN on SST and SLP data to forecast
SST data in the tropical Pacific. Tang et al. [149, 148] used SLP and wind stress
separately as predictors in an NN to forecast tropical Pacific SST. Also using NNs,
Wu et al. [178] forecasted SST data from SLP data in the tropical Pacific. Fitting
an NN requires minimising a loss or cost function, and in any application there may
be multiple minima. Another issue with neural networks is they can be difficult to
interpret. Hsieh and Tang [86] suggest using a phase-space interpretation on the
hidden layer and then spectral analysis to understand the nonlinear relations found
from NN models. Guo et al. [74] used a self-organising maps (a NN-based clustering
method) to extract SST anomaly patterns. NNs can be used to perform nonlinear
regression, nonlinear CCA, or nonlinear PCA, and they can be implemented within
a Bayesian framework [40, 86].

2.6 Dimension Reduction

Large datasets are becoming increasingly prevalent as our ability to collect data
grows, often outpacing analyses of these datasets using traditional statistical meth-
ods. Thus, statistical methods are in demand that can, in a computationally efficient manner, model and analyse large datasets. A common approach is to “project” the data into a smaller-dimensional space, which is termed dimension reduction. This can reduce the number of parameters required to be estimated from the available data and improve estimation of these parameters.

A number of the methods discussed above can be used to perform dimension reduction. For example, in PCA/EOF analysis, where there are \( p \) orthogonal functions (principal components), projecting the data onto the \( q \) (\( q < p \)) orthogonal functions that correspond to the \( q \) largest eigenvalues gives the most efficient representation of the dataset in a \( q \)-dimensional space. We intend to use EOFs later in our analysis of SST, but first we explore the spatial mean and spatial standard deviation of SST in the Niño 3.4 region.
Chapter 3

Mean-Standard Deviation Dependence

The majority of previous analyses of Pacific SST focus on the SST anomalies, obtained by subtracting the appropriate monthly averages from the data. This removes seasonal effects, but some of our preliminary data analysis showed that not only spatial means but also spatial variances varied with month of the year. We conjectured that the monthly spatial variances might be related to the monthly spatial means, which led to this study. In this thesis, we conduct a spatio-temporal analysis of the raw SST data and introduce diagnostic plots which identify there appears be a negative trend: when the Niño 3.4 region has a higher spatial mean temperature it has lower spatial standard deviation. By going directly to anomalies, important nonlinear behaviour in the raw data may be overlooked.

3.1 Spatial Mean

Define the spatial mean of a spatio-temporal dataset \( \{z(s, t) : s \in D_s, t \in D_t\} \) for each time as,

\[
\bar{z}(t) \equiv \frac{1}{|D_s|} \sum_{s \in D_s} z(s, t) ; \quad t \in D_t ,
\]

where \( |D_s| \) denotes the number of pixels (here, 500) in \( D_s \) [51]. Figure 3.1 shows a time series of the spatial mean over the Niño 3.4 region of the SST dataset over our time period of interest, \( D_t \). The monthly variability of the mean temperature in the region can be seen clearly. Figure 3.2 shows three boxplots of the same spatial
mean values, grouped by the three ENSO phases (as defined by the NOAA ONI El Niño definition). As expected, the El Niño months have a higher average spatial mean temperature than the other two phases. The neutral months have the largest range, which is not surprising as these months are when the ocean and atmosphere are transitioning between El Niño and La Niña events.

Figure 3.1: Spatial mean, $z(t)$, of the SST dataset in the Niño 3.4 region as defined in Equation 3.1. Units on the vertical axis are °C.
Figure 3.2: Spatial mean, $\bar{\tau}(t)$, of the SST dataset in the Niño 3.4 region grouped by ENSO phase (as defined by the NOAA ONI and event definitions; see Section 1.1.2). Units on the vertical axis are °C.
3.2 Spatial Variance and Spatial Standard Deviation

We calculated the corrected (unbiased) spatial variance of the data \( \{ z(s, t) : s \in D_s, t \in D_t \} \), for each time as, \[ S_z(t)^2 = \frac{1}{|D_s| - 1} \sum_{s \in D_s} (z(s, t) - \bar{z}(t))^2 ; t \in D_t. \] (3.2)

We define the spatial standard deviation as the square root of the spatial variance and denote it as \( S_z(t) \). Figure 3.3 shows a time series of the spatial variance. Two periods, June–August 1998 (months 200–202) and May and July 1988 (months 79 and 81), have noticeably larger spatial variance than the rest; each month in these two periods has spatial variance above 2.75(°C)^2. Figure 3.4 shows the spatial variance values grouped by ENSO phase (as defined by the NOAA ONI El Niño definition). The El Niño months have the smallest average spatial variance. It can also be seen that the months with the largest spatial variances were during La Niña and neutral ENSO phases.

3.3 Diagnostic plots of Mean-Standard Deviation Dependence

We plotted the spatial standard deviation versus the spatial mean in the Niño 3.4 region for all \( t \in D_t \); see Figure 3.5. There appears be a negative trend: when the Niño 3.4 region has a higher spatial mean temperature it has lower spatial standard deviation. Diagnostic plots of this form for SST data were introduced in [51].

Recall from Section 1.2 the definitions of the monthly temporal indices, \( D_t^{Jan}, D_t^{Feb}, \ldots, D_t^{Dec} \). We repeated the plot of the spatial standard deviation versus the spatial mean in the Niño 3.4 region given in Figure 3.5, but used different colours to represent the different seasons; see Figure 3.6. While there is a range of points across the seasons, there does appear to be substantial between-season variability. For example, the small-spatial-mean and small-spatial-standard-deviation region of the figure is predominantly austral summer months (i.e., December, January, and February) and, in general, the austral spring months (i.e., September, October, and November) have high spatial standard deviations.
Figure 3.3: Spatial variance, $S_z(t)^2$, of the SST dataset in the Niño 3.4 region as defined in Equation 3.2. Units on the vertical axis are $(^\circ C)^2$.

Figure 3.4: Spatial variance, $S_z(t)^2$, of the SST dataset in the Niño 3.4 region grouped by ENSO phase (as defined by the NOAA ONI and event definitions; see Section 1.1.2). Units on the vertical axis are $(^\circ C)^2$.

To understand this variability, we plotted the spatial standard deviation versus the spatial mean in the Niño 3.4 region for all $t \in D_{t,Jan}^t, D_{t,Feb}^t, \ldots, D_{t,Dec}^t$; see the twelve
Figure 3.5: Spatial standard deviation, $S_z(t)$, versus spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_t$. Units on both axes are °C.

plots in Figure 3.7. Combining some of the months together would increase the number of observations in each group, potentially improving the power of hypothesis tests we intend to apply. However, we need to be careful not to introduce bias by combining months that are dissimilar. Further, we will lose the opportunity to see monthly seasonal behaviours and dependencies on other environmental factors.

The negative slope identified in Figure 3.5 also appears when the data is separated into the twelve individual months. We would like to fit a straight line to the mean-standard deviation relationships. If the slope of the line is significantly different from zero, this could indicate that tropical Pacific Ocean SST should be transformed to a different scale, where homoskedasticity and additivity of components of variation hold (Chapter 7.2). We have not been able to find any mention in the literature of the relationships apparent in Figures 3.5–3.7 prior to our published results in [51]. Figure 3.7 shows a consistent negative slope, albeit different in strength, across months. This could possibly be caused by coupled dynamics with other ocean and atmospheric variables (e.g., ocean currents or surface winds). It is generally accepted that the ENSO phenomenon is nonlinear [82, 24, 95], and it would appear that
Figure 3.6: Spatial standard deviation, $S_z(t)$, versus spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_t$, where each season has a different colour. The OLS-fitted lines for each season are superimposed (DJF line is red; MAM line is orange; JJA line is blue; SON line is green). Units on both axes are °C.

The relationships shown in Figure 3.7 represent another way to describe this. The dependence that we observe between the spatial mean and the spatial standard deviation has implications for further modelling and forecasting of the dynamics of Pacific SST; see Chapters 8–9.

To obtain the slope of the linear relationships by month (Figure 3.7), it might seem natural to use an ordinary-least-squares (OLS) estimator. However, outliers are clearly present in the data, including May 1988 (which was the start of a La Niña event), June 1998 (which was a neutral month after the end of an El Niño event), November–December 1982 (which were in the middle of an El Niño event), and November–December 1997 (which were in the middle of an El Niño event). The presence of outliers suggests that a robust-regression method for fitting $y = S_z(t)$ versus $x = \bar{z}(t)$ would be more appropriate. In the next chapter, a robust statistical method is presented for simple linear regression; implementing it will provide some assurance that our analysis of variability is not dominated by the outliers.
Figure 3.7: Spatial standard deviation, $S_z(t)$, versus the spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_{t}^{\text{Jan}}, D_{t}^{\text{Feb}}, \ldots, D_{t}^{\text{Dec}}$. Units on both axes are °C.
Chapter 4

Robust Linear Regression

Linear regression is a statistical methodology that models a dependent variable \( y \) as a linear combination of explanatory variables \( x \) plus an error term \( \varepsilon \). Estimating the regression coefficients is most commonly done by ordinary least squares (OLS). However, in the presence of strong outliers, OLS does not give accurate estimates of the regression parameters, and other, more robust, forms of regression are recommended. In this chapter, we shall consider simple linear regression, which involves just one explanatory variable, \( x \). This is the case needed when we apply regression to the apparent linear dependence of the spatial standard deviation on the spatial mean (for tropical Pacific SST) seen at the end of the previous chapter.

4.1 Ordinary Least Squares for Simple Linear Regression

The simple-linear-regression equation based on \( n \) data, \((x_1, y_1), \ldots, (x_n, y_n)\), is typically written as,

\[
y_i = \alpha + \beta x_i + \varepsilon_i ,
\]

where \( y_i \) represents the dependent variable, \( x_i \) is the explanatory variable, and \( \varepsilon_i \) is the error, for \( i = 1, \ldots, n \).

Six assumptions underpin the optimality of OLS estimation [47, 118, 139].

1. There is a linear relationship between the variables, \( y \) and \( x \).
2. The measurement error in $x$ is negligible.

3. The values of the $\{x_i : i = 1, \ldots, n\}$ are not all the same.

4. The errors are mutually uncorrelated:

\[
\text{cov}(\varepsilon_i, \varepsilon_j) = 0 ; i \neq j .
\] (4.2)

5. The expected value of the error is zero:

\[
\mathbb{E}(\varepsilon_i) = 0 ; i = 1, \ldots, n .
\] (4.3)

6. The errors have constant variance (homoskedasticity):

\[
\text{var}(\varepsilon_i) = \sigma^2 ; i = 1, \ldots, n .
\] (4.4)

Estimates for $\alpha$ and $\beta$, namely $\hat{\alpha}$ and $\hat{\beta}$, can be derived by minimising the sum of squares of the residuals:

\[
(\hat{\alpha}, \hat{\beta}) = \arg \min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .
\] (4.5)

The solution $(\hat{\alpha}, \hat{\beta})$ satisfies:

\[
\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} ,
\] (4.6)

where $\hat{\beta}$ is:

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} .
\] (4.7)

4.1.1 Gauss-Markov Theorem

The Gauss-Markov Theorem implies that for a statistical model with constant mean, constant variance, and uncorrelated errors (i.e., Equations 4.2, 4.3, and 4.4), the OLS estimate is the best linear unbiased estimate of the mean [140]. That is, the OLS estimate is unbiased and has the smallest variance amongst the class of all linear unbiased estimators of the mean. We use this result to motivate various estimates in the simple-linear regression setting. [17]
Let observations, \( u_1, \ldots, u_n \), be independent random variables. Consider non-negative corresponding weights, \( w_1, \ldots, w_n \), and define the weighted average as

\[
\frac{\sum_{i=1}^{n} u_i w_i}{\sum_{i=1}^{n} w_i},
\]

(4.8)

where it is assumed that \( \sum_i w_i > 0 \).

The Gauss-Markov Theorem also implies that the optimal (i.e., minimum variance) weights for the weighted average of \( \{ u_i : i = 1, \ldots, n \} \), are proportional to the inverses of the respective variances. That is,

\[
w_i \propto \frac{1}{\text{var}(u_i)} ; \quad i = 1, \ldots, n.
\]

(4.9)

### 4.1.2 Alternative Derivation of the OLS Regression Estimates

Here we give an non-standard derivation of the OLS regression estimates given in Equations 4.6 and 4.7. Consider two points \((x_i, y_i)\) and \((x_j, y_j)\), where \(x_i \neq x_j\); without loss of generality, assume that \(x_i > x_j\). The slope, \(s_{i,j}\), between these two points is

\[
s_{i,j} = \frac{y_i - y_j}{x_i - x_j}.
\]

(4.10)

It is easy to see that \(\mathbb{E}(s_{i,j}) = \beta\), for all \(i, j\).

Since the variances of \(\{ y_i : i = 1, \ldots, n \} \) are constant, equal to \(\sigma^2\), the variance of a slope \(s_{i,j}\) can be calculated as follows:

\[
\text{var}(s_{i,j}) = \frac{1}{(x_i - x_j)^2} \text{var}(y_i - y_j)
\]

\[= \frac{2\sigma^2}{(x_i - x_j)^2}
\]

\[\propto \frac{1}{(x_i - x_j)^2}.
\]

(4.11)

Thus, the Gauss-Markov Theorem suggests that a weighted average of the slopes \(\{s_{i,j}\}\) might use weights equal to the squared difference in \(x\)-values. Note that the \(\{s_{i,j}\}\) are statistically dependent in general, but we shall invoke the Gauss-Markov Theorem, nonetheless.
Now, for any weights $w_{i,j}$ such that $\sum_{i,j} w_{i,j} > 0$,

$$E \left( \frac{\sum_{i,j} w_{i,j} s_{i,j}}{\sum_{i,j} w_{i,j}} \right) = \beta,$$

which results in the unbiased estimator;

$$\beta^* = \frac{\sum_{i,j} w_{i,j} s_{i,j}}{\sum_{i,j} w_{i,j}}.$$  (4.13)

When $x_i = x_j$, $w_{i,j} = 0$, and that term in the sum in the numerator of Equation 4.13 is interpreted as zero.

The Gauss-Markov theorem suggests weights,

$$w_{i,j} \propto \frac{1}{\text{var}(s_{i,j})} \propto (x_i - x_j)^2.$$  (4.14)

Upon substituting these weights into Equation 4.13, we obtain

$$\beta^* = \frac{\sum_{i,j} \frac{y_i - y_j}{x_i - x_j} (x_i - x_j)^2}{\sum_{i,j} (x_i - x_j)^2} = \frac{\sum_{i,j} (y_i - y_j)(x_i - x_j)}{\sum_{i,j} (x_i - x_j)^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$  (4.15)

which is precisely the OLS expression for $\hat{\beta}$ given in Equation 4.7. It is straightforward to show that

$$\frac{\sum_{i<j} (x_i - x_j)(y_i - y_j)}{\sum_{i<j} (x_i - x_j)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$  (4.16)

The OLS estimate $\hat{\alpha}$ can also be written as a (trivial) weighted average involving all pairs $(x_i, y_i)$ and $(x_j, y_j)$. Define

$$\alpha^* = \frac{\sum_{i,j} \left( \frac{1}{2}(y_i + y_j) - \frac{1}{2}\hat{\beta}(x_i + x_j) \right) w_{i,j}}{\sum_{i,j} w_{i,j}},$$  (4.17)

where here the weights are assumed constant, namely $w_{i,j} = 1$, for all $i, j$. Then,
\[ \alpha^* = \frac{\sum_{i,j} \left( \frac{1}{2} (y_i + y_j) - \frac{1}{2} \hat{\beta}(x_i + x_j) \right)}{n^2} \]
\[ = \frac{n \sum_i y_i + n \sum_j y_j - \hat{\beta}(n \sum_i x_i + \sum_j x_j)}{2n^2} \]
\[ = \frac{2n^2 \bar{y} - \hat{\beta}(2n^2 \bar{x})}{2n^2} \]
\[ = \bar{y} - \hat{\beta}\bar{x}, \]  
(4.18)

which is the OLS estimate, \( \hat{\alpha} \), given in Equation 4.6. Note that setting \( w_{i,j} = 1 \) for all \( i \) and \( j \) in Equation \( 4.17 \) is appropriate because,

\[ \text{var}\left( \frac{1}{2} (y_i + y_j) - \frac{\beta}{2} (x_i + x_j) \right) = \sigma^2, \]  
(4.19)

and \( \text{var}(\hat{\beta}) = O(1/n) \).

### 4.1.3 Statistical Properties of OLS Estimates

**Expectation**

OLS estimates for linear regression are unbiased. That is, \( \mathbb{E}(\hat{\beta}) = \beta \) and \( \mathbb{E}(\hat{\alpha}) = \alpha \) (e.g., [170]).

**Variance**

Under the homoskedasticity assumption that \( \text{var}(\varepsilon_i) = \sigma^2 \) for all \( i \), and the mutual-uncorrelatedness assumption that \( \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \) for all \( i \neq j \), the variances of the estimates are (e.g., [170]):

\[ \text{var}(\hat{\alpha}) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right), \]  
(4.20)

and,

\[ \text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}. \]  
(4.21)
Variance – an alternative expression

Following Cressie and Keightley [46, 47], define $\mathcal{S}(a)$ as the sum of squared deviations of $\{a_i : i = 1, \ldots, n\}$ about its mean:

$$\mathcal{S}(a) = \sum_i (a_i - \bar{a})^2 .$$  \hfill (4.22)

Also define $\varrho_n$ as the correlation between $x_1, \ldots, x_n$ and 1, \ldots, $n$:

$$\varrho_n = \frac{\sum_i (x_i - \bar{x}) \left\{ i - \frac{1}{2} n(n + 1) \right\}}{\{\mathcal{S}(x), \mathcal{S}(1, \ldots, n)\}^{1/2}} .$$  \hfill (4.23)

If the mutually uncorrelated errors, $\{\varepsilon_i : i = 1, \ldots, n\}$, have pdf $g$, the variance of an error term is:

$$\text{var}(\varepsilon_i) = \sigma^2(g) = \int_y \left[ y - \left\{ \int_z z g(z) \, dz \right\} \right]^2 g(y) \, dy ,$$  \hfill (4.24)

which can be estimated unbiasedly by

$$(\hat{\sigma})^2 = \frac{\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n - 2} .$$  \hfill (4.25)

The denominator of Equation 4.25 is $n - 2$ because two parameters have been estimated ($\alpha$ and $\beta$), and unbiased estimation is achieved by dividing the numerator by $n$ minus the number of estimated parameters. [69]

The variances of the OLS estimates are given by [46, 47]:

$$\text{var}(\hat{\alpha}) = \frac{\sigma^2(g) \sum_i x_i^2}{n \mathcal{S}(x)}$$  \hfill (4.26)

$$\text{var}(\hat{\beta}) = \frac{\sigma^2(g)}{\mathcal{S}(x)} .$$  \hfill (4.27)

Clearly, Equations 4.21 and 4.27 are equivalent expressions:

$$\text{var}(\hat{\beta}) = \frac{\sigma^2(g)}{\mathcal{S}(x)} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} .$$  \hfill (4.28)
It is also true that Equations 4.20 and 4.26 are equal:

\[
\text{var}(\hat{\alpha}) = \sigma^2 \left( \frac{1}{n} + \frac{x^2}{\sum_i (x_i - \bar{x})^2} \right) = \sigma^2 \left( \frac{\sum_i (x_i - \bar{x})^2}{n \sum_i (x_i - \bar{x})^2} + \frac{n \bar{x}^2}{n \sum_i (x_i - \bar{x})^2} \right) = \sigma^2 \left( \frac{\sum_i (x_i^2 - 2x_i \bar{x} + \bar{x}^2) + n \bar{x}^2}{n \sum_i (x_i - \bar{x})^2} \right) = \sigma^2 \left( \frac{\sum_i (x_i^2 - 2n \bar{x}^2 + n \bar{x}^2 + n \bar{x}^2)}{n \sum_i (x_i - \bar{x})^2} \right) = \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2} = \frac{\sigma^2(g) \sum_i x_i^2}{n \mathcal{F}(x)}. \tag{4.29}
\]

**Hypothesis Testing**

In the regression setting, it is common to test for the dependence of \(y\) on \(x\) through the slope \(\beta\) being nonzero, so the null hypothesis is that of no dependence; that is,

\[H_0: \beta = 0,\]

versus the alternative hypothesis,

\[H_1: \beta \neq 0.\]

A commonly used test statistic for this hypothesis test is:

\[
T = \frac{\hat{\beta}}{\hat{\sigma}} \sqrt{\sum_i (x_i - \bar{x})^2}, \tag{4.30}
\]

where \(\hat{\sigma}\) is given in Equation 4.25. If the errors are assumed to have a Gaussian distribution (i.e., \(\varepsilon_i\) has the probability density, \(g(\varepsilon) = (2\pi \sigma^2)^{-1/2} \exp(-\varepsilon^2/2\sigma^2)\) for all \(i\)), then the test statistic, \(T\), follows a \(t\)-distribution with \(n - 2\) degrees of freedom.

For a chosen significance level of 5% (say), the null hypothesis is rejected if \(T > t_{97.5,n-2}\) or if \(T < -t_{97.5,n-2}\). The value \(t_{97.5,n-2}\) can be found by looking up \(t\)-tables, which can be found in many statistical software packages. These tables solve for \(t_{\gamma,m}\) in the equation:

\[
\Pr(T_m \leq t_{\gamma,m}) = \gamma/100, \tag{4.31}
\]

where \(T_m\) is a \(t\)-distributed random variable with \(m\) degrees of freedom, and \(0 \leq \gamma \leq 100.\)
Confidence Intervals

A 95% two-sided confidence interval for $\beta$ is (e.g. [170]):

$$
(\hat{\beta} - t_{97.5,n-2} \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}, \hat{\beta} + t_{97.5,n-2} \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}).
$$

(4.32)

Notice that the null hypothesis, $H_0$, is rejected if zero is not contained in the confidence interval.

4.1.4 Example

To illustrate different regression methods and the effect of outliers, a subset of the famous Fisher’s Iris data has been chosen [64]. We shall analyse the sepal and petal lengths of the Iris virginica data. The data with the OLS fitted line superimposed ($\hat{\alpha}_{\text{ORIG}} = 1.06$ and $\hat{\beta}_{\text{ORIG}} = 1.00$) are shown in Figure 4.1. The fit seems to be very good, but notice that one data point appears to be an outlier (bottom left of plot).

![Iris virginica dataset](image)

Figure 4.1: Iris virginica dataset \{(x_i, y_i) : i = 1, \ldots, n\} with OLS fitted line superimposed; $n = 50$. 

37
Historically, once outliers were identified, they were removed from the model-fitting procedure. Figure 4.2 shows the effect of removing the one outlier from this dataset. The green solid line ($\hat{\alpha} = 1.04$ and $\hat{\beta} = 0.95$) has been estimated via OLS without the outlier (i.e., $n = 49$) and has a slightly shallower slope compared to the original black dashed line (recall, $\hat{\alpha}_{\text{ORIG}} = 1.06$ and $\hat{\beta}_{\text{ORIG}} = 1.00$) estimated from all 50 points.

Figure 4.2: *Iris virginica* dataset with two OLS fitted lines superimposed. The green solid line is fitted to the data without the point in the bottom-left corner; $n = 49$. The black dashed line is fitted to all data and is identical to that in Figure 4.1; $n = 50$.

Outliers are often found in datasets. Measurement or experimental errors are always possible. In this example, the researcher who collected the data could have rushed towards the end of the day and measured some of the flowers inaccurately. Alternatively, some of the flowers may have been incorrectly identified as *Iris virginica* and be from a different Iris species. To simulate these scenarios, six of the 50 original points have been altered; Figure 4.3 shows how: the six circled points have been replaced with the points indicated by the plus symbols.

The altered data with the OLS fitted line superimposed ($\hat{\alpha} = 1.82$ and $\hat{\beta} = 0.86$) are shown in Figure 4.4. Altering six of the fifty observations resulted in the estimated slope changing by 20%. This demonstrates how strongly outliers can affect the OLS
Figure 4.3: Altered *Iris virginica* dataset. The six circled points have been replaced with the six points indicated by the red plus symbols.

estimates.
Figure 4.4: Altered *Iris virginica* (see Figure 4.3) dataset with OLS fitted line (magenta dotted) superimposed. The original OLS fitted line (black dashed), fitted to the unaltered data, is also superimposed.
4.2 Robust Regression Methods

Detecting and rejecting outliers can involve subjectivity or difficult simultaneous inference, and robust methods provide an automatic way to deal with them. Three main classes of robust estimators are M-, L-, and R-estimators. M-estimators, which are based on maximum likelihood estimation, were introduced by Huber [88]. Rather than minimising the squared residuals, a different function (e.g., the absolute value of the residuals) is minimised, typically through an iterative procedure [108, 114]. M-estimators have been extended to modified maximum likelihood type estimators, MM-estimators [181], and to weighted M-estimators [44]. L-estimators, which are linear combinations of order statistics, were introduced by Jaekel [89]. R-estimators are based on the ranks of residuals [89, 90, 96].

Each of these classes of robust estimators yields an estimate of the slope in a simple-linear regression analysis. Rousseeuw [135] proposed least median of squares regression, which minimises the median of the squared residuals. This has been generalised to least quantile regression. A more recently developed robust regression method is that of Verdoolaege [162], who proposed geodesic least-squares regression. This involves minimising the Rao geodesic distance on a probabilistic manifold, compared to OLS that minimises the Euclidean distance. All of these robust-regression methods minimise some function of the residuals, however, we use the alternative expressions for the OLS estimates of slope and intercept given by Equations 4.13 and 4.17, and we robustify these.

4.3 Theil-Sen Method

We noted in Section 4.1.2 that \( E(s_{i,j}) = \beta \) for all \( i, j \), since

\[
E(s_{i,j}) = E\left( \frac{y_i - y_j}{x_i - x_j} \right) = \frac{E(y_i - y_j)}{(x_i - x_j)} = \frac{(\alpha + \beta x_i) - (\alpha + \beta x_j)}{(x_i - x_j)} = \beta . \tag{4.33}
\]
Hence, a natural estimator of the slope $\beta$ is,

$$\hat{\beta}_{AV} \equiv \overline{s} = \text{average} \left\{ \left( \frac{y_i - y_j}{x_i - x_j} \right) : x_i \neq x_j, \ i < j \right\}.$$  \hspace{1cm} (4.34)

One way to robustify this estimate of the slope would be to replace the average with the median. Hence consider,

$$\hat{\beta}_{TS} \equiv \text{median} \left\{ \left( \frac{y_i - y_j}{x_i - x_j} \right) : x_i \neq x_j, \ i < j \right\},$$  \hspace{1cm} (4.35)

which is known as the Theil-Sen estimator (also known as Sen’s estimator) [141, 151]. The estimate, $\hat{\beta}_{TS}$, is an unbiased estimator of the true slope, $\beta$, of a simple linear regression of $y$ on $x$ [47, 166].

Since,

$$\mathbb{E} \left( \frac{1}{2}(y_i + y_j) - \frac{1}{2}\hat{\beta}(x_i + x_j) \right) = \frac{(\alpha + \beta x_i) + (\alpha + \beta x_j)}{2} - \frac{\beta(x_i + x_j)}{2} = \frac{\alpha}{2},$$  \hspace{1cm} (4.36)

a natural estimator of the intercept $\alpha$ is

$$\hat{\alpha}_{av} = \text{average} \left\{ \frac{1}{2}(y_i + y_j) - \frac{1}{2}\hat{\beta}(x_i + x_j) : i \leq j \right\}. $$  \hspace{1cm} (4.37)

Thus, a similar way to robustify this estimate of the intercept $\alpha$ is:

$$\hat{\alpha}_{TS} = \text{median} \left\{ \frac{1}{2}(y_i + y_j) - \frac{1}{2}\hat{\beta}_{TS}(x_i + x_j) : i \leq j \right\}. $$  \hspace{1cm} (4.38)

This is the estimate of $\alpha$ used in Theil-Sen estimation of parameters in a simple-linear regression.

4.3.1 Statistical Properties of Theil-Sen Estimates

Expectation

Sen [141] showed that the distribution of $\hat{\beta}_{TS}$ is symmetrical about $\beta$, which implies that $\mathbb{E}(\hat{\beta}_{TS}) = \beta$. Adichie [10] showed that $\mathbb{E}(\hat{\alpha}_{TS}) = \alpha$. 

42
Variance

From Cressie and Keightley [46, 47],

\[
\begin{align*}
\text{var}(\hat{\beta}_{TS}) & \sim \left\{12\rho_n^2 \mathcal{I}(x)I(g)^2\right\}^{-1}, \\
\text{var}(\hat{\alpha}_{TS}) & \sim \text{var}(\hat{\beta}_{TS}) \left\{\frac{1}{n}\rho_n^2 \mathcal{I}(x) + \pi^2\right\},
\end{align*}
\]  

(4.39)

(4.40)

where \( g \) is the density of the errors \( \{\varepsilon_i\} \); \( I(g) = \int g(y) \, dy \); and \( LHS_n \sim RHS_n \) means that

\[
\frac{LHS_n}{RHS_n} \xrightarrow{n \to \infty} 1.
\]  

(4.41)

Hypothesis Testing

From Cressie and Keightley [46], an asymptotic 95% confidence intervals can be constructed for \( \hat{\beta}_{TS} \) as:

\[
\left(\hat{\beta}_{TS} - z_{97.5} \sqrt{\text{var}(\hat{\beta}_{TS})}, \hat{\beta}_{TS} + z_{97.5} \sqrt{\text{var}(\hat{\beta}_{TS})}\right),
\]

(4.42)

where \( z_{97.5} \) is found from standard normal tables. The null hypothesis, \( H_0 : \beta = 0 \), is rejected at the 5% level of significance if zero is not contained in this confidence interval.

Example

The Theil-Sen fitted line for the altered Iris data is obtained from \( \hat{\alpha}_{TS} = 1.48 \) and \( \hat{\beta}_{TS} = 0.93 \). Figure 4.5 shows the altered Iris data plotted with the Theil-Sen (blue dashed line) and OLS (magenta dotted line) fitted lines superimposed. The OLS line fitted to the original data is also shown. The Theil-Sen line fitted to the altered data is more similar to the OLS line fitted to the original data than to the altered data. This is expected, since the Theil-Sen estimates are less affected by outliers.

The sums of differences squared is a measure of fit of an estimated model. This can be used to compare the OLS line on the altered dataset \( (\hat{\alpha}_{OLS}, \hat{\beta}_{OLS}) \) and the Theil-Sen line on the altered dataset \( (\hat{\alpha}_{TS}, \hat{\beta}_{TS}) \) to the OLS estimated line on the original dataset \( (\hat{\alpha}_{ORIG}, \hat{\beta}_{ORIG}) \):

\[
SS_{OLS} = \sum_{i=1}^n (\hat{y}_{OLS}(x_i) - \hat{y}_{ORIG}(x_i))^2 = 0.36,
\]

(4.43)
Figure 4.5: Altered Iris virginica dataset with the Theil-Sen (blue dashed) and OLS (magenta dotted) fitted lines superimposed. The OLS fitted line (black dashed) for the original data is also shown.

and

\[ SS_{TS} = \sum_{i=1}^{n} (\hat{y}_{TS}(x_i) - \hat{y}_{ORIG}(x_i))^2 = 0.18. \]  

(4.44)

Here the values of \( \{x_i : i = 1, \ldots, n\} \) are from the altered Iris dataset. The fact that \( SS_{TS} < SS_{OLS} \) indicates that the Theil-Sen estimated line is modelling the data better than the OLS estimated line, in the presence of known outliers.

4.4 Weighted Median

The weighted median as an estimate of a location parameter was first proposed by F.Y. Edgeworth in 1888 [58]. Given observations, \( u_1, \ldots, u_n \), attach weights \( w_1, \ldots, w_n \), where \( w_k > 0 \) and \( \sum w_k > 0 \). [138]

Then a distribution function \( G(q) \) can be defined as follows:

\[ G(q) = \sum_i \frac{w_i}{\sum_j w_j} 1_{[u_i \leq q]} \]  

(4.45)
where
\[ I_{[u_i \leq q]} = \begin{cases} 
1 & \text{if } u_i \leq q \\
0 & \text{otherwise}.
\end{cases} \tag{4.46} \]

Now find,
\[ m_1 = \inf \{ q : G(q) \geq 0.5 \} \]
\[ m_2 = \sup \{ q : G(q) \leq 0.5 \}; \]
then the weighted median \( m_w \) is defined as:
\[ m_w = \frac{m_1 + m_2}{2}. \tag{4.47} \]

### 4.5 Weighted-Theil-Sen Method

We noted earlier that \( \hat{\beta} = \bar{\pi} \) is a natural estimator of \( \beta \) (see Equation 4.34), since it is unbiased. Recall the alternately derived OLS estimate of \( \beta \) from Equation 4.13:
\[ \beta^* = \text{weighted average} \left\{ \left( \frac{y_i - y_j}{x_i - x_j} \right) : x_i \neq x_j, \ i < j ; w \right\}, \tag{4.48} \]
where the weights \( w_{i,j} = (x_i - x_j)^2 \).

Since \( \beta^* \) is the OLS estimator (see Equation 4.15), it is unbiased, and it will be a more efficient estimator of \( \beta \) than the unweighted average of the slopes, \( \bar{\pi} \).

In Section 4.3, we robustified the estimate \( \hat{\beta} = \bar{\pi} \) by replacing the average with a sample median. In this section, we shall robustify the estimate and make it more efficient by using weights to obtain the median.

Consider,
\[ \hat{\beta}_{WTS} = \text{weighted median}\{s_{ij} : x_i < x_j ; w\}, \tag{4.49} \]
for some \( w = \{w_{i,j}\} \). This estimator is called the weighted Theil-Sen estimator [47, 138, 144]. Sievers [144], Scholz [138], and Cressie and Kightely [47] suggest weighting by \( w_{ij} = |x_i - x_j| \). Cressie and Kightely [47] showed this to be an asymptotically optimal choice, giving more weight to the slopes between points that are further apart in the explanatory variable. Note that if \( w_{i,j} = 1 \) in Equation 4.13, then the usual Theil-Sen estimator is obtained.
Henceforth, we shall use,

\[ \hat{\beta}_{WTS} = \text{weighted median}\{s_{ij} : x_i < x_j; w\}, \quad (4.50) \]

where \( w = (|x_i - x_j| : i < j) \).

The weights, \( w = (|x_i - x_j| : i < j) \), are the square-roots of the weights used previously to derive the OLS estimate of slope. This matches the results given in [44], where the asymptotically optimal weights for weighted medians are shown to be inversely proportional to the standard deviation.

Recall the alternatively derived OLS estimate of \( \alpha; \)

\[ \hat{\alpha} = \text{weighted average} \left\{ \frac{1}{2}(y_i + y_j) - \frac{1}{2} \beta^*(x_i + x_j) : i \leq j ; w \right\}, \quad (4.51) \]

where \( w_{i,j} = 1 \); see Equation 4.17. If we use the median in place of the average, then the robust intercept estimator in the weighted Theil-Sen method is an unweighted median, namely [47]:

\[ \hat{\alpha}_{WTS} = \text{median}\left\{ \frac{1}{2}(y_i + y_j) - \frac{1}{2} \hat{\beta}_{WTS}(x_i + x_j); : i \leq j \right\}. \quad (4.52) \]

4.5.1 Statistical Properties of Weighted Theil-Sen (WTS) Estimates

Expectation

From Sievers [144], \( n^{1/2}(\hat{\beta}_{WTS} - \beta) \) has a limiting (as \( n \to \infty \)) normal distribution with mean zero. That is, asymptotically, \( \hat{\beta}_{WTS} \) is unbiased.

Since \( \hat{\alpha}_{WTS} = \hat{\alpha}_{TS} \), Adichie [10] gives the conditions under which \( E(\hat{\alpha}_{WTS}) = \alpha \).

Variance

From Cressie and Keightley [47],

\[ \text{var}(\hat{\beta}_{WTS}) \sim \left\{ 12 \mathcal{I}(x)I(g)^2 \right\}^{-1}, \quad (4.53) \]

\[ \text{var}(\hat{\alpha}_{WTS}) \sim \text{var}(\hat{\beta}_{WTS}) \left\{ \frac{1}{n} \mathcal{I}(x) + \bar{x}^2 \right\}, \quad (4.54) \]
where $g$ is the probability density function of the errors $\{\varepsilon_i\}$; 
$I(g)^2 = \int g^2(y) dy = 1/(4\pi\sigma^2)$, for $g$ a normal density with mean $\mu$ and variance $\sigma^2$ and recall that 
$LHS_n \sim RHS_n$ means 
\[ \frac{LHS_n}{RHS_n} \to 1 \quad \text{as} \quad n \to \infty. \] (4.55)

**Hypothesis Testing and Confidence Intervals**

An analogous discussion to that given in Section 4.3.1 yields $p$-values and confidence intervals based on asymptotic normality rather than on an exact $t$-distribution.

**Example**

The weighted Theil-Sen fitted line for the altered Iris data has estimates $\hat{\alpha} = 1.54$ and $\hat{\beta} = 0.91$. Figure 4.6 shows the altered iris data plotted, with the weighted Theil-Sen, Theil-Sen, and OLS fitted lines superimposed. This shows visually that the Theil-Sen and weighted Theil-Sen fitted lines for the altered data are very similar, but the measures of fit show a difference:

\[ SS_{WTS} = \sum_{i=1}^{n} (\hat{y}_{WTS}(x_i) - \hat{y}_{ORIG}(x_i))^2 = 0.15, \] (4.56)

which is less than $SS_{TS} = 0.16$ and $SS_{OLS} = 0.36$. The altered Iris data plotted with the weighted Theil-Sen fitted line superimposed and the original OLS fitted line (fitted to the unaltered data) also superimposed are shown in Figure 4.7. This shows visually that the two lines are quite similar.

The sums of squared differences show that the weighted Theil-Sen fitted line for the altered data is fitting the data best, since $SS_{WTS} < SS_{TS} < SS_{OLS}$. In what follows, we use weighted Theil-Sen method as our robust regression method exploring the dependence of the spatial standard deviation on the spatial mean of SST in the Niño 3.4 region.
Figure 4.6: Altered *Iris virginica* dataset with the weighted Theil-Sen (red solid), Theil-Sen (blue dashed), and OLS (magenta dotted) lines superimposed.
Figure 4.7: Altered *Iris virginica* dataset with the weighted Theil-Sen (red solid) fitted line superimposed. The original OLS fitted line (black dashed), fitted to the unaltered data, is also superimposed.
Chapter 5

Further Exploration of the Mean-Standard Deviation Dependence

In Chapter 3 we observed a negative linear relationship between the spatial standard deviation and the spatial mean in the Niño 3.4 region for all-months-combined \( t \in D_t \) and for all \( t \in D_t^{Jan}, D_t^{Feb}, \ldots, D_t^{Dec} \). It was also noted there that a linear relationship between \( y = S_z(t) \) and \( x = \bar{z}(t) \) implies that a variance-stabilising transformation might be applied to the raw data. In what follows, we estimate the parameters of this linear relationship using OLS and WTS.

5.1 Ordinary Least Squares (OLS) Estimation in the Diagnostic Plots

In Figure 5.1, we plotted the spatial standard deviation versus the spatial mean in the Niño 3.4 region for all \( t \in D_t^{Jan}, D_t^{Feb}, \ldots, D_t^{Dec} \), and we superimposed the individual fitted OLS lines. We also plotted the spatial standard deviation versus the spatial mean in the Niño 3.4 region for all \( t \in D_t \) and superimposed the fitted OLS line in Figure 5.2. Each of the fitted lines in the twelve individual plots and the all-months-combined plot clearly have negative slope. The OLS slope estimates and the associated 95% confidence intervals for each month and all-months-combined are listed in Table 5.1. Section 4.1.3 described the \( t \)-test for testing the null hypothesis that the regression slope is zero in the plot of \( y = S_z(t) \) versus \( x = \bar{z}(t) \). The
*p*-values from this *t*-test for each month and for all-months-combined are also listed in Table 5.1. This shows that using OLS estimates and assuming a *t*-distribution, at the 0.05 level, February is the only month without significant linear dependence between the spatial standard deviation and the spatial mean, although February’s *p*-value of 0.06 is still small. The presence of outliers suggests that perhaps a robust regression method should be used in place of OLS; see Section 5.2 for an analysis based on the robust WTS approach. It should be noted that the outliers are not contradicting the observed linear relationship, rather they appear to be extreme observations caused by unusual SST conditions that were discussed in Section 3.3.

Figure 5.1: Spatial standard deviation, $S_z(t)$, versus spatial mean, $\tau(t)$, in the Niño 3.4 region, for all $t \in D_{Jan}^{t}, D_{Feb}^{t}, \ldots, D_{Dec}^{t}$, with the OLS-fitted lines superimposed. Units on both axes are °C.
Figure 5.2: Spatial standard deviation, $S_z(t)$, versus spatial mean, $z(t)$, in the Niño 3.4 region, for all $t \in D_t$, with the OLS-fitted line superimposed. Units on both axes are °C.

5.2 Weighted Theil-Sen (WTS) Estimation in the Diagnostic Plots

The plots in Figures 5.1–5.2 show that there are some noticeable outliers. Hence, we repeat the straight-line fitting given in the previous section using robust regression. In Figure 5.3, we plot the spatial standard deviation versus the spatial mean in the Niño 3.4 region, for all $t \in D_{t_{Jan}}, D_{t_{Feb}}, \ldots, D_{t_{Dec}}$, but now WTS-fitted lines are superimposed. Figure 5.4 gives the analogous plot for all $t \in D_t$. We presented both of these plots in Davies and Cressie [51].

Recall from Section 4.5.1 that the asymptotic variance of the WTS slope estimate was given, which depends on $\sigma^2$; see Equation 4.53. To maintain robust inferences, we use the median absolute deviation (MAD) to estimate $\sigma$ [75];

$$
\tilde{\sigma} = \frac{1}{0.6745} \text{median}_i \left\{ \left| (y_i - \hat{\alpha} - \hat{\beta} x_i) - \left( \text{median}_j (y_j - \hat{\alpha} - \hat{\beta} x_j) \right) \right| \right\} .
$$

(5.1)

The square of the MAD is a robust estimate of $\sigma^2$, which should be compared
<table>
<thead>
<tr>
<th>Month</th>
<th>OLS slope estimate</th>
<th>95% confidence interval for slope</th>
<th>p-value for slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-0.06019</td>
<td>(-0.1150, -0.005364)</td>
<td>0.03</td>
</tr>
<tr>
<td>February</td>
<td>-0.04706</td>
<td>(-0.09521, 0.001084)</td>
<td>0.06</td>
</tr>
<tr>
<td>March</td>
<td>-0.06662</td>
<td>(-0.1110, -0.02221)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>April</td>
<td>-0.107</td>
<td>(-0.1871, -0.03436)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>May</td>
<td>-0.2473</td>
<td>(-0.3695, -0.1251)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>June</td>
<td>-0.2695</td>
<td>(-0.3973, -0.1416)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>July</td>
<td>-0.2153</td>
<td>(-0.3321, -0.09849)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>August</td>
<td>-0.1656</td>
<td>(-0.2396, -0.09159)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>September</td>
<td>-0.1154</td>
<td>(-0.1884, -0.04237)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>October</td>
<td>-0.1002</td>
<td>(-0.1527, -0.04775)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>November</td>
<td>-0.1285</td>
<td>(-0.1800, -0.07693)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>December</td>
<td>-0.1073</td>
<td>(-0.1586, -0.05604)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>All months</td>
<td>-0.1632</td>
<td>(-0.1892, -0.1372)</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

Table 5.1: Results from a linear regression of the spatial standard deviation, $S_z(t)$, on the spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_t^{\text{Jan}}, D_t^{\text{Feb}}, \ldots, D_t^{\text{Dec}}$, and $t \in D_t$. Shown are the OLS-estimated slope coefficients for each month; and the associated 95% confidence intervals and $p$-values from a two-sided $t$-test.

to the (bias-corrected) average of squared differences given in Equation 4.25 [136]. Table 5.2 gives WTS slope estimates for each month with the corresponding asymptotic 95% confidence intervals and $p$-values from a two-sided normal test using this variance calculation and the standard normal tables. The confidence interval for January contains zero so we are unable to reject the null hypothesis of independence between the spatial mean and the spatial standard deviation for those two months. However, the confidence interval for the other individual months and the all-months-combined do not contain zero, so we conclude that there is linear dependence between the spatial mean and the spatial standard deviation for these regressions. In the next chapter we consider approaches to obtain the $p$-values and confidence intervals without making distributional assumptions.
Figure 5.3: Spatial standard deviation, $S_z(t)$, versus spatial mean, $z(t)$, in the Niño 3.4 region, for all $t \in D_{\text{Jan}}, D_{\text{Feb}}, \ldots, D_{\text{Dec}}$, with the WTS-fitted lines superimposed. Units on both axes are $^\circ C$. 
Figure 5.4: Spatial standard deviation, $S_z(t)$, versus spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_t$, with the WTS-fitted line superimposed. Units on both axes are °C.

<table>
<thead>
<tr>
<th>Month</th>
<th>WTS slope estimate</th>
<th>95% confidence interval for slope</th>
<th>p-value for slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-0.04215</td>
<td>(-0.1077, 0.02341)</td>
<td>0.21</td>
</tr>
<tr>
<td>February</td>
<td>-0.04385</td>
<td>(-0.08695, -0.0007475)</td>
<td>0.046</td>
</tr>
<tr>
<td>March</td>
<td>-0.05894</td>
<td>(-0.09394, -0.02395)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>April</td>
<td>-0.1086</td>
<td>(-0.1789, -0.03825)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>May</td>
<td>-0.1726</td>
<td>(-0.2437, -0.1014)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>June</td>
<td>-0.2343</td>
<td>(-0.3395, -0.1292)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>July</td>
<td>-0.2201</td>
<td>(-0.3355, -0.1048)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>August</td>
<td>-0.1653</td>
<td>(-0.2390, -0.09162)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>September</td>
<td>-0.1094</td>
<td>(-0.1506, -0.06820)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>October</td>
<td>-0.08275</td>
<td>(-0.1223, -0.04321)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>November</td>
<td>-0.08901</td>
<td>(-0.1254, -0.05261)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>December</td>
<td>-0.07786</td>
<td>(-0.1262, -0.02953)</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>All months</td>
<td>-0.1692</td>
<td>(-0.1983, -0.1401)</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

Table 5.2: Results from a linear regression of the spatial standard deviation, $S_z(t)$, on the spatial mean, $\bar{z}(t)$, in the Niño 3.4 region, for all $t \in D_{\text{Jan}}$, $D_{\text{Feb}}$, ..., $D_{\text{Dec}}$, and $t \in D_t$. Shown are the WTS-estimated slope coefficients for each month, and the associated asymptotic 95% confidence intervals and p-values from a two-sided normal test.
Chapter 6

Bootstrap for Regression

The results given in Sections 5.1 and 5.2 make distributional assumptions to obtain the \( p \)-values and confidence intervals. A resampling approach that is distribution free, called the bootstrap, was first developed by Efron [59]. Here we describe two bootstrap algorithms for regression.

6.1 Standard Bootstrap Algorithm

In bootstrapping, the observed data can be treated as a population from which samples are randomly drawn with replacement. The parameter or statistic of interest is then calculated from these resampled data, and the process is repeated a large number of times. The general algorithm is as follows [59, 60]:

**Algorithm 6.1.1. General Bootstrapping Algorithm**

The standard deviation of an estimated \( \hat{\theta} \), can be estimated as follows.

1. Observed data, \( x_1, \ldots, x_n \), is given. Sample, with replacement, \( n \) indices \( I = (i_1, \ldots, i_n) \) from \( \{1, \ldots, n\} \), where each index has equal probability of being sampled, namely \( 1/n \).

2. Use the bootstrap sample, \( \{x_i^* : i \in I\} \), to calculate \( \hat{\theta}^* \).

3. Repeat Steps 1 and 2, \( B \) times, resulting in bootstrap replicates, \( \hat{\theta}_1^*, \ldots, \hat{\theta}_B^* \).
4. The standard deviation of the bootstrap replicates can then be calculated and used as an estimate of the true standard deviation of $\hat{\theta}$,

$$S_{\hat{\theta}} = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}_b^* - \overline{\hat{\theta}})^2}{B - 1}},$$

(6.1)

where $\overline{\hat{\theta}}$ is the mean of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

Alternative measures of the uncertainty of the estimate can be calculated from the bootstrap replicates, including the bias, the mean-squared prediction error and a confidence interval.

The confidence interval referred to above could be calculated several ways. First, it could be calculated using tables from the standard normal distribution or the $t$-distribution as in Equation 4.32. Second, the empirical quantiles of the estimated parameter could be used, which results in a bootstrap percentile confidence interval. To calculate a bootstrap percentile confidence interval, the $B$ bootstrap replicates are sorted: $\hat{\theta}_1^* \leq \hat{\theta}_2^* \leq \cdots \leq \hat{\theta}_B^*$. The lower bound of a two-sided $(1 - \alpha) \times 100$ percentile confidence interval is $\hat{\theta}_{[Ba/2]}^*$ and the upper bound is $\hat{\theta}_{[B(1-\alpha/2)]}^*$. Here the square brackets, $[c]$, represent rounding $c$ to the nearest integer [65, 66]. Third, a bias-corrected, accelerated percentile interval can be calculated, which is an extension of the bootstrap percentile interval [61].

### 6.2 Alternative Bootstrap Algorithm

Bootstrapping can be used for uncertainty quantification in regression. We use bootstrapping as an alternative way to obtain estimates of the variance of the slope estimator, whether the estimates were obtained by OLS or WTS. We draw randomly with replacement from the ($x,y$) pairs where $x$ is the spatial mean and $y$ is the spatial standard deviation, and then we estimate the slope parameter from these random samples. An alternative bootstrap procedure is to re-sample the residuals from the fitted model. This treats the $x$-values as fixed and assumes that the errors are homoskedastic. In our case, the spatial mean $\overline{z}(t)$ has much less variability than the spatial standard deviation $S_z(t)$, and the number of pixels contributing to the spatial moments are the same for each calculation. Hence, the bootstrap assumptions for regression are not unreasonable in our case. The bootstrap algorithm we use is as follows [51, 66]:
Algorithm 6.2.1. Bootstrapping regression residuals

1. Given observed data, \((x_1, y_1), \ldots, (x_n, y_n)\), and fitted line, \(y = \hat{\alpha} + \hat{\beta}x\), calculate \(e_1, \ldots, e_n\), as follows:

\[
e_i = y_i - \hat{\alpha} - \hat{\beta}x_i; \quad i = 1, \ldots, n.
\]  

2. Sample, with replacement, \(n\) indices \(I = (i_1, \ldots, i_n)\) from \(\{1, \ldots, n\}\), where each index has equal probability of being sampled, namely \(1/n\).

3. Use the bootstrap sample, \(\{(x_i, e_i^*) : i \in \{1, \ldots, n\}, i^* \in I\}\) to calculate \(y_i^*\):

\[
y_i^* = \hat{\alpha} + \hat{\beta}x_i + e_i^*.
\]  

Fit a linear regression to the \(n\) pairs \(\{(x_i, y_i^*) : i = 1, \ldots, n\}\), to obtain \(\hat{\alpha}^*\) and \(\hat{\beta}^*\).

4. Repeat Steps 2 and 3, \(B\) times, resulting in the bootstrap replicates of the slope coefficient, \(\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*\).

5. Calculate the desired measures of uncertainty of \(\hat{\beta}\) from \(\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*\).

We have calculated the mean and standard deviation of the bootstrap replicates of the regression slope, \(\beta\), along with 95% bootstrap percentile confidence intervals, using \(B = 1000\) bootstrap replicates for OLS and WTS regression.

The mean and standard deviation calculations are defined as follows:

\[
\text{mean } (\hat{\beta}^*) = \frac{\sum_{b=1}^{1000} \hat{\beta}_b^*}{1000}, \quad \text{(6.4)}
\]

\[
\text{sd } (\hat{\beta}^*) = \sqrt{\frac{\sum_{b=1}^{1000} \left( \hat{\beta}_b^* - \text{mean } (\hat{\beta}^*) \right)^2}{999}}. \quad \text{(6.5)}
\]

The bootstrap percentile confidence intervals are based on quantiles of the bootstrap replicates. Sort \(\hat{\beta}_1^*, \ldots, \hat{\beta}_{1000}^*\) so that \(\hat{\beta}_{(1)}^* \leq \hat{\beta}_{(2)}^* \leq \ldots \leq \hat{\beta}_{(1000)}^*\). A bootstrap 95% percentile confidence interval uses as the lower and upper limits the 0.025 and 0.975 bootstrap quantiles.
6.3 Bootstrap Results

We plotted the 1000 bootstrap replicates of the regression slope, $\{\hat{\beta}_b^* : b = 1, \ldots, 1000\}$, for each month and all-months-combined for OLS-fitted (Figures 6.1 and 6.2), and WTS-fitted (Figures 6.3 and 6.4) regressions. The resampling distributions of the $\{\hat{\beta}_b^*\}$ are unimodal, and they have somewhat comparable variability between the 12 months. This evident from the standard deviation results given in Tables 6.1 and 6.2.

![Histograms of bootstrap replicates for each month](image1)

Figure 6.1: Histogram of the 1000 bootstrap values $\{\hat{\beta}_b^* : b = 1, \ldots, 1000\}$, for OLS-fitted slopes from regressing $S_z(t)$ on $z(t)$, in the Niño 3.4 region, for all $t \in D_{\text{Jan}}^1, D_{\text{Feb}}^1, \ldots, D_{\text{Dec}}^1$.

The model-based and bootstrap standard deviation values for OLS are quite similar; the model-based value is slightly larger in all cases. The model-based WTS standard deviation values are also all quite similar to the bootstrap standard deviation values. However, for some of the months the WTS bootstrap standard deviation is larger than the model-based standard deviation.
Figure 6.2: Histogram of the 1000 bootstrap values \( \{ \hat{\beta}^*_b : b = 1, \ldots, 1000 \} \), for OLS-fitted slopes from regressing \( S_z(t) \) on \( \tau(t) \), in the Niño 3.4 region, for all \( t \in D_t \).

According to the 95% bootstrap percentile confidence intervals, for OLS estimation, all of the months individually and combined have non-zero slopes at the (pointwise) 5% significance level (i.e., mean-standard deviation dependence). For WTS estimation, January and February have 95% bootstrap percentile confidence intervals that contain zero, but for March–December months individually and for all-months-combined the 95% bootstrap percentile confidence intervals do not contain zero, and hence we conclude dependence between the spatial standard deviation and the spatial mean at the 0.05 level of significance. In the next chapter we will use WTS regression with bootstrapping to define a variance-stabilising transformation for tropical SST data in the Niño 3.4 region.
Figure 6.3: Histogram of the 1000 bootstrap values \( \{ \hat{\beta}^*_b : b = 1, \ldots, 1000 \} \), for WTS-fitted slopes from regressing \( S_z(t) \) on \( z(t) \), in the Niño 3.4 region, for all \( t \in D_{t}^{\text{Jan}}, D_{t}^{\text{Feb}}, \ldots, D_{t}^{\text{Dec}} \).
Figure 6.4: Histogram of the 1000 bootstrap values \( \{ \hat{\beta}_b^* : b = 1, \ldots, 1000 \} \), for WTS-fitted slopes from regressing \( S_z(t) \) on \( \pi(t) \), in the Niño 3.4 region, for all \( t \in D_t \).

<table>
<thead>
<tr>
<th>Month</th>
<th>Model-based OLS</th>
<th>Bootstrap OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std Dev.</td>
<td>Std Dev.</td>
</tr>
<tr>
<td>January</td>
<td>0.02688</td>
<td>0.02505</td>
</tr>
<tr>
<td>February</td>
<td>0.02361</td>
<td>0.02294</td>
</tr>
<tr>
<td>March</td>
<td>0.02177</td>
<td>0.02141</td>
</tr>
<tr>
<td>April</td>
<td>0.03745</td>
<td>0.03540</td>
</tr>
<tr>
<td>May</td>
<td>0.05991</td>
<td>0.05824</td>
</tr>
<tr>
<td>June</td>
<td>0.06269</td>
<td>0.06102</td>
</tr>
<tr>
<td>July</td>
<td>0.05726</td>
<td>0.05432</td>
</tr>
<tr>
<td>August</td>
<td>0.03628</td>
<td>0.03505</td>
</tr>
<tr>
<td>September</td>
<td>0.03580</td>
<td>0.03518</td>
</tr>
<tr>
<td>October</td>
<td>0.02573</td>
<td>0.02476</td>
</tr>
<tr>
<td>November</td>
<td>0.02530</td>
<td>0.02468</td>
</tr>
<tr>
<td>December</td>
<td>0.02518</td>
<td>0.02422</td>
</tr>
<tr>
<td>All months</td>
<td>0.01321</td>
<td>0.01290</td>
</tr>
</tbody>
</table>

Table 6.1: Bootstrap (using 1000 replicates) measures of uncertainty for the OLS-fitted slopes from regressing \( S_z(t) \) on \( \pi(t) \), in the Niño 3.4 region, for all \( t \in D_t \).
<table>
<thead>
<tr>
<th>Month</th>
<th>Model-based WTS</th>
<th>Bootstrap WTS</th>
<th>95% Percentile Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>0.03345</td>
<td>0.02761</td>
<td>(-0.09865, 0.01589)</td>
</tr>
<tr>
<td>February</td>
<td>0.02199</td>
<td>0.02266</td>
<td>(-0.09315, 0.0003808)</td>
</tr>
<tr>
<td>March</td>
<td>0.01785</td>
<td>0.01842</td>
<td>(-0.09332, -0.01649)</td>
</tr>
<tr>
<td>April</td>
<td>0.03589</td>
<td>0.03210</td>
<td>(-0.1767, -0.04454)</td>
</tr>
<tr>
<td>May</td>
<td>0.03631</td>
<td>0.04564</td>
<td>(-0.2644, -0.08023)</td>
</tr>
<tr>
<td>June</td>
<td>0.05366</td>
<td>0.05114</td>
<td>(-0.3392, -0.1349)</td>
</tr>
<tr>
<td>July</td>
<td>0.05884</td>
<td>0.05566</td>
<td>(-0.3380, -0.1156)</td>
</tr>
<tr>
<td>August</td>
<td>0.03759</td>
<td>0.03321</td>
<td>(-0.2328, -0.1020)</td>
</tr>
<tr>
<td>September</td>
<td>0.02102</td>
<td>0.03123</td>
<td>(-0.1754, -0.04365)</td>
</tr>
<tr>
<td>October</td>
<td>0.02017</td>
<td>0.01982</td>
<td>(-0.1248, -0.04110)</td>
</tr>
<tr>
<td>November</td>
<td>0.01857</td>
<td>0.01799</td>
<td>(-0.1268, -0.05184)</td>
</tr>
<tr>
<td>December</td>
<td>0.02465</td>
<td>0.02102</td>
<td>(-0.1246, -0.03677)</td>
</tr>
<tr>
<td>All months</td>
<td>0.01484</td>
<td>0.01339</td>
<td>(-0.1965, -0.1429)</td>
</tr>
</tbody>
</table>

Table 6.2: Bootstrap (using 1000 replicates) measures of uncertainty for the WTS-fitted slopes from regressing $S_z(t)$ on $z(t)$, in the Niño 3.4 region, for all $t \in \mathbb{D}_t^{Jan}, \mathbb{D}_t^{Feb}, \ldots, \mathbb{D}_t^{Dec}$, and $t \in \mathbb{D}_t$. 
Chapter 7

Variance Stabilising Transform

In Chapter 5 we used OLS and WTS to estimate the parameters of the linear relationship between the spatial standard deviation and the spatial mean in the Niño 3.4 region for all-months-combined $t \in D_t$ and for all $t \in D_{Jan}^t, D_{Feb}^t, \ldots, D_{Dec}^t$. In Chapter 6 we concluded that when outliers are present, WTS is a better regression estimation method than OLS. In what follows, we demonstrate the periodicity in this linear relationship through an exploratory plot and use the WTS estimated parameters to define a variance-stabilising transformation for the tropical SST data.

7.1 Dependence of the Spatial Standard Deviation on the Spatial Mean

The slope coefficient in a regression analysis of $S_z(t)$ versus $\overline{z}(t)$ captures a mean-standard deviation dependence for tropical Pacific SST. Figure 7.1 summarises the results of Sections 5.1–6 through plots of slope estimates as a function of month, from December, January, February, \ldots, November, December, January, where the months of December and January are repeated to emphasise the periodicity of the results: the dotted lines show the upper and lower limits of the point-wise 95% bootstrap percentile confidence intervals. The horizontal blue solid line shows the slope estimate for all-months-combined.

The estimated slopes for each month from OLS and WTS methods are all negative and the austral autumn and winter months (May, June, July, and August) have the steepest slopes. For some of the months (July, August, and September) the
Figure 7.1: Time sequences from December, January, February, . . . , November, December, January, showing slope estimates, upper and lower limits from point-wise 95% bootstrap percentile confidence intervals. Also shown as the horizontal blue solid line is the slope estimate for all-months-combined: (a) OLS-fitted estimates, (b) WTS-fitted estimates.

estimated slopes from OLS and WTS are very similar; see Figure 7.2. On closer inspection of Figure 5.1 or 5.3, it can be seen that these months do not have strong outliers. In the months with outliers, OLS and WTS slope estimates differ more.

These plot can be repeated for the intercept estimates. Figure 7.3 summarises the results of Sections 5.1–6. The estimated intercepts for each month from OLS and WTS methods are all positive and the austral autumn and winter months (May, June, July, and August) have the highest intercept estimates. For some of the months (July, August, and September) the estimated intercepts from OLS and WTS are very similar; see Figure 7.4.

In the months with strong outliers the OLS and WTS estimates for regression slope and intercept differ more than other months. The OLS slopes estimates are also steeper for most months. We conclude that the WTS regression estimates are better as they are less influence by the outliers. We will use the WTS regression estimates to define a variance stabilising transform for the tropical Pacific SST values.

### 7.2 Variance-Stabilising Transform

Consider a random variable $X$ such that $\mathbb{E}(X^2) < \infty$; then a variance-stabilising transformation, $f$, satisfies

$$\text{var}(f(X)) \simeq c,$$

(7.1)
for a constant \( c \) that does not depend on \( \mathbb{E}(f(X)) \). As will be apparent below, the approximation in Equation 7.1 relies on \( \sigma_x^2 \equiv \text{var}(X) \) being small relative to
Figure 7.4: Time sequences from December, January, February, ..., November, December, January, showing OLS- and WTS-fitted intercept estimates.

\[ \mu_x \equiv \mathbb{E}(X). \]

The delta method can be used to identify a variance-stabilising transformation as follows. Assuming that \( f \) is twice differentiable, we can write a first-order Taylor series expansion of \( f(x) \) about a real value \( \nu \) [77],

\[
f(x) = f(\nu) + f'(\nu)(x - \nu) + O((x - \nu)^2) .
\]  

(7.2)

Then put \( \nu = \mu_x \), and hence an approximation to the variance of \( f(X) \) is given by,

\[
\text{var}(f(X)) \approx (f'(\mu_x))^2 \text{var}(X) .
\]  

(7.3)

Let \( \text{var}(X) = \sigma_x^2 \) be some function of the mean \( \mu_x \), which we write as \( h(\mu_x) \). To find a function \( f \) that satisfies Equation 7.1, we rewrite Equation 7.3 as follows,

\[
f'(\mu_x) \approx \sqrt{\frac{c}{h(\mu_x)}} .
\]  

(7.4)
This implies that for any function $h(\mu_x)$, where $1/\sqrt{h(\mu_x)}$ is integrable with respect to $\mu_x$, the relationship,

$$f(x) \propto \int_x^x \frac{1}{\sqrt{h(\mu_x)}} d\mu_x,$$

produces an $f$ that satisfies (7.1); that is, $f$ given by Equation 7.5 is a variance-stabilising transformation [23]. In what follows, we are particularly interested in the case, $h(\mu_x) = (\alpha + \beta \mu_x)^2$.

We have established that the spatial standard deviation of SST has an approximately linear relationship with the spatial mean of SST, in the Niño 3.4 region. That is,

$$S_z(t) \simeq \alpha + \beta \bar{z}(t).$$

(7.6)

If we now replace the empirical (i.e., spatial) moments of $z$ with theoretical moments in Equation 7.5, we obtain,

$$f(x) \propto \int_x^x \frac{1}{\sqrt{(\alpha + \beta \mu)^2}} d\mu = \ln(\alpha + \beta x),$$

(7.7)

modulo an additive and multiplicative constant, and provided $\alpha + \beta x > 0$. That is, in the domain $\alpha + \beta x > 0$,

$$f(x) = \ln(\alpha + \beta x),$$

(7.8)

is a variance-stabilising transformation. For the SST data, we write

$$u(s, t) \equiv \ln(\alpha + \beta z(s, t)),$$

(7.9)

for constants $\alpha$ and $\beta$ such that $(\alpha + \beta z(s, t)) > 0$. The slope $\beta$ and the intercept $\alpha$ in Equation 7.9 were obtained by WTS estimation; see Table 7.1.

### 7.2.1 All-months-combined

We plotted the monthly spatial standard deviation versus the monthly spatial mean of the transformed data \{u(s, t) : s \in D_s, t \in D_t\} from Equation 7.9, where the same $\hat{\alpha} = 5.559$ and $\hat{\beta} = -0.1692$ estimates were used to transform each month; see Figure 7.5 where a WTS fitted line is also superimposed. For this choice of $\hat{\alpha}$ and $\hat{\beta}$, $(\alpha + \beta z(s, t)) > 0$ for all \{z(s, t)\}.

We separated the points plotted in Figure 7.5 into their respective months and again plotted the spatial standard deviation versus the spatial mean of the transformed
Table 7.1: WTS slope and intercept estimates by month and for all-months-combined, for a regression of spatial standard deviation, $S_z(t)$, on spatial mean, $z(t)$, for $z(s,t)$ in the Niño 3.4 region.

<table>
<thead>
<tr>
<th>Month</th>
<th>WTS intercept estimate $\hat{\alpha}$</th>
<th>WTS slope estimate $\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>2.117</td>
<td>-0.04215</td>
</tr>
<tr>
<td>February</td>
<td>1.937</td>
<td>-0.04385</td>
</tr>
<tr>
<td>March</td>
<td>2.252</td>
<td>-0.05894</td>
</tr>
<tr>
<td>April</td>
<td>3.622</td>
<td>-0.1086</td>
</tr>
<tr>
<td>May</td>
<td>5.497</td>
<td>-0.1726</td>
</tr>
<tr>
<td>June</td>
<td>7.319</td>
<td>-0.2343</td>
</tr>
<tr>
<td>July</td>
<td>7.094</td>
<td>-0.2201</td>
</tr>
<tr>
<td>August</td>
<td>5.687</td>
<td>-0.1653</td>
</tr>
<tr>
<td>September</td>
<td>4.174</td>
<td>-0.1094</td>
</tr>
<tr>
<td>October</td>
<td>3.487</td>
<td>-0.08275</td>
</tr>
<tr>
<td>November</td>
<td>3.640</td>
<td>-0.08901</td>
</tr>
<tr>
<td>December</td>
<td>3.230</td>
<td>-0.07786</td>
</tr>
<tr>
<td>All months</td>
<td>5.559</td>
<td>-0.1692</td>
</tr>
</tbody>
</table>

SSTs $\{u(s,t)\}$ from Equation 7.9 for all $t \in D_t^{Jan}, D_t^{Feb}, \ldots, D_t^{Dec}$; see Figure 7.6. We again used the same estimates $\hat{\alpha} = 5.559$ and $\hat{\beta} = -0.1692$ for each month and fitted WTS lines to the points in each month, these lines are also shown in Figure 7.6. Negative dependence appears to be present in most of the months. It was hoped that this single transformation will remove the mean-standard deviation dependence in the individual months, January, February, . . . , December. In fact, it does not.

The analogous plot to Figure 7.1 for the transformed data, $\{u(s,t)\}$, is given in Figure 7.7. In Figures 7.6 and 7.7 the slopes still show a strong pattern, although they now oscillate around zero; see Table 7.2, whereas on the untransformed data all the slopes were negative; see Table 7.1. Our goal is to transform the data so that there is no mean-standard deviation dependence; that is, for each month (and for all-months-combined), the confidence interval for the slope $\beta$ should contain 0. In Figure 7.7, two regimes are present; March, April, May, and June where the slope increases noticeably from one month to the next and the other months of the year where the slope is more similar. This motivates the transform we consider in the next section.
Figure 7.5: Spatial standard deviation, $S_u(t)$, versus the spatial mean, $\bar{u}(t)$, of the transformed data $\{u(s,t)\}$, in the Niño 3.4 region, for all $t \in D_t$ with the WTS-fitted line superimposed.

Table 7.2: WTS model-based slope estimate and point-wise 95% bootstrap percentile confidence interval by month and for all-months-combined, for spatial standard deviation versus the spatial mean of the transformed data, $\{u(s,t)\}$, in the Niño 3.4 region.

<table>
<thead>
<tr>
<th>Month</th>
<th>WTS Slope estimate $\hat{\beta}$</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-0.1122</td>
<td>(-0.1495, -0.07483)</td>
</tr>
<tr>
<td>February</td>
<td>-0.06545</td>
<td>(-0.09344, -0.03747)</td>
</tr>
<tr>
<td>March</td>
<td>-0.06222</td>
<td>(-0.08641, -0.03803)</td>
</tr>
<tr>
<td>April</td>
<td>-0.03192</td>
<td>(-0.07655, 0.01272)</td>
</tr>
<tr>
<td>May</td>
<td>0.01213</td>
<td>(-0.03160, 0.05587)</td>
</tr>
<tr>
<td>June</td>
<td>0.04809</td>
<td>(-0.01816, 0.1143)</td>
</tr>
<tr>
<td>July</td>
<td>0.001304</td>
<td>(-0.08005, 0.08266)</td>
</tr>
<tr>
<td>August</td>
<td>-0.04324</td>
<td>(-0.08643, -0.00004002)</td>
</tr>
<tr>
<td>September</td>
<td>-0.1020</td>
<td>(-0.1338, -0.07019)</td>
</tr>
<tr>
<td>October</td>
<td>-0.1229</td>
<td>(-0.1436, -0.1022)</td>
</tr>
<tr>
<td>November</td>
<td>-0.1272</td>
<td>(-0.1496, -0.1048)</td>
</tr>
<tr>
<td>December</td>
<td>-0.1294</td>
<td>(-0.1509, -0.1079)</td>
</tr>
<tr>
<td>All months</td>
<td>0.007827</td>
<td>(-0.01824, 0.03389)</td>
</tr>
</tbody>
</table>
Figure 7.6: Spatial standard deviation, $S_u(t)$, versus the spatial mean, $\bar{u}(t)$, of the transformed data \{\(u(s, t)\}\}, in the Niño 3.4 region, for all $t \in D_{Jan}^t, D_{Feb}^t, \ldots, D_{Dec}^t$ with the WTS-fitted line superimposed.
Figure 7.7: Time sequences from December, January, February, . . . , November, December, January, showing WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{ u(s, t) \} \). The horizontal dotted black line is the zero line. The horizontal solid blue line is the estimate for all-months-combined.
7.2.2 ENSO Spring Barrier

It was noted in Figure 7.7 that there appeared to be two regimes, where the boreal summer months had different behaviour to the other months. Based on this an attempt was made to use what is known about the ENSO spring barrier to forecasting SSTs and El Niño/La Niña events [102]. The ENSO spring barrier is a phrase used to describe the breakdown in ENSO prediction skill during boreal spring. A number of hypotheses for the causes of this barrier have been postulated; see [34, 56, 102, 169, 179, 183]. The data from contiguous months were combined, but a pattern of estimated slopes remained.

By combining March, April, May, and June into one group and the other months into a second group it was possible to define a transform so that when a straight line was fitted (using WTS) to each month, zero was in the point-wise 95% bootstrap percentile confidence interval of the slope for each month; see Figure 7.8. However, there was still a pattern to the estimated slopes. Based on this data analysis, we concluded that each month has its own individual mean-standard deviation dependence that should be respected.

![Figure 7.8](image)

Figure 7.8: Time sequences from December, January, February, ..., November, December, January, showing WTS-estimated slope and intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data when two different transforms are used (one for March, April, May and June, and one for the other months). The horizontal dotted black line is the zero line.
7.2.3 Individual Months

To illustrate the next stage of our analysis, consider the month of January. The top left-hand panel of Figure 5.1 shows a plot of \( y = S_z(t) \) versus \( x = \tau(t) \), for all \( t \in D_{\text{Jan}} \). We used robust regression to obtain a WTS-fitted line, \( y = \hat{\alpha}_{\text{Jan}} + \hat{\beta}_{\text{Jan}} x \), where from Table 7.1, \( \hat{\alpha}_{\text{Jan}} = 2.117 \) and \( \hat{\beta}_{\text{Jan}} = -0.04215 \). Following the transformation given by Equation 7.9, but now just for the data in the month of January, we define for \( t \in D_{\text{Jan}} \),

\[
\nu_{\text{Jan}}(s, t) \equiv \ln \left( \hat{\alpha}_{\text{Jan}} + \hat{\beta}_{\text{Jan}} z(s, t) \right).
\]  

(7.10)

Similarly, we define \( \nu_{\text{Feb}}(s, t) \) for \( t \in D_{\text{Feb}} \), \ldots, and \( \nu_{\text{Dec}}(s, t) \) for \( t \in D_{\text{Dec}} \). All arguments of the log transformation were positive for all months. Finally, we combine these individual-month definitions to define a transformation of all the data, \( \{z(s, t)\} \), as:

\[
\nu(s, t) \equiv \nu^M(s, t), \text{ for } t \in M \text{ and } M \in \{\text{Jan, Feb, \ldots, Dec}\}.
\]  

(7.11)

The analogous plot to Figure 7.1, for the transformed data \( \{\nu(s, t)\} \), is given in Figure 7.10, and the analogous plot to Figure 7.6 for the transformed data \( \{\nu(s, t)\} \), is given in Figure 7.9. Clearly, the transformation given in Equation 7.11 has successfully removed the strong pattern in the estimated slopes, leaving a zero slope inside all twelve confidence intervals.

With the mean-standard deviation dependence removed, it is now meaningful to look at the intercept estimates. These show how the spatial standard deviations change from month-to-month, but they are now unconfounded with the level of warming or cooling in the Niño 3.4 region. In other words, on the transformed scale, the data \( \{\nu(s, t)\} \) show us the pure spatial variability of tropical Pacific SSTs by month. Like Figure 7.10, which was for the estimated slope estimates based on the transformed data, Figure 7.11 shows the estimated intercept estimates based on the transformed data with the associated point-wise bootstrap percentile confidence intervals. The ENSO spring barrier between April and May stands out. Here we recognise it as the month-to-month transition where the standard deviation increases the most.
Figure 7.9: Spatial standard deviation, $S_\nu(t)$, versus the spatial mean, $\nu(t)$, of the transformed data $\{\nu(s, t)\}$ (as defined by Equation 7.11), in the Niño 3.4 region, for all $t \in D_{\text{Jan}}$, $t \in D_{\text{Feb}}$, ..., $t \in D_{\text{Dec}}$ with the WTS-fitted line superimposed.
Figure 7.10: Time sequences from December, January, February, . . . , November, December, January, showing WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{\nu(s,t)\} \). The horizontal dotted black line is the zero line.
Figure 7.11: Time sequences from December, January, February, . . . , November, December, January, showing WTS-estimated intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \{\nu(s,t)\}. The horizontal dotted black line is the zero line. The intercept estimates track the spatial variability in the Niño 3.4 region.
Table 7.3: WTS model-based slope and intercept estimates and point-wise 95\% bootstrap percentile confidence interval of the slope for all \( t \in D_t^{\text{Jan}}, t \in D_t^{\text{Feb}}, \ldots, t \in D_t^{\text{Dec}}, \) and \( t \in D_t, \) for spatial standard deviation versus the spatial mean of the transformed data, \( \{v(s,t)\}, \) in the Niño 3.4 region.

<table>
<thead>
<tr>
<th>Month</th>
<th>WTS intercept est.</th>
<th>WTS slope est.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>0.04230</td>
<td>0.002129</td>
<td>(-0.05112, 0.05512)</td>
</tr>
<tr>
<td>February</td>
<td>0.04647</td>
<td>-0.008964</td>
<td>(-0.05777, 0.03870)</td>
</tr>
<tr>
<td>March</td>
<td>0.05860</td>
<td>0.001761</td>
<td>(-0.03504, 0.04512)</td>
</tr>
<tr>
<td>April</td>
<td>0.1014</td>
<td>0.01747</td>
<td>(-0.05158, 0.08556)</td>
</tr>
<tr>
<td>May</td>
<td>0.1613</td>
<td>0.03100</td>
<td>(-0.05602, 0.1181)</td>
</tr>
<tr>
<td>June</td>
<td>0.2203</td>
<td>0.03913</td>
<td>(-0.05101, 0.1336)</td>
</tr>
<tr>
<td>July</td>
<td>0.2186</td>
<td>0.02930</td>
<td>(-0.07663, 0.1362)</td>
</tr>
<tr>
<td>August</td>
<td>0.1624</td>
<td>-0.006191</td>
<td>(-0.07269, 0.06173)</td>
</tr>
<tr>
<td>September</td>
<td>0.1081</td>
<td>-0.001398</td>
<td>(-0.06887, 0.06687)</td>
</tr>
<tr>
<td>October</td>
<td>0.08262</td>
<td>0.0009334</td>
<td>(-0.04519, 0.04001)</td>
</tr>
<tr>
<td>November</td>
<td>0.08952</td>
<td>0.003934</td>
<td>(-0.03017, 0.04095)</td>
</tr>
<tr>
<td>December</td>
<td>0.07915</td>
<td>0.008073</td>
<td>(-0.03216, 0.04523)</td>
</tr>
<tr>
<td>All months</td>
<td>0.1693</td>
<td>-0.007837</td>
<td>(-0.02486, 0.009886)</td>
</tr>
</tbody>
</table>

7.2.4 Chosen Transform

Note that the transform defined in Equations 7.10–7.11 is a monotonic decreasing transform. Therefore for easier interpretability of the geophysical variable we shall make a trivial adjustment and use the following transform

\[
v^M(s,t) \equiv -\ln\left(\hat{\alpha}^M + \hat{\beta}^M z(s,t)\right), \quad \text{for \( t \in M \) and \( M \in \{\text{Jan, Feb, \ldots, Dec}\},\ }
\]

(7.12)

where \( \hat{\alpha}^M \) and \( \hat{\beta}^M \) are defined in Table 7.1.

The analogous plot to Figure 7.1, for the transformed data \( \{v(s,t)\}, \) is given in Figure 7.13, and the analogous plot to Figure 7.6 for the transformed data \( \{v(s,t)\}, \) is given in Figure 7.12. Clearly, the transformation given in Equation 7.12 has also successfully removed the strong pattern in the estimated slopes, leaving a zero slope inside all twelve confidence intervals.

Table 7.3 is based on transformation \( v \) and summarises the WTS-estimates on the transformed data and gives point-wise 95\% bootstrap percentile confidence intervals for the slope.

In what follows we will analyse this transformed SST data, \( \{v(s,t)\}. \) Visualisations of the transforms effect on the data are given in Figures 7.15–7.17. Figures 7.15 and 7.16 are spatial maps of the Niño 3.4 region for January 1983 and July 1988.
Figure 7.12: Spatial standard deviation, \( S_v(t) \), versus the spatial mean, \( \bar{v}(t) \), of the transformed data \( \{v(s, t)\} \) (as defined by Equation 7.12), in the Niño 3.4 region, for all \( t \in D_{\text{Jan}} \), \( t \in D_{\text{Feb}} \), ..., \( t \in D_{\text{Dec}} \) with the WTS-fitted line superimposed.

respectively. These figures show that the transform maintains the spatial structure of the original data. Figure 7.17 shows that differences between the transforms on the different seasons (and months) of the data.
Figure 7.13: Time sequences from December, January, February, . . . , November, December, January, showing WTS-estimated slope coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{ v(s, t) \} \). The horizontal dotted black line is the zero line.
Figure 7.14: Time sequences from December, January, February, ..., November, December, January, showing WTS-estimated intercept coefficients, and upper and lower limits from point-wise 95% bootstrap percentile confidence intervals for the transformed data \( \{v(s, t)\} \). The horizontal dotted black line is the zero line.
Figure 7.15: Comparison of the raw SST data, $z(s,t)$, to the variance-stabilised transformed data, $v(s,t)$, for January 1983 in the Niño 3.4 region. Units on the upper colour scale are °C, while the units on the lower colour scale are log(°C).
Figure 7.16: Comparison of the raw SST data, $z(s,t)$, to the variance-stabilised transformed data, $v(s,t)$, for July 1988 in the Niño 3.4 region. Units on the upper colour scale are °C, while the units on the lower colour scale are $\log(°C)$. 
Figure 7.17: Comparison of the raw SST data, $z(s,t)$, to the variance-stabilised transformed data, $v(s,t)$, for all $t \in D_t$ in the Niño 3.4 region. The values for each season have different colours (December, January, and February are red; March, April, and May are yellow; June, July, and August are blue; September, October, and November are green).
Chapter 8

Forecast Model Development for SSTs

In this chapter we fit autoregressive time series models to the spatial averages of the original SST data \( \{z(s, t)\} \), and to the spatial averages of the transformed SST data \( \{v(s, t)\} \), as defined by Equation 7.12 in Section 7.2.4. In Section 8.1, we give a theoretical introduction to autoregressive processes and some examples of their previous applications to SST data. Then in Section 8.2 we fit autoregressive processes to the spatial averages of the original SST data and to the spatial averages of the transformed SST data. In Section 8.3 we derive the equations required to forecast values on the original SST scale and on the transformed scale. To control for bias, we propose a parametric bootstrap to transform the forecasts made on the transformed scale back to the original SST scale. We use the first-order autoregressive process where the necessary means and variances are known in closed form to show that the parametric-bootstrap approach is appropriate.

8.1 Autoregressive Processes

An autoregressive process is a statistical model of time series data, where the current observation depends linearly on the previous observations plus an independent component that is independent and identically distributed (iid) across time. Autoregressive processes can be considered as discretisations of continuous-time ordinary or partial differential equations which are often used to model processes evolving over time [109, 155, 164].
A constant-mean Gaussian autoregressive process of order \( p \), denoted \( AR(p) \), can be written as follows:

\[
Y(t) - \mu = \phi_1[Y(t-1) - \mu] + \phi_2[Y(t-2) - \mu] + \cdots + \phi_p[Y(t-p) - \mu] + W(t),
\]

(8.1)

for \( t = p, p+1, \ldots \), where \( Y(t) \) is an observation at time \( t \); \( E[Y(t)] = \mu \) for all \( t \); \( W(t) \) is a mean-zero Gaussian process with variance \( \omega^2_Y \) that is independent of \( Y(t-1), \ldots, Y(t-p) \); and \( \{\phi_j : j = 1, \ldots, p\} \) are fixed but unknown parameters to be estimated [45]. The starting value \( Y(0) \) may be considered fixed or random in the model.

A temporal stochastic process with finite variance is called weakly (or wide-sense) stationary if its mean is constant and the covariance between any pair of observations is a function only of their lag (i.e. their temporal separation). Stationarity is important as it implies that aspects of the process are invariant through time. In contrast to weak stationarity, strict stationarity implies time-invariance of the whole process [159]. A straightforward calculation based on the coefficients \( \{\phi_1, \ldots, \phi_p\} \) of the autoregressive process (Equation 8.1) will determine whether the process is weakly stationary. If the \( p \) roots \( d_1, \ldots, d_p \) of the equation,

\[
f(d) \equiv d^p - \phi_1 d^{p-1} - \cdots - \phi_{p-1} d - \phi_p = 0 ,
\]

(8.2)

all lie within the unit circle of the complex plane, that is if \(|d_i| < 1\) for all \( i = 1, \ldots, p \), then the autoregressive process (Equation 8.1) is weakly stationary [106].

There are a number of ways to estimate the parameters of the \( AR(p) \) process from the time series data \( \{Y(1), \ldots, Y(T)\} \). Assume henceforth that the process is at least weakly stationary. Then typically \( \mu \) is estimated with the empirical mean of the time series:

\[
\hat{\mu} \equiv \frac{1}{T} \sum_{t=1}^{T} Y(t) .
\]

(8.3)

From Equation 8.1, the other parameters of the model (\( \omega^2_Y \) and \( \{\phi_j\} \)) can be obtained from an estimating equation based on the stationary autocovariance function, \( C(\tau) \equiv \text{cov}(Y(t), Y(t+\tau)) \), specifically the Yule-Walker equations [45, 182, 165]:

\[
\begin{align*}
C(0) &= \phi_1 C(1) + \cdots + \phi_p C(p) + \omega^2_Y \quad \text{(8.4)} \\
C(k) &= \phi_1 C(k-1) + \cdots + \phi_p C(k-p) \ , \ k = 1, \ldots, p , \quad \text{(8.5)}
\end{align*}
\]

where it should be remembered that \( C(h) = C(-h) \). Now, suppose \( C(\cdot) \) is estimated
with the empirical covariance function;

\[ \hat{C}(\tau) \equiv \frac{1}{T} \sum_{t=1}^{T-\tau} (Y(t + \tau) - \hat{\mu})(Y(t) - \hat{\mu}), \tau = 0, 1, \ldots, T - 1; \quad (8.6) \]

then solving Equations 8.4 and 8.5 based on \( \hat{C}(\cdot) \) yields estimates \( \hat{\omega}_Y^2 \) and \( \{\hat{\phi}_j\} \).

8.1.1 Autocorrelation and Partial Autocorrelation Functions

The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of a time series capture the dependence in the time series as a function of lag. For a time series \( \{Y(t) : t = 1, \ldots\} \), the ACF at lag \( h \) is defined as

\[ \text{ACF}(h) \equiv \text{corr}(Y(t), Y(t - h)), \quad h = 1, \ldots, t - 1. \quad (8.7) \]

The PACF at lag \( h \) is the correlation between \( Y(t) \) and \( Y(t - h) \), after adjusting for the intermediate observations, \( Y(t - h + 1), \ldots, Y(t - 1) \). The PACF at lag \( h \) is defined as follows: \( \text{PACF}(1) \equiv \text{corr}(Y(t), Y(t - 1)) \) and

\[ \text{PACF}(h) \equiv \text{corr} \left( Y(t) - \hat{Y}(t), Y(t - h) - \hat{Y}(t - h) \right), \quad h \geq 2, \quad (8.8) \]

where \( \hat{Y}(t) \) is the best (in a mean-squared-error sense) linear predictor of \( Y(t) \) based on the observations \( Y(t - h + 1), \ldots, Y(t - 1) \). Similarly, \( \hat{Y}(t - h) \) is the best (in a mean-squared-error sense) linear predictor of \( Y(t-h) \) based on \( Y(t-h+1), \ldots, Y(t-1) \) [28]. If we write \( \phi_{p,j} \) to represent the \( j \)-th fitted coefficient in an \( AR(p) \) process then, assuming \( \mu \) is known, these predictors can be written as [28]:

\[
\begin{align*}
\hat{Y}(t) & = \mu + \phi_{h-1,1}(Y(t - 1) - \mu) + \phi_{h-1,2}(Y(t - 2) - \mu) + \\
& \quad \cdots + \phi_{h-1,h-1}(Y(t - h + 1) - \mu), \quad (8.9) \\
\hat{Y}(t - h) & = \mu + \phi_{h-1,1}(Y(t - h + 1) - \mu) + \phi_{h-1,2}(Y(t - h + 2) - \mu) + \\
& \quad \cdots + \phi_{h-1,h-1}(Y(t - 1) - \mu). \quad (8.10)
\end{align*}
\]

The PACF can be estimated through fitting (by ordinary least squares) successively more complex AR processes of order 1, 2, 3, \ldots and keeping the estimates of the last autoregressive coefficients \( \phi_{1,1}, \phi_{2,2}, \phi_{3,3}, \ldots \) in each model [28]. Approximations are also possible, using, for example, the Levinson-Durbin recursion algorithm [57, 98], the details of which are given in an appendix of [28].
The PACF can be used as a diagnostic to determine whether an AR process (and more generally one of a number of diagnostics whether an autoregressive moving average, or ARMA process) is appropriate and, if so, what order model should be used. In particular, for an AR($p$) process the PACF will have its initial $p$ partial autocorrelations non-zero and the rest will be zero. That is, for $h > p$, the estimated PACF($h$) ≃ 0. Also, if the process is autoregressive, the ACF will taper to zero for increasing $h$ [28]. Using the ACF and PACF to determine the appropriateness and order, respectively, of an AR (ARMA) process is referred to as the Box-Jenkins approach to AR-order (ARMA-order) selection [28]. In Section 8.2, we shall use the PACF to determine the order of a fitted AR process.

### 8.1.2 Akaike Information Criterion

We have just seen that an inspection of where the PACF becomes approximately zero is an informal way to determine the order of an AR process. An alternative method for determining the AR’s order is the Akaike information criterion (AIC). Akaike specifically discussed goodness-of-fit testing for AR models in his seminal 1974 paper [14]. The AIC is a widely used model-selection criterion that balances goodness of fit and model complexity. Let the parameters in a model be represented by a vector $\theta$; the AIC of $\theta$ is given by

$$AIC(\hat{\theta}) = -2 \ln(ML) + 2k \quad (8.11)$$

where $ML$ represents the maximum with respect to $\theta$ of the likelihood function of the model, and $k$ is the number of independently adjusted parameters in the maximum likelihood estimate (MLE) $\hat{\theta}$ [14]. From a set of candidate models, the model chosen using AIC is the one that gives the minimum AIC value.

Akaike [12, 13] showed that if the error terms $\{W(t)\}$ are iid from a Gaussian distribution with mean zero and variance $\omega^2_Y$ in a stationary AR($p$), then the AIC can be obtained as

$$AIC(p) = T \ln(\hat{\omega}_Y^2) + 2(p + 2) \quad (8.12)$$

where $T$ is the number of observations, $\hat{\omega}_Y^2$ is the MLE of $\omega_Y^2$, and $(p + 2)$ takes into account estimating the mean $\mu$, the variance $\omega_Y^2$, and the AR coefficients $\phi_1, \ldots, \phi_p$ [15, 53, 76]. Shibata [142] gave the approximate MLE of $\hat{\omega}_Y^2$ in a mean-zero AR($p$)
process as
\[
\hat{\omega}^2_Y = \frac{1}{T} \sum_{t=p+1}^{T} (Y(t) - \hat{\phi}_1 Y(t-1) - \cdots - \hat{\phi}_p Y(t-p))^2, \tag{8.13}
\]
where \(\{\hat{\phi}_i : i = 1, \ldots, p\}\) are the MLEs of \(\{\phi_i : i = 1, \ldots, p\}\) for each given \(p \geq 1\). For an \(AR(p)\) with a known non-zero mean, \(\mu\), this estimate would be
\[
\hat{\omega}^2_Y = \frac{1}{T} \sum_{t=p+1}^{T} (\bar{Y}(t) - \hat{\phi}_1 \bar{Y}(t-1) - \cdots - \hat{\phi}_p \bar{Y}(t-p))^2, \tag{8.14}
\]
where \(\bar{Y}(t) \equiv Y(t) - \mu\). In practice, \(\bar{Y}\) is substituted for \(\mu\), and this gives a pseudo maximum likelihood estimate (see [70] for details on pseudo maximum likelihood estimation). Note that Equation 8.14 is also an empirical mean squared prediction error of the one-step ahead forecasts, \(\{\hat{Y}(t)\}\), although the presence of \(T\) instead of \((T - p)\) in the denominator makes it a biased estimator of \(\omega^2_Y\).

There are a number of other information criteria based on model-selection approaches, including the Bayes information criterion (e.g., [143, 53]) and the deviance information criterion (e.g., [29, 101]). In what follows we shall use AIC along with the PACF to select time series models for forecasting.

### 8.1.3 Autoregressive Processes in SST modelling

Autoregressive processes are useful for describing variability of a current observation in terms of past observations, and hence they can be used to forecast future behaviour. We shall fit an autoregressive process of order \(p\) determined by the SST data in the Niño 3.4 region (5S–5N, and 120W–170W) of the tropical Pacific Ocean.

Some authors have applied AR processes directly to SST data with varying levels of success. Reynolds [126] fit \(AR(1)\) models to \(5^\circ \times 5^\circ\) regions of SST anomalies in the mid-latitude North Pacific. Approximately half, (54%), of these regions could be represented as an \(AR(1)\) process, and those regions were mainly in the central part of the North Pacific.

Other authors extended the use of AR processes to incorporate additional oceanographic and atmospheric components or to capture more complex dynamics in the data. Newman et al. [112] modelled the annual average Pacific Decadal Oscillation (calculated from SST anomalies) as an \(AR(1)\). When the authors included
an ENSO index as a covariate in the SST model, they found the model performed better. Zwiers and von Storch [184] modelled an index of the El Niño Southern Oscillation using an extension of an AR process called a regime-dependent autoregressive process, where external variables control which parameters are used. They used a one-month-shifted version of the usual grouping of the months into seasons, namely FMA (February, March, April), MJJ (May, June, July), ASO (August, September, October), and NDJ (November, December, January), but they did not explain why. Another approach taken was to calculate the EOFs of the region of interest and then model the coefficients of those EOFs using AR processes (e.g. [94]). Berliner et al. [24] combined this idea with regime switching.

First-order Markov processes are a generalisation of AR(1) processes, where the conditional distribution of the future given the entire past depends only on the immediate past [137]. Xue et al. [180] applied first-order Markov processes to EOFs of tropical Pacific SSTs. They found that using Markov processes with seasonality resulted in better fits to the data, and their forecasts had substantially higher skill than using Markov processes without seasonality. The authors treated the months individually and had different transition matrices for each.

In what follows, we shall use autoregressive processes to forecast spatial averages of the original SST data and spatial averages of the transformed SST data (using the variance-stabilising transform developed in Chapter 7), averaged over the Niño 3.4 region. In the latter case, those forecasts need to be converted to unbiased forecasts on the original SST scale (in degrees Celsius).

### 8.2 Fitting Autoregressive Processes to Time Series of Tropical Pacific SSTs

In this section, we fit autoregressive processes to both the original SST time series data and the transformed SST time series data. Specifically, we fit an autoregressive model of order \( p \) (\( AR(p) \)), namely

\[
(Y(t) - \mu) = \phi_1(Y(t - 1) - \mu) + \phi_2(Y(t - 2) - \mu) + \cdots + \phi_p(Y(t - p) - \mu) + G_t(0, \omega_Y^2),
\]

and part of our fitting is to select the order \( p \). In Equation 8.15, \( \{G_t(0, \omega_Y^2)\} \) is a Gaussian process of iid Gaussian random variables with mean 0 and variance \( \omega_Y^2 \). We first identify outliers in each of the time series and then use PACF and AIC to choose \( p \) for data on the individual scales.
8.2.1 Original SST Data: Possible Outliers

Recall that in Section 1.2 we defined the spatial region of interest (Niño 3.4 region) as \( D_s = \{s_1, \ldots, s_{500}\} \) and the temporal period of interest (November 1981 – December 2014) as \( D_t = \{1, \ldots, 398\} \). In Chapter 3 we calculated the spatial means, \( \tau(t) \), of the original SST data, \( \{z(s,t)\} \), as

\[
\tau(t) \equiv \frac{1}{|D_s|} \sum_{s \in D_s} z(s,t) ; \quad t \in D_t .
\]  

We use these means to define the monthly-centred data as

\[
z_c(t) \equiv \tau(t) - \bar{z}^M ; \quad t \in D_t^M \quad \text{and} \quad M \in \{\text{Jan}, \text{Feb}, \ldots, \text{Dec}\},
\]  

where \( \bar{z}^M \equiv \frac{\sum_{t \in D_t^M} \tau(t)}{\sum_{t \in D_t^M} 1} \). Also recall that \( D_t^M \) was defined in Section 1.2.2; \( D_t^\text{Jan} \) corresponds to \{January 1982, January 1983, \ldots, January 2014\}, and similarly for the months February, \ldots, December, respectively. Notice that Equation 8.17 does not define anomalies from a pre-determined anomaly period (e.g., 30 years from 1970 on), but the centred data will be close to the anomalies. Here, similar to Xue et al. [180], we are treating each month separately, however, they used anomalies and we centred using data from the period of interest (November 1981–December 2014). The centred time series, \( \{z_c(t)\} \), is shown in Figure 8.1. All figures in this chapter and the chapter following were produced using the \texttt{R} software \texttt{ggplot2} package [171].

![Figure 8.1](image)

Figure 8.1: The centred time series, \( \{z_c(t)\} \), defined by Equation 8.17, where the index \( t \) ranges from 1 (November 1981) to 398 (December 2014).

From initial examination of Figure 8.1, some extreme observations were identified as potential outliers that could cause problems when fitting a stationary process such as
an AR\((p)\) to the centred data. We also plotted a histogram of the values of the time series; see Figure 8.2. Denote \(Q_1\) and \(Q_3\) as the lower and upper quartiles of the data. Tukey [160] suggested observations outside \([Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]\) be considered as outliers, where these boundaries were referred to as fences. If the data are a random sample from a normal distribution, the lower fence occurs at the 0.4 percentile and the upper fence at the 99.6 percentile. Putting aside for the moment the temporal dependence in the time series, these fences indicated that the following observations from \(\{\bar{z}_c(t)\}\) are potential outliers [7, 8]:

- November 1982 to February 1983 (El Niño);
- November 1988 (La Niña); and

The 1982–3 El Niño was very strong and, as a consequence, Australia had below average rainfall from April 1982 to February 1983 [7]. Similarly, the 1997–8 El Niño was strong to very strong, and below-average rainfall was recorded across most of eastern Australia from April 1997 to March 1998 [7]. For both of these El Niños, the outlying observations occurred during the austral summer towards the end of the El Niño event. There was also one outlier beyond the lower fence, namely November 1988, which was during a strong La Niña [8]. The collection of potential outliers are indicated by the colour blue in Figures 8.1 and 8.2. To be cautious, we shall exclude these observations when fitting the AR process.

Here we have identified global outliers, where the outlying value is unusual with regard to the entire dataset. We note there are other types of outliers, such as
contextual outliers, also referred to as conditional outliers. In a time series, these outliers are unusual values with regard to other data points with a similar temporal context. For example, in our analysis this could be a February observation that is substantially different from the collection of other February observations, or it could be an observation that is substantially different from its preceding and subsequent observations. Another type of outlier is collective outliers, where a group of values, rather than single values, are identified as substantially different from the entire dataset.

As previously discussed in Section 4.2, there are two ways to account for outliers in modelling, either omitting them from the model fitting and resulting inference or using robust methods. In this chapter we have chosen the first approach, however it should be noted that the second approach can also be undertaken for robust estimation of an AR process. Maronna et. al [103] provide details on different robust-estimation options for time series, including AR processes.

8.2.2 Transformed SST Data: Possible Outliers

On the transformed scale, as defined by Equation 7.12, we define the spatial means, \( \bar{v}(t) \) of the transformed SST data, \( \{v(s, t)\} \), as

\[
\bar{v}(t) \equiv \frac{1}{|D_s|} \sum_{s \in D_s} v(s, t); \quad t \in D_t, \quad (8.18)
\]

and the monthly-centred (and re-scaled) data as

\[
\bar{v}_c(t) \equiv \frac{\bar{v}(t) - \bar{v}_M}{I_v^M}; \quad t \in D_t^M \text{ and } M \in \{\text{Jan, Feb, \ldots, Dec}\}, \quad (8.19)
\]

where \( \{I_v^M\} \) are the intercepts by month of a linear regression of spatial standard deviations on spatial means; these intercept estimates were given in Table 7.3. The values \( \{\bar{v}^M\} \) for each month are defined as \( \bar{v}^M \equiv \sum_{t \in D_t^M} \bar{v}(t) / \sum_{t \in D_t^M} 1 \). In Table 8.1 we give the values for \( \{\bar{v}^M\} \) and \( \{I_v^M\} \).

The time series \( \{\bar{v}_c(t)\} \) defined by Equation 8.19 is shown in Figure 8.3. In Figure 8.4, we also plotted a histogram of the values of the time series and, as for the original data, we used the Tukey fences to identify outliers. These fences indicated that the following observations from \( \{\bar{v}_c(t)\} \) are potential outliers [7, 8]:

- January to March 1983 (El Niño);
Table 8.1: The values \( \{ \bar{v}^M \} \) and \( \{ I_v^M \} \) in Equation 8.19 for each month \( M \in \{ \text{Jan, Feb, \ldots, Dec} \} \).

<table>
<thead>
<tr>
<th>Month</th>
<th>( \bar{v}^M )</th>
<th>( I_v^M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>0.0011</td>
<td>0.04230</td>
</tr>
<tr>
<td>February</td>
<td>0.2663</td>
<td>0.04647</td>
</tr>
<tr>
<td>March</td>
<td>0.4354</td>
<td>0.05860</td>
</tr>
<tr>
<td>April</td>
<td>0.5113</td>
<td>0.1014</td>
</tr>
<tr>
<td>May</td>
<td>0.3988</td>
<td>0.1613</td>
</tr>
<tr>
<td>June</td>
<td>0.2275</td>
<td>0.03100</td>
</tr>
<tr>
<td>July</td>
<td>-0.0582</td>
<td>0.2203</td>
</tr>
<tr>
<td>August</td>
<td>-0.2035</td>
<td>0.1624</td>
</tr>
<tr>
<td>September</td>
<td>-0.2134</td>
<td>0.1081</td>
</tr>
<tr>
<td>October</td>
<td>-0.2405</td>
<td>0.08262</td>
</tr>
<tr>
<td>November</td>
<td>-0.2308</td>
<td>0.08952</td>
</tr>
<tr>
<td>December</td>
<td>-0.1436</td>
<td>0.07915</td>
</tr>
</tbody>
</table>

Figure 8.3: The time series, \( \{ v_c(t) \} \), defined by Equation 8.19, where the index \( t \) ranges from 1 (November 1981) to 398 (December 2014).

- February to April 1992 (El Niño); and

Recall that the 1982–3 El Niño and the 1997–8 El Niño were both strong and impacted large regions of Australia; some of their months were also identified as having potentially outlying values on the original SST scale. The 1991-2 El Niño was also strong (and had the greatest impact in Queensland, Australia); during February to April 1992, the effects were strongest in the northern, tropical parts of Australia [7]. These potential outliers in the time series \( \{ v_c(t) \} \) are indicated by the colour blue in Figures 8.3 – 8.4, and in the next sub-section are combined with the
outliers from the time series $\{z_c(t)\}$.

### 8.2.3 Combined Outliers

Some of the observations identified as potential outliers in the time series $\{v_c(t)\}$ were also identified as potential outliers in the first time series $\{z_c(t)\}$. Table 8.2 summarises the observations identified as outliers in one or both time series.

Table 8.2: The observations identified as outliers in the time series, $\{z_c(t)\}$ and $\{v_c(t)\}$.

<table>
<thead>
<tr>
<th>${v_c(t)}$ outlier</th>
<th>${z_c(t)}$ outlier</th>
<th>Not ${z_c(t)}$ outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>${v_c(t)}$ outlier</td>
<td>January – February 1983</td>
<td>March 1983</td>
</tr>
<tr>
<td></td>
<td>December 1997 – February 1998</td>
<td>February – April 1992</td>
</tr>
<tr>
<td>Not ${v_c(t)}$ outlier</td>
<td>November – December 1982</td>
<td></td>
</tr>
<tr>
<td></td>
<td>November 1988</td>
<td></td>
</tr>
<tr>
<td></td>
<td>September – November 1997</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\mathbb{D}_t$</td>
<td></td>
</tr>
</tbody>
</table>

To allow a consistent use of the data for model fitting we exclude all potential outliers. We define $\mathbb{D}_t$ to be the subset of $D_t$ without the observations identified as outliers in Table 8.2, and hence we fit AR models robustly on both the original scale and the transformed scale using only the observations in $\mathbb{D}_t$. When forecasting in Sections 9.1 and 9.2, we assess in-sample and out-of-sample forecasting using $D_t$.  

---

Figure 8.4: Histogram of the time series, $\{v_c(t)\}$, defined by Equation 8.19, where potential outliers are shown in blue.
8.2.4 Original SST Data: AR Fitting

In Figure 8.5, we show the PACF for the centred time series \( \{ \tau_c(t) : t \in D_t \} \) given by Equation 8.17, which suggests that an autoregressive process of order 3 would be best. Using the \texttt{arima} function in the \texttt{stats} package in the statistical software \texttt{R} [121], we calculated AIC for fitting an \( AR(p) \), where \( p = 1, \ldots, 8 \). The AIC values are shown in Figure 8.6, which for orders 3 (AIC= 42.9), 4 (AIC= 43.6), 5 (AIC= 42.2), and 6 (AIC= 41.0) are all very similar. Based on the PACF and AIC results, we chose \( p = 3 \) and fitted an \( AR(3) \) process.

Figure 8.5: Partial autocorrelation function (PACF) plot for the centred time series \( \{ \tau_c(t) \} \) defined by Equation 8.17.

Figure 8.6: AIC plot for fitting an \( AR(p) \) to the centred time series \( \{ \tau_c(t) \} \) defined by Equation 8.17.

The \( AR(3) \) process was fitted using the \texttt{ar} function in the R \texttt{stats} package, re-
sulting in:

\[(\bar{z}_c(t) + 0.0219) = 1.291(\bar{z}_c(t - 1) + 0.0219) - 0.238(\bar{z}_c(t - 2) + 0.0219) - 0.1335(\bar{z}_c(t - 3) + 0.0219) + G_t(0, 0.062), \quad (8.20)\]

where recall that the last term comes from a Gaussian process of iid Gaussian random variables with mean 0 and variance 0.062. Notice that because \{\bar{z}(t)\} has been centred, the fitted constant mean of the centred data \{z_c(t)\} is small (\(\hat{\mu} = -0.0219\)).

The residuals from the fitted AR\( (p)\) model are defined as:

\[
\bar{z}_c(t)_{\text{res}} = \bar{z}_c(t) - \hat{\mu} - \sum_{i=1}^{3} \hat{\phi}_i(\bar{z}_c(t - i) - \hat{\mu}) \; ; \; t = 4, \ldots, T , \quad (8.21)
\]

where \(\hat{\mu}\) is the fitted mean and \{\(\hat{\phi}_i : i = 1, \ldots, p\)\} are the fitted AR coefficients.

The residuals from the fitted model given in Equation 8.20 are shown in Figures 8.7 (time series) and 8.8 (histogram). There is no obvious trend to the residuals and few outliers. From the histogram in Figure 8.7, three residual values were identified to be outside Tukey’s fences, which is what is expected if the residuals are Gaussian distributed. These are indicated by the colour blue in Figure 8.7. These residuals correspond to May and October 1988 (during a La Niña event) and October 1991 (during an El Niño event). Further, the ACF and PACF of \{\bar{z}_c(t)_{\text{res}}\} shown in Figures 8.9 and 8.10, respectively, indicate no remaining autocorrelation in the residuals. Taken together, these analyses indicate that the AR(3) process given by Equation 8.20 is a good fit to the centred time series, \{\bar{z}_c(t)\}.

![Figure 8.7: Residuals, \{\bar{z}_c(t)_{\text{res}}\} given by Equation 8.21, from the AR(3) process fitted to \{\bar{z}_c(t)\}.](Image)
Figure 8.8: Histogram of the residuals, $\{z_c(t)_{\text{res}}\}$ given by Equation 8.21, from the $AR(3)$ process fitted to $\{z_c(t)\}$.

Figure 8.9: Plot of the autocorrelation function (ACF) for $\{z_c(t)_{\text{res}}\}$ given by Equation 8.21, the residuals from the $AR(3)$ process fitted to $\{z_c(t)\}$. 
Figure 8.10: Plot of the partial autocorrelation function (PACF) for \( \{z_c(t)_{\text{res}}\} \) given by Equation 8.21, the residuals from the \( AR(3) \) process fitted to \( \{z_c(t)\} \).
8.2.5 Transformed SST Data: AR Fitting

In Figure 8.11, we show the PACF for the centred time series \( \{ \bar{v}_c(t) : t \in D_t \} \) given by Equation 8.19, which suggests that an autoregressive process of order 2, should be considered for \( \{ \bar{v}_c(t) \} \). Again using the \texttt{arima} function in the \texttt{R} \texttt{stats} package [121], we calculated AIC for fitting an \( AR(p) \), where \( p = 1, \ldots, 8 \). The AIC values are shown in Figure 8.12. The smallest AIC value was 159.4 for \( p = 2 \), but for \( p = 3 \) and 4 it was not very much larger.

![Partial ACF plot](image1)

Figure 8.11: Partial autocorrelation function (PACF) plot for the centred time series \( \{ \bar{v}_c(t) \} \) defined by Equation 8.19.

![AIC plot](image2)

Figure 8.12: AIC plot for fitting an \( AR(p) \) to the centred time series \( \{ \bar{v}_c(t) \} \) defined by Equation 8.19.

Residual analysis showed that in fact an \( AR(2) \) had a less-than-satisfactory fit, so
we fitted an AR(3) model. The fitted AR(3) model using the ar function is,

\[
(\hat{v}_c(t) + 0.0203) = 1.304(\hat{v}_c(t - 1) + 0.0203) - 0.347(\hat{v}_c(t - 2) + 0.0203) \\
- 0.0437(\hat{v}_c(t - 3) + 0.0203) + G_t(0, 0.0846),
\] (8.22)

where again the last term comes from a Gaussian process of iid Gaussian random variables with mean 0 and variance 0.0846.

We define the residuals from the model given in Equation 8.22 as,

\[
\hat{v}_c(t)_{\text{res}} = \hat{v}_c(t) - \hat{\mu} - \sum_{i=1}^{3} \hat{\phi}_i(\hat{v}_c(t - i) - \hat{\mu}); \ t = 4, \ldots, T.
\] (8.23)

These residuals, \(\{\hat{v}_c(t)_{\text{res}}\}\) from Equation 8.23, are shown in Figures 8.13 (time series) and 8.14 (histogram). There is no obvious trend to the residuals, however there are more outlying residuals than were identified from the AR fit to the original data. Tukey’s fences identified 10 outliers, which are indicated by the colour blue in Figure 8.13.

Some of these outliers (June 1982 and June 1987) are within El Niño events, and three (April and May 1988, and October 1988) are either in La Niña events or just before a La Niña event. Note, both May and October 1988 were also outliers in \(\hat{v}_c(t)_{\text{res}}\). Recall the “ENSO spring barrier” is a term used to describe the breakdown in ENSO prediction skill during the boreal spring (i.e. March, April, and May). The other outliers are April and June from 1993 and 2003, which could be related to the spring barrier. These could also be collective outliers not previously identified as we only identified and excluded global outliers. The transformed scale has allowed us to identify unusual observations that were not obvious on the original scale.

Further, the ACF of the residuals, Figure 8.15, shows unexpectedly large values at lags 7 and 16, which indicates there is some variability in the data that the AR(3) model is not capturing. It also suggests that the transformed scale allows us to identify high-lag periodicities in the data that we were unable to identify on the original scale. We fitted higher-order AR(p) models, considered more complex autoregressive processes, and different standardisation methods, however, we were unable to find a model that had fewer residual outliers than the AR(3) fitted to \(\{v_c(t)\}\).
Figure 8.13: Residuals, $\{\tau_c(t)_{\text{res}}\}$ given by Equation 8.23, from the AR(3) process fitted to $\{\tau_c(t)\}$.

Figure 8.14: Histogram of the residuals, $\{\tau_c(t)_{\text{res}}\}$ given by Equation 8.23, from the AR(3) process fitted to $\{\tau_c(t)\}$.
Figure 8.15: Plot of the autocorrelation function (ACF) for the residuals, \( \{ \tau_c(t)_{\text{res}} \} \) given by Equation 8.23, from the AR(3) process fitted to \( \{ \tau_c(t) \} \).

Figure 8.16: Plot of the partial autocorrelation function (PACF) for the residuals, \( \{ \tau_c(t)_{\text{res}} \} \) given by Equation 8.23, from the AR(3) process fitted to \( \{ \tau_c(t) \} \).
8.3 Forecasting Equations

Forecasting is the prediction of a future value based on past and present values. After fitting an AR process, we used the estimated mean, variance, and autoregression coefficients in the forecast as if they were known. That is, these additional sources of variability in the forecast are not accounted for when claiming that we have found a minimum mean-squared-prediction-error (i.e., optimal) forecast. This is in line with what is often common practice in time series and corresponds to the so-called EBLUP (empirical best linear unbiased prediction) in small-area estimation [123].

The optimal one-step-ahead forecast for an \( AR(p) \) process \( \{Y(t)\} \) (as described in Equation 8.1) is,

\[
E[Y(t+1)|Y(1), \ldots, Y(t)] = \mu + \sum_{i=1}^{p} \phi_i (Y(t+1-i) - \mu), \quad t = p, p+1, \ldots, \tag{8.24}
\]

where the parameters \( \mu \) and \( \{\phi_i : i = 1, \ldots, p\} \) are assumed known.

8.3.1 Forecasting Equations: Original SST

Forecasting is straightforward on the original scale for the process \( \{z_c(t)\} \). At time \( t \), \( z_c(t) \), \( z_c(t-1) \), and \( z_c(t-2) \) were used to forecast \( \hat{z}_c(t+\tau) \), for \( \tau = 1, \ldots, 7 \) using Equation 8.20. While the possible outliers we identified in Sections 8.2.1–8.2.3 were excluded from the AR fitting, for the purpose of forecasting we consider all the observations (i.e. \( D_t \)). The optimal forecast from our fitted AR(3) is:

\[
\hat{z}_c(t + \tau) = E[z_c(t + \tau)|z_c(t-2), z_c(t-1), z_c(t)]. \tag{8.25}
\]

We then rearrange Equation 8.17 to obtain forecasts on the original degrees-Celsius scale and denote these forecasts as \( \hat{z}(t + \tau) ; \tau = 1, \ldots, 7 \). They can be calculated as follows:

\[
\hat{z}(t + \tau) = \hat{z}_c(t + \tau) + \bar{z}^{M_{t+\tau}} ; \quad t \in D_t^M, \tau = 1, \ldots, 7, \tag{8.26}
\]

where \( t \) is a time point corresponding to month \( M \) and \( M_{t+\tau} \) is the month that is \( \tau \) months beyond month \( M \). Hence, \( \bar{z}^{M_{t+\tau}} \) is the relevant mean for month \( M_{t+\tau} \), and \( D_t^M \) is made of all the time points in \( D_t \) that are in month \( M \), and \( M \in \{\text{Jan, Feb, \ldots, Dec}\} \).
8.3.2 Forecasting Equations: Transformed SST

At time \( t \), \( \bar{v}_c(t) \), \( \bar{v}_c(t-1) \), and \( \bar{v}_c(t-2) \) were used to forecast \( \bar{v}_c(t+\tau) \), for \( \tau = 1, \ldots, 7 \) using Equation 8.22; we denote \( \hat{v}_c(t+\tau) \) as the forecast. The optimal forecast on the transformed scale is:

\[
\hat{v}_c(t + \tau) \equiv \mathbb{E}[\bar{v}_c(t + \tau)|\bar{v}_c(t-2), \bar{v}_c(t-1), \bar{v}_c(t)] .
\] (8.27)

In what is to follow, we back-transform to obtain an unbiased forecast on the original scale, where it is a non-linear forecast. We shall denote this forecast as \( \tilde{z}(t+\tau) \).

Recall \( \hat{z}_c(t + \tau) \) was the optimal forecast from the centred data on the original scale. Hence \( \{\hat{z}(t + \tau)\} \) is unbiased on the original scale for predicting \( z(t) \). Notice though that \( \{\hat{z}(t + \tau)\} \) is a linear forecast on the original scale. Both forecasts, \( \{\hat{z}(t + \tau)\} \) and \( \{\tilde{z}(t + \tau)\} \), are constructed under different assumptions, and we shall use their in-sample and out-of-sample forecasting performance to determine which is preferable.

First, we need to transform \( \hat{v}_c(t + \tau) \) to the original scale. Recall from Equation 8.19 that we calculated monthly-centred data on the "\( v \)-scale". To back-transform to the "\( z \)-scale", we first use Equation 8.19, to obtain a possibly biased forecast on the \( v \)-scale, namely

\[
\hat{v}(t + \tau) \equiv v^{M_{t+\tau}} + \hat{v}_c(t + \tau) \times I^{M_{t+\tau}} ; \ t \in D^{M_{t+\tau}}, \tau = 1, \ldots, 7 ,
\] (8.28)

where recall that \( M_{t+\tau} \) is the month, \( \tau \) months beyond month \( M \). We define \( v^{M_{t+\tau}} \equiv \sum_{t \in D^{M_{t+\tau}}} v(t) / \sum_{t \in D^{M_{t+\tau}}} 1 \), and recall that \( I^{M_{t+\tau}} \) is the intercept of a linear regression of spatial standard deviation against spatial mean for month \( M_{t+\tau} \); see Table 8.1.

We want a predictor \( \tilde{z}(t + \tau) \) based on \( \hat{v}(t + \tau) \) in Equation 8.28 such that it is unbiased; that is,

\[
\mathbb{E} [\tilde{z}(t + \tau)] = \mathbb{E}[z(t + \tau)] .
\] (8.29)

In Chapter 7 we defined the transformation to obtain \( v(s, t) \) from \( z(s, t) \) in Equation 7.11, where \( \alpha^M \) is the relevant monthly intercept and the \( \beta^M \) is the relevant monthly slope from Table 7.1. From that we can write

\[
z(t) = \frac{1}{\beta^M} \sum_{s \in D_s} \exp(-v(s, t)) - \frac{\alpha^M}{\beta^M} ; \ t \in D^M_t, \tau = 1, \ldots, 7 ,
\] (8.30)

where recall \( D_s \) is the 500 1 \times 1 degree ocean pixels in the Niño 3.4 region.
If we use a first-order Taylor series expansion, namely \( \exp(-x) \approx 1 - x \), and write

\[
\begin{align*}
\overline{z}(t) & \approx \frac{1}{\beta^M} \sum_{s \in D_s} \frac{1 - v(s, t)}{|D_s|} - \frac{\alpha^M}{\beta^M} \\
& = \frac{1}{\beta^M} \left( 1 - \overline{v}(t) \right) - \frac{\alpha^M}{\beta^M} \\
& \approx \frac{1}{\beta^M} \exp(-\overline{v}(t)) - \frac{\alpha^M}{\beta^M} \\
& = \frac{\exp(-\overline{v}(t)) - \alpha^M}{\beta^M} \quad (8.31)
\end{align*}
\]

We use this result to motivate an unbiased predictor as defined by Equation 8.29.

We note that if we substitute \( t + \tau \) for \( t \) in Equation 8.31 it is of the form

\[
\overline{z}(t + \tau) \approx c_1 + c_2 \exp(-\overline{v}(t + \tau)) \quad (8.32)
\]

and so we seek to find \( k_1 \) and \( k_2 \) in

\[
\overline{\tilde{z}}(t + \tau) = k_1 + k_2 \exp \left( -\overline{v}(t + \tau) \right) ; \ t \in D_t^M \ , \ \tau = 1, \ldots , 7 \ , \quad (8.33)
\]

such that Equation 8.29 is satisfied (i.e., \( \overline{\tilde{z}}(t + \tau) \) is unbiased). Note that \( k_1 \) and \( k_2 \) depend on \( M \) and \( \tau \) and on occasions we write them as \( k_1(M, \tau) \) and \( k_2(M, \tau) \), respectively. In what follows we use a parametric bootstrap and simple linear regression to obtain estimates for \( k_1(M, \tau) \) and \( k_2(M, \tau) \).

On the transformed scale, it is assumed that \( \{v_c(t)\} \) follows a Gaussian process due to the averaging of \( v(s, t) \) over \( s \in D_s \) and the transformation to make variances constant. Hence, \( \overline{v}(t + \tau) \) is Gaussian distributed and consequently \( \exp \left( -\overline{v}(t + \tau) \right) \) is log-Gaussian distributed. Therefore, to calculate \( \overline{E} \left[ \exp \left( -\overline{v}(t + \tau) \right) \right] \) in Equation 8.33 requires the calculation of the mean and variance of \( \overline{v}(t + \tau) \). For an AR(3) process, the mean and variance calculations are relatively straightforward for \( \tau = 1 \). However, for an AR(p) process, \( p > 1 \), as \( \tau \) increases the formulas for means and variances become more and more complicated and are only available as recursive formulas (see, for example, [37]). In the next section, we instead use parametric bootstrapping to find \( k_1 \) and \( k_2 \) in Equation 8.33. Then in the following section we implement it on an AR(1) to show that our alternative approach matches very well the exactly unbiased forecasts.
8.3.3 Parametric Bootstrap

Bootstrapping was discussed in Chapter 6 in the context of estimating regression parameters. There we performed non-parametric bootstrapping by resampling from the data. Here we use parametric bootstrapping by simulating data from the fitted model \[81\]. This approach obtains Monte Carlo approximations of the coefficients \(k_1\) and \(k_2\) for Equation 8.33. For an AR(3) our procedure is as follows: For \(b = 1, \ldots, 1600\), where 1600 is the bootstrap sample size, carry out the following steps.

1. Based on the AR(3) process defined by Equation 8.22 with \(\mu = -0.0203, \phi_1 = 1.304, \phi_2 = -0.347, \phi_3 = -0.0437\) and \(\sigma^2 = 0.0846\), generate a time series of length 398, and denote this time series as \(\{v_{ct}^{(b)}(t) : t = 1, \ldots, 398\}\).

2. Calculate the corresponding values, \(\bar{z}^{(b)}(t)\), using Equation 8.30.

3. Forecast for \(t = 1, \ldots, 398 - \tau\), from \(v_{ct}^{(b)}(t)\) based on the AR(3) in Equation 8.22 to obtain \(\hat{v}_{ct}^{(b)}(t + \tau)\), for \(\tau = 1, \ldots, 7\).

4. Calculate \(\hat{v}_{ct}^{(b)}(t + \tau)\) using

\[
\hat{v}_{ct}^{(b)}(t + \tau) \equiv \bar{v}_{M_t + \tau} + \hat{v}_{ct}^{(b)}(t + \tau) \times I_{M_t + \tau} ; t \in D_t^M, \tau = 1, \ldots, 7 .
\]

Then calculate \(\exp \left( -\hat{v}_{ct}^{(b)}(t + \tau) \right) ; t \in D_t^M, \tau = 1, \ldots, 7 .\)

For each month \(M\) and each lag \(\tau\), we use the corresponding bootstrap replicates to define

\[ x \equiv \frac{1}{1600} \sum_{b=1}^{1600} \exp \left( -\hat{v}_{ct}^{(b)}(t + \tau) \right) \text{ and } y \equiv \frac{1}{1600} \sum_{b=1}^{1600} \hat{z}^{(b)}(t + \tau) ; t \in D_t^M, \tau = 1, \ldots, 7 .\]

Then, motivated by Equation 8.31 and the variability in the bootstrap estimates, we use linear regression of \(y\) versus \(x\) to estimate \(k_1\) and \(k_2\) in:

\[ y = k_1 + k_2 x + \text{error}, \]

for each month \(M\) and each lag \(\tau\).

In the following simulation study (Section 8.4) we found that when fitting \(y\) to \(x\) in Equation 8.36, using Weighted Theil-Sen (WTS) regression parameter estimates to obtain the values of \(k_1\) and \(k_2\) results in better forecasts than using OLS estimates.
The WTS fitted values obtained from the bootstrap algorithm given above applied to the SST data are given in Tables 8.3 and 8.4. Figure 8.17 provides examples of the $12 \times 7 = 84$ WTS fitted lines; shown are the fitted lines for $(M, \tau) \in \{\text{Feb, Oct}\} \times \{1, 3, 7\}$.

Table 8.3: The WTS fitted value of $k_1$ for each $M \in \{\text{Jan, Feb,\ldots, Dec}\}$ and $\tau = 1, \ldots, 7$.

<table>
<thead>
<tr>
<th>Month</th>
<th>$\tau = 1$</th>
<th>$\tau = 2$</th>
<th>$\tau = 3$</th>
<th>$\tau = 4$</th>
<th>$\tau = 5$</th>
<th>$\tau = 6$</th>
<th>$\tau = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>44.66</td>
<td>39.10</td>
<td>34.30</td>
<td>32.83</td>
<td>32.26</td>
<td>33.58</td>
<td>36.82</td>
</tr>
<tr>
<td>February</td>
<td>38.89</td>
<td>34.17</td>
<td>32.83</td>
<td>32.27</td>
<td>34.09</td>
<td>35.74</td>
<td>39.98</td>
</tr>
<tr>
<td>March</td>
<td>33.63</td>
<td>32.60</td>
<td>32.17</td>
<td>33.87</td>
<td>36.89</td>
<td>41.12</td>
<td>45.04</td>
</tr>
<tr>
<td>April</td>
<td>32.07</td>
<td>31.71</td>
<td>33.02</td>
<td>35.60</td>
<td>41.07</td>
<td>44.94</td>
<td>43.64</td>
</tr>
<tr>
<td>May</td>
<td>31.46</td>
<td>32.61</td>
<td>34.72</td>
<td>38.91</td>
<td>43.75</td>
<td>42.00</td>
<td>42.31</td>
</tr>
<tr>
<td>June</td>
<td>63.22</td>
<td>32.57</td>
<td>35.95</td>
<td>42.37</td>
<td>50.02</td>
<td>47.93</td>
<td>48.15</td>
</tr>
<tr>
<td>July</td>
<td>55.95</td>
<td>50.02</td>
<td>34.61</td>
<td>39.82</td>
<td>45.40</td>
<td>43.25</td>
<td>44.44</td>
</tr>
<tr>
<td>August</td>
<td>57.49</td>
<td>50.18</td>
<td>43.67</td>
<td>38.53</td>
<td>43.82</td>
<td>42.77</td>
<td>44.19</td>
</tr>
<tr>
<td>September</td>
<td>55.81</td>
<td>50.41</td>
<td>43.22</td>
<td>35.10</td>
<td>43.26</td>
<td>42.29</td>
<td>43.86</td>
</tr>
<tr>
<td>October</td>
<td>57.29</td>
<td>52.68</td>
<td>44.92</td>
<td>37.16</td>
<td>34.87</td>
<td>41.84</td>
<td>42.93</td>
</tr>
<tr>
<td>November</td>
<td>56.25</td>
<td>50.12</td>
<td>42.47</td>
<td>36.04</td>
<td>33.50</td>
<td>33.54</td>
<td>42.17</td>
</tr>
<tr>
<td>December</td>
<td>50.62</td>
<td>46.28</td>
<td>39.82</td>
<td>34.47</td>
<td>32.63</td>
<td>31.79</td>
<td>33.23</td>
</tr>
</tbody>
</table>

Table 8.4: The WTS fitted value of $k_2$ for each $M \in \{\text{Jan, Feb,\ldots, Dec}\}$ and $\tau = 1, \ldots, 7$.

<table>
<thead>
<tr>
<th>Month</th>
<th>$\tau = 1$</th>
<th>$\tau = 2$</th>
<th>$\tau = 3$</th>
<th>$\tau = 4$</th>
<th>$\tau = 5$</th>
<th>$\tau = 6$</th>
<th>$\tau = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>-23.43</td>
<td>-18.35</td>
<td>-10.80</td>
<td>-7.27</td>
<td>-5.61</td>
<td>-5.88</td>
<td>-8.07</td>
</tr>
<tr>
<td>February</td>
<td>-18.02</td>
<td>-10.57</td>
<td>-7.26</td>
<td>-5.61</td>
<td>-6.34</td>
<td>-7.20</td>
<td>-10.66</td>
</tr>
<tr>
<td>April</td>
<td>-6.12</td>
<td>-4.88</td>
<td>-5.32</td>
<td>-7.06</td>
<td>-11.53</td>
<td>-14.31</td>
<td>-13.44</td>
</tr>
<tr>
<td>June</td>
<td>-36.74</td>
<td>-4.86</td>
<td>-7.32</td>
<td>-12.55</td>
<td>-18.28</td>
<td>-16.82</td>
<td>-18.63</td>
</tr>
<tr>
<td>July</td>
<td>-29.47</td>
<td>-30.43</td>
<td>-6.22</td>
<td>-10.49</td>
<td>-14.65</td>
<td>-13.11</td>
<td>-15.43</td>
</tr>
<tr>
<td>August</td>
<td>-31.00</td>
<td>-30.64</td>
<td>-25.43</td>
<td>-9.45</td>
<td>-13.41</td>
<td>-12.73</td>
<td>-15.20</td>
</tr>
<tr>
<td>September</td>
<td>-29.32</td>
<td>-30.94</td>
<td>-24.72</td>
<td>-12.15</td>
<td>-12.96</td>
<td>-12.34</td>
<td>-14.91</td>
</tr>
<tr>
<td>October</td>
<td>-30.79</td>
<td>-33.89</td>
<td>-27.34</td>
<td>-15.57</td>
<td>-10.32</td>
<td>-11.98</td>
<td>-14.10</td>
</tr>
<tr>
<td>December</td>
<td>-24.12</td>
<td>-25.54</td>
<td>-19.46</td>
<td>-11.09</td>
<td>-6.99</td>
<td>-5.03</td>
<td>-5.57</td>
</tr>
</tbody>
</table>
Figure 8.17: Shown are the WTS fitted lines for $M = \text{Feb}$ (left panels) and $M = \text{Oct}$ (right panels), $\tau = 1$ (top panels), $\tau = 3$ (middle panels) and $\tau = 7$ (bottom panels); in the parametric bootstrap, $x = \frac{1}{1600} \sum_{b=1}^{1600} \exp \left( -\hat{v}^{(b)}(t + \tau) \right)$ (horizontal axis) and $y = \frac{1}{1600} \sum_{b=1}^{1600} \bar{z}^{(b)}(t + \tau)$ (vertical axis).
8.4 AR(1) Simulation Study

This simulation study is intended to explain the parametric bootstrap approach and ensure that we fully understand how it transforms forecasts from the transformed $v$-scale back to the raw SST $z$-scale unbiasedly. We use the closed-form expressions for mean and variance of forecasts from an AR(1) to benchmark the performance of our approach.

We fitted an AR(1) process to the $v$-scale monthly-centred data, to use as the basis of our simulation study:

$$v_c(t) - \mu = \phi(v_c(t - 1) - \mu) + \varepsilon(t),$$  \hspace{1cm} (8.37)

where $\varepsilon(t)$ is a Gaussian process with mean zero and variance $\sigma^2$. The fitted values were $\hat{\mu} = -0.0955$, $\hat{\phi} = 0.929$, and $\hat{\sigma}^2 = 0.1037$. This initial calibration ensures that we will be simulating from an AR(1) process whose realisations will have temporal variability that will be on a similar scale to what is seen in the real data.

Our simulation procedure for this study is as follows: For $l = 1, \ldots, 10000$, we carry out the following steps.

1. Based on the AR(1) process defined by Equation 8.37 (with $\mu = \hat{\mu} = 0.0955$, $\phi = \hat{\phi} = 0.929$, $\sigma^2 = \hat{\sigma}^2 = 0.1037$), generate a time series of length 398 (the length of the time series of the data) and denote it by $\{v_c^{(l)}(t) : t = 1, \ldots, 398\}$.

2. Calculate the corresponding $z$-values:

$$z^{(l)}(t) = \frac{\exp\left(-v_c^{(l)}(t) \times I_v^M - \bar{v}^M\right) - \alpha^M}{\beta^M}; t \in D_t^M,$$  \hspace{1cm} (8.38)

where $t$ is in month $M \in \{\text{Jan}, \ldots, \text{Dec}\}$. This formula is derived from Equations 8.19 and 8.30 in Chapter 8. We treat $I_v^M$, $\bar{v}^M$, $\alpha^M$, and $\beta^M$ as known constants for this simulation. In Chapter 7, $\alpha^M$ and $\beta^M$ are defined in Table 7.1 and $I_v^M$ is the intercept estimates given in Table 7.3. Table 8.1 in Chapter 8 gives the values of $\bar{v}^M$.

3. For $t = 1, \ldots, 398 - \tau$, forecast $\tau$ months ahead based on $v_c^{(l)}(t)$ to obtain $\hat{v}_c^{(l)}(t + \tau)$, for $\tau = 1, \ldots, 7$. Here, for the AR(1), the optimal forecast is $\hat{v}_c^{(l)}(t + \tau) = \phi^\tau v_c^{(l)}$. 


4. Calculate the re-scaled quantity,

$$\hat{v}^{(l)}(t + \tau) \equiv \overline{v}^{M_{t + \tau}} + \hat{v}_c^{(l)}(t + \tau) \times I_v^{M_{t + \tau}}, \quad \text{(8.39)}$$

and then also calculate $\exp(-\hat{v}^{(l)}(t + \tau))$, for each simulated and forecast value, $\hat{v}_c^{(l)}(t + \tau)$. Recall $M_{t + \tau}$ denotes the month $\tau$ months after $M$.

From the law of large numbers we know the following should be true:

$$\frac{1}{L} \sum_{l=1}^{L} g \left( \overline{v}_c^{(l)}(t + \tau) \right) \rightarrow \mathbb{E}\left[ g \left( \overline{v}_c(t + \tau) \right) \right], \quad \text{(8.40)}$$

as $L \to \infty$ where $g$ is a continuous function. For $g(x) = x$, the left-hand side of Equation 8.40 converges to

$$\mathbb{E}\left[ \overline{v}_c(t + \tau) \right] = \mathbb{E}\left[ \mathbb{E}\left[ \overline{v}_c(t + \tau) | v_a(t) \right] \right] = \mathbb{E}\left[ \overline{v}_c(t + \tau) \right] = \hat{\mu} = -0.0955. \quad \text{(8.41)}$$

Also, as $L \to \infty$,

$$\frac{1}{L} \sum_{l=1}^{L} \left( \overline{v}_c^{(l)}(t + \tau) \right)^2 \rightarrow \mathbb{E}\left[ \left( \overline{v}_c(t + \tau) \right)^2 \right], \quad \text{(8.42)}$$

and

$$\left( \frac{1}{L} \sum_{l=1}^{L} \overline{v}_c^{(l)}(t + \tau) \right)^2 \rightarrow \left( \mathbb{E}(\overline{v}_c(t + \tau)) \right)^2. \quad \text{(8.43)}$$

Therefore,

$$\frac{1}{L} \sum_{l=1}^{L} \left( \overline{v}_c^{(l)}(t + \tau) \right)^2 - \left( \frac{1}{L} \sum_{l=1}^{L} \overline{v}_c^{(l)}(t + \tau) \right)^2 \rightarrow \var\left( \overline{v}_c(t + \tau) \right), \quad \text{(8.44)}$$

and we have calculated elsewhere (Equation 8.57) that

$$\var\left( \overline{v}_c(t + \tau) \right) = \frac{\hat{\phi}^2 \tilde{\sigma}^2}{1 - \hat{\phi}^2}. \quad \text{(8.45)}$$

We determined (not shown here) that for each $\tau$, the Monte-Carlo-based predictive means and variances are within Monte Carlo sampling error of the theoretical values given by the right-hand sides of Equations 8.41 and 8.45, which gives us confidence in the choice of $L = 10000$ and in the correctness of the R script we wrote to carry out the preceding steps.
As established in Equation 8.33, we want to find \( \{k_1(M,\tau)\} \) and \( \{k_2(M,\tau)\} \) such that \( \tilde{\tau}(t + \tau) \) is unbiased, i.e., \( \mathbb{E}[\tilde{\tau}(t + \tau)] = \mathbb{E}[\tau(t + \tau)] \). In what follows, we shall calculate these values theoretically and through a parametric bootstrap and compare them.

### 8.4.1 Theoretical Calculation of \( k_1 \) and \( k_2 \)

We need to solve for \( k_1 \) and \( k_2 \) in

\[
\mathbb{E}[k_1 + k_2 \exp(-\tilde{\tau}(t + \tau))] = \mathbb{E}[\tau(t + \tau)] ; \tag{8.46}
\]

that is,

\[
k_1 + k_2 \mathbb{E}[\exp(-\tilde{\tau}(t + \tau))] = \mathbb{E}\left[\frac{\exp(-\tau(t + \tau)) - \alpha^{M_{t+r}}}{\beta^{M_{t+r}}}\right]
\equiv d_1 + d_2 \mathbb{E}[\exp(-\tau(t + \tau))] , \tag{8.47}
\]

where \( d_1 = -\alpha^{M_{t+r}}/\beta^{M_{t+r}} \) and \( d_2 = 1/\beta^{M_{t+r}} \).

Start with the right-hand side of Equation 8.47 and recall that \( \tau_c(t) \) is an AR(1) process,

\[
\tau_c(t) = \mu + \phi(\tau_c(t - 1) - \mu) + \varepsilon(t) ; t = 1, \ldots, T, \tag{8.48}
\]

where \( \varepsilon(t) \) is a Gaussian random variable with mean zero and variance \( \sigma^2 \).

Then \( \mathbb{E}[\tau_c(t)] = \mu \), for all \( t = 1, \ldots, T \).

Let the stationary variance, \( \text{var}(\tau_c(t)) = \omega^2 \), for all \( t \). Then for \( \phi^2 < 1 \), the AR(1) process is stationary and hence,

\[
\omega^2 = \phi^2 \omega^2 + \sigma^2
\]

\[
\omega^2 = \frac{\sigma^2}{1 - \phi^2}. \tag{8.49}
\]

Also, the forecasts from this AR(1) are:

\[
\tilde{\tau}_c(t + \tau) = \mathbb{E}[\tau_c(t + \tau)|\tau_c(t)] = \mu + \phi^\tau (\tau_c(t) - \mu) . \tag{8.50}
\]

Further, as \( \tau_c(t + \tau) \) is Gaussian distributed with mean \( \mu \) and variance \( \omega^2 = \frac{\sigma^2}{1 - \phi^2} \), and

\[
\bar{\tau}(t + \tau) = \bar{\nu}^{M_{t+r}} + \tau_c(t + \tau)I_{\nu}^{M_{t+r}} , \tag{8.51}
\]
so \( \bar{v}(t + \tau) \) is also Gaussian distributed with mean \( \bar{v}^{M+t+\tau} + \mu I_v^{M+t+\tau} \) and variance \( \frac{\sigma^2}{1-\phi^2} (I_v^{M+t+\tau})^2 \), and hence \( \exp(-\bar{v}(t + \tau)) \) is log-Gaussian distributed. This means that

\[
E[\exp(-\bar{v}(t + \tau))] = \exp(-E[\bar{v}(t + \tau)] + \text{var}[\bar{v}(t + \tau)]/2) \\
= \exp \left( -\bar{v}^{M+t+\tau} - \mu I_v^{M+t+\tau} + \left( I_v^{M+t+\tau} \right)^2 \frac{\sigma^2}{2(1-\phi^2)} \right). \tag{8.52}
\]

Now consider the left-hand side of Equation 8.47. We know

\[
\hat{\bar{v}}(t + \tau) \equiv \bar{v}^{M+t+\tau} + \hat{v}_c(t + \tau) \times I_v^{M+t+\tau}, \tag{8.53}
\]

and hence

\[
E \left[ \exp \left( -\hat{\bar{v}}(t + \tau) \right) \right] = \exp(-\bar{v}^{M+t+\tau}) E \left[ \exp \left( -\hat{v}_c(t + \tau) \times I_v^{M+t+\tau} \right) \right]. \tag{8.54}
\]

Recall that \( \hat{v}_c(t + \tau) \) is normally distributed, and hence \( \exp(-I_v^{M+t+\tau} \hat{v}_c(t + \tau)) \) is lognormally distributed. Consequently,

\[
E \left[ \exp \left( -I_v^{M+t+\tau} \hat{v}_c(t + \tau) \right) \right] = \exp(-I_v^{M+t+\tau} E \left[ \hat{v}_c(t + \tau) \right] + \left( I_v^{M+t+\tau} \right)^2 \text{var}[\hat{v}_c(t + \tau)]/2) \tag{8.55}
\]

Hence we calculate

\[
E \left[ \hat{v}_c(t + \tau) \right] = E \left[ E[\bar{v}_c(t + \tau)|\bar{v}_c(t)] \right] = E [\mu + \phi^\tau (\bar{v}_c(t) - \mu)] = \mu. \tag{8.56}
\]

Also, we calculate

\[
\text{var} \left[ \hat{v}_c(t + \tau) \right] = \text{var} \left[ E[\bar{v}_c(t + \tau)|\bar{v}_c(t)] \right] = \text{var} [\mu + \phi^\tau (\bar{v}_c(t) - \mu)] = \phi^{2\tau} \sigma^2 \frac{1}{1-\phi^2}. \tag{8.57}
\]

Thus, Equation 8.54 is

\[
E \left[ \exp(-\hat{\bar{v}}(t + \tau)) \right] = \exp(-\bar{v}^{M+t+\tau}) \times \exp \left( -\mu I_v^{M+t+\tau} + \left( I_v^{M+t+\tau} \right)^2 \frac{\phi^{2\tau} \sigma^2}{2(1-\phi^2)} \right). \tag{8.58}
\]
Finally, substituting Equations 8.58 and 8.52 into Equation 8.47 gives

\[ k_1 + k_2 \exp \left( -\bar{v}^{M_t+\tau} - \mu I_v^{M_t+\tau} + \left(I_v^{M_t+\tau}\right)^2 \frac{\phi^{2\tau} \sigma^2}{2(1 - \phi^2)} \right) \]

\[ = d_1 + d_2 \exp \left( -\bar{v}^{M_t+\tau} - \mu I_v^{M_t+\tau} + \left(I_v^{M_t+\tau}\right)^2 \frac{\sigma^2}{2(1 - \phi^2)} \right). \]

(8.59)

We put

\[ k_1 = d_1 \equiv -\alpha^{M_t+\tau}/\beta^{M_t+\tau}, \]

(8.60)

so that

\[ k_2 \exp \left( -\mu I_v^{M_t+\tau} + \left(I_v^{M_t+\tau}\right)^2 \frac{\phi^{2\tau} \sigma^2}{2(1 - \phi^2)} \right) = d_2 \exp \left( -\mu I_v^{M_t+\tau} + \left(I_v^{M_t+\tau}\right)^2 \frac{\sigma^2}{2(1 - \phi^2)} \right), \]

(8.61)

where \( d_2 = \frac{1}{\beta^{M_t+\tau}} \). Hence, solving for \( k_2 \) yields

\[ k_2 = \frac{1}{\beta^{M_t+\tau}} \exp \left( \left(I_v^{M_t+\tau}\right)^2 \frac{\sigma^2(1 - \phi^{2\tau})}{2(1 - \phi^2)} \right). \]

(8.62)

In conclusion, \( k_1 \) given by Equation 8.60 and \( k_2 \) given by Equation 8.62 yields an unbiased predictor,

\[ \tilde{z}(t + \tau) \equiv k_1 + k_2 \exp \left( -\bar{v}(t + \tau) \right) ; \ t = 1, \ldots, T - \tau, \]

(8.63)

of \( \bar{z}(t + \tau) \). Importantly, \( k_1 \) and \( k_2 \) are functions of \( M_t+\tau \) and \( \tau \) (equivalently of \( M_t \) and \( \tau \)) and are not functions of \( \mu \).

To further investigate the predictor \( \tilde{z}(t + \tau) \), we define the difference,

\[ d_L(t, \tau) \equiv \frac{1}{L} \sum_{l=1}^{L} z^{(l)}(t + \tau) - \frac{1}{L} \sum_{l=1}^{L} \bar{z}^{(l)}(t + \tau) ; \ t \in D^M_t, \ \tau \in \{1, \ldots, 7\}. \]

(8.64)

To demonstrate that the predictions from \( \tilde{z}(t + \tau) \) are unbiased, we made histograms of the difference given by Equation 8.64 by varying \( t = 1, \ldots, 398 - \tau \), for each \( \tau = 1, \ldots, 7 \); see Figure 8.18. The histograms are centred around zero, implying that the predictions, \( \tilde{z}(t + \tau) \), are unbiased. Note also, that as \( \tau \) increases the histograms exhibit longer tails.

Further, we broke down the histograms in Figure 8.18 by month, \( M \), for \( \tau = 1, \tau = 3, \) and \( \tau = 7 \); see Figure 8.19. Note that there are only around 33 values in
each of these histograms. For $\tau = 1$ and most months (except August), there is a strong central peak to the histograms. This is less pronounced for June and July, while in August the histogram exhibits slight skewness. For $\tau = 3$, the histograms are all still centred around zero; however, for April through November, the central peak appears less pronounced. For $\tau = 7$, the histograms are all still centred around zero, with similar general behaviour that we saw for $\tau = 3$. For each month, the histograms exhibit longer tails as $\tau$ increases.

Figure 8.20 gives the 5%, 25%, 50%, 75% and 95% quantiles of $d_L(t, \tau)$ for all months $M$, $\tau = 1$, $\tau = 3$, and $\tau = 7$. The median values ($q_{.5}$) are close to zero, which reflects that the histograms are centred around zero. For each month, the $q_{.05}$ values for $\tau = 7$ are less than those for $\tau = 1$ and the $q_{.95}$ values for $\tau = 7$ are greater than
those for \( \tau = 1 \). This matches our observation that the histograms exhibit longer tails as \( \tau \) increases.

Figure 8.19: Difference of \( \tilde{z} \) and \( \hat{z} \) given by Equation 8.64, for \( \tau = 1 \) (left panels), \( \tau = 3 \) (centre panels) and \( \tau = 7 \) (right panels) by month, \( M \in \{\text{Jan}, \ldots, \text{Dec}\} \).
Figure 8.20: Quantiles ($q_{0.05}, q_{0.25}, q_{0.5}, q_{0.75}, q_{0.95}$) of $d_L(t, \tau)$ for all months $M$, $\tau = 1$ (top panel), $\tau = 3$ (centre panel), and $\tau = 7$ (bottom panel).
8.4.2 Simulation Study Parametric Bootstrap for $k_1$ and $k_2$

Similar to Section 8.3.3 our procedure is as follows: For $b = 1, \ldots, 1600$, where 1600 is the bootstrap sample size,

1. Based on the AR(1) process defined by Equation 8.37 with $\mu = 0.0955$, $\phi = 0.929$, and $\sigma^2 = 0.1037$, generate a time series of length 398, and denote this time series as $\{v_c^{(b)}(t) : t = 1, \ldots, 398\}$.

2. Calculate the corresponding values, $\bar{z}^{(b)}(t)$, using Equation 8.30.

3. Forecast for $t = 1, \ldots, 398 - \tau$, from $\bar{v}_c^{(b)}(t)$ based on the AR(1) process in Equation 8.37 to obtain $\hat{v}_c^{(b)}(t + \tau)$, for $\tau = 1, \ldots, 7$.

4. Calculate $\hat{v}^{(b)}(t + \tau)$ using

$$\hat{v}^{(b)}(t + \tau) = \bar{v}^{M_t+\tau} + \bar{v}_c^{(b)}(t + \tau) \times I_{v}^{M_t+\tau} ; t \in D_t^M, \tau = 1, \ldots, 7 . \quad (8.65)$$

Then calculate $\exp(-\hat{v}^{(b)}(t + \tau)) ; t \in D_t^M, \tau = 1, \ldots, 7$.

We set $x = \frac{1}{1600} \sum_{b=1}^{1600} \exp(-\hat{v}^{(b)}(t + \tau))$ and $y = \frac{1}{1600} \sum_{b=1}^{1600} \bar{z}^{(b)}(t + \tau)$. We use OLS and WTS (weighted Theil-Sen from Chapter 4) in a simple linear regression to solve for $k_1$ and $k_2$, for each $M$ and $\tau$. Figure 8.21 provides examples of the OLS and WTS fitted lines compared to THL, the line obtained using the theoretical values.

We considered $M = \text{Feb}$ and $M = \text{Oct}$; and $\tau = 1$, $\tau = 3$, and $\tau = 7$. As $\tau$ increases, the three lines look progressively more different. When $\tau = 7$, in both October and February, the slopes of the WTS lines look quite different to the OLS and theoretical lines, though at a nominal 5% level the WTS slopes are not significantly different from the OLS slopes or the THL slopes.
Figure 8.21: Comparison of the OLS (red line), WTS (blue line) and THL (black line) lines for $M$ equal to February (left panels) and October (right panels); $\tau = 1$ (top panels), $\tau = 3$ (middle panels) and $\tau = 7$ (bottom panels);

$x = \frac{1}{1600} \sum_{b=1}^{1600} \exp(-\hat{v}^{(b)}(t + \tau))$ and $y = \frac{1}{119} \sum_{b=1}^{1600} \hat{z}^{(b)}(t + \tau)$. 
8.4.3 Comparison of Theoretical and Parametric Bootstrap $k_1$ and $k_2$ values

We compare the parametric bootstrap values (OLS and WTS) for $k_1$ and $k_2$, to the theoretical values through line plots of the estimate versus $\tau$ for each month $M$; see Figures 8.22 (for $k_1$) and 8.23 (for $k_2$). Within each month, the three methods have similar dependence on $\tau$, and generally the OLS values for $k_1$ and $k_2$ are closer than the WTS values to the theoretical values. The “curves” shown for months within a season are similar to each other (for both $k_1$ and $k_2$); for example, the austral autumn (March, April, and May) curves look similar. Further, for August, September, October, and November, the WTS $k_1$ and $k_2$ values are closer to zero than the corresponding theoretical and OLS values. However, for December, January, February, March, and April the WTS $k_1$ and $k_2$ values are further from zero than the corresponding theoretical and OLS values. While these trends are interesting to observe, we are more interested in the relative performance of the $\tilde{y}(t + \tau)$, which we shall consider in the next section.
Figure 8.22: Comparison of the OLS and WTS values for $k_1$ and the corresponding theoretical values, for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ and $\tau \in \{1, \ldots, 7\}$.
Figure 8.23: Comparison of the OLS and WTS values for $k_2$ and the corresponding theoretical values, for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ and $\tau \in \{1, \ldots, 7\}$. 
8.4.4 Forecasting using Theoretical and Parametric Bootstrap $k_1$ and $k_2$ values

Using the WTS and OLS values for $k_1$ and $k_2$, and our simulation results we have the forecast,

$$\tilde{z}(t + \tau) = k_1 + k_2 \frac{1}{L} \sum_{l=1}^{10000} \exp \left( -\tilde{v}(t + \tau) \right),$$

(8.66)

and we denote the respective forecasts as $\tilde{z}_{OLS}(t + \tau)$ and $\tilde{z}_{WTS}(t + \tau)$. We also calculated $\tilde{z}_{THL}(t + \tau)$ using the theoretical calculations of $k_1$ and $k_2$, to serve as the “gold standard”.

These forecasts of $\tilde{z}(t + \tau)$ are compared via boxplots in Figure 8.24, where for each $t + \tau$ the observed $\tilde{z}(t + \tau)$ (which is available to us from the simulation) has been subtracted from the forecast. We call this difference the forecast error. The forecast errors for THL are mostly centred around zero, and thus the forecasts are unbiased for most $M$ and $\tau$, as expected. Also, as $\tau$ increases the interquartile range (IQR) for THL generally increases for each month. The median forecast errors for THL are also closer to zero than the median forecast errors for OLS and WTS.

The median, first, and third quantile forecast errors for WTS are also generally closer to zero (and thus closer to being unbiased) than the respective OLS forecast errors. For some months, the IQR of the WTS forecast errors is also noticeably smaller than the IQR of the OLS forecast errors; for example, $M = \text{Oct}, \tau = 2$ and $M = \text{Jun}, \tau = 4$. This suggests the WTS values for $k_1$ and $k_2$ are producing better forecasts than the OLS values.

The mean squared prediction error (MSPE) for each forecast value is defined as:

$$MSPE(M, \tau) \equiv \frac{1}{|D^M_t|} \sum_{t \in D^M_t} \left( \tilde{z}(t + \tau) - \tilde{z}(t + \tau) \right)^2 ; t \in D^M_t, \tau \in \{1, \ldots, 7\},$$

(8.67)

where $|D^M_t|$ is the number of elements in $D^M_t$.

We use the MSPE to compare the OLS, WTS, and THL forecasts through line plots of MSPE versus $\tau$ for each month, $M$, in Figure 8.25. The MSPE for each month when $\tau = 1$ is very similar for THL and the two bootstrap methods OLS and WTS. As $\tau$ increases, there is generally more difference in the MSPE between the three methods, particularly for January through to July. Also, for the majority of months and $\tau$, the MSPE for OLS is the largest. The MSPE for WTS is almost always closer to the MSPE for THL than to the MSPE for OLS.
Figure 8.24: Boxplots comparing of $\tilde{z}_{OLS}(t+\tau) - \bar{z}(t+\tau)$, $\tilde{z}_{WTS}(t+\tau) - \bar{z}(t+\tau)$, and $\tilde{z}_{THL}(t+\tau) - \bar{z}(t+\tau)$, for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ and $\tau \in \{1, \ldots, 7\}$.

We also use the MSPE to calculate a relative skill between the different forecasts:

$$RS_{A/B}(M, \tau) = \frac{MSPE_A(M, \tau)}{MSPE_B(M, \tau)} \ ; t \in D^M_t, \tau \in \{1, \ldots, 7\},$$

where A and B are two methods chosen from OLS, WTS, and THL. A relative skill greater than one implies that method B gives better forecasts. Figure 8.26 shows the relative skills of OLS to THL, WTS to THL, and OLS to WTS for each $M$ ranging across the months and $\tau = 1, \ldots, 7$. As expected, the relative skill values are mostly greater than one for OLS compared to THL and for WTS compared to THL. Further, for $\tau = 1$ the RS of OLS to WTS is centred around 1, indicating that the two methods have similar performance. However, for $\tau > 1$, the comparison between OLS and WTS shows without question that $\tilde{z}_{WTS}(t+\tau)$ is a better forecast.
Figure 8.25: Comparison of MSPE as defined in Equation 8.67 for $\tilde{z}_{OLS}(t + \tau)$, $\tilde{z}_{WTS}(t + \tau)$, and $\tilde{z}_{THL}(t + \tau)$, for all $M \in \{\text{Jan, \ldots, Dec}\}$ and $\tau \in \{1, \ldots, 7\}$.

This simulation study shows that for an AR(1) process a statistical computational approach (parametric bootstrapping) can be used to obtain values of $k_1$ and $k_2$ that unbiasedly transform forecasts from the transformed scale back to the original SST scale. Further, we have shown that using WTS to obtain the values of $k_1$ and $k_2$ results in better forecasts than using OLS.
Figure 8.26: Comparison of relative skill (RS) as defined in Equation 8.68 for OLS/THL, WTS/THL, and OLS/WTS, for all $M \in \{\text{Jan,\ldots,Dec}\}$ and $\tau \in \{1,\ldots,7\}$. 
Chapter 9

Forecasting SSTs

In this chapter we use the forecasting equations derived in Chapter 8 to forecast SST data and then compare the results from the original SST scale and the transformed scale. In Section 9.1 we undertake in-sample forecasting and use the equations to forecast “future” values on the original SST scale and on the transformed scale. While the possible outliers we identified in Sections 8.2.1–8.2.3 were excluded from the AR fitting, for the purpose of forecasting we consider all the observations (i.e. $D_t$) to undertake a complete comparison of the two forecasting methods. In Section 9.2 we use the same approaches to obtain out-of-sample forecasts, using the data from January 2015–December 2017. In both sections, based on the forecast errors, we compare the two forecasts.

9.1 In-Sample Forecasting

In-sample forecasts are when the model is trained on the whole set of data observed at $\{1, \ldots, T\}$ and then used to predict a “future”, $\tau$-step-ahead value based on “past” and “present” values, where the “present” ranges over $\{1, \ldots, T - \tau\}$. In other words, the data used to train the model is also used in forecasting. Since our goal is to compare forecasts, it is reasonable to make the comparison this way, although, in Section 9.2 we compare the two using out-of-sample forecasts.

9.1.1 In-Sample Forecasting: Original SST Data

We use Equation 8.26 from Section 8.3.1 to obtain the forecasts $\hat{z}(t+\tau)$, for $t \in D_t^M$ and $\tau = 1, \ldots, 7$. 

We define the forecast error of \( \{ \tilde{z}(t + \tau) \} \) as;

\[
e(t + \tau) \equiv \tilde{z}(t + \tau) - z(t + \tau) ; \quad t \in D_t, \tau = 1, \ldots, 7 ,
\]

(9.1)

where for in-sample forecasting \( z(t + \tau) \) is in the original dataset. However, since only data up to and including time \( t \) is used in the forecast, it represents a future datum. The forecast errors for the in-sample forecasts are compared via boxplots in Figure 9.1. All of the boxes defined by the lower and upper quartiles contain zero, indicating that overall the months and lags produce unbiased forecasts. For most months there is more mass in the boxplots less than zero, so the forecasts are slightly skewed, which means the model tends to slightly under predict the observed values. Unexpectedly, for some months (particularly August and September) the IQR is narrower for \( \tau = 7 \) than some of the smaller values of \( \tau \) for that month. This implies that forecast performance is a function of month \( M \) and lag \( \tau \).

The quantile-quantile plots (Q-Q plots) in Figures 9.2, 9.3 and 9.4 compare the forecast error to a standard Gaussian distribution (mean 0 and variance 1) for \( \tau = 1 \) and \( \tau = 7 \), respectively. In Figure 9.2, where \( \tau = 1 \), in most months the points follow a straight line very closely, indicating the forecast errors within each month follow a Gaussian distribution. In August, September, and November the Q-Q plots suggest the forecast errors have a heavier tail than a Gaussian distribution. That is, there are more extreme values in the forecast errors than would be expected if they followed a Gaussian distribution.

In Figure 9.3, where \( \tau = 3 \), the Q-Q plots suggest that the forecast errors for April still approximately follow a standard Gaussian distribution. The plots also suggest that the forecast errors for January, February, March, June, September, and December are all slightly skewed right of a standard Gaussian distribution and May, July, and November are slightly skewed left of a standard Gaussian distribution. Finally, for August and October the Q-Q plots suggest the forecast errors have a heavier tail than a Gaussian distribution.

In Figure 9.4, where \( \tau = 7 \), there are some months (July and September) where the Q-Q plots indicate the forecast errors still follow a Gaussian distribution. However, for this larger lag in more months the Q-Q plots indicate the forecast errors do not strictly follow a Gaussian distribution. The Q-Q plots indicate the forecast errors for February, March, October, and November are slightly right skewed (the model tends to over predict the observations), and August they are slightly left skewed (the model tends to under predict the observations). Finally, in January, April,
Figure 9.1: Shown are boxplots of forecast errors, \( \{ e(\hat{z}(t+\tau)) \} \), given by Equation 9.1, broken down by month \( M \) and lag \( \tau = 1, \ldots, 7 \).

May, June, and December the forecast errors have a distribution that has a heavier tail than the standard Gaussian distribution.

We define the bias of \( \{ \hat{z}(t+\tau) \} \) as;

\[
b\left(\hat{z}(t+\tau); M, \tau \right) \equiv \frac{1}{|D^M_t|} \sum_{t \in D^M_t} e(\hat{z}(t+\tau)) ;
\]

where recall that \( e(\hat{z}(t+\tau)) \) is the forecast error defined in Equation 9.1. Figure 9.5 shows plots of the bias broken down by month \( M \) and lag \( \tau \). The bias values are all very small, within \((-0.03^\circ C, 0.015^\circ C)\). There are some unexpected patterns,
Figure 9.2: Shown are Q-Q plots of forecast errors, $\left\{ \epsilon \left( \hat{z}(t+\tau) \right) \right\}$, given by Equation 9.1, broken down by month $M$ and lag $\tau = 1$. The forecast errors are shown on the vertical axis, and standardized normal values are shown on the horizontal axis.

for example January, February, and March the bias values decrease as $\tau$ increases; October, November, and December all have a U-shape, where there are a couple of lags where the bias values on either side are higher.
Figure 9.3: Shown are Q-Q plots of forecast errors, \( \{ e\left(\tilde{z}(t + \tau)\right) \} \), given by Equation 9.1, broken down by month \( M \) and lag \( \tau = 3 \). The forecast errors are shown on the vertical axis, and standardized normal values are shown on the horizontal axis.
Figure 9.4: Shown are Q-Q plots of forecast errors, \( \{ e(\tilde{x}(t + \tau)) \} \), given by Equation 9.1, broken down by month \( M \) and lag \( \tau = 7 \). The forecast errors are shown on the vertical axis, and standardized normal values are shown on the horizontal axis.
Figure 9.5: Shown are plots of bias, \( b \left( \mathcal{Z}(t + \tau); M, \tau \right) \) given by Equation 9.2 (vertical axis) versus \( \tau = 1, \ldots, 7 \) (horizontal axis), broken down by month \( M \), units on the vertical axis are degrees Celsius.
9.1.2 In-Sample Forecasting: Transformed SST Data

We use the parametric bootstrap results and the back-transform derived in Chapter 8 to obtain forecasts from the transform scale. In Equation 8.33, we use the WTS-fitted values of $k_1$ and $k_2$ given in Tables 8.3 and 8.4 to obtain the forecasts $\tilde{z}(t+\tau)$, for $t \in D_t^M$ and $\tau = 1, \ldots, 7$.

We define a forecast error of $\{\tilde{z}(t+\tau)\}$ as

$$e(\tilde{z}(t+\tau)) \equiv \tilde{z}(t+\tau) - z(t+\tau) ; t \in D_t, \tau = 1, \ldots, 7,$$

which are shown in boxplots in Figure 9.6. The boxplots are mostly centred around zero, indicating that the forecasts are mostly unbiased. The magnitude of the square of the bias as a percentage of mean squared prediction error will be analysed later, in Section 9.1.3. Recall that Figure 9.1 showed boxplots of the forecast errors for $\{\tilde{z}(t+\tau)\}$. The upper limit of the vertical axis in that figure is 2, while the upper limit of the vertical axis in Figure 9.6 is 3.5. The boxplots in Figure 9.6 indicate larger forecast errors and have more outliers than the boxplots in Figure 9.1. For each given month, the general trend in the size of the boxes in the boxplots across $\tau \in \{1, 2, \ldots, 7\}$ in Figure 9.6 is very similar to that in Figure 9.1. For July, August, and September the boxes increase in size as $\tau$ increases from 1 to 4 or 5, and then decrease again for larger $\tau$. For October and November, the boxes have similar size across all lags. Particularly in December, January, February, and March there are some large forecast error values for large $\tau$ which may effect bias calculations.

Figures 9.7, 9.8, and 9.9, give Q-Q plots that compare the forecast error to a standard Gaussian distribution for $\tau = 1, \tau = 3$, and $\tau = 7$, respectively. In Figure 9.7 ($\tau = 1$), for January, February, and July the points follow a straight line very closely, indicating the forecast errors within each month follow a standard Gaussian distribution. For most of the months (June, August, September, October, November, and December) the forecast errors have heavier tails than a standard Gaussian distribution. For March, April, and May, the Q-Q plot indicates that the forecast errors have a slight right skew (the model tends to over predict the observations). In Figure 9.2 we similarly saw that the Q-Q plots for August, September, and November suggested the forecast errors $\hat{z}(t+\tau)$ for have a heavier tail than a Gaussian distribution.

In Figure 9.8 ($\tau = 3$), the Q-Q plots for May, June, and December suggest the forecast errors of $\tilde{z}(t+\tau)$ for those months approximately follow a standard Gaussian
### Figure 9.6: Shown are boxplots of forecast errors, $\{e(\tilde{z}(t+\tau))\}$ given by Equation 9.3, broken down by month $M$ and lag $\tau = 1, \ldots, 7$.

The Q-Q plot for November indicates the forecast errors of $\tilde{z}(t+\tau)$ are slightly skewed left of a Gaussian distribution which is consistent with what we saw for $\hat{z}(t+\tau)$ in Figure 9.3. In Figure 9.8 the Q-Q plots for March and April suggest the forecast errors of $\tilde{z}(t+\tau)$ are slightly skewed right of a Gaussian distribution, from Figure 9.3 we concluded that the forecast errors for $\hat{z}(t+\tau)$ in March were also skewed right. Finally, in Figure 9.8 the Q-Q plots for January, February, July, August, September, and October suggest that in each month the forecast errors for $\tilde{z}(t+\tau)$ have a heavier tail than a standard Gaussian distribution. For August and October this is consistent with the observed behaviour of the forecast errors of $\tilde{z}(t+\tau)$ in Figure 9.3.
In Figure 9.9 ($\tau = 7$), there are no months where visual inspection of the Q-Q plots indicate the forecast errors follow a standard Gaussian distribution. This is different to Figure 9.4, where for $\tau = 7$ and $\tilde{z}(t+\tau)$, we found there are some months (July and September) where the Q-Q plots indicate the forecast errors still follow a Gaussian distribution. In Figure 9.9 the Q-Q plots indicate that the forecast errors of $\tilde{z}(t+\tau)$ for January, February, October, and November are slightly right skewed, and for July and August they are slightly left skewed. In Figure 9.4 we also concluded the forecast errors of $\tilde{z}(t+\tau)$ for February, October, and November were slightly right skewed and August they were slightly left skewed. Finally, in Figure 9.9 for March, April, May, June, September, and December the forecast errors of $\tilde{z}(t+\tau)$ have a distribution that has a heavier tail than the standard Gaussian distribution. For April, May, June, and December this is consistent with what we concluded for the forecast errors of $\tilde{z}(t+\tau)$.

We define the bias of $\{\tilde{z}(t+\tau)\}$ as

$$b \left( \tilde{z}(t+\tau); M, \tau \right) = \frac{1}{|D^M|} \sum_{t \in D^M} e \left( \tilde{z}(t+\tau) \right);$$

$$t \in D^M, \tau = 1, \ldots, 7, M \in \{\text{Jan, \ldots, Dec}\}, \quad (9.4)$$

where $e \left( \tilde{z}(t+\tau) \right)$ is the forecast error defined in Equation 9.3. Figure 9.10 shows plots of the bias broken down by month $M$ and lag $\tau$. We note that all of the bias values are positive. We noticed in the boxplots of forecast error that particularly in December, January, February, and March there are some large forecast error values for large $\tau$ which may be affecting the respective bias calculations.

Despite our efforts to derive an unbiased forecast, there is clearly still a small bias in the forecasts. Further, as seen in the original scale forecasts, there is a seasonal pattern to the bias values. This pattern will be investigated fully in the next subsection, where we compare the forecasts $\{\hat{z}(t+\tau)\}$ with $\{\tilde{z}(t+\tau)\}$ through bias, variance, and mean squared prediction error.
Figure 9.7: Shown are Q-Q plots of forecast errors, \( \left\{ e \left( \tilde{z}(t+\tau) \right) \right\} \), given by Equation 9.3, broken down by month \( M \) and lag \( \tau = 1 \).
Figure 9.8: Shown are Q-Q plots of forecast errors, \( \{ e(\bar{z}(t+\tau)) \} \), given by Equation 9.3, broken down by month \( M \) and lag \( \tau = 3 \).
Figure 9.9: Shown are Q-Q plots of forecast errors, \( \left\{ e \left( \bar{z}(t + \tau) \right) \right\} \), given by Equation 9.3, broken down by month \( M \) and lag \( \tau = 7 \).
Figure 9.10: Shown are plots of bias, $b\left(\tilde{z}(t + \tau)\right)$, given by Equation 9.4 (vertical axis) versus $\tau = 1, \ldots, 7$ (horizontal axis), broken down by month $M$, units on the vertical axis are degrees Celsius.
9.1.3 Comparison of In-sample Forecasts

In this section we directly compare the forecasts \( \{ \tilde{z}(t + \tau) \} \) and \( \{ \bar{z}(t + \tau) \} \) through their forecast error, its mean (bias), its variance, and the mean squared prediction error. The errors and bias are defined in Section 9.1.1 and 9.1.2. We now give formula for MSPE and variance of the errors.

The MSPE for the forecast \( \tilde{z}(t + \tau) \) from Equation 8.26 is defined as:

\[
\text{MSPE} \left( \tilde{z}; M, \tau \right) \equiv \frac{1}{|D_t^M|} \sum_{t \in D_t^M} \left( \tilde{z}(t + \tau) - \bar{z}(t + \tau) \right)^2 ;
\]

\( t \in D_t^M, M \in \{\text{Jan, \ldots, Dec}\}, \tau = 1, \ldots, 7, \quad (9.5) \)

where \(|D_t^M|\) is the number of elements in \( D_t^M \), and recall that \( \bar{z}(t + \tau) \) is the value at \( t + \tau \) from the original SST dataset. Likewise, we define the MSPE using the forecasted values \( \tilde{z}(t + \tau) \) from Equation 8.33 as:

\[
\text{MSPE} \left( \tilde{z}; M, \tau \right) \equiv \frac{1}{|D_t^M|} \sum_{t \in D_t^M} \left( \tilde{z}(t + \tau) - \bar{z}(t + \tau) \right)^2 ;
\]

\( t \in D_t^M, M \in \{\text{Jan, \ldots, Dec}\}, \tau = 1, \ldots, 7. \quad (9.6) \)

We define the variance of \( \{ e \left( \tilde{z}(t + \tau) \right) \} \) as:

\[
\text{var} \left( e(\tilde{z}(t + \tau)); M, \tau \right) = \frac{1}{|D_t^M| - 1} \sum_{t \in D_t^M} \left( e \left( \tilde{z}(t + \tau) \right) - \frac{1}{|D_t^M|} \sum_{t \in D_t^M} e \left( \tilde{z}(t + \tau) \right) \right)^2 ;
\]

\( t \in D_t^M, \tau = 1, \ldots, 7, M \in \{\text{Jan, \ldots, Dec}\}, \quad (9.7) \)

and the variance of \( \{ e \left( \tilde{z}(t + \tau) \right) \} \) as:

\[
\text{var} \left( e(\tilde{z}(t + \tau)); M, \tau \right) = \frac{1}{|D_t^M| - 1} \sum_{t \in D_t^M} \left( e \left( \tilde{z}(t + \tau) \right) - \frac{1}{|D_t^M|} \sum_{t \in D_t^M} e \left( \tilde{z}(t + \tau) \right) \right)^2 ;
\]

\( t \in D_t^M, \tau = 1, \ldots, 7, M \in \{\text{Jan, \ldots, Dec}\}. \quad (9.8) \)

We consider first a comparison of the respective errors. In Figures 9.11–9.13 we directly compare the forecast errors \( \{ e \left( \tilde{z}(t + \tau) \right) \} \) from Equation 9.1 to the forecasts \( \{ e \left( \tilde{z}(t + \tau) \right) \} \) from Equation 9.3, broken down by month \( M \), and for \( \tau = 1, \ldots, 7 \).
\( \tau = 3 \), and \( \tau = 7 \), respectively. In each figure the red lines represent the bias of each dimension for that month and \( \tau \). As previously mentioned in Sections 9.1.1 and 9.1.2, we note how close to zero the bias values are.

In Figure 9.11, where \( \tau = 1 \), February, March, and July through November all look similar and the points fall approximately on a straight 45\(^\circ\) line of equal forecast error. This is also the case for July through October in Figure 9.12, and May through December in Figure 9.13. In the other months in these figures the points still mostly fall on or near the 45\(^\circ\) line of equality, there is just more variability in them. Our parametric bootstrap approach is resulting in forecasts with similar error values to those on the original scale.
Figure 9.11: Plots of \( \{ e(\tilde{z}(t + \tau)) \} \) from Equation 9.3 versus \( \{ e(\tilde{z}(t + \tau)) \} \) from Equation 9.1, broken down by month \( M \), where lag \( \tau = 1 \). The dashed black line is where the two forecast errors are equal. The red lines represent the bias (which is the average) of each axis.
Figure 9.12: Plots of \( \{ e \left( \hat{z}(t + \tau) \right) \} \) from Equation 9.3 versus \( \{ e \left( \bar{z}(t + \tau) \right) \} \) from Equation 9.1, broken down by month \( M \), where lag \( \tau = 3 \). The dashed black line is where the two forecast errors are equal. The red lines represent the bias (which is the average) of each axis.
Figure 9.13: Plots of \( \{ e \left( \hat{z}(t + \tau) \right) \} \) from Equation 9.3 versus \( \{ e \left( \tilde{z}(t + \tau) \right) \} \) from Equation 9.1, broken down by month \( M \), where lag \( \tau = 7 \). The dashed line is where the two forecast errors are equal. The red lines represent the bias (which is the average) of each axis.
Next we consider a comparison of the respective biases. The bias values were shown as red lines in Figures 9.11–9.13, however, at that scale it was not easy to compare them to each other. Figure 9.14 directly compares the bias from the two forecast methods through line plots; \( b\left(\tilde{z}(t + \tau); M, \tau\right) \) is represented by the black dotted line, and \( b\left(\hat{z}(t + \tau); M, \tau\right) \) by the solid grey line. For each month the two bias lines have similar patterns, however, \( \tilde{z}(t + \tau) \) has larger (more positive bias) than \( \hat{z}(t + \tau) \), implying that \( \tilde{z}(t + \tau) \) is more likely to over-predict the observed values. The yellow band represents the spring barrier, that is which forecasts are April and May. Figure 9.15 gives an alternate comparison of the bias values, through raster plots of \( b\left(\tilde{z}(t + \tau); M, \tau\right)^2 \) and \( b\left(\hat{z}(t + \tau); M, \tau\right)^2 \). This highlights that the magnitude of the bias is less for \( b\left(\tilde{z}(t + \tau); M, \tau\right) \) than \( b\left(\hat{z}(t + \tau); M, \tau\right) \). Both figures show that the largest \( b\left(\tilde{z}(t + \tau); M, \tau\right) \) values occur forecasting across the spring barrier into June.
Figure 9.14: Shown are comparison plots of bias broken down by month $M$ and lag $\tau = 1, \ldots, 7$. $b \left( \tilde{z}(t + \tau); M, \tau \right)$ is represented by the black dotted line, and $b \left( \hat{z}(t + \tau); M, \tau \right)$ by the solid grey line. The yellow band highlights the lags $\tau$ that forecast into April and May.
Figure 9.15: Raster plots of bias squared by month $M$ and lag $\tau = 1, \ldots, 7$, upper panel: $b\left(\hat{z}(t + \tau); M, \tau\right)^2$ and lower panel: $b\left(\tilde{z}(t + \tau); M, \tau\right)^2$. 
Next, we consider a comparison of the respective MSPEs. The bias values are small and considering the bias of forecasts is important, however we are often most interested in the mean squared prediction error (MSPE) of the forecasts. Bias is a key component of MSPE as it can be written as a function of bias squared and variance.

In Figure 9.16, we compare the MSPEs of forecasts $\tilde{z}(t+\tau)$ and $\hat{z}(t+\tau)$ through line plots of their respective MSPE, versus $\tau$ for each month, $M$. Recall, the formulas for the two MSPE were given in Equations 9.5 and 9.6. For the latter half of the year, July through December, within each month the two forecasts have similar MSPE across the lags. For the earlier half of the year, the MSPE of two forecasts both increase as the lag increases. The spring barrier (April–May as represented by the yellow bands) marks the start of a period of increased MSPE.

To better understand the relative performance of the two methods, we use the MSPE to calculate a relative skill between the two forecasts:

$$RS_{\text{MSPE}}(M, \tau) = \frac{\text{MSPE}(\hat{z}; M, \tau)}{\text{MSPE}(\tilde{z}; M, \tau)}; \quad M \in \{\text{Jan}, \ldots, \text{Dec}\}, \tau = 1, \ldots, 7, \quad (9.9)$$

where a relative skill greater than one implies that the transformed scale gives better forecasts.

Figure 9.17 shows the relative skills of the forecast $\hat{z}$ relative to the forecast $\tilde{z}$, for each $\tau = 1, \ldots, 7$, and for $M$ ranging across the months. Figure 9.18 shows the same values as a raster plot. Both figures show that for July through October the relative skill values are close to one. For November through February there is a period where the relative skill values are noticeably greater than one indicating that the forecasts on the transformed scale have better performance. This period leads up to or coincides with the spring barrier (April–May). For March through June there is a period following the spring barrier where the relative skill values are less than one indicating that the original scale gives better forecasts.
Figure 9.16: Comparison of $MSPE\left(\hat{z}; M, \tau\right)$ (black dotted lines) and $MSPE\left(\tilde{z}; M, \tau\right)$ (dark grey solid lines) for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ (12 panels) and $\tau \in \{1, \ldots, 7\}$ (horizontal axis). The yellow band highlights the lags $\tau$ that forecast into April and May.
Figure 9.17: Relative skill, $RS_{MSPE}$, as defined by Equation 9.9, for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ and $\tau \in \{1, \ldots, 7\}$. The relative skill of 1 is given by the solid red line. The yellow band highlights the lags $\tau$ that forecast into April and May.
Figure 9.18: Relative skill, $RS_{MSPE}$, as defined by Equation 9.9, for all $M \in \{\text{Jan}, \ldots, \text{Dec}\}$ and $\tau \in \{1, \ldots, 7\}$. The colour scale is linear in the log of the relative skill. Blue colours indicate that the forecasts on the original scale have better performance and red colours indicate that the forecasts on the transformed scale have better performance.
It is seen in Sections 9.1.2 and 9.1.3 that some bias in $\tilde{z}(t + \tau)$, while small, remains and that it has structure. Therefore we seek to understand whether that small amount of bias is important, and what impact it is having on the MSPE. We noted earlier in Section 9.1.3 that MSPE can be written as a function of bias squared and variance. For $\hat{z}(t + \tau)$,

$$\text{MSPE} \left( \hat{z}; M, \tau \right) \equiv b \left( \hat{z}(t + \tau); M, \tau \right)^2 + \text{var} \left( e(\tilde{z}(t + \tau); M, \tau) \right) ;$$

$$M \in \{\text{Jan}, \ldots, \text{Dec}\}, \quad \tau = 1, \ldots, 7,$$ (9.10)

that is the MSPE is the bias squared plus the variance of the error. A similar expression to Equation 9.10 can be given for $\tilde{z}(t + \tau)$.

Therefore to carefully examine this relationship, and to understand whether bias is having an effect on MSPE, we consider bias squared as a percentage of mean squared prediction error:

$$\frac{b \left( \tilde{z}(t + \tau); M, \tau \right)^2}{\text{MSPE} \left( \tilde{z}; M, \tau \right)} \times 100$$ (9.11)

for $\hat{z}(t + \tau)$ and similarly for $\tilde{z}(t + \tau)$. Figure 9.19 shows this percentage as line plots. For $\hat{z}(t + \tau)$, bias squared is a very small percentage of the MSPE. For the transformed scale, $\tilde{z}(t + \tau)$, the bias squared is also a small percentage of the MSPE for lags greater than 1 but is larger generally than that of $\hat{z}(t + \tau)$. Overall, while there is some bias and even some pattern to the bias, the forecasting method based on untransforming from the transformed scale does appear to be controlling for the bias. Next, we focus next on the other component of MSPE, the variance of errors.
Figure 9.19: Shown are plots of bias squared as a percentage of MSPE, broken down by month $M$ and lag $\tau = 1, \ldots, 7$; the values for $\tilde{z}(t + \tau)$ are represented by the black dotted line, and $\hat{z}(t + \tau)$ by the solid grey line. The yellow band highlights the lags $\tau$ that forecast into April and May.
Recall Equations 9.7 and 9.8 gave the formula for the variances of the forecast errors. Figure 9.20 compares the variances of the errors from the two forecast methods through line plots; \( \text{var} \left( e(\tilde{z}(t + \tau)); M, \tau \right) \) is represented by the black dotted line, and \( \text{var} \left( e(\hat{z}(t + \tau)); M, \tau \right) \) by the solid grey line. Clearly, the variances of the forecast errors are a function of both month \( M \) and lag \( \tau \). Figure 9.21 gives an alternate comparison of the variances, through raster plots of \( \text{var} \left( e(\tilde{z}(t + \tau)); M, \tau \right) \) and \( \text{var} \left( e(\hat{z}(t + \tau)); M, \tau \right) \). For the later half of the year the two methods appear to have very similar variance. Forecasting into the spring barrier (April and May) represented by the yellow bands in Figure 9.20 appears to be the period when the methods are starting to have difficulty forecasting as there is larger forecast-error variance when forecasting into months that follow the spring barrier. We note the similarity in patterns between the forecast error variance and MSPE plots since the percentage bias is small (Figure 9.19).
Figure 9.20: Shown are plots of forecast-error variance, broken down by month $M$ and lag $\tau = 1, \ldots, 7$; $\text{var} \left( e(\tilde{z}(t+\tau)); M, \tau \right)$ is represented by the black dotted line, and $\text{var} \left( e(\tilde{z}(t+\tau)); M, \tau \right)$ by the solid grey line. The yellow band highlights the lags $\tau$ that forecast into April and May.
Figure 9.21: Raster plots of forecast error variance by month $M$ and lag $\tau = 1, \ldots, 7$, upper panel: $\text{var} \left(e(\tilde{z}(t + \tau)); M, \tau \right)$ and lower panel: $\text{var} \left(e(\tilde{z}(t + \tau)); M, \tau \right)$. 
Figures 9.20 and 9.21 show that the forecast error variances are similar for the two methods for most months so we could use the two variances to define a relative skill between the two forecasts:

$$\text{RSVAR}(M, \tau) = \frac{\text{var}(e(\tilde{z}(t + \tau)); M, \tau)}{\text{var}(e(\tilde{z}(t + \tau)); M, \tau)} ; \quad M \in \{\text{Jan}, \ldots, \text{Dec}\}, \quad \tau = 1, \ldots, 7,$$

(9.12)

where a relative skill greater than one implies the forecast errors on the original scale have larger variance and hence is a more uncertain forecast. Figure 9.22 shows a raster plot of these relative skill values for each $\tau = 1, \ldots, 7$, and for $M$ ranging across the months. This figure allows us to directly compare the variances and consider what the relative forecasting performance of the two methods would be if we could completely remove bias. Figure 9.22 highlights that when forecasting to March and April from December, January, and February the forecasts on the original scale have larger error variance. Also, when forecasting from the first half of the year (across the spring barrier) into July, August, and September, the forecasts on the transformed scale have larger error variance. We note that Figure 9.22 showing $\text{RSVAR}$ has a very similar pattern to Figure 9.18 showing $\text{RSMSPE}$, and hence we rely on the MSPEs to compare the forecasts.

Davies and Cressie (2016) gives a comparison of two similar forecasts in the journal, *Advances in Statistical Climatology, Meteorology and Oceanography* [51]. There are
a few differences from this initial study to what is presented here in the thesis. First, the transform in the paper (described in Section 7.2.3) was a monotonic decreasing transform. For easier interpretability of the geophysical variable in the thesis we made a trivial adjustment to the transform by making it monotonic increasing (described in Section 7.2.4). Second, in the paper, AR(2) models were fitted to both the original SST data and transformed SST data. In the thesis, we used PACF and AIC to determine that AR(3) models were more appropriate. Third, in the paper, we used relative absolute deviation (RAD),

\[
RAD(\hat{z}(t+\tau); M, \tau) \equiv \sqrt{|\hat{z}(t; \tau) - z(t+\tau)|} ;
\]

\[ t \in D^M_t, \tau = 1,\ldots,7, M \in \{\text{Jan},\ldots,\text{Dec}\}\]

as a measure of the predictive skills of the forecasts. In the thesis, due to the presence of some bias, we have chosen instead to use MSPE (Equation 9.10) as a measure of predictive skill. Our conclusions from fitting AR(2) models in the paper were that, when predicting from December, January, and February, the forecasts made on the transformed scale were better. For the months March—September, the two forecasts had comparable forecast, while for October and November, the forecasts on the transform scale did not perform as well as those on the original scale. Similar patterns are seen here in the thesis (in terms of MSPE), predicting from December—February and July—September for relative forecasting skill; see Figure 9.18.

The paper motivated us to undertake a more complete study, given here in the thesis, with a thorough comparison between the two forecasts that further stratifies the variability and forecasting performance of the two forecasts. We also analyse our findings in the context of the (boreal) spring barrier and introduce raster plots that can be used to gain an improved understanding of the influences of the spring barrier on the (relative) forecast performance. In the following section, we use the same approach to compare the out-of-sample forecasting performance of the two forecast methods.
9.2 Out-of-Sample Forecasting

Out-of-sample forecasts are made when the model is trained (i.e., fitted) only on the data observed up to time \( t \), that is at times \( \{1, \ldots, t\} \). Then that fitted model is used to forecast future values at \( t + \tau \), for \( \tau = 1, 2, \ldots \). We shall use the same models (trained on November 1981 – December 2014) that we used for in-sample forecasting to forecast with the out-of-sample values January 2015 to December 2017. As forecasting from January and February 2015 with our AR(3) models would involve using some of the training data, our forecasting time period of interest is March 2015 to December 2017. We write

\[
D_t \equiv \text{March 2015 to December 2017.} \tag{9.14}
\]

The observed SSTs for the months April 2017–June 2018 are then used to compute forecast errors and their summaries. We are cognisant that the out-of-sample forecasts only involve three years. However, we wanted to use as much data as possible for fitting the model. Also, this three-year period includes one El Niño which lasted for 13 months. This El Niño is likely to dominate the period and thus we may not see all the same patterns in the out-of-sample forecasting that we saw over approximately 30 years of in-sample forecasting.

9.2.1 Out-of-Sample Forecasts: Original SST Data

Similar to Section 9.1.1 at time \( t \in D_t \), \( \bar{z}_c(t) \), \( \bar{z}_c(t-1) \), and \( \bar{z}_c(t-2) \) were used to forecast \( \hat{z}_c(t+\tau) \), for \( \tau = 1, \ldots, 7 \), using the fitted AR(3) process given by Equation 8.20. We then used Equation 8.26 to obtain \( \{\hat{z}(t+\tau)\} \), the forecasts on the original degrees-Celsius scale.

For the out-of-sample forecasts, the forecast errors, given by Equation 9.1, are compared in Figure 9.23. There are some large negative forecast errors, particularly for larger \( \tau \) in March through July 2015. These correspond to the April 2015–April 2016 El Niño, which the model appears to have difficulty forecasting. The in-sample forecast errors, \( e \left( \hat{z}(t+\tau) \right) \) were given as boxplots in Figure 9.1, showing that forecasting from September and October appears easy as there are relatively small forecast errors across all lags for both in-sample and out-of-sample forecasting.
9.2.2 Out-of-Sample Forecasts: Transformed Data

Similar to Section 9.1.2, at time $t \in D_t$ we forecast $\hat{v}(t + \tau)$ on the $v$-scale for $\tau \in \{1, \ldots, 7\}$ and use those in Equation 8.33 to obtain $\{\tilde{z}(t + \tau)\}$.

The out-of-sample forecast errors, $e(\tilde{z}(t + \tau))$ are shown in Figure 9.24. Similar to Figure 9.23, there are some large-negative forecast-error values, particularly for larger $\tau$ in March through July 2015. These correspond to the April 2015–April 2016 El Niño event. This suggests this model also has difficulty forecasting in this El Niño event. Within each month, the out-of-sample error figures for both methods have similar patterns between the years across the lags. The in-sample forecast errors for $\tilde{z}(t + \tau)$ were given in Figure 9.6 where there are also some large-negative forecast-error values for larger $\tau$ in March and July. A comparison of the two forecast methods for out-of-sample forecasting is given in the next subsection.
Figure 9.23: Shown are plots of out-of-sample forecast errors, \( \{ e(\hat{z}(t + \tau)) \} \) given by Equation 9.1, broken down by month \( M \) and lag \( \tau = 1, \ldots, 7 \).
Figure 9.24: Shown are plots of out-of-sample forecast errors, \( \{ e(\tilde{z}(t + \tau)) \} \), broken down by month \( M \) and lag \( \tau = 1, \ldots, 7 \).
9.2.3 Comparison of Out-of-Sample Forecasts

In this section we directly compare the out-of-sample forecasts \( \{ \hat{z}(t + \tau) \} \) and \( \{ \tilde{z}(t + \tau) \} \) through their bias and mean squared prediction error.

Recall from Section 9.1 that bias is the average of the forecast errors. Figure 9.25 compares the bias from the two forecast methods through line plots, which is the average of the lines shown in Figures 9.23 and 9.24. For July through December the bias values are very similar for the two forecast methods, as we saw for the in-sample forecasts. Both methods have largest bias when forecasting from January and February across the (boreal) spring barrier into July, August, and September. The overall shape of the bias patterns across lag within each month is similar to those seen for in-sample forecasting (see Sections 9.2.1 and 9.2.2), however, the bias values are considerably larger for the out-of-sample forecasts. This is likely, at least partially, due to the out-of-sample bias values being an average of at most three values while the in-sample bias values are averages of at least 33 values.

As for the in-sample forecasts, we are interested in the bias-variance relationship and the contribution of bias to MSPE through bias squared as a percentage of mean squared prediction error. We wish to determine if these larger bias values for out-of-sample forecasting also correspond to bias squared being a larger percentage of the MSPE. For the in-sample forecasts, bias was at most 17% of the MSPE, however, for the out-of-sample forecasts, bias squared is at times a much larger percentage of the MSPE (up to 99.9%) and both forecast methods had multiple month and lag combinations where bias squared was a substantial percentage of the MSPE. However, the sample size is small so this out-of-sample forecasting study is limited. We note that the shape of the patterns within each month across the lags is also different for in-sample and out-of-sample forecasting. For this small out-of-sample forecasting study neither method appears to have control of the bias.
Figure 9.25: Out-of-sample forecasts: shown are comparison plots of bias broken down by month $M$ and lag $\tau = 1, \ldots, 7$. $b(\hat{z}(t+\tau); M, \tau)$ is represented by the black dotted line, and $b(\tilde{z}(t+\tau); M, \tau)$ by the solid grey line. The yellow band highlights the lags $\tau$ that forecast into April and May.
We have seen that the bias values are large and that for some months and lags the out-of-sample forecasting MSPE is dominated by bias. Nonetheless, we shall consider MSPE as a forecasting performance measure. We compare the $MSPE(\tilde{z}; M, \tau)$ and $MSPE(\hat{z}; M, \tau)$ through line plots in Figure 9.26. The MSPE for each month when $\tau = 1$ is very similar for the two methods. For some months, July through December, the MSPE for the two methods is very similar across all lags. For January through June the MSPE for the two methods have different behaviour across the $\tau$ values. For January, February, and March for values of $\tau$ larger than the spring barrier (April–May as represented by the yellow bands in the figures) the MSPE for $\tilde{z}(t+\tau)$ is larger than the MSPE for $\hat{z}(t+\tau)$. However, for April, May, and June and larger values of $\tau$, the reverse is true, the MSPE for $\tilde{z}(t+\tau)$ is less than the MSPE for $\hat{z}(t+\tau)$. These are the same patterns that were observed in the in-sample forecast MSPE values, however, the MSPE values are larger for out-of-sample forecasting.

As we saw similar patterns in Figure 9.26 to what we saw for in-sample forecasting we also consider the MSPE relative skill from Equation 9.9. Figure 9.27 shows the MSPE relative skills of the the forecasts for each $\tau = 1, \ldots, 7$, and $M$ ranging across the months. This plot has very similar diagonal patterns to the equivalent in-sample forecasting plot; Figure 9.18. For October through February there is a period where the relative skill values are noticeably greater than one indicating that the forecasts on the transformed scale have better performance. This period leads up to or coincides with the spring barrier (April–May). Also, when forecasting from May, June, and July into the later part of the year, the relative skill values are larger than one indicating that the forecasts derived from the transformed scale are better. However, when forecasting from December into the January and February; and from January through April, across the (boreal) spring barrier into May, June, and July, the relative skill values are all less than one indicating that the forecast derived on the original scale is better.

Neither forecast method had control of the bias for out-of-sample forecasting. Despite this, we saw the same seasonal patterns of MSPE relative skill for out-of-sample forecasting as we saw for in-sample forecasting. Specifically, when forecasting across the (boreal) spring barrier the forecast derived on the original scale had better performance (in a MSPE sense). However, when forecasting up to the spring barrier the forecasts made on the transformed scale had better performance. Hopefully in future as more data becomes available, perhaps once the period includes a La Niña event, more of the patterns seen for in-sample forecasting would become apparent in the out-of-sample forecasts.
Figure 9.26: Comparison of out-of-sample forecast $MSPE(\hat{z}(t + \tau); M, \tau)$ (black dotted line) and $MSPE(\tilde{z}(t + \tau); M, \tau)$ (grey solid line) broken down by month $M$ and lag $\tau = 1, \ldots, 7$. The yellow band highlights the lags $\tau$ that forecast into April and May.
Figure 9.27: Relative skill, $R_{MSPE}$, as defined by Equation 9.9, for all $M \in \{\text{Jan, \ldots, Dec}\}$ and $\tau \in \{1, \ldots, 7\}$. The colour scale is linear in the log of the relative skill. Blue colours indicate that the forecasts on the original scale have better performance and red colours indicate that the forecasts on the transformed scale have better performance.
Chapter 10

Discussion and Conclusions

The first part of this thesis provides an exploratory data analysis of tropical Pacific SSTs during the period November 1981 – December 2014. Most analyses and models found in the literature work directly with the SST anomalies. We demonstrated here that there is structure in the raw data that these analyses miss, namely a spatial mean-variability relationship that suggests a nonlinear transformation of the data would facilitate inference. This finding is also consistent with the generally accepted nonlinearity of the ENSO phenomenon [18, 82].

Working with anomalies implies that large-scale seasonal processes and smaller-scale processes are additive. The approach based on anomalies subtracts the seasonal component leaving behind a residual component (made up of the anomalies) that is modelled. Because there is a mean-standard deviation relationship, this needs to be respected first, before anomalies are considered and dynamical models are built.

In this thesis, we give a statistical methodology that removes the mean-standard deviation relationship through transforming the SST data. The transformation is empirically driven, but it is based on 33 years of data for which a consistent pattern is seen; outliers are apparent for less than 3% of the 398 months that we considered. In order that the outliers did not overly affect the patterns we inferred, we used robust methods to derive our transformation.

The transformation we derived is logarithmic, so it is monotonic and nonlinear, and it respects the variability seen in SSTs from month-to-month during the year. It is strictly increasing, which allows for easy interpretation of the geophysical variable. At the very least, the estimated parameters of the transformation offer another
characterisation of the enigmatic patterns of SSTs that lead to El Niño and La Niña events, specifically the spring barrier.

In the latter part of this thesis, we develop a forecast methodology by fitting an autoregressive process to the monthly-centred data on the original scale and the transformed scale. In fitting the autoregressive process on the transformed scale, we noticed that it is able to identify a high-lag periodicity in the data, which we were unable to identify on the original scale. This furthers our case that there is structure in the data that analyses of the raw data have missed.

Forecasting based on autoregressive processes is straightforward. We use a parametric bootstrap to transform the forecasts made on the transformed scale back to the original SST scale, in degrees Celsius, to enable direct comparison between the two forecasting methods. We compare the two forecasting methods via in-sample forecasting using the data the model was trained on (November 1981 – December 2014), and then we compare them via out-of-sample forecasting using the latest three years of the data that were not used in fitting the model.

We undertook a careful examination of the mean squared prediction error, separately exploring the squared-bias and variance components of the forecasts and their relative contributions towards the mean squared prediction error. Our in-sample forecasting results indicated that both methods had controlled for the bias (the original forecasts a little better) and that the forecasts on the transformed scale performed better when forecasting up to or into the boreal spring, while the forecasts on the original scale performed better when forecasting across and from boreal spring into summer. We also provided visualisations of the forecast-error bias, variance, and mean squared prediction error, which could be used to better identify and understand the (boreal) spring barrier.

When transforming the SST data, each month was transformed differently, and it was conjectured that an autoregressive process would be able to capture the dynamics for the whole process on the transformed scale, where we hoped those dynamics would be simpler. Unfortunately, this was not the case, and an AR(3) was the best model on both scales. Further, the model on the transformed scale had difficulty with the dynamics caused by the spring barrier, as did the model on the original scale.

Our out-of-sample results are less definitive than our in-sample results since only the last three years of data were considered, which included one El Niño period over a
period of 13 months. The squared-bias components are larger, and neither forecasting approach appears to have control of the bias. However, in terms of MSPE, the same relative forecasting-performance behaviour holds, where the forecasts on the transformed scale performed better when forecasting into the boreal spring, while the forecasts on the original scale performed better when forecasting across and from the boreal spring into summer. As more data become available through time, and out-of-sample forecasting can be performed on a larger dataset (which might include a La Niña event), it would be interesting to see if both methods would prove to have better control of the bias and whether the relative forecasting performance patterns still hold.

We note that our statistical-computing approach to back-transform the forecasts made on the transformed scale relies on approximations of the exponential function and linear regression. While we demonstrated our approach holds for forecasting up to lag 7 from an AR(1) process, it could break down at higher lags and potentially as the order of the autoregressive process increases.

We also note that no measurement error was accounted for, and hence any forecast at time $t + \tau$ relied only on the current and the immediate-past data at times $t$, $t - 1$, and $t - 2$. Our approach might be enhanced by embedding it in a hierarchical framework. Extensions of this work could also include using robust fitting of the autoregressive processes and determining if the observed mean-standard deviation relationship in the Niño 3.4 region is also present in other parts of the ocean.

The transformed scale captures structure and variability not seen on the raw scale, respects the spatial mean-variability relationship identified, has comparable control of the bias, and sometimes has better forecasting performance than on the original scale. The extra computation involved in calculating the transformed scale and back-transforming appears worthwhile, particularly when forecasting into the boreal spring. There is also the potential in this work to understand more clearly various geophysical phenomena involving tropical Pacific SSTs, such as the spring barrier and interconnected oceanic and atmospheric effects, by transforming the SSTs onto a different scale.
Bibliography


