Multivariate statistical models for seasonal climate prediction and climate downscaling

by

Alex Jason Cannon

B.Sc., The University of British Columbia, 1995
Dipl. Meteorology, The University of British Columbia, 1996
M.Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Atmospheric Science)

The University Of British Columbia

December, 2008

© Alex Jason Cannon 2008
Abstract

This dissertation develops multivariate statistical models for seasonal forecasting and downscaling of climate variables. In the case of seasonal climate forecasting, where record lengths are typically short and signal-to-noise ratios are low, particularly at long lead-times, forecast models must be robust against noise. To this end, two models are developed. Robust nonlinear canonical correlation analysis, which introduces robust cost functions to an existing model architecture, is outlined in Chapter 2. Nonlinear principal predictor analysis, the nonlinear extension of principal predictor analysis, a linear model of intermediate complexity between multivariate regression and canonical correlation analysis, is developed in Chapter 3. In the case of climate downscaling, the goal is to predict values of weather elements observed at local or regional scales from the synoptic-scale atmospheric circulation, usually for the purpose of generating climate scenarios from Global Climate Models. In this context, models must not only be accurate in terms of traditional model verification statistics, but they must also be able to replicate statistical properties of the historical observations. When downscaling series observed at multiple sites, correctly specifying relationships between sites is of key concern. Three models are developed for multi-site downscaling. Chapter 4 introduces nonlinear analog predictor analysis, a hybrid model that couples a neural network to an analog model. The neural network maps the original predictors to a lower-dimensional space such that predictions from the analog model are improved. Multivariate ridge regression with negative values of the ridge parameters is introduced in Chapter 5 as a means of performing expanded downscaling, which is a linear model that constrains the covariance matrix of model predictions to match that of observations. The expanded Bernoulli-gamma density network, a nonlinear probabilistic extension of expanded downscaling, is introduced in Chapter 6 for multi-site precipitation downscaling. The single-site model is extended by allowing multiple predictands and by adopting the expanded downscaling covariance constraint.
# Table of Contents

Abstract .............................................................. ii

Table of Contents .................................................. iii

List of Tables ....................................................... vii

List of Figures ....................................................... ix

Acknowledgements .................................................. xiv

Dedication ........................................................... xv

Statement of Co-Authorship .......................................... xvi

## 1 Introduction .................................................... 1

1.1 Overview ......................................................... 1

1.2 Linearity and nonlinearity ...................................... 2

1.3 Linear multivariate statistical models ....................... 3

  1.3.1 Multivariate regression .................................. 3

  1.3.2 Canonical correlation analysis and maximum covariance
       analysis ...................................................... 3

  1.3.3 Principal predictor analysis and redundancy analysis .... 4

1.4 Nonlinear multivariate statistical models .................... 6

  1.4.1 General methods .......................................... 6

  1.4.2 Nonlinear multivariate regression ....................... 6

  1.4.3 Nonlinear canonical correlation analysis ................ 7

1.5 Seasonal climate prediction .................................. 8

  1.5.1 Linear methods .......................................... 9

  1.5.2 Nonlinear methods ...................................... 9

1.6 Climate downscaling ........................................ 10

  1.6.1 Linear methods ........................................ 11

  1.6.2 Nonlinear methods .................................... 12

  1.6.3 Hybrid methods ........................................ 13

1.7 Research goals ................................................ 14
# Table of Contents

## 2 Robust nonlinear canonical correlation analysis

2.1 Introduction ................................................................................. 21
2.2 Method ....................................................................................... 22
  2.2.1 NLCCA .................................................................................. 22
  2.2.2 Biweight midcorrelation ......................................................... 25
  2.2.3 $L_p$-norm ............................................................................ 29
  2.2.4 Robust NLCCA ................................................................. 29
2.3 Synthetic test problem ............................................................... 29
  2.3.1 Data .................................................................................... 29
  2.3.2 Training and testing procedure ........................................... 31
  2.3.3 Model performance ............................................................ 31
2.4 Seasonal prediction .................................................................... 33
  2.4.1 Data .................................................................................... 33
  2.4.2 Training and testing procedure ........................................... 33
  2.4.3 Skill for models with one mode .......................................... 35
  2.4.4 Skill for models with two modes .......................................... 41
2.5 Conclusion .................................................................................. 42

## 3 Nonlinear principal predictor analysis

3.1 Introduction ................................................................................. 46
3.2 Method ....................................................................................... 47
3.3 Application to the Lorenz system .............................................. 54
3.4 Discussion and conclusion ......................................................... 59

## 4 Nonlinear analog predictor analysis

4.1 Introduction ................................................................................. 67
4.2 Method ....................................................................................... 69
  4.2.1 Analog model ....................................................................... 69
  4.2.2 Nonlinear Principal Predictor Analysis .................................. 70
  4.2.3 Nonlinear Analog Predictor Analysis ..................................... 72
4.3 Synthetic benchmark tasks ......................................................... 73
  4.3.1 Linear benchmark .................................................................. 73
  4.3.2 Nonlinear benchmark .......................................................... 73
  4.3.3 Multicollinear benchmark ...................................................... 74
  4.3.4 Multiplicative benchmark ...................................................... 74

## Bibliography

................................................................. 15
................................................................. 43
................................................................. 65
................................................................. 67
# Table of Contents

4.3.5 Synthetic benchmark results ........................................ 74  
4.4 Multivariate downscaling ................................................. 76  
4.4.1 Synoptic-scale atmospheric circulation dataset .................. 77  
4.4.2 Whistler temperature and precipitation dataset .................. 77  
4.4.3 Multi-site precipitation dataset ................................. 81  
4.5 Discussion and conclusion ............................................. 82  

Bibliography ........................................................................ 84  

5 Multivariate ridge regression with negative ridge parameters .......... 87  
5.1 Introduction ................................................................. 87  
5.2 Data .......................................................................... 88  
5.3 Univariate ridge regression .............................................. 88  
5.4 Multivariate ridge regression .......................................... 89  
5.5 Univariate ridge regression with negative $k$ ......................... 90  
5.6 Multivariate ridge regression with negative elements in $K$ .......... 92  
5.6.1 Model optimization .................................................. 92  
5.6.2 Temperature downscaling ......................................... 92  
5.6.3 Precipitation downscaling ....................................... 96  
5.7 Computational cost ..................................................... 98  
5.8 Discussion and conclusions .......................................... 100  

Bibliography ........................................................................ 102  

6 Expanded Bernoulli-gamma density network ............................... 104  
6.1 Introduction ................................................................. 104  
6.2 Datasets and benchmarks .............................................. 107  
6.2.1 Precipitation data .................................................... 107  
6.2.2 Synoptic-scale predictor data ..................................... 107  
6.2.3 Benchmark model .................................................... 110  
6.3 Background ................................................................. 110  
6.3.1 Expanded downscaling ............................................ 110  
6.3.2 Bernoulli-gamma distribution ..................................... 111  
6.4 Method ...................................................................... 113  
6.4.1 Bernoulli-gamma density network (BDN) ....................... 113  
6.4.2 Conditional simulation of precipitation series .................... 115  
6.4.3 Expanded Bernoulli-gamma density network (EBDN) ............ 116  
6.5 Results ................................................................. 120  
6.6 Discussion and conclusion .......................................... 126  

Bibliography ........................................................................ 128
# Table of Contents

7 Conclusion .................................................................................................................. 132
  7.1 Summary .................................................................................................................. 132
  7.2 Seasonal climate prediction ...................................................................................... 132
    7.2.1 Robust nonlinear canonical correlation analysis ............................................. 132
    7.2.2 Nonlinear principal predictor analysis ............................................................ 133
  7.3 Climate downscaling ............................................................................................... 134
    7.3.1 Nonlinear analog predictor analysis ................................................................. 134
    7.3.2 Multivariate ridge regression with negative ridge parameters ..................... 135
    7.3.3 Expanded Bernoulli-gamma density network ................................................. 135
  7.4 Overall recommendations ....................................................................................... 136

Bibliography .................................................................................................................... 138
## List of Tables

3.1 A qualitative summary of the main attributes of nonlinear multivariate models.  
4.1 Root MSE (RMSE) of analog model predictions on test cases of the four synthetic benchmark tasks described in Section 4.3. The bold font indicates the best performing method and underlining indicates the worst performing method.  
4.2 Performance of analog models on test cases for the Whistler downscaling dataset described in Section 4.4. Model statistics are RMSE for temperature ($T$), rainfall ($R$), and snowfall ($S$) predictions; percentage explained variance (EV) for daily temperature, rainfall, and snowfall predictions; ratios of predicted to observed number of temperature events $< 0^\circ$C (FREEZE); ratios of predicted to observed number of wet days (WET); and ratios of predicted to observed number of snow days (SNOW). Values for the best performing model are shown in bold and values for the worst performing model are underlined.  
4.3 Daily precipitation observing stations  
4.4 Performance of analog models on test cases for the multi-site daily precipitation downscaling dataset described in Section 4.4. Model statistics are RMSE of predictions; percentage explained variance of predictions; ratios of predicted to observed number of wet days; and ratios of predicted to observed number of days with more than 100 mm of precipitation (EXTREME). Results are mean values over the ten sites. Values for the best performing model are shown in bold and values for the worst performing model are underlined.  
5.1 Performance statistics for model predictions of temperature.  
5.2 Performance statistics for model predictions of precipitation.  
6.1 Daily precipitation observing stations
6.2 Cross-validated model performance statistics for EBDN and TreeGen downscaling models. NLPD is the mean negative log predictive density; rmse is the root mean squared error of the conditional mean series; rmse$_s$ is the root mse of the spatial mean series; and $r_s$ is the correlation between the observed and predicted spatial mean series. The spatial mean is defined as the mean of the 10 station values. TSS, HIT, POD, FAR, and BIAS are, respectively, the true skill statistic, hit-rate, probability of detection, false alarm rate, and bias ratio for binary predictions of daily precipitation occurrence. Values corresponding to the better performing model are italicized.  

6.3 Cross-validated model rmse statistics for median and inter-quartile range (IQR) of monthly number of wet days (NWD), maximum 5-day precipitation (PX5D), maximum number of consecutive wet days (WRUN), and maximum number of consecutive dry days (DRUN) for EBDN and TreeGen downscaling models at stations C4 and C9. Results summarize Figures 6.10 and 6.11. Values corresponding to the better performing model are italicized.
List of Figures

1.1 Diagram showing a multivariate regression model with $l = 4$ predictors and $m = 3$ predictands. Lines represent model parameters and circles represent model inputs or outputs. For simplicity, bias parameters are not illustrated. ................................... 4
1.2 Diagram showing a CCA model with $l = 4$ and $m = 3$. ........................................ 5
1.3 Diagram showing a PPA model with $l = 4$ and $m = 3$. ........................................ 5
1.4 Diagram showing a nonlinear multivariate regression model by a neural network with $l = 4, q = 2$, and $m = 3$. ................................................................. 7
1.5 Diagram showing a NLCCA model with $l = 4, q = 2$, and $m = 3$. .................. 8
1.6 Ternary diagram showing the three categories of statistical downscaling approaches. The shaded area indicates the realm of methods considered in this dissertation. .................................................. 12
2.1 Neural network architecture used to perform NLCCA. ............................................. 23
2.2 Empirical comparison between the Pearson correlation (cor) and the biweight midcorrelation (bicor) on random variables $x$ and $y$, each with samples drawn from a standard normal distribution, and $x'$ and $y$, where $x'$ is the same as $x$ but with one case replaced with an outlier. Panel (a) shows sample time series of $x'$ (solid) and $y$ (dashed); (b) compares cor($x$, $y$) and bicor($x$, $y$); (c) compares cor($x$, $y$) and cor($x'$, $y$); and (d) compares bicor($x$, $y$) and bicor($x'$, $y$). Plots are for 1000 randomly generated datasets. .................................................. 26
2.3 Time series of $u$ with (a) $p = 1$ and (b) $p = 9$; outliers in $v$ occur at the same time as those in $u$. (c) The effect on cor($u$, $v$) (solid line) and bicor($u$, $v$) (dashed line) from increasing the separation between common outlier and non-outlier points by increasing $p$. .................................................. 28
2.4 Synthetic test datasets used to evaluate the performance of the standard and robust NLCCA models. The first mode is shown by the thick black curve and the second mode is shown by the thin black curve. Test samples with added noise are shown as asterisks. .................................................. 30
2.5 Boxplots showing the distribution of rmse between the first synthetic mode and the first mode extracted by NLCCA models for (a) x and (b) y with different combinations of non-robust and robust cost functions over 50 trials. Boxes extend from the 25th to 75th percentiles, with the line indicating the median. Whiskers represent the most extreme data within ±1.5 times the interquartile range (i.e., the box height); values outside this range are plotted as dots. The dashed line indicates a rmse equal to one. The ordinate is log-scaled to accommodate the large range in rmse. ........................................... 32
2.6 Diagram showing how data were split into training (light gray) and validation (dark gray) segments for the first (CV1) and second (CV2) cross-validation procedures. ........................................... 34
2.7 Cross-validated correlation skill for NLCCA models trained with cor/mse, bicor/mse, and bicor/mae cost functions. Weight penalty was applied to the model denoted cor/mse(p). Bars show the mean correlation over the spatial domain, averaged over the 10 trials. Vertical lines extend from the minimum to the maximum spatial mean correlation from the 10 trials. Horizontal lines show correlation skill from the CCA model. The ordinate is limited to showing positive cross-validated skill. ........................................... 35
2.8 Plots of (a) PC scores from the leading SST (solid line) and SLP (dashed line) PCA modes; (b) the canonical variate u for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and one with bicor/mse cost functions (solid line); and (c) canonical variate v for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and bicor/mse cost functions (solid line). ........................................... 37
2.9 (a) Plots of the first SST mode for CCA (thin line) and NLCCA with bicor/mae cost functions (thick line) in the planes of the PC1-PC2 and PC1-PC3 scores at 0, 3, 6, 9, and 12-month lead times. Spatial patterns for (b) PC1, (c) PC2, and (d) PC3, all normalized to unit norm. ........................................... 39
2.10 Spatial correlation skill at 0-month lead time for (a) NLCCA with bicor/mae cost functions and (b) CCA. Panel (c) shows NLCCA skill minus CCA skill. Panels (d) to (f) as in (a) to (c) but for 12-month lead time. ........................................... 40
2.11 As in Figure 2.7, but for NLCCA models with two modes. ........................................... 41
3.1 Neural network architecture for NLPPA. The multi-layer neural network at the top consists of $I$ input nodes and $K$ output nodes, which correspond to the predictors and predictands respectively. Between the input and output nodes are three sets of internal nodes (starting with $J$ hidden nodes, then a single bottleneck node, and finally $L$ hidden nodes). The neural network at the bottom consists of one input node, which corresponds to the principal predictor extracted from the bottleneck node, $J$ hidden nodes, and $I$ output nodes corresponding to the predictors. Each line represents a parameter in the model.

3.2 Noisy synthetic data and the theoretical mode in (a) $x$ space and (b) $y$ space.

3.3 Noisy synthetic data and the first two linear modes (where the first is represented by a thin line and the second is represented by a thick line) modelled by PPA in (a) $x$ space and (b) $y$ space. Noisy synthetic data and the first nonlinear mode modelled by NLPPA in (c) $x$ space and (d) $y$ space.

3.4 MSE values for model predictions at each lead time and noise level.

3.5 Percent improvement in MSE between NLPPA and NLCCA for each lead time and noise level. Positive values indicate that NLPPA performed better than NLCCA.

3.6 MSE values associated with the addition of each mode at the 9 step lead time and $n = 0.3$.

3.7 Lorenz data (gray dots) and the first (thick black line), second (medium black line) and third (thin black line) modes from NLPCA shown projected onto each of the 2-D planes in $y$ space and as a 3-D scatterplot.

3.8 As in Figure 3.7 but for NLPPA. Predictand modes were extracted from the 9 step lead time data with 30% added random noise.

3.9 As in Figure 3.8 but for NLCCA.

3.10 Boxplots of MSE values for predictand modes extracted by NLPPA and NLCCA from the noisy data with respect to NLPCA modes extracted from the noise free data over the 10 trials.

4.1 ANN architecture for NLPPA. The multi-layer ANN consists of $I$ input nodes and $K$ output nodes, which correspond to the predictors and predictands respectively. Between the input and output nodes are three sets of internal nodes starting with the $J$ hidden nodes, then $L$ bottleneck nodes, and finally $J$ hidden nodes. Each line represents a parameter in the model.

4.2 Schematic diagram of a NLAPA model with $I = 4$ predictors, $J = 3$ hidden-layer nodes, and $L = 2$ analog variables.

4.3 Map showing the extent of the synoptic-scale fields from the NCEP-DOE AMIP-II Reanalysis, with the small dots representing centers of the model grid points. The square shows the location of Whistler, B.C. The large dots show locations of the ten daily precipitation observing stations.
List of Figures

4.4 Plots of predicted versus observed quantiles of daily temperature, rain amounts, and snow amounts for NLAPA, analog/PCA, and analog/CCA models. ............... 80

4.5 Plots of predicted versus observed inter-site daily precipitation correlations for NLAPA, analog/PCA, and analog/CCA models. ................................. 82

5.1 Variances and mean squared errors of surface temperature predictions at station T1 plotted against scaled values of the ridge parameter $k/n$ from univariate ridge regression models. Stars indicate values for the model with no variance inflation ($k/n = 0$); solid dots correspond to the model in which predicted and observed variances are equal ($k/n = -0.08$). __________________ 91

5.2 Effect of the magnitude of the penalty coefficient $\lambda$ on the mean squared error and the covariance constraint term (i.e., the sum of squared deviations between elements of the observed and predicted covariance matrices) in Eq. 5.10. ....... 93

5.3 Observed versus predicted elements of the predictand covariance matrix for (a) the multivariate regression model, (b) the inflated multivariate regression model, (c) the expanded downscaling model, and (d) the multivariate ridge regression model. Linear fits to the data are shown by dashed lines. .............. 94

5.4 A graphical representation of the optimized ridge matrix $K$ based on 1959-1988 temperature training data. Shades represent values of $K/n$. .................. 95

5.5 As in Figure 5.3 but for the precipitation covariance matrix. ...................... 97

5.6 As in Figure 5.4 but for the precipitation ridge matrix. ............................ 98

5.7 Time required to optimize the multivariate ridge regression cost function (Eq. 5.9), expressed as a ratio with respect to the amount of time required to optimize the corresponding expanded downscaling model. Ratios in excess of one mean that the multivariate ridge regression model take longer to converge than the expanded downscaling model. Values are for different numbers of model predictands $q$. .......................... 99

6.1 Map showing the extent of the ECMWF ERA-40 re-analysis domain used in the study. Grid-points are numbered and circles show locations of the daily precipitation observing stations. ........................... 108

6.2 Multivariate regression tree used for predictor selection in the EBDN model. Splits in the tree are labelled by meteorological field and grid-point number from Figure 6.1. At each node in the tree, bar plots show the relative magnitude of daily precipitation amounts at the 10 stations. ...................... 109

6.3 Histograms of observed daily precipitation (grey shading) and fitted Bernoulli-gamma pdfs (dashed line for positive amounts and asterisk for the probability at zero) at (a) station C9 and (b) station C10, along with quantile-quantile plots for (c) station C9 and (d) station C10. .................. 112
6.4 Flowchart showing the steps involved in (a) training and (b) evaluating the EBDN model. Relevant sections of the text are given in parentheses.

6.5 Sample schematic diagram of a Bernoulli-gamma density ANN for modeling conditional Bernoulli-gamma distribution parameters at two sites.

6.6 Time series of (a) conditional Bernoulli-gamma parameters where the probability of precipitation $p(t)$ is shown by the dashed line, the shape $\alpha(t)$ by the solid line, and the scale $\beta(t)$ by the dotted line; (b) AR(1) cumulative probabilities $z(t)$; (c) conditional mean (solid line) and the range from the conditional 5th to 95th percentiles (gray shading); and (d) synthetic daily precipitation (vertical lines). Observed daily precipitation values are shown as dots in (d). Results shown are for station C10 during 1989.

6.7 Observed correlations versus boxplots of predicted correlations from 30 synthetic daily precipitation series following (a) the initialization of the correction algorithm and (b) the second and final iteration of the correction algorithm.

6.8 Quantile-quantile plots of daily station mean precipitation based on 100 synthetic precipitation series from (a) the EBDN model and (b) the TreeGen model.

6.9 Observed covariances versus boxplots of predicted covariances from 100 synthetic precipitation series from (a) the EBDN model and (b) the TreeGen model. The line inside each box represents the median, boxes extend from the 25th to 75th percentiles, whiskers extend to points within 1.5 times the inter-quartile range, and outliers are shown as circles.

6.10 Interannual distributions of monthly observed (white), TreeGen modeled (light gray), and EBDN modeled (dark grey) (a) number of wet days (NWD), (b) maximum 5-day precipitation amounts (PX5D), (c) maximum number of consecutive wet days (WRUN), and (d) maximum number of consecutive dry days (DRUN) at station C4.

6.11 As in Figure 6.10 but for station C9.
Acknowledgements

First and foremost, I would like to thank my research advisor, Professor William Hsieh for his superb guidance, enthusiasm, and intellectual support. I would also like to thank Professors Ian McKendry and R. D. Moore for serving on my dissertation committee.

My deepest gratitude goes to Paul Whitfield at Environment Canada, who encouraged me to pursue my academic studies, and who has been an excellent colleague and mentor over the years. Without his personal support, and the support of Environment Canada as an organization, I would not have been able to complete my studies. I am also very grateful to have been inspired, challenged, and supported by extraordinary groups of researchers at both UBC and Environment Canada.

Finally, I would like to thank my family for their love, patience, and care, especially my parents Karen and Richard, and my wife Trina, to whom this thesis is dedicated.
To Trina, for your support, and love.
Statement of Co-Authorship

The results presented in Chapter 2 originate from research conducted within the context of my Ph.D. studies. The co-author of this paper, Professor William Hsieh, identified the need for a robust variant of nonlinear canonical correlation analysis, provided oversight of the research project in his role as my supervisor, and assisted in the editing of the final text.
Chapter 1

Introduction

1.1 Overview

The field of multivariate statistics is concerned with the use of statistical models in the analysis of multivariate datasets, that is datasets consisting of more than one variable. In the context of climatology, multivariate datasets could, for example, be made up of surface air temperatures observed at multiple stations, geopotential heights from different pressure levels of a numerical model, or indices characterizing modes of variability in the global atmospheric circulation. Multivariate statistical models can be separated into two classes: (i) those that are unsupervised, where the goal is to find structure in a set of variables without reference to a set of target data; and (ii) those that are supervised, where the goal is to analyze relationships between two sets of variables, with one set (or, in some cases, both sets) acting as target data. This dissertation is concerned with the application of supervised multivariate statistical models to climate datasets.

Supervised models may be further divided into three classes depending on the ultimate goal of the analysis, although these divisions are markedly less distinct than the one between supervised and unsupervised methods: (i) exploratory models, which are designed to find relationships between datasets that are not known a priori by the investigator, for example determining if lead-lag relationships exist between Eurasian snow cover and Northern Hemisphere surface air temperatures; (ii) simulation models, which are designed to replicate the statistical properties of one set of variables based on those of a second set, for example randomly generating series that resemble observed station wind speeds based on upstream surface pressure observations; and (iii) predictive models, which are designed to predict, in some optimal way, values of one set of variables (called the predictands) from another set (called the predictors), for example forecasting winter precipitation amounts in western Canada from fall sea surface temperatures in the equatorial Pacific Ocean. The main focus of this dissertation is on predictive models, although portions will touch on exploratory and simulation models as well.

Multivariate statistical models have a rich history of use in the climate sciences. Mirroring the course of development of multivariate statistics in general, climatology has progressed from relying almost exclusively on linear methods to a broader acceptance and more widespread use of nonlinear methods (Hsieh, 2004). Reflecting this trend, linear and nonlinear models for multivariate prediction will first be reviewed in turn. This will be followed by an overview of these models – in particular ones based on the multilayer perceptron neural network – in the fields of seasonal climate prediction and climate downscaling, the two application areas
addressed by this dissertation.

In this context, the primary goal will be to develop multivariate statistical models that are suited to climate prediction in the midlatitudes, in particular British Columbia, Canada. In the case of seasonal prediction, this means developing models that are robust, that is models in which parameter estimates are less affected by outliers (i.e., observations that lie outside the “expected” range of the signal) and that can handle weak signals in the short, noisy extratropical climate records. In the case of climate downscaling, the focus will be on the challenges involved with modelling mixed discrete-continuous variables, such as daily precipitation amounts, at multiple sites. As British Columbia is a region with complex terrain, involving the interface between ocean and land, along with mountain/valley systems that interact with the dominant flow, precipitation is extremely variable over short spatial scales, making the region an ideal benchmark for testing multi-site downscaling algorithms.

1.2 Linearity and nonlinearity

Mathematically, a function $f$ is said to be nonlinear if and only if it obeys the inequality

$$f(a_1x_1 + a_2x_2) \neq a_1f(x_1) + a_2f(x_2).$$

(1.1)

In other words, the response of the function is non-additive. Nonlinearity is a fundamental characteristic of the physical climate system, appearing, for example, in the primitive equations – the set of nonlinear differential equations that govern the flow of the atmosphere and ocean. Other examples of nonlinear behaviour in climatology can be found in Rial et al. (2004). The reader is referred to Yuval and Hsieh (2002) for a discussion of the detectability of nonlinear empirical relationships in the climate system.

Given the mathematical definition of nonlinearity, a further distinction is made between nonlinear statistical models that are linear in terms of the parameters to be estimated and those that are nonlinear in terms of the parameters to be estimated. In the former case, the functional relationship described by the statistical model is nonlinear, i.e. its response is not additive, but derivatives of the function with respect to its parameters do not contain said parameters. In this case, the nonlinear relationship can be made linear by applying a suitable transformation to the predictor (and/or predictand) variables. As an example, the function

$$f(x) = a \exp(x)$$

(1.2)

is nonlinear in the sense of Eq. 1.1 but is linear in terms of the parameter $a$. The response can be made linear by taking the natural logarithm of $x$. On the other hand, the function

$$f(x) = \frac{1}{1 + \exp(ax)}$$

(1.3)

is nonlinear both in the sense of Eq. 1.1 and in terms of the parameter $a$. See Clarke (1973)
for applied examples of mathematical and statistical nonlinearity in the context of hydrological modelling.

### 1.3 Linear multivariate statistical models

#### 1.3.1 Multivariate regression

The objective of multiple linear regression is to determine how a predictand variable $y$ with $t = 1, \ldots, n$ samples is linearly related to the predictor variables $x = \{x_i(t)\}, i = 1, \ldots, l$. Grouping the predictors into the column vector $x(t) = [x_i(t), \ldots, x_l(t)]^T$, the regression equation can be written as

$$y'(t) = w \cdot x(t) + b, \quad \epsilon(t) = y(t) - y'(t) \quad (1.4)$$

where $y'$ are model predictions, $\epsilon$ are the residuals, $w$ is a column vector of weight parameters of length $l$, and $b$ is a bias or intercept parameter. Values of the model parameters $w$ and $b$ are found by minimizing the least squares cost function, i.e., the sum of $\epsilon^2$. (Note: the terms residuals and errors are used synonymously throughout the remainder of the thesis.)

Multivariate regression is a multiple linear regression model with $m$ predictand variables $y(t) = [y_1(t), \ldots, y_m(t)]^T$. Predictions of the $j$th predictand are given by

$$y_j'(t) = [Wx(t) + b]_j, \quad \epsilon_j(t) = y_j(t) - y_j'(t) \quad (1.5)$$

where $W$ is an $m \times l$ matrix of weight parameters and $b$ is a column vector of bias parameters of length $m$. Model parameters are set by minimizing the sum of $\epsilon_j^2$ over the $j = 1, \ldots, m$ predictands, where it is assumed that residuals for the $m$ predictands are linearly independent. In this case, multivariate regression reduces to a series of $m$ multiple linear regression equations and model parameters do not depend on the fact that there are multiple predictands (Hastie et al., 2001). Relaxing the independence assumption leads to, among others, a class of multivariate linear models, usually referred to as multivariate ridge regression (Brown and Zidek, 1980; Haitovsky, 1987), in which the least squares cost function is augmented by a penalty term and the between regressions covariance matrix is given explicit consideration in the regression estimator. This form of model will be mentioned briefly later in the chapter in the context of multivariate climate downscaling.

As a basis for subsequent comparison with other types of models, a diagram illustrating multivariate regression is shown in Figure 1.1.

#### 1.3.2 Canonical correlation analysis and maximum covariance analysis

Canonical correlation analysis (CCA) (Hotelling, 1936) and maximum covariance analysis (MCA) (Tucker, 1958) both extend the standard multivariate regression model. CCA seeks linear com-
Figure 1.1: Diagram showing a multivariate regression model with \( l = 4 \) predictors and \( m = 3 \) predictands. Lines represent model parameters and circles represent model inputs or outputs. For simplicity, bias parameters are not illustrated.

\[
\begin{align*}
  u(t) &= a \cdot x(t), \\
  v(t) &= b \cdot y(t)
\end{align*}
\]  

such that the Pearson product-moment correlation between the derived canonical variates \( u \) and \( v \) is maximized. The datasets \( x \) and \( y \) are then linearly regressed onto \( u \) and \( v \), and the next leading canonical variates are extracted from the regression residuals. MCA differs from CCA in that the goal is to find patterns that successively maximize the covariance between the datasets rather than the correlation (Jolliffe, 2002). Refer to Figure 1.2 for a diagram of the CCA model.

Whereas multivariate regression seeks relationships between each predictand variable \( y_j \) and the predictors \( x \) independently, CCA and MCA consider how the entire set of predictands \( y \) is related to \( x \) (and \textit{vice versa}). Both models extract coupled modes of variability from the two datasets. Additionally, the role of the two datasets is interchangeable in CCA and MCA. That is \( x \) and \( y \) both act simultaneously as predictors and predictands. Multivariate regression is asymmetric. If the roles of the predictor and predictand variables are switched, then the resulting model parameters will be different.

1.3.3 Principal predictor analysis and redundancy analysis

CCA and MCA are both symmetric models. In some situations, a clear set of predictors and predictands may be obvious. Principal predictor analysis (PPA) (Thacker, 1999) and redundancy analysis (RA) (van den Wollenberg, 1977) restore the asymmetry of multivariate regression, but, like CCA and MCA, sequentially extract dominant modes of variability from the datasets.

In the PPA model, which is shown schematically in Figure 1.3, the goal is to find the linear combination of variables...
Chapter 1. Introduction

Figure 1.2: Diagram showing a CCA model with $l = 4$ and $m = 3$.

Figure 1.3: Diagram showing a PPA model with $l = 4$ and $m = 3$. 
such that the principal predictor $p$ maximizes the sum of squared correlations over all of the predictand variables $y_j$, $j = 1, .., m$. As in CCA, the datasets $x$ and $y$ are then linearly regressed onto $p$, and the next leading principal predictor is extracted from the regression residuals. Subsequent principal predictors are constrained to be uncorrelated with those previously extracted. The goal of redundancy analysis is similar, except that successive derived variates are correlated, while the corresponding vectors of weights $a$ are constrained to be orthogonal (Jolliffe, 2002).

1.4 Nonlinear multivariate statistical models

1.4.1 General methods

The methods described in Section 1.3 are all linear, and, as a result, are incapable of describing nonlinear relationships between predictors and predictands without a priori specification of normalizing/linearizing transformations, which can be problematic (Shacham et al., 1993). In practice, this means that the resulting approximations of the climate system by linear models may be too simplistic, leading to poor predictive performance (Hsieh, 2004). As an alternative, one can apply nonlinear statistical models that are capable of explicitly modelling nonlinear relationships. Hastie et al. (2001) provide a comprehensive overview of nonlinear statistical models. Here, the focus is on one particular approach, the multi-layer perceptron neural network, which is a flexible nonlinear statistical model that has been widely applied in the atmospheric sciences (Gardner and Dorling, 1998; Hsieh and Tang, 1998). While other kinds of nonlinear statistical models have been applied to seasonal climate prediction and climate downscaling tasks, including alternative neural network architectures, e.g. recurrent neural networks (Coulibaly et al., 2000), radial basis functions (Harpham and Wilby, 2005), adaptive neuro-fuzzy inference systems (Faucher et al., 1999) etc., the multi-layer perceptron forms the basis for many nonlinear variants of standard multivariate linear techniques (Hsieh, 2004).

1.4.2 Nonlinear multivariate regression

In its simplest form, a neural network performs nonlinear multivariate regression. The basic architecture is shown in Figure 1.4. First, $x(t)$, the input column vector of length $l$, is mapped to the hidden-layer of the network containing $q$ nodes, which are represented by the column vector $h(t)$ with elements

$$h_k(t) = \tanh \left( \left[ W^{(1)} x(t) + b^{(1)} \right]_k \right),$$

where $W^{(1)}$ is an $q \times l$ weight matrix, $b^{(1)}$ is a column vector of length $q$ containing bias
parameters, and \( k = 1, \ldots, q \). The hidden-layer is then mapped to the output-layer, giving predictions

\[
y_j'(t) = \left[ W^{(2)} h(t) + b^{(2)} \right]_j, \quad \epsilon_j(t) = y_j(t) - y_j'(t),
\]

where \( W^{(2)} \) is an \( m \times q \) weight matrix, and \( b^{(2)} \) is a column vector of length \( m \) containing bias parameters. Model parameters are set by minimizing the sum of \( \epsilon_j^2 \) over the \( j = 1, \ldots, m \) predictands.

Unlike in multivariate regression, where the number of model parameters is fully specified by the number of predictors \( l \) and predictands \( m \), the complexity of a neural network depends on the number of hidden-layer nodes \( q \). This is evident by comparing Figures 1.1 and 1.4. This leads to one of the main drawbacks of the neural network, overfitting, which means fitting to noise in the finite training sample rather than the underlying signal of interest. If \( q \), and, as a result, the number of model parameters, is too large, then the generalization ability of the model, its ability to accurately make predictions when presented with new samples, will suffer due to overfitting. Overviews of methods used to avoid overfitting in neural networks are given in Gardner and Dorling (1998), Hsieh and Tang (1998), and Hsieh (2004).

Furthermore, note that the parameters from the nonlinear multivariate regression model described above will, in general, not be the same as those obtained from a sequence of \( m \) univariate nonlinear regression models. As illustrated in Figure 1.4, each of the \( m \) outputs in a multivariate neural network are fed by common hidden-layer nodes. This sharing of information between outputs can lead to improved performance relative to univariate models (Caruana, 1997). As shown in Section 1.3.1, this behaviour is not shared by linear regression. The multivariate regression model and the corresponding multiple linear regression models are equivalent.

### 1.4.3 Nonlinear canonical correlation analysis

Chapter 1. Introduction

Figure 1.5: Diagram showing a NLCCA model with $l = 4$, $q = 2$, and $m = 3$.

The goal of NLCCA is the same as that of CCA, namely to compress the variables $x$ and $y$ to the canonical variates $u$ and $v$ such that the Pearson product-moment correlation between $u$ and $v$ is maximized. Whereas the compression in CCA is linear, obtained via weighted sums of the original variables (Eq. 1.6), the compression in NLCCA is performed by neural networks (Eqs. 1.8 and 1.9) and is nonlinear. Once the nonlinear canonical variates have been identified, $x$ and $y$ are nonlinearly regressed onto $u$ and $v$ respectively, again by neural networks. In CCA, the regression step results in straight lines through predictor/predictand space, whereas the nonlinearity of the NLCCA model means that the responses in predictor/predictand space can be curved.

As pointed out by Hsieh (2001), it does not appear to be possible to implement a nonlinear variant of MCA by neural networks.

1.5 Seasonal climate prediction

The goal of seasonal climate prediction is to forecast the evolution of the climate system on time scales from months to seasons, in other words at lead times beyond the limit of predictability of the instantaneous state of the atmosphere (Zwiers and Von Storch, 2004). Current approaches involve physically based general circulation models, statistical models, or hybrid physical-statistical models (Zwiers and Von Storch, 2004). Regardless of the method, predictive skill usually derives from teleconnections between the atmosphere and slowly varying boundary conditions such as sea surface temperatures (Quan et al., 2006), soil moisture (Douville, 2003),
or snow cover (Qian and Saunders, 2003).

Because of the long lead times and averaging time scales involved in seasonal climate prediction, the number of observational samples available for training statistical models is usually short. In addition, low signal-to-noise ratios outside of the Tropics means that producing skillful seasonal forecasts in these areas may be inherently difficult (George and Sutton, 2006). Furthermore, averaging, for example to monthly or seasonal time scales, will tend to make nonlinear relationships difficult to detect in the short, potentially noisy seasonal climate records (Yuval and Hsieh, 2002).

As the focus of this dissertation is on multivariate statistics, the following sections review, in turn, the use of linear and nonlinear multivariate statistical models for seasonal climate prediction.

1.5.1 Linear methods

The first efforts at seasonal climate forecasts by univariate linear regression and correlation analyses were made in the early 1900s by Walker (Katz, 2002). Multivariate regression methods remained dominant (see, for example, the review by Barnett and Hasselmann, 1979) until CCA was introduced for seasonal prediction of surface air temperatures in the United States by Barnett and Preisendorfer (1987). CCA is now common and has been used to make skillful seasonal forecasts of El Niño-Southern Oscillation episodes (Barnston and Ropelewski, 1992), surface air temperatures and precipitation amounts in Canada (Shabbar and Barnston, 1996), surface air temperatures in northern Europe (Johansson et al., 1998), summer precipitation in South Africa (Landman and Mason, 1999), and surface air temperatures and precipitation amounts in eastern Asia (Hwang et al., 2001). Similarly, MCA, which was popularized as a method for analyzing coupled modes of climate variability by Bretherton et al. (1992) and Wallace et al. (1992), has since been used to predict seasonal maximum surface air temperatures in South Africa (Klopper et al., 1998), 850-hPa air temperature anomalies over the North Atlantic (SanchezGomez and OrtizBevia, 2003), and 500-hPa geopotential heights over the North Pacific (Liu et al., 2006).

As pointed out by Mason and Baddour (2008), the symmetry of methods like CCA and MCA means that the derived modes may fail to explain as much variance as possible in the dataset for which seasonal forecasts are desired. RA, which is explicitly designed to maximize explained variance in a predefined predictand dataset, is instead recommended. A similar argument has been put forward by Legendre and Legendre (1998) in the context of ecological prediction. RA was used by Wang and Zwiers (2001) as a means of post-processing predictions from a physically based seasonal forecast system.

1.5.2 Nonlinear methods

Nonlinearities in the climate system are well documented (Rial et al., 2004) and, as a result, adoption of nonlinear statistical models could result in improvements in seasonal forecast skill
over linear methods. As mentioned previously, however, time averaging may obscure nonlinear relationships and the potential exists for nonlinear models to overfit short, noisy records.

Most neural network-based seasonal forecast models have been developed for a single predictand. Notable examples include prediction of average sea surface temperature anomalies in the equatorial Pacific Ocean (Tangang et al., 1997, 1998a,b), where results relative to linear models have been mixed. For instance, Tang et al. (2000) found no significant differences in forecast skill between neural network and linear models, whereas Wu et al. (2006) found small improvements by neural networks at lead times longer than three months.

Nonlinear multivariate regression by a multivariate neural network (i.e., with multiple outputs) was used by Mishra and Desai (2006) to forecast the evolution of drought conditions in India. Comparisons were made with results from linear models and neural networks with a single predictand. Nonlinear multivariate regression outperformed the other models at lead times longer than one month. Krasnopolsky and Fox-Rabinovitz (2006) compared the ability of neural networks with univariate and multivariate predictands to emulate radiative transfer parameterizations in a coupled general circulation model used for seasonal prediction. In situations where correlations were present between predictands, the multivariate model performed best. Results are consistent with those anticipated by Caruana (1997).

NLCCA has been applied to seasonal prediction of sea surface temperatures in the tropical Pacific Ocean (Wu and Hsieh, 2002) and the Indian Ocean (Collins et al., 2004). In both cases, results were compared with CCA. Wu and Hsieh (2002) found small improvements by the nonlinear model, while Collins et al. (2004) noted the importance of nonlinearity in the development of cold anomalies in the Indian Ocean in response to La Niña events. Hsieh (2001), in an analysis of coupled ocean-atmosphere variability in the tropical Pacific, found that the neural networks used to derive the canonical variates in NLCCA are susceptible to overfitting due to lack of robustness against outliers. As a result, a spuriously large Pearson product-moment correlation between the canonical variates can occur if the mappings are allowed to become excessively nonlinear. The propensity of NLCCA to overfit led, in part, to the development of nonlinear projection, a variant of multivariate nonlinear regression with a single predictor (Wu and Hsieh, 2004).

A simple nonlinear extension of RA was developed by Makarenkov and Legendre (2002). Rather than adopting a truly nonlinear modelling paradigm, nonlinear RA (NLRA) relies instead on polynomial transformations of the predictor variables to account for nonlinear relationships. A variant of RA or PPA that is nonlinear in its parameters has yet to be developed.

### 1.6 Climate downscaling

Climate downscaling refers to a set of techniques used to predict local-scale climate variables from the synoptic-scale atmospheric circulation. The most common application of downscaling is as a tool for generating climate scenarios at high temporal and spatial resolution based on ou-
puts from global climate models (GCMs). For example, downscaling can provide meteorological inputs necessary to drive a watershed, crop, or local ecosystem model (Wilby and Wigley, 1997; Xu, 1999). These application areas require climate data to be defined at very fine spatial scales, often on the order of kilometers or at specific sites. As GCMs operate with grid spacings on the order of hundreds of kilometers, climate predictions cannot be directly used as predictors in process models. Instead, climate downscaling techniques are used to bridge the gap between the synoptic- and local-scales to provide the necessary site-specific climate information. Synoptic downscaling techniques may be physically based, for example using limited-area numerical weather prediction models, which provide dynamically consistent fields but have high computational costs, or they may be statistically based, relying instead on observed historical records to build empirical relationships between scales and variables (Hewitson and Crane, 1996).

In practice, most statistical downscaling models have focused on univariate predictions, that is predictions of a single predictand at a single site. If values of multiple predictands are required at many sites, then models are usually developed for each separately. This affords the user little control over the consistency of predictions between sites and variables, perhaps only through the specification of a common set of synoptic-scale predictors. Maintaining realistic relationships between sites and variables is particularly important in hydrological modelling. Streamflow depends strongly on the spatial distribution of precipitation in a watershed, and on interactions between temperature and precipitation that determine whether precipitation falls as rain or snow (Lindstrom et al., 1997). The development of multivariate downscaling models is thus of key concern for hydrological applications (Fowler et al., 2007).

Traditionally, approaches to statistical downscaling have been separated into three basic categories: (i) regression models, (ii) weather typing methods, and (iii) stochastic weather generators (Wilby and Wigley, 1997). Each has different strengths and weaknesses, and, in practice, hybrid methods combining aspects of each are common. As a result, it is perhaps best to think of models in the space of a ternary diagram (Figure 1.6), with each approach forming one of the points of the triangle. In the context of the multivariate statistical methods outlined in Sections 1.3 and 1.4, the main focus here is on methods falling near the top point. A review of multivariate linear and nonlinear regression-based methods, including hybrid approaches that combine aspects of weather-typing methods and stochastic weather generators, follows.

It bears noting that this review is concerned with methods that have been explicitly designed for multi-site or multivariate predictand datasets. A more general review of downscaling methods, including those designed for single sites, is given by Fowler et al. (2007).

1.6.1 Linear methods

Multivariate linear models such as CCA and MCA have been applied to downscaling tasks, in particular those involving multiple sites. For example, Huth (2002) compared the ability of five linear methods, including CCA and MCA, to downscale winter daily mean temperatures in central Europe. The spatial structure of the temperature field was best captured by
CCA. Zorita and von Storch (1999) used CCA to downscale monthly and daily winter rainfall amounts on the Iberian peninsula. They found that CCA performed well in comparison to an analog model at the monthly time scale. Due to the non-normality of daily precipitation amounts, direct application of CCA was not recommended at the shorter time scale. In an application of MCA to downscaling of 12-hr precipitation amounts in Japan, Uvo et al. (2001) applied a shifted logarithmic transformation to precipitation amounts to linearize the predictor/predictand relationship. While the timing of extreme precipitation events was well captured by the model, extremes were still underpredicted, leading the authors to recommend use of a nonlinear model.

Expanded downscaling, a multivariate linear model with similarities to CCA and MCA, was developed by Bürger (1996). Expanded downscaling extends the multivariate regression model by imposing a constraint on the model parameters such that modelled covariances between sites are forced to match observed covariances. The resulting model is thus able to accurately simulate inter-site relationships. As pointed out by Bürger (2002), linearizing transformations are necessary when applying expanded downscaling to non-normal variables, such as daily precipitation amounts.

1.6.2 Nonlinear methods

CCA, MCA, and expanded downscaling are all linear methods and, as mentioned before, may fail when nonlinear relationships are present. Downscaling of precipitation amounts on sub-
monthly time scales, in particular, may benefit from the application of nonlinear models (Yuval and Hsieh, 2002). In general, precipitation is difficult to model on short time scales due to temporal and spatial intermittency, along with its highly non-normal distribution.

Most applications of neural networks to downscaling have focused on single sites, with performance relative to linear methods depending on the variable of interest. For instance, Schoof and Pryor (2001) found that neural network models for daily temperature performed better than their linear counterparts, but that neural networks for daily precipitation performed poorly when dry days at a site were modelled at the same time as wet days. Sequential neural network models for precipitation occurrence and wet-day precipitation amounts can be built to avoid this problem (Olsson et al., 2004). A different approach was adopted by Williams (1998), who successfully described seasonal variations in daily precipitation using a neural network to conditionally model parameters of a mixed Bernoulli-gamma distribution. Haylock et al. (2006) used the same model formulation to downscale precipitation at single sites in the United Kingdom. The Bernoulli-gamma distribution can be fit to precipitation series that include both dry and wet days, which means that precipitation occurrence and wet-day precipitation amounts can be specified by the same model. A multivariate extension of the Bernoulli-gamma conditional density neural network has yet to be developed.

In fact, few studies have applied neural networks with multiple outputs to multivariate downscaling tasks. One notable exception is the study by Harpham and Wilby (2005), in which multivariate neural networks were developed for multi-site downscaling of daily precipitation occurrences and amounts. The multivariate neural networks captured the influence of the synoptic-scale forcing, leading to reasonable performance for precipitation occurrence and averaged accumulations, but, at the same time, they overestimated inter-site correlations of daily precipitation amounts. Additional constraints, following expanded downscaling (Bürger, 1996), may thus be needed to ensure that spatial relationships in multivariate nonlinear regression models are accounted for in a realistic manner. A complementary method for dealing with temporal and spatial intermittency would also be required.

### 1.6.3 Hybrid methods

The most common form of weather typing scheme used in multivariate downscaling is the analog model (Zorita and von Storch, 1999), a non-parametric method more commonly referred to in statistics as the $k$-nearest neighbour model (Hastie et al., 2001). The analog model can be thought of in terms of conditional resampling, whereby predictand samples from the historical record that are deemed “close” in terms of some predictor dataset are chosen to make predictions for new predictor samples. The fundamental concern in analog modelling is how to best measure “closeness” between predictor samples. Linear models, including regression and CCA, have been used to define distance measures for analog downscaling, thereby creating hybrid regression-weather typing downscaling methods (Fernandez and Saenz, 2003; Wilby et al., 2003). Nonlinear regression models have not yet been used for this purpose.
Chapter 1. Introduction

Stochastic weather generators are simulation models whose goal is to simulate weather element series such that the statistical properties of the randomly generated series match those of observations (Wilks and Wilby, 1999). In the context of regression-based downscaling, weather generators can be used as a means of simulating regression residuals. The variance of predictions from a regression-based downscaling model will be smaller than the variance of the observed series, due in part to the influence of small scale phenomena that are not represented in the regression model. The addition of random noise is thus required for the variance of the predicted series to match that of observations (von Storch, 1999). For instance, Bürger and Chen (2005) used a multivariate noise model to supplement predictions from a multivariate linear regression model. Another instance of a hybrid regression-weather generator model can be found in Cawley et al. (2007), where a Bernoulli-gamma conditional density network was used as a generator model for randomly simulating precipitation series. Again, a multivariate extension of this method has yet to be developed.

1.7 Research goals

This dissertation develops multivariate statistical models that are suitable for seasonal forecasting and downscaling of climate variables. In the case of seasonal climate forecasting, where record lengths are typically short and signal-to-noise ratios are low, particularly at long lead-times, forecast models must be robust against noise. From a statistical standpoint, this means that parameter estimates should not be unduly affected by outliers. To this end, two new nonlinear models are developed. (i) Based on the observation by Hsieh (2001) concerning the propensity of NLCCA to overfit in the presence of outliers, a more robust model is introduced that adds robust cost functions to the existing NLCCA architecture. (ii) To date, a truly nonlinear variant of RA or PPA has not been developed. To fill this gap, NLPPA, a nonlinear extension of PPA, is introduced as a nonlinear model of lesser complexity than NLCCA.

In the case of climate downscaling, models must not only be accurate in terms of traditional model verification statistics, but they must also be able to replicate statistical properties of the historical observations. Three models are developed for multi-site downscaling. (i) First, nonlinear analog predictor analysis (NLAPA), a hybrid model that couples a neural network to an analog model, is developed as a means of nonlinearly extending methods by Fernandez and Saenz (2003) and Wilby et al. (2003). The neural network maps the original predictors to a lower-dimensional space such that predictions from the analog model are improved. (ii) Next, a multivariate ridge regression model that allows negative values of the ridge parameters is introduced as an alternative means of performing expanded downscaling (Bürger, 1996). (iii) Finally, the expanded Bernoulli-gamma density network (EBDN) is introduced for multi-site precipitation downscaling. The EBDN model extends the single-site neural network model of Williams (1998) by allowing multiple predictands and by adopting the covariance constraint from expanded downscaling.
Bibliography


Bibliography


Chapter 2

Robust nonlinear canonical correlation analysis

2.1 Introduction

Canonical correlation analysis (CCA) is a multivariate linear model used to find the modes of maximum correlation between two sets of variables (von Storch and Zwiers, 1999). CCA was first popularized as a tool for prediction in the atmospheric sciences by Glahn (1968) and has since been used extensively in climatology, particularly for seasonal forecasting (Barnett and Preisendorfer, 1987; Barnston and Ropelewski, 1992; Shabbar and Barnston, 1996).

CCA is a linear model and is thus unable to describe nonlinear relationships between datasets. To nonlinearly generalize CCA, various approaches, based on artificial neural network and kernel methods, have been proposed (Hsieh, 2000; Lai and Fyfe, 1999, 2000; Melzer et al., 2003; Shawe-Taylor and Cristianini, 2004; Suykens et al., 2002). For instance, Hsieh (2000) used three feed-forward (multi-layer perceptron) neural network mappings to perform nonlinear CCA (NLCCA). This method has been applied to climate research, for analyzing the structure of the El Niño-Southern Oscillation (ENSO) (Hsieh, 2001; Wu and Hsieh, 2002) and its interdecadal changes (Wu and Hsieh, 2003), and for determining the midlatitude atmospheric response to tropical Pacific sea surface temperature (SST) variability (Wu et al., 2003). Operational NLCCA forecasts of SST in the equatorial Pacific Ocean are also made available by the Climate Prediction Group of the University of British Columbia (see http://www.ocgy.ubc.ca/projects/clim.pred/ for more details).

While able to describe coupled nonlinear variability, this rather complicated NLCCA model is prone to overfitting (i.e., fitting to the noise rather than the signal), particularly when applied to the short, noisy datasets common in climate studies. This prompted the development of simpler multivariate nonlinear models such as nonlinear projection (Wu and Hsieh, 2004), which maps a univariate predictor to a multivariate predictand dataset, and nonlinear principal predictor analysis (Cannon, 2006), which maps a multivariate predictor dataset to a multivariate

Chapter 2. Robust nonlinear canonical correlation analysis

predictand dataset. While mitigating the influence of short, noisy datasets on model overfitting by reducing the number of neural networks in the model, neither nonlinear projection nor nonlinear principal predictor analysis are as general as NLCCA.

The main goal of this paper is the development of a robust version of NLCCA that can successfully operate on datasets with low signal-to-noise-ratios. The basic model architecture chosen by Hsieh (2000) is kept intact. Instead, the cost functions used to set the model parameters are replaced with more robust versions. A cost function based on the biweight mid-correlation replaces one based on the Pearson (product-moment) correlation and cost functions based on the $L_1$-norm (i.e., mean absolute error, mae) replace ones based on the $L_2$-norm (i.e., mean squared error, mse). Robust variants of NLCCA are demonstrated on a synthetic dataset and are used to forecast SSTs in the tropical Pacific Ocean based on sea level pressure (SLP) data.

### 2.2 Method

#### 2.2.1 NLCCA

Consider a dataset \( \{x_i(t)\} \) with \( i \) variables and another dataset \( \{y_j(t)\} \) with \( j \) variables, where each dataset has \( t = 1, ..., N \) samples. The variables \( \{x_i(t)\} \) can be grouped to form the vector \( \mathbf{x}(t) \) and the variables \( \{y_j(t)\} \) can be grouped to form the vector \( \mathbf{y}(t) \). CCA looks for the linear combinations

\[
  u(t) = \mathbf{a} \cdot \mathbf{x}(t), \quad v(t) = \mathbf{b} \cdot \mathbf{y}(t) \tag{2.1}
\]

such that the Pearson correlation between the canonical variates \( u \) and \( v \), i.e., \( \text{cor}(u, v) \), is maximized. If, for example, \( \mathbf{x} \) is a gridded SLP dataset and \( \mathbf{y} \) is a gridded SST dataset, then the vectors \( \mathbf{a} \) and \( \mathbf{b} \) represent correlated spatial patterns corresponding to the SLP and SST fields respectively. Unlike linear regression, which looks for relationships between a predictor dataset (e.g., \( \mathbf{x} \)) and each of the predictands (e.g., \( \mathbf{y}_j \)) separately, CCA takes a holistic approach and looks for relationships between each of the sets of variables in their entirety. No distinction is made between the two fields; each can act interchangeably as predictors or predictands.

In NLCCA, the nonlinear analog of linear CCA, the linear mappings in Eq. 2.1 are replaced with nonlinear mappings performed by neural networks. The neural network architecture for NLCCA is shown in Figure 2.1. The double-barreled network on the left-hand side nonlinearly maps \( \mathbf{x} \) to \( u \) and \( \mathbf{y} \) to \( v \) by

\[
  h_k^{(x)} = \tanh [(W^{(x)} x + b^{(x)})_k], \quad u = \tilde{w}^{(x)} \cdot h^{(x)} + \tilde{b}^{(x)}
\]

\[
  h_l^{(y)} = \tanh [(W^{(y)} y + b^{(y)})_l], \quad v = \tilde{w}^{(y)} \cdot h^{(y)} + \tilde{b}^{(y)}
\]

\( \tag{2.2} \)
Chapter 2. Robust nonlinear canonical correlation analysis

Figure 2.1: Neural network architecture used to perform NLCCA.
where \( h_k^{(x)} \) and \( h_l^{(y)} \) are the hidden-layer nodes; \( \tanh(\cdot) \) is the hyperbolic tangent function; \( W^{(x)} \) and \( W^{(y)} \) are the hidden-layer weight matrices; \( b^{(x)} \) and \( b^{(y)} \) are the hidden-layer bias vectors; \( \tilde{w}^{(x)} \) and \( \tilde{w}^{(y)} \) are the output-layer weight vectors; \( \tilde{b}^{(x)} \) and \( \tilde{b}^{(y)} \) are the output-layer biases; and \( k \) and \( l \) are indices of the vector elements. The number of hidden-layer nodes controls the overall complexity of the network; the hidden-layer must contain more than one node (\( k = 1, ..., K, K \geq 2 \) and \( l = 1, ..., L, L \geq 2 \)) to obtain a nonlinear solution (Hsieh, 2001).

Weight and bias parameters in the double-barreled network are set by minimizing the cost function

\[
C_1 = -\operatorname{cor}(u, v) + \langle u \rangle^2 + \langle v \rangle^2 + \left( \left\langle u^2 \right\rangle^{\frac{1}{2}} - 1 \right)^2 + \left( \left\langle v^2 \right\rangle^{\frac{1}{2}} - 1 \right)^2 + P_1 \left[ \sum_k \left( W_{k}^{(x)} \right)^2 + \sum_l \left( W_{l}^{(y)} \right)^2 \right]
\] (2.3)

where \( \langle \cdot \rangle \) denotes the sample or temporal mean. The first term maximizes the correlation between the canonical variates \( u \) and \( v \); the second, third, fourth, and fifth terms are normalization constraints that force \( u \) and \( v \) to have zero mean and unit variance; the sixth term is a weight penalty whose relative magnitude is controlled by the parameter \( P_1 \). Larger values of \( P_1 \) lead to smaller weights (i.e., fewer effective model parameters), which results in a more linear model. If \( \tanh(\cdot) \) is replaced by the identity function, then Eq. 2.2 reduces to Eq. 2.1 and the network performs linear CCA.

Once the canonical variates \( u \) and \( v \) have been found, the inverse mappings to \( \hat{x} \) and \( \hat{y} \) are given by the two neural networks on the right-hand side of Figure 2.1:

\[
h_k^{(u)} = \tanh[\{w^{(u)}u + b^{(u)}\}_k], \quad \hat{x} = \tilde{W}^{(u)}h^{(u)} + \tilde{b}^{(u)}
\] (2.4)

\[
h_l^{(v)} = \tanh[\{w^{(v)}v + b^{(v)}\}_l], \quad \hat{y} = \tilde{W}^{(v)}h^{(v)} + \tilde{b}^{(v)}.
\] (2.5)

Weight and bias parameters in these two networks are found by minimizing the cost functions

\[
C_2 = \left\langle \| \hat{x} - x \|^2 \right\rangle + P_2 \sum_k \left( W_{k}^{(x)} \right)^2
\] (2.6)

\[
C_3 = \left\langle \| \hat{y} - y \|^2 \right\rangle + P_3 \sum_l \left( W_{l}^{(y)} \right)^2
\] (2.7)

respectively, where \( \| \cdot \|^2 \) is the square of the \( L_2 \)-norm, with the \( L_p \)-norm given by

\[
L_p(e) = (\| e \|^p)^{1/p} = \left( \sum_i |e_i|^p \right)^{1/p}.
\] (2.8)

\( C_2 \) and \( C_3 \) thus give the mse between the neural network predictions and the observed \( x \) and \( y \).
variables subject to weight penalty terms whose magnitudes are controlled by the parameters \( P_2 \) and \( P_3 \). Once the first mode has been extracted from the data, the next leading mode can be extracted from the model residuals, and so on for higher modes.

For seasonal climate prediction tasks, where the goal is to predict values of a multivariate predictand dataset from a multivariate predictor dataset, e.g., \( \hat{y} = f(x) \), values of the canonical variate \( v \) must be predicted from values of the canonical variate \( u \). For canonical variates normalized to unit variance and zero mean, the linear least-squares regression solution is given by

\[
\hat{v} = u \text{cor}(u, v) \quad (2.9)
\]


### 2.2.2 Biweight midcorrelation

The Pearson correlation is not a robust measure of association between two variables, as its estimates can be affected by the presence of a single outlier (Wilcox, 2004). For short, noisy datasets the cost function \( C_1 \) [Eq. 2.3] in the NLCCA model may lead to overfitting as the model attempts to maximize the correlation between the canonical variates by generating mappings between \( x \) and \( u \) and \( y \) and \( v \) that are more complicated than necessary due to outliers. Rather than using the Pearson correlation, a more robust measure of association could instead be adopted in the cost function to avoid this problem.

Robust correlation coefficients, including commonly used functions like the Spearman rank correlation, are reviewed by Wilcox (2004). Trials with the Spearman rank correlation resulted in models with poor convergence properties and inconsistent performance on real-world climate datasets. Instead, the biweight midcorrelation (Wilcox, 2004) was selected as a robust alternative to the Pearson correlation. The biweight midcorrelation is calculated in the same manner as the Pearson correlation coefficient, except non-robust measures (the mean, expected deviation from the mean, and covariance) are replaced by robust measures. The biweight midcorrelation can also be used to predict \( v \) from \( u \) (and vice versa) in a manner similar to Eq. 2.9 for the standard NLCCA model, which is not possible with the Spearman rank correlation.

To calculate the biweight midcorrelation function \( \text{bicor}(x, y) \), first rescale \( x \) and \( y \) by

\[
p = \frac{x - M_x}{9 \text{MAD}_x}, \quad q = \frac{y - M_y}{9 \text{MAD}_y}
\]

(2.10)

where \( M_x \) and \( M_y \) are the median values of \( x \) and \( y \) respectively and \( \text{MAD}_x \) and \( \text{MAD}_y \) are the median values of \( |x - M_x| \) and \( |y - M_y| \) respectively. Next, the sample biweight midcovariance is given by

\[
\text{bicov}(x, y) = \frac{N \sum_t a(t)b(t)c(t)^2d(t)^2(x(t) - M_x)(y(t) - M_y)}{(\sum_t a(t)c(t)(1 - 5p(t)^2))((\sum_t b(t)d(t)(1 - 5q(t)^2))}
\]

(2.11)
Chapter 2. Robust nonlinear canonical correlation analysis

Figure 2.2: Empirical comparison between the Pearson correlation (cor) and the biweight midcorrelation (bicor) on random variables $x$ and $y$, each with samples drawn from a standard normal distribution, and $x'$ and $y$, where $x'$ is the same as $x$ but with one case replaced with an outlier. Panel (a) shows sample time series of $x'$ (solid) and $y$ (dashed); (b) compares $\text{cor}(x, y)$ and $\text{bicor}(x, y)$; (c) compares $\text{cor}(x, y)$ and $\text{cor}(x', y)$; and (d) compares $\text{bicor}(x, y)$ and $\text{bicor}(x', y)$. Plots are for 1000 randomly generated datasets.

where $a(t) = 1$ if $-1 \leq p(t) \leq 1$, otherwise $a(t) = 0$; $b(t) = 1$ if $-1 \leq q(t) \leq 1$, otherwise $b(t) = 0$; $c(t) = (1-p(t))^2$; and $d(t) = (1-q(t))^2$. The biweight midcorrelation is then given by

$$\text{bicor}(x, y) = \frac{\text{bicov}(x, y)}{\sqrt{\text{bicov}(x, x) \text{bicov}(y, y)}}.$$  \hspace{1cm} (2.12)

The biweight midcorrelation, like the Pearson correlation, ranges from -1 (negative association) to +1 (positive association).

Figure 2.2 shows estimates of the Pearson correlation and the biweight midcorrelation between normally distributed random variables $x \sim N(0, 1)$ and $y \sim N(0, 1)$ and between $x'$ and $y$, where $x'$ is the same as $x$ but with one case replaced by an outlier (Figure 2.2a). On the outlier-free dataset, both $\text{cor}(x, y)$ and $\text{bicor}(x, y)$ give approximately equal estimates of the strength of association between the variables (Figure 2.2b). Estimates of $\text{cor}(x', y)$ are strongly affected...
by the outlier, showing almost no association between values calculated with and without the
outlying data point (Figure 2.2c), whereas estimates of bicor\(x', y\) are essentially unaffected
by the outlier (Figure 2.2d).

NLCCA with the Pearson correlation cost function may fail when outliers occur simulta-
neously in both datasets. To illustrate, consider two identical sinusoidal series, each with a
common outlier

\[ x(t) = y(t) = \sin(0.5t) + \delta(t), \quad \text{where} \]
\[ \delta(t) = \begin{cases} 
6 & \text{at } t = 150 \\
0 & \text{elsewhere}
\end{cases} \]

where \(t = 1, 2, ..., 300\). Next, create new series \(x'\) and \(y'\) by adding noise drawn from \(N(0, 0.5)\)
to \(x\) and \(y\). The expected values of cor\((x', y')\) and bicor\((x', y')\) are found to be 0.71 and 0.68
respectively. Now, consider values of cor\((u, v)\) and bicor\((u, v)\), where \(u = x'^p, v = y'^p\), and \(p\)
is an odd integer (Figure 2.3a). Increasing the value of \(p\) effectively increases the separation
between the outlier and the non-outliers (Figure 2.3b). Values of cor\((u, v)\) and bicor\((u, v)\) for
values of \(p\) from 1 to 9 are shown in Figure 2.3c. The Pearson correlation can be increased
simply by increasing \(p\), whereas the biweight midcorrelation decreases as \(p\) increases. This case
illustrates how increasing the nonlinearity of the mapping functions \(u\) and \(v\) (by increasing \(p\))
can lead to very high Pearson correlation.

In the context of NLCCA, spuriously high values of cor\((u, v)\) can be found by the double-
barreled network when the nonlinear neural network mapping greatly magnifies an outlier in
both \(x\) and \(y\). This artifact can be particularly dangerous when NLCCA is applied to datasets
that are affected by strong, concurrent climate signals, for example those with large El Niño
or La Niña anomalies, as shown by Hsieh (2001). NLCCA performed worse than CCA when
weight penalty terms were not used to reduce the nonlinearity of the double-barreled network.
Based on results shown in Figure 2.3, adopting bicor in the cost function should prevent this
artifact.

When NLCCA models are used for multivariate prediction, a regression model is needed
to estimate \(v\) from \(u\) (and vice versa). For the standard NLCCA model, the linear least
squares estimate for the regression coefficient is given by Eq. 2.9. Similarly, the biweight
midcorrelation is associated with a robust regression model that can be used to predict values
of one canonical variate from the other. Following Lax (1985) and Hou and Koh (2004), the
biweight midregression solution is given by

\[ \hat{v} = u \text{ bicor}(u, v) \]  \( (2.14) \)

for canonical variates normalized to unit variance and zero mean.
Figure 2.3: Time series of $u$ with (a) $p = 1$ and (b) $p = 9$; outliers in $v$ occur at the same time as those in $u$. (c) The effect on $\text{cor}(u, v)$ (solid line) and $\text{bicor}(u, v)$ (dashed line) from increasing the separation between common outlier and non-outlier points by increasing $p$. 
2.2.3 $L_p$-norm

Now consider the inverse mapping from $u$ and $v$ back to $\hat{x}$ and $\hat{y}$ (i.e., the networks on the right hand side of Figure 2.1). The $L_p$-norm given in Eq. 2.8 forms the basis for a class of cost functions used in regression models (Bishop, 1995). Of these, the $L_2$-norm, which leads to the mse cost function, is commonly used in statistical models. Models that minimize the mse are optimal if the data are generated from a deterministic function corrupted by a normally distributed noise process with constant variance. However, a potential problem exists with cost functions based on the $L_2$-norm (e.g., $C_2$ and $C_3$ defined in Eqs. 2.6 and 2.7). Samples with the greatest errors exert disproportionately large influence on the cost function. Thus, a small number of outliers can come to dominate the solution. Adopting the $L_1$-norm, which leads to the mae cost function, reduces the influence of outliers. 

Bishop (1995, Sec. 6.1-6.2) showed that in the limit of infinite samples and with a flexible enough model (e.g., a neural network with enough hidden nodes), the model converges to the conditional mean if the mse is used and the conditional median if the mae is used. The median is robust to outliers whereas the mean is not.

2.2.4 Robust NLCCA

Robust variants of NLCCA use the model architecture shown in Figure 2.1 but with the cost functions $C_1$, $C_2$, and $C_3$ replaced by the robust versions described in Sections 2.2.2 and 2.2.3. The biweight midcorrelation replaces the Pearson correlation in $C_1$ and the mae replaces the mse in $C_2$ and $C_3$.

2.3 Synthetic test problem

2.3.1 Data

To illustrate the effect of the changes to the NLCCA cost functions, consider the three dimensional synthetic test problem used by Hsieh (2000) to introduce the standard NLCCA model. The first correlated mode ($x$ and $y$) is given by

$$x_1 = t - 0.3t^2, \quad x_2 = t + 0.3t^2, \quad x_3 = t^2$$

$$y_1 = t^3, \quad y_2 = -t + 0.3t^3, \quad y_3 = t + 0.3t^2$$

where $t$ is a uniformly distributed random number in $[-1, 1]$. The second correlated mode ($x'$ and $y'$) is given by

$$x'_1 = -s - 0.3s^2, \quad x'_2 = s - 0.3s^3, \quad x'_3 = -s^4$$

(2.17)
Figure 2.4: Synthetic test datasets used to evaluate the performance of the standard and robust NLCCA models. The first mode is shown by the thick black curve and the second mode is shown by the thin black curve. Test samples with added noise are shown as asterisks.
Chapter 2. Robust nonlinear canonical correlation analysis

\[ y_1' = \text{sech}(4s), \quad y_2' = s + 0.3s^3, \quad y_3' = s - 0.3s^2 \]  
\hspace{1cm} (2.18)

where \( s \) is a uniformly distributed random number in \([-1, 1]\). The shapes described by \( x \) and \( x' \) are shown in Figure 2.4a and those described by \( y \) and \( y' \) are shown in Figure 2.4b.

To test the performance of the NLCCA models, 50 training and test datasets, each with 500 samples, were randomly generated from Eqs. 2.15-2.18. The signal in each dataset was produced by adding the second mode to the first mode, with the variance of the second equal to one third that of the first. Normally distributed random noise with standard deviation equal to 50\% of the signal standard deviation was added to the data. The variables were then standardized to zero mean and unit standard deviation.

2.3.2 Training and testing procedure

NLCCA models with different combinations of the non-robust (cor and mse) and robust (bicor and mae) cost functions were developed on the training datasets and applied to the test datasets. Following Hsieh (2000), all neural networks had three nodes in their hidden-layers and were trained without weight penalty terms. A quasi-Newton nonlinear optimization scheme with finite-difference approximations of the gradient and Hessian was used to minimize the cost functions. While the \( L_1 \) norm is not continuously differentiable, convergence problems were not noted during optimization. The \( L_1 \) norm can, however, be approximated by the Huber norm, which is continuously differentiable, if issues with convergence are found (Panayiotis et al., 2006). To avoid local minima in the error surface, each network in Figure 2.1 was trained 30 times, each time starting from a different randomly selected set of initial weights and biases. The network with the lowest value of its associated cost function was then selected for use and applied to the test data.

2.3.3 Model performance

Root mse (rmse) values between the first synthetic mode and the first mode extracted by NLCCA models with different combinations of non-robust and robust cost functions are shown in Figure 2.5 for the 50 test datasets. On average, all models performed approximately the same, although, for the leading NLCCA mode of the \( x \) dataset, NLCCA with bicor/mse cost functions yielded the lowest median rmse (0.44), followed by NLCCA with bicor/mae (0.45) and NLCCA with cor/mse (0.45). NLCCA with cor/mae performed worst with a median rmse of 0.47. Median rmse values and relative rankings of the models were the same for the leading NLCCA mode of the \( y \) dataset.

Of the four models, NLCCA with the robust cost functions (bicor/mae) was the most stable. No trial yielded an rmse in excess of the series standard deviation of one, with the maximum value under 0.6 for the \( x \) mode. The other models had at least one trial with an rmse value greater than one, which is indicative of severe overfitting. Maximum values for the \( x \) mode
Figure 2.5: Boxplots showing the distribution of rmse between the first synthetic mode and the first mode extracted by NLCCA models for (a) x and (b) y with different combinations of non-robust and robust cost functions over 50 trials. Boxes extend from the 25th to 75th percentiles, with the line indicating the median. Whiskers represent the most extreme data within ±1.5 times the interquartile range (i.e., the box height); values outside this range are plotted as dots. The dashed line indicates a rmse equal to one. The ordinate is log-scaled to accommodate the large range in rmse.
Chapter 2. Robust nonlinear canonical correlation analysis

ranged from 1.8 for NLCCA with bicor/mse, to 47.4 for NLCCA with cor/mse, and 49.6 for cor/mae. NLCCA with bicor/mae performed similarly for the y mode, although two trials with rmse greater than 20 were found for NLCCA with bicor/mse cost functions.

Overall, results for the synthetic dataset suggest that replacing the cor/mse cost functions in NLCCA with bicor/mae cost functions leads to a more stable model that was less susceptible to overfitting and poor test performance. All models were run without weight penalty terms in this comparison. In practice, the non-robust models will need weight penalty terms to reduce overfitting, as is done in our next test, where NLCCA models are applied to a real-world climate prediction problem.

2.4 Seasonal prediction

2.4.1 Data

As the primary goal of the study is to investigate the effect of the robust cost functions on performance, and not to build an operational forecast model, predictor/predictand fields were selected as in Hsieh (2001).

SST data were obtained from the second version of the NOAA Extended Reconstructed SST (ERSST.v2) dataset (Smith and Reynolds, 2004). Monthly data on a 2° × 2° grid were extracted for a spatial domain covering the tropical Pacific Ocean (22°S–22°N, 122°E–288°E) for the time period 1948 to 2003. The climatological seasonal cycle was removed, data were detrended, and a 3-month running mean filter was applied. Principal component analysis (PCA) was applied to the data; the first 6 modes accounting for 73% of the total SST variance were retained for further analysis. PC scores (i.e., the time series for the leading PCA modes) were scaled according to the amount of variance explained by each mode.

SLP data from the NCEP/NCAR Reanalysis (Kalnay et al., 1996) were obtained for the same region and period. Data on a 2.5° × 2.5° grid had the climatological seasonal cycle removed, the data were detrended, and then smoothed by a 3-month running mean filter. PCA was applied to the data and the first 6 modes, accounting for 80% of the total SLP variance, were retained for further analysis.

2.4.2 Training and testing procedure

Three variants of the NLCCA model were applied to the SST and SLP datasets. The first, representing the standard NLCCA model, incorporated both non-robust cost functions (cor/mse). The second and third used the bicor cost function to train the double-barreled network and either the mae or mse cost function to train the inverse mapping networks. For brevity, the model with cor/mae cost functions was dropped from consideration.

To assess the usefulness of the three variants of NLCCA for seasonal forecasting, models were validated on the basis of their forecast performance. PC scores from the 6 leading PCs of the SLP dataset were used to predict PC scores from the 6 leadings PCs of the SST dataset at
lead times of 0, 3, 6, 9, and 12-months. (Lead times are defined as the number of months from the predictor observation to the predictand observation, e.g., a forecast with a 3-month lead time from January would be for April.) Taking $x$ to be historical values of the SLP PC scores and $y$ to be historical values of the SST PC scores, forecasts for a new case $\hat{y}(t_n)$ at time $t_n$ were made as follows. First, the double-barreled network was trained with $x$ and $y$ as inputs and the resulting values of $u$ and $v$ were used to train the inverse mapping networks. Given a new SLP data point $x(t_n)$, a new value of the canonical variate $u(t_n)$ was obtained from the double-barreled network. Regression equations [Eq. 2.9 or Eq. 2.14] were then used to predict a new value of $\hat{v}(t_n)$. Finally, $\hat{v}(t_n)$ was entered into the appropriate inverse mapping network to give $\hat{y}(t_n)$. For the second and higher NLCCA modes, the same procedure was followed using residuals from the previous mode as inputs.

Following Hsieh (2001), neural networks were trained both with and without weight penalty terms using two hidden-layer nodes. A two-stage cross-validation procedure was used to set the weight penalty coefficients and to estimate forecast performance. For reference, a schematic diagram showing how data were split into training/validation segments is shown in Figure 2.6.

To avoid overfitting in models trained with weight penalty, values of the coefficients $P_1$, $P_2$, and $P_3$ in Eqs. 2.3, 2.6, and 2.7 were determined via 10-fold cross-validation on the training dataset (CV1 in Figure 2.6). The training record was split into 10 contiguous segments. Models were trained on 9 of the 10 segments using weight penalties from the set $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Forecasts on the remaining segment were then recorded for each weight penalty coefficient. While fixing the weight penalties, these steps were repeated 9 times, each time making forecasts on a different segment. Finally, forecasts for all 10 segments were combined and validated against observations. Weight penalties that minimized the aggregated cross-validation error were recorded, neural networks were retrained on all 10 segments combined using these penalties. Ten models were trained in this manner to assess sensitivity to initial weights and biases.

A second round of cross-validation was used to estimate out-of-sample forecast performance of the models (CV2 in Figure 2.6). The historical record was split into 5 contiguous segments (each approximately 11 years in length). Models were trained on 4 of the 5 segments using
Chapter 2. Robust nonlinear canonical correlation analysis

Figure 2.7: Cross-validated correlation skill for NLCCA models trained with cor/mse, bicor/mse, and bicor/mae cost functions. Weight penalty was applied to the model denoted cor/mse(p). Bars show the mean correlation over the spatial domain, averaged over the 10 trials. Vertical lines extend from the minimum to the maximum spatial mean correlation from the 10 trials. Horizontal lines show correlation skill from the CCA model. The ordinate is limited to showing positive cross-validated skill.

the cross-validation procedure outlined above. Forecasts on the remaining segment were then recorded. These steps were repeated 4 times, each time making forecasts on a different segment. Finally, forecasts for all 5 segments were combined and compared with observations.

2.4.3 Skill for models with one mode

Results from NLCCA models with one mode are shown in Figure 2.7. For reference, results from linear CCA models are also shown. Cross-validated Pearson correlation skill is averaged over the entire domain following reconstruction of the SST field from the predicted SST PC scores. Values of rmse were also calculated, but are not shown as relative performance between models was the same as for correlation skill. Results with weight penalty are only given for the NLCCA model with cor/mse cost functions as the addition of penalty terms to models with the bicor cost function did not generally lead to significant changes in skill.

Without weight penalty, the NLCCA model with cor/mse cost functions performed poorly, exhibiting mean skills worse than CCA at all lead times. Even with concurrent predictor/predictand
fields, the mean correlation skill was lower than 0.2. NLCCA with bicor/mse cost functions and bicor/mae cost functions performed much better, with mean correlation skills exceeding 0.5 at the 0-month lead time. Over the 10 trials, minimum skills from models incorporating the bicor cost function were higher than maximum skills from the corresponding cor/mse models without weight penalty.

For NLCCA with cor/mse cost functions, minimum correlations were lower than zero (i.e., no cross-validation skill) for 6, 9, and 12-month lead times. All NLCCA models with bicor/mse and bicor/mae cost functions, even those at a 12-month lead time, showed positive skill. In general, NLCCA models with bicor exhibited the least variability in skill between repeated trials. In no case was the range between minimum and maximum skill greater than 0.2. For NLCCA with cor/mse cost functions, the range in skill exceeded 0.2 at all lead times, indicating a very unstable model.

Little difference in skill was evident between bicor/mse and bicor/mae models, which suggests that the switch from cor to bicor in the double-barreled network cost function was responsible for most of the increase in skill relative to the standard NLCCA model. Inspection of the canonical variates shows that this was due to the insensitivity of the bicor cost function to the common outlier artifact described in Section 2.2.2 and illustrated in Figure 2.3.

Plots of the canonical variates $u$ and $v$ for the first mode of NLCCA models with cor/mse and bicor/mse cost functions at the 0-month lead time are shown in Figure 2.8 along with PC scores from the leading SST and SLP PCA modes. For these series, values of $\text{cor}(u, v)$ and $\text{bicor}(u, v)$ were 1.00 and 0.96 respectively. The high correlation between $u$ and $v$ for the NLCCA model with the cor cost function was driven almost exclusively by the common outliers present during 1997-1998. With the 1997-1998 outliers removed, $\text{cor}(u, v)$ dropped to 0.28. On the other hand, the high correlation between $u$ and $v$ for the NLCCA model with the bicor cost function was indicative of the strong relationship between the SLP and SST series, as evidenced by the Pearson correlation of 0.91 between the leading SST and SLP PCs.

Results discussed to this point have been for NLCCA models without weight penalty. Hsieh (2001) found that the addition of weight penalty to the standard NLCCA model led to improvements in performance, due in part to the avoidance of the common outlier artifact. Addition of weight penalty to the standard NLCCA model resulted in improvements in mean correlation skill, although performance still lagged behind NLCCA with the bicor cost function at 9 and 12-month lead times. At 0, 3, and 6-month lead times, maximum skill over the 10 trials did, however, exceed the mean level of skill of the bicor-based models, which suggests that an appropriate amount of weight penalty can result in a good performing model. Inspection of the time series of $u$ and $v$ for the best performing runs suggests that improved performance was due to avoidance of the common outlier artifact. However, the wide range in performance over the 10 trials (e.g., at 0 and 6-month lead times) reflects the instability of the training and cross-validation steps needed to choose the weight penalty coefficients. In practice, it may be difficult to consistently reach the performance level of the robust model by relying solely on
Figure 2.8: Plots of (a) PC scores from the leading SST (solid line) and SLP (dashed line) PCA modes; (b) the canonical variate $u$ for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and one with bicor/mse cost functions (solid line); and (c) canonical variate $v$ for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and bicor/mse cost functions (solid line).
weight penalty to control overfitting of the standard NLCCA model.

Returning to the NLCCA models with bicor/mse and bicor/mae cost functions, little difference in skill between the models is apparent from Figure 2.7. At short lead times (0 and 3-months), when the signal is strongest, the bicor/mse model performed slightly better than the bicor/mae model, whereas at the longest lead time (12-months), when the signal is weakest, the bicor/mae model performed best (and with less run-to-run variability in skill).

NLCCA models with the bicor/mse and bicor/mae cost functions tended to perform slightly better than CCA. For the bicor/mae model, the small improvement in performance was significant (i.e., minimum skill over the 10 trials exceeded CCA skill) at 0, 3, 6, and 12-month lead times, while the same was true of the bicor/mse model at 0 and 3-month lead times.

To investigate the differences between the linear and nonlinear models, plots of the first CCA and NLCCA (bicor/mae) SST modes projected onto the PC1-PC2 and PC1-PC3 planes are shown in Figure 2.9a for one of the ensemble models. Spatial loading patterns associated with each PC are shown in Figures 2.9b-d. For the NLCCA mode at short lead times, a quadratic response was present in the PC1-PC2 plane. Negative values of PC2 occurred when values of PC1 were both negative and positive, which, from the spatial loading patterns, means that the predicted SST response at the minimum/maximum values of $v$ (which, at left/right extremes of the curve, correspond to La Niña/El Niño states respectively) exhibited east/west asymmetry. The curve rotated counter clockwise and straightened with increasing lead time. At short lead times, the leading CCA mode was driven mainly by PC1.

Conversely, the NLCCA curve in the PC1-PC3 plane displayed increased nonlinearity with lead time. Predicted values of PC3 were typically positive when values of PC1 were both negative and positive, which, from the spatial loading pattern of PC3, indicates differences in the contrast between predicted SST anomalies along the equator and off the equator during La Niña and El Niño states. Observed asymmetry in the spatial patterns and magnitudes of SST anomalies associated with La Niña and El Niño are present in the observational record and have previously been detected by nonlinear methods (Hsieh, 2001; Monahan, 2001; Wu and Hsieh, 2002).

To this point, reported skills have been averaged over the entire spatial domain. For reference, Figure 2.10 shows spatial patterns of ensemble averaged correlation skill for NLCCA models with bicor/mae cost functions at lead times of 0 and 12-months respectively. For comparison, correlation skills from CCA models are also plotted. Spatially at 0-month lead time, skill was highest in the central equatorial Pacific Ocean, with a secondary maximum to the northeast of Papua New Guinea and east of Australia. Somewhat similar spatial patterns are seen at the other lead times. Differences in skill between NLCCA and CCA are generally small, with the largest improvements by NLCCA occurring along the boundary separating the two skill maxima.
Figure 2.9: (a) Plots of the first SST mode for CCA (thin line) and NLCCA with bicor/mae cost functions (thick line) in the planes of the PC1-PC2 and PC1-PC3 scores at 0, 3, 6, 9, and 12-month lead times. Spatial patterns for (b) PC1, (c) PC2, and (d) PC3, all normalized to unit norm.
Chapter 2. Robust nonlinear canonical correlation analysis

Figure 2.10: Spatial correlation skill at 0-month lead time for (a) NLCCA with bicor/mae cost functions and (b) CCA. Panel (c) shows NLCCA skill minus CCA skill. Panels (d) to (f) as in (a) to (c) but for 12-month lead time.
2.4.4 Skill for models with two modes

Results reported in the previous section were from NLCCA models with a single nonlinear mode. Inclusion of the second NLCCA mode may improve forecast performance in the tropical Pacific Ocean (Wu and Hsieh, 2002). To investigate, results from NLCCA models with two modes are shown in Figure 2.11.

Model skill with two modes improved relative to NLCCA with a single mode at short lead times. For instance, mean correlation skill for the NLCCA model with bicor/mae went from 0.55 with one mode to 0.65 with two modes at a 0-month lead time, and from 0.52 to 0.59 at a 3-month lead time. At longer lead times performance dropped, even to a level below CCA at 6-months, which is indicative of overfitting. However, the same was also true of the CCA model where, at 9 and 12-month lead times, skill decreased relative to the model with a single mode. Results are somewhat at odds with those reported by Wu and Hsieh (2002), who found the largest improvements in model performance to occur at longer lead times. However, cross-validation was not employed by Wu and Hsieh (2002), which means that overfitting may have caused inflated skill estimates at these lead times.

As pointed out by Hsieh (2001), nonlinearity in the tropical Pacific is strongest in the leading SST mode and is much weaker (or even not evident) in higher modes. As a result, using a nonlinear model, even one that can be estimated robustly, to extract the second or
higher modes may not be warranted and could lead to poor forecast performance. When the skill improvement of NLCCA over CCA is minimal, as is the case here even at short lead times, it may be more appropriate to apply CCA to residuals from the first NLCCA mode. This approach is currently used in operational NLCCA forecast models run by the Climate Prediction Group at the University of British Columbia (A. Wu, 2007, personal communication).

2.5 Conclusion

NLCCA based on multi-layer perceptron neural networks is a flexible model capable of non-linearly generalizing linear CCA (Hsieh, 2000). However, the complicated model architecture and use of non-robust cost functions means that overfitting is difficult to avoid, particularly when dealing with the short, noisy datasets that are common in seasonal climate forecasting problems. To make NLCCA more robust, non-robust cost functions in the model are replaced by robust cost functions: the Pearson correlation in the double-barreled network is replaced by the biweight midcorrelation, while the mse in the inverse mapping network can be replaced by the mae.

Through analysis of a synthetic dataset and a real-world climate dataset, adoption of the biweight midcorrelation is shown to result in large improvements in model stability, mainly by avoiding the common outlier artifact noted by Hsieh (2001). Replacing the mse by the mae leads to improved performance on the synthetic dataset, but little improvement on the climate dataset, except at the longest lead time where the signal-to-noise ratio is smallest.

Based on these results, it is recommended that the biweight midcorrelation replace the Pearson correlation in the NLCCA model. Choosing the mse or mae cost function appears to be more problem dependent, and should be considered as part of the model selection process. Other cost functions, for example those based on the $L_p$ norm with $1 < p < 2$ (Hanson and Burr, 1988), might also be viable, depending on the prediction task. More research is needed to determine the most appropriate cost function for the inverse mapping networks.

Development of a robust NLCCA model for operational prediction of SSTs in the equatorial Pacific Ocean is currently underway. To maximize skill, additional predictors, for example lagged SSTs (Wu et al., 2006), upper ocean heat content, and Madden-Julian oscillation indices (McPhaden et al., 2006), are being investigated. Model performance may also be improved by specifying corrections on predictions when the model extrapolates beyond the limits of the training data, as suggested by Wu et al. (2007).
Bibliography


Bibliography


Chapter 3

Nonlinear principal predictor analysis

3.1 Introduction

Multivariate statistical models are commonly used in climatology to investigate relationships between two sets of variables (Jolliffe, 2002). For example, canonical correlation analysis (CCA) (Glahn, 1968) is used to find modes of maximum correlation between two sets of variables, while maximum covariance analysis (von Storch and Zwiers, 1999; Wallace et al., 1992) seeks modes of maximum covariance. In each of these methods the two sets of variables are treated equally, i.e., there is no distinction between predictors and predictands. The symmetric nature of methods like CCA is considered a limitation in some fields of science. In numerical ecology, for example, multivariate models like redundancy analysis (von Storch and Zwiers, 1999) or principal predictor analysis (PPA) (Thacker, 1999), which focus on minimizing error in a set of predictands given a separate set of predictors, are recommended (Legendre and Legendre, 1998).

Multivariate models traditionally define new sets of variables that are linear combinations of the original sets of variables. These methods will lead to suboptimal results when datasets contain nonlinear structure. Building on earlier work on nonlinear principal component analysis (NLPCA) (Kramer, 1991; Monahan, 2000), nonlinear CCA (NLCCA), which is also based on a neural network architecture, was developed by Hsieh (2000). NLCCA has been used to investigate linked ocean-atmosphere variability in the tropical Pacific (Hsieh, 2001; Wu et al., 2002) and North America (Wu and Hsieh, 2003), and has been reviewed in Hsieh (2004) along with other nonlinear multivariate models.

One drawback of neural network-based NLCCA is that it requires parameters in three neural networks to be optimized separately, which can lead to overfitting and poor performance on noisy datasets. To avoid this problem, a simpler model, nonlinear projection, was used by Wu and Hsieh (2004) to investigate the influence of El Niño-Southern Oscillation (ENSO) on the wintertime Northern Hemisphere atmosphere. Unlike NLCCA, which considers multivari-
Chapter 3. Nonlinear principal predictor analysis

ate predictors/predictands, nonlinear projection maps a univariate predictor, e.g., the ENSO climate index, onto predictands via a neural network. The model is thus less susceptible to overfitting on noisy problems, but is also unable to easily describe the influence of multiple climate predictors on the predictands.

A nonlinear variant of PPA or redundancy analysis would fall between nonlinear projection and NLCCA in terms of complexity. PPA and redundancy analysis both have similar goals (the main difference being that derived variables in PPA are uncorrelated while vectors of loadings in redundancy analysis are orthogonal; Jolliffe, 2002). This paper focuses on extending the PPA algorithm because its close ties to PCA and CCA (Thacker, 1999) make implementation via neural networks straightforward. A nonlinear PPA model would be able to account for the influence of multiple predictors while containing fewer parameters than NLCCA. To date such a model has not been developed and applied to problems in climatology.

The main goals of this study are (1) to develop a model for NLPPA and (2) to compare its performance with CCA and NLCCA. (Note: the NLCCA algorithm used in this chapter is not the robust version of Chapter 2.) To meet the first goal, a simple neural network architecture, which is based on nonlinear models for PCA and CCA, is introduced to perform NLPPA. To meet the second goal, NLPPA is used to predict solutions from the Lorenz system of equations and the model’s performance is compared with CCA and NLCCA.

3.2 Method

PPA is used to find relationships between two sets of variables, with the main goal being the discovery of dominant linear modes of variability in a set of predictands from a set of predictors (Thacker, 1999). Specifically, PPA finds the coefficient vector \( \mathbf{a} \) such that the resulting linear combination of the predictor variables \( \mathbf{x}(t) \equiv [x_1, \ldots, x_I] \) (where each variable \( x_i \) has \( N \) samples labelled by the index \( t \), which is typically time for climate datasets),

\[
p(t) = \mathbf{a} \cdot \mathbf{x}(t),
\]

which is referred to as the principal predictor \( p \), has the maximum sum of squared correlations with the predictand variables \( \mathbf{y}(t) \equiv [y_1, \ldots, y_K] \). If the predictands are standardized, this is equivalent to minimizing the sum of squared errors between \( \mathbf{y} \) and \( \hat{\mathbf{y}} \), which are the predictions obtained by mapping \( p \) onto \( \mathbf{y} \). \( \hat{\mathbf{y}} \) represents a straight line through predictand space. The linear representation of \( p \) in predictor space \( \hat{\mathbf{x}} \) is found by mapping \( p \) onto \( \mathbf{x} \). Once the first principal predictor mode has been identified, the second mode is found by searching for linear combinations of the model residuals, and so on for higher modes. In practice, all modes are typically extracted simultaneously by solving an eigen equation. However, it should be possible to extract principal predictors sequentially, using a variational approach, in the same way that principal components (Kramer, 1991; Monahan, 2000) and canonical variates (Hsieh, 2000) may be extracted.
Figure 3.1: Neural network architecture for NLPPA. The multi-layer neural network at the top consists of $I$ input nodes and $K$ output nodes, which correspond to the predictors and predictands respectively. Between the input and output nodes are three sets of internal nodes (starting with $J$ hidden nodes, then a single bottleneck node, and finally $L$ hidden nodes). The neural network at the bottom consists of one input node, which corresponds to the principal predictor extracted from the bottleneck node, $J$ hidden nodes, and $I$ output nodes corresponding to the predictors. Each line represents a parameter in the model.
PPA represents the predictand response by a straight line in the predictand space. The goal of NLPPA is to generalize PPA to allow data to be approximated by a curve instead of a straight line. Like PPA, which reduces to PCA if the predictors/predictands are the same (Thacker, 1999), by extension NLPPA should also reduce to NLPCA. Consequently, the neural network architecture for NLPPA follows naturally from the one used to represent NLPCA (Kramer, 1991; Monahan, 2000), except that the mapping is not auto-associative. Following this reasoning, NLPPA can be represented using the two multi-layer perceptron neural networks (Gardner and Dorling, 1998) shown in Figure 3.1.

The core component of the NLPPA architecture, pictured at the top of Figure 3.1, is a neural network with layers arranged the same as NLPCA. This network can be regarded as two mappings. The first mapping, from $\mathcal{R}^{J} \rightarrow \mathcal{R}^{1}$, is represented by the portion of the network passing from the predictors $x$ through the $J$ hidden nodes to the bottleneck node, and gives the 1-D principal predictor $p$

$$p = \sum_{j=1}^{J} f \left( \sum_{i=1}^{I} x_i w_{ji}^{(1)} + b_j^{(1)} \right) w_j^{(2)} + b_j^{(2)}, \quad (3.2)$$

where $f$ is the node transfer function; the adjustable model parameters are each hidden-layer weight $w_{ji}^{(1)}$, hidden-layer bias $b_j^{(1)}$, output-layer weight $w_j^{(2)}$, and output-layer bias $b_j^{(2)}$; and bracketed superscripts keep track of layers. The same naming conventions are used for model parameters in subsequent equations. Dimension reduction is achieved by passing the multivariate predictors through the single bottleneck node.

The second mapping, from $\mathcal{R}^{1} \rightarrow \mathcal{R}^{K}$, is represented by the portion of the network passing from $p$ through the layer of $L$ hidden nodes to the output layer. This part of the model maps $p$ onto the standardized predictands $y$. Predicted values of the $k^{th}$ predictand $\hat{y}_k$ are given by

$$\hat{y}_k = \sum_{l=1}^{L} f \left( p w_l^{(3)} + b_l^{(3)} \right) w_{kl}^{(4)} + b_k^{(4)}. \quad (3.3)$$

If $f$ is the identity function, then the network is linear and Equation 3.2 for $p$ becomes the sum of two linear regression equations: (1) a multivariate linear regression from $x$ to the hidden layer, and (2) a multiple linear regression from the hidden layer to $p$. The two linear transformations can be collapsed down to the inner product between $x$ and a vector of weight coefficients, which is equivalent to Equation 3.1 for linear PPA. (Note: the bias parameters are zero if the predictors, predictands, and principal predictor have zero mean.) The second mapping maps $p$ linearly onto the predictands. In this case, the network represents a straight line through the data points in predictand space. If, however, the transfer function is nonlinear, the network is capable of accounting for nonlinear relationships and the network represents the NLPPA model. For NLPPA, $f$ is usually set to the hyperbolic tangent tanh, a commonly used transfer function for neural networks (Hsieh, 2004). In this case, the network represents a continuous 1-D curve through the data points in predictand space.
Parameters in the first neural network are set by minimizing the cost function

\[
P_1 = \frac{1}{K} \sum_{k=1}^{K} \langle (y_k - \hat{y}_k)^2 \rangle + \langle p \rangle + (\langle p^2 \rangle - 1)^2 + \lambda^{(1)} \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{I} (w_{ji}^{(1)})^2 + \lambda^{(3)} \frac{1}{L} \sum_{l=1}^{L} (w_{i}^{(3)})^2 \]  

(3.4)

where \( \langle \cdot \rangle \) denotes the sample or temporal mean. The first term is the mean squared error (MSE) between the model predictions and the predictands. As the scaling of \( p \) has no effect on the MSE, the second and third terms are included to force the mean and variance of \( p \) to zero and one respectively. The last two terms penalize large absolute weights in the nonlinear layers of the model. As the function tanh is approximately linear when its input is near zero, large values of the weight penalty coefficients will force the weights to be smaller, thus leading to a less nonlinear model. Weight penalty terms (also known as regularization or smoothing terms) are often included in neural network models to help prevent model overfitting (Hsieh and Tang, 1998). Both the number of nodes and the weight penalty coefficients determine the overall complexity of the model; increasing the number of nodes or decreasing the weight penalty coefficients will increase the nonlinear modelling capability of the network. The optimal values are typically chosen to minimize an estimate of generalization error, i.e., the error for cases not used to optimize the values of the model parameters, for example through split-sample or cross-validation.

Unlike NLPCA, which can be represented using just the one neural network, NLPPA requires a second neural network to determine the 1-D representation of \( p \) in predictor space. Therefore, once \( p \) has been extracted by the first neural network, the network at the bottom of Figure 3.1 maps from \( \mathcal{R}^1 \to \mathcal{R}^I \) by mapping \( p \) onto the predictors \( x \). The same type neural network is used by NLCCA to map the canonical variates back to predictor/predictand space (Hsieh, 2000). Predicted values of the \( i^{th} \) predictor \( \hat{x}_i \) are given by

\[
\hat{x}_i = \sum_{j=1}^{J} f(p w_{j}^{(5)} + b_{j}^{(5)}) w_{ij}^{(6)} + b_{i}^{(6)},
\]  

(3.5)

and lie on a continuous curve through predictor space. Parameters in this network are trained to minimize the cost function

\[
P_2 = \frac{1}{I} \sum_{i=1}^{I} \langle (x_i - \hat{x}_i)^2 \rangle + \lambda^{(5)} \frac{1}{J} \sum_{j=1}^{J} (w_{j}^{(5)})^2 .
\]  

(3.6)

A mode in NLPPA consists of the 1-D principal predictor \( p \) along with the corresponding curves represented by \( \hat{y} \) and \( \hat{x} \). Each value of the principal predictor maps to a coordinate on the curve defined in predictand space and a coordinate on the curve defined in predictor space. As in NLPCA, modes in NLPPA can be interpreted “cinematically” (Monahan, 2001). For example, if the predictors and predictands are PCs of gridded sea surface temperatures (SST) and sea level pressures (SLP) respectively, values of \( p \) correspond to different values of the PC
scores, which represent different spatial fields of SST and SLP. As the value of $p$ changes from its minimum to its maximum, e.g. moving from one end of the curves in SST and SLP space to the other, the spatial fields will also change. As NLPPA can represent curves in predictor and predictand space, both the magnitude and pattern of the fields are free to change. In contrast, PPA, which can only represent straight lines in predictor and predictand space, results in a fixed spatial pattern; only the magnitude is free to change as the principal predictor ranges from its minimum to maximum.

Once the first principal predictor mode has been extracted from the data, another set of neural networks can be trained on the model residuals $x - \hat{x}$ and $y - \hat{y}$ to extract the next leading mode, and so on for higher modes.

The NLPPA architecture described above allows for different numbers of nodes in the compression/regression layers of the neural networks, as well as different weight penalty coefficients. For many problems, especially when the numbers of predictors/predictands are similar, the neural networks may be simplified by setting the numbers of nodes and the weight penalty coefficients to be the same in all layers. This approach was taken by Hsieh (2000) for NLCCA and is adopted in this paper.

To illustrate the difference between PPA and NLPPA, consider a 2-D predictor dataset $x = [t+m_1, -0.7t^2-0.7+m_2]$ and a 2-D predictand dataset $y = [1.6\tanh(t)+n_1, 3.2\tanh(t)^2-1.2+n_2]$ (shown in Figures 3.2a and 3.2b respectively), where $t$ is drawn from a Gaussian distribution with zero mean and unit standard deviation and the noise terms $m_{1,2}$ and $n_{1,2}$ are drawn from a Gaussian distribution with zero mean and a standard deviation of 0.2. The first two linear modes linking the $x$ and $y$ datasets, shown in Figures 3.3a and 3.3b, were extracted by PPA while the first nonlinear mode, shown in Figures 3.3c and 3.3d, was extracted by NLPPA. PPA exhibits the problem noted by Hsieh (2004) for linear models applied to datasets with nonlinear structure. In this case, variability is spread out over two linear modes, even though the theoretical mode is a single curve. The first two modes extracted by PPA are straight lines along the horizontal and vertical axes, accounting for 43% and 30% of variability in the noisy predictand dataset respectively, while the theoretical mode is quadratic, accounting for 96% of variability in the noisy predictand dataset. NLPPA is able to extract the quadratic structure with a single mode, accounting for 91% of the variability in the predictand dataset, which is 95% of the theoretical maximum.

The main attributes of NLPPA are summarized in Table 3.1, which includes a comparison with nonlinear projection and NLCCA, the two nonlinear methods to which it is most closely related. The main difference between nonlinear projection and NLPPA/NLCCA is that the former is restricted to a single predictor whereas the latter are not. Despite the simplicity of nonlinear projection, this constraint limits its usefulness and it will not be discussed further in this paper. In terms of complexity, NLPPA contains fewer parameters than NLCCA (a quantitative comparison is given in the next section), but requires one dataset to be defined as the predictors and one as the predictands. If the roles of the two datasets are reversed,
Figure 3.2: Noisy synthetic data and the theoretical mode in (a) $x$ space and (b) $y$ space.
Figure 3.3: Noisy synthetic data and the first two linear modes (where the first is represented by a thin line and the second is represented by a thick line) modelled by PPA in (a) $x$ space and (b) $y$ space. Noisy synthetic data and the first nonlinear mode modelled by NLPPA in (c) $x$ space and (d) $y$ space.
Table 3.1: A qualitative summary of the main attributes of nonlinear multivariate models.

<table>
<thead>
<tr>
<th>Model</th>
<th># Inputs</th>
<th># Outputs</th>
<th># Parameters</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL Projection</td>
<td>univariate</td>
<td>multivariate</td>
<td>low</td>
<td>predictor → predictands</td>
</tr>
<tr>
<td>(Wu and Hsieh, 2004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLPPA</td>
<td>multivariate</td>
<td>multivariate</td>
<td>medium</td>
<td>predictors → predictands</td>
</tr>
<tr>
<td>(Hsieh, 2000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLCCA</td>
<td>multivariate</td>
<td>multivariate</td>
<td>high</td>
<td>predictors → predictands</td>
</tr>
<tr>
<td>(Hsieh, 2000)</td>
<td></td>
<td></td>
<td></td>
<td>predictands → predictors</td>
</tr>
</tbody>
</table>

The modes that are extracted will be different. NLCCA makes no a priori distinction between predictors/predictands, which means than one can be predicted from the other and vice versa, although, as is pointed out by Legendre and Legendre (1998), this symmetry is unwarranted in situations where the predictors force the predictands but the reverse association does not make physical sense. In these cases, NLPPA provides a more direct path from predictors to predictands.

### 3.3 Application to the Lorenz system

The Lorenz system (Lorenz, 1963) of coupled nonlinear ODEs is given by

\[
\begin{align*}
\frac{du_1}{dt} &= -\sigma u_1 + \sigma u_2, \\
\frac{du_2}{dt} &= -u_1 u_3 + r u_1 - u_2, \\
\frac{du_3}{dt} &= u_1 u_2 - b u_3,
\end{align*}
\]  

(3.7) 
(3.8) 
(3.9)

where \( t \) is time and parameter values are set to \( \sigma = 10, r = 28, \) and \( b = 8/3 \). Solutions of the system of equations (which provide an approximate description of 2-D convection in a tank of fluid heated from below) lie on what has become known as the Lorenz attractor and exhibit chaotic behaviour. The Lorenz system was previously used to evaluate NLPCA (Monahan, 2000) because of the attractor’s nonlinearity and well known structure. Given the strong links between NLPCA and NLPPA, the same set of equations are used to test the NLPPA model, albeit from a predictive rather than descriptive standpoint.

To test the performance of NLPPA, datasets were created by integrating Equations 3.7-3.9 using a fourth-order Runge-Kutta routine. Solutions were sampled in the range \( t = [0, 100] \) with \( \Delta t = 0.02 \) resulting in 10000 cases. Five sets of data were created, each with different amounts of Gaussian noise (standard deviations equal to 0.1, 0.2, 0.3, 0.4, and 0.5 times the signal standard deviation) added to the variables. Predictors \( x \) for the multivariate models
were the solutions at time \( t \), standardized to zero mean and unit standard deviation, and the predictands \( y \) were the solutions at 3, 6, 9, and 12 \( \Delta t \) time steps into the future.

A set of benchmark predictions were made by integrating the Lorenz equations forward in time using the noisy predictors at each time step as initial conditions. Predictions from PPA/CCA were also made for test purposes. (Note: results for PPA and CCA are reported together as the performance of the two models was similar for this test problem.) Models were trained on the first 2000 cases; model performance was tested on the remaining 8000 cases. In all trials, three modes were extracted from the dataset.

To find the best architecture for NLPPA and NLCCA, split-sample tests were carried out on the 2000 training cases. Eighty percent of the data (1600 cases) were selected for training the models and the remaining twenty percent (400 cases) were used to validate model performance. The number of hidden nodes and values of the weight penalty coefficients were identified through a systematic search aimed at minimizing the MSE on the validation sample. Because the NLPPA and NLCCA models are nonlinear in their parameters, an iterative nonlinear optimization is required to minimize the cost functions and set the model weights and biases. A quasi-Newton optimization scheme was used in this paper. As the optimization may become trapped in a local minimum (which may not be close to the global minimum), multiple restarts from random initial conditions are necessary to ensure good performance. Training runs were repeated 10 times, each time starting from a different random initialization of the model parameters (drawn from a uniform distribution from -0.5 to 0.5). Weights and biases from the run with the smallest validation error were selected for use. All models contained two nodes in the hidden layers \( (J = L = 2) \). Values of the weight penalty coefficients varied depending on the model and level of noise. Once the appropriate levels of model complexity were determined, models were retrained on the full set of 2000 training cases, again using random initializations from different starting parameters to avoid shallow local minima.

MSE values for the model predictions on the test datasets are shown in Figure 3.4. Rankings of the models, based on MSE, were usually the same across the noise levels and lead times, with numerical integration usually performing best, followed by NLPPA, NLCCA, and CCA/PPA. When signal-to-noise ratios were low (e.g., 9 and 12 step lead times with 40% and 50% added noise) NLPPA equalled or slightly exceeded the performance of numerical integration. For reference, percent improvements in MSE between NLPPA and NLCCA for each lead time and noise level are plotted in Figure 3.5. NLPPA consistently outperformed NLCCA, with the exception of the three step lead time with 10% added noise. For a given level of added noise, the gap between NLPPA and NLCCA typically increased with increasing lead time. At the longest lead times, NLPPA performed between 10% (with 50% added noise) and 40% (with 10% added noise) better than NLCCA.

MSE values resulting from the addition of each extracted mode are given in Figure 3.6 for the 9 step lead time with 30% added noise, which, amongst the datasets tested, is moderately noisy. Other lead times and noise levels exhibited similar relative changes in error and are not
Figure 3.4: MSE values for model predictions at each lead time and noise level.
Figure 3.5: Percent improvement in MSE between NLPPA and NLCCA for each lead time and noise level. Positive values indicate that NLPPA performed better than NLCCA.
Figure 3.6: MSE values associated with the addition of each mode at the 9 step lead time and
$n = 0.3$. 
Chapter 3. Nonlinear principal predictor analysis

shown. For NLPPA and NLCCA, most of the improvement in skill relative to the linear models is due to the first mode; the second mode contributes a small reduction in error, while the contribution of the third mode is minimal. At low signal-to-noise ratios, the first and second NLPPA modes are necessary to reduce MSE below the level of numerical integration with noisy initial conditions.

The low dimensionality of the Lorenz system means that the nonlinearity of the extracted modes can be visualized easily. To evaluate how well NLPPA reproduced features of the Lorenz attractor, results are compared with those from NLPCA (Monahan, 2000) applied to the noise-free predictand dataset. As stated in the previous section, NLPCA is equivalent to NLPPA with identical predictor and predictand datasets. The extracted mode from NLPCA on the noise-free dataset therefore represents an estimate of the “optimal” 1-D representation of the Lorenz attractor by a neural network, and can be used as a benchmark to evaluate NLPPA.

Results from the NLPCA model are shown in Figure 3.7. NLPPA models were applied to the 9 step lead time dataset with 30% added noise. To test the stability and robustness of the nonlinear mode extraction, models were built 10 times, each time with different randomly generated noise. Results from one of the runs are shown in Figure 3.8. The corresponding predictand modes extracted by NLCCA are shown in Figure 3.9 for comparison. A summary of MSE values for the leading NLPPA/NLCCA predictand modes, taken with respect to the corresponding NLPCA modes, over the 10 trials is shown in Figure 3.10.

Both NLPPA and NLCCA models captured the main features of the leading NLPCA mode, which exhibits relatively strong nonlinearity, although MSE values were lower and exhibited less spread for NLPPA than for NLCCA. The second and third NLPCA modes show no evidence of nonlinearity. NLPPA was able to extract them both consistently and with modest levels of error, due mainly to a clockwise rotation of the lines on the $y_3$ and $y_1$ and $y_3$ and $y_2$ planes. Similar rotation is present in the NLCCA modes, but results are generally poorer than for NLPPA, with larger errors and more variability between trials. Overall NLPPA performed better and extracted each mode more robustly than NLCCA.

The NLPPA model requires fewer parameters than the corresponding NLCCA model (Table 3.1). For models with two nodes, as used here for the Lorenz system, NLPPA contains 37 parameters (see Figure 3.1), whereas NLCCA, with an extra neural network (Hsieh, 2000), contains 48 parameters. In the test problem, models were implemented from a common code base and shared the same optimization routine. The mean training time to extract each mode roughly scaled with the number of parameters, with NLPPA requiring 23% less computational time than NLCCA.

3.4 Discussion and conclusion

Results from previous work on nonlinear multivariate models were used to develop a neural network architecture capable of performing NLPPA, a method which fits between nonlinear
Figure 3.7: Lorenz data (gray dots) and the first (thick black line), second (medium black line) and third (thin black line) modes from NLPCA shown projected onto each of the 2-D planes in $y$ space and as a 3-D scatterplot.
Figure 3.8: As in Figure 3.7 but for NLPPA. Predictand modes were extracted from the 9 step lead time data with 30% added random noise.
Figure 3.9: As in Figure 3.8 but for NLCCA.
Figure 3.10: Boxplots of MSE values for predictand modes extracted by NLPPA and NLCCA from the noisy data with respect to NLPCA modes extracted from the noise free data over the 10 trials.
projection (Wu and Hsieh, 2004) and NLCCA (Hsieh, 2000) in terms of the complexity of its architecture. Unlike NLCCA, which makes no distinction between predictors and predictands, NLPPA is designed to predict values of a set of predictands given a set of predictors. In this regard, NLPPA is similar to nonlinear projection, except that NLPPA makes use of multiple predictors instead of a single predictor. In contrast to the symmetric formulation of NLCCA, NLPPA requires variables to be assigned either as predictors and predictands a priori. However, the attendant reduction in size and complexity of the architecture means that NLPPA can be trained more quickly (and with less opportunity for overfitting) than NLCCA.

While PPA was originally conceived as a tool for compressing large datasets using many principal predictors (Thacker, 1999), NLPPA is geared more towards extracting a small number of modes suitable either for interpretation or prediction. The nonlinear optimization of the model parameters, along with the reduction in signal-to-noise ratio as each successive mode is extracted, makes obtaining stable estimates of higher nonlinear principal predictors difficult. However, NLPPA may be better at handling noisy problems than NLCCA due to NLPPA’s simpler model architecture. Measures of model performance on the Lorenz system of equations suggest that NLPPA is capable of outperforming NLCCA when datasets are corrupted with noise. Also, as pointed out by Hsieh (2001), nonlinear methods, including NLPPA, can be used to extract the first mode or two of variability where the nonlinear structure is often most obvious and easily modelled, leaving remaining variability to be extracted via standard linear methods, which are more robust. The similarity of NLPPA to NLPCA and NLCCA means that it shares some of these models’ limitations. For example, NLPPA cannot model closed curves (e.g., a circle) or discontinuous functions (e.g., a step function can only be approximated by a curve with a steep gradient), although the former problem can be solved by by employing a circular bottleneck node, as in Hsieh (2004) for geophysical datasets.

The main areas of focus for this paper were (1) the development of the NLPPA neural network architecture and (2) comparisons between NLPPA and NLCCA on a simple, synthetic, nonlinear test problem. Work is currently being undertaken to apply NLPPA to real-world climate datasets, in particular focusing on the extratropical atmospheric response to low frequency modes of climate variability (e.g., ENSO and the Arctic Oscillation). As with all nonlinear multivariate statistical techniques, NLPPA is empirical in nature. However, as shown by Hsieh (2001) and Wu and Hsieh (2004) among others, techniques such as nonlinear projection and NLCCA can, if applied judiciously, be used to gain insight into the workings of the climate system. NLPPA, with a less restrictive architecture than nonlinear projection and a simpler formulation than NLCCA, provides a valid alternative to these techniques.
Bibliography


Chapter 4

Nonlinear analog predictor analysis

4.1 Introduction

Synoptic downscaling models are used to predict local-scale climate variables (e.g., station temperatures or precipitation amounts) from synoptic-scale atmospheric circulation data (e.g., sea-level pressure or geopotential height fields), often as a means of creating climate scenarios at high temporal and spatial resolution, e.g., for assessing the impacts of climate change on a watershed, crop, or ecosystem (Wilby and Wigley, 1997; Xu, 1999). Synoptic downscaling techniques may be physically based, for example using limited-area numerical weather prediction models, or they may be statistically based, for example using linear or nonlinear regression models, stochastic weather generators, or analog models. Statistical models are attractive because they require few computational resources and are capable of performing as well as more complicated physically based approaches (Hellstrom et al., 2001; Murphy, 1999, 2000; Wood et al., 2004).

Linear regression models offer a simple and direct way to link the synoptic-scale forcing with the local climatic response, and, as a result, they are commonly used in statistical climate downscaling. Most regression-based downscaling models are developed for a variable, e.g., precipitation or temperature, at an observing site. If many variables are required at many sites, different regression equations are usually developed for each separately. Using this approach, the model developer has little control over the consistency of outputs between sites and variables, perhaps only through the specification of a common set of synoptic-scale predictors. Maintaining realistic relationships between sites and variables is particularly important in hydroclimatological applications, as streamflow depends strongly on the spatial distribution of precipitation in a watershed and on interactions between precipitation and temperature that determine whether precipitation falls as rain or snow.

Recent studies have addressed multi-site downscaling by moving away from classical multiple regression models to multivariate linear models, e.g., canonical correlation analysis, CCA, or maximum covariance analysis, MCA (Uvo et al., 2001). Others have extended multivariate linear models by adopting cost functions that force the covariance matrix of the model predictions to match observations, e.g., expanded downscaling (Bürger, 1996, 2002; Bürger and

---

Chapter 4. Nonlinear analog predictor analysis

Chen, 2005). While these methods may be better suited to spatial downscaling than standard multiple regression models, they may fail when nonlinear relationships are present, for example when trying to predict daily precipitation amounts (Yuval and Hsieh, 2002).

Flexible nonlinear models, such as artificial neural networks (ANNs), may perform better than the classical linear approaches when nonlinear relationships are present. Comparisons between multivariate ANNs (i.e., ANNs with multiple outputs) and ANNs with a single output have also demonstrated the potential of the multivariate approach, due in part to the sharing of weights between outputs of the multivariate neural network (Caruana, 1997). It is conceivable that inter-site correlations would be modeled more accurately using this approach. However, constraints similar to the one used in expanded downscaling would likely still be needed to ensure close correspondence between modeled and observed spatial relationships. Modeling a variable like precipitation can also be a challenge, as the amount of precipitation is conditional on the occurrence of precipitation. Occurrences and amounts thus need to be modeled sequentially, which adds an additional layer of complexity to the modeling process.

Instead, analog (i.e, k-nearest neighbor) models have often been used for downscaling tasks, particularly for the prediction of precipitation at multiple sites (Zorita and von Storch, 1999). The analog model is nonlinear, and, due to the sampling process, is capable of preserving spatial correlations between sites and handling conditional variables like precipitation. Analog models are, however, prone to the ‘curse of dimensionality’ (Hastie et al., 2001), which means that they may perform poorly when the dimensionality of the dataset is high and the number of cases is low. To mitigate this problem, predictor scaling algorithms have been developed that give more weight to relevant variables than irrelevant or redundant variables, thereby reducing the overall size of the predictor space (Fraedrich and Ruckert, 1998; Mehrhotra and Sharma, 2006). Alternatively, dimensionality reduction can be undertaken by linear models that create a smaller set of analog predictors from the original predictors. For example, principal component analysis (PCA) (Fernandez and Saenz, 2003; Zorita and von Storch, 1999), and multiple linear regression (Wilby et al., 2003) have been used to reduce the dimensionality of the predictors, with outputs from the linear models acting as inputs to the analog model. Results have shown this to be a promising avenue for multi-site downscaling research.

Given the improved performance of multivariate ANNs relative to multivariate linear models, and the recent work into nonlinear versions of classical multivariate linear models, such as nonlinear PCA (NLPCA) and nonlinear CCA (NLCCA) (Hsieh, 2004), a logical next step would be to combine an ANN with an analog model in a manner similar to what has been done previously for linear models. In the current study, an approach that couples an ANN to an analog model is developed. The coupled model, referred to as nonlinear analog predictor analysis (NLAPA), is based on nonlinear principal predictor analysis (NLPPA) (Cannon, 2006), which is a nonlinear multivariate model that is closely related to both NLPCA and NLCCA. Rather than creating ANN and analog models separately, the analog model in NLAPA is instead coupled to an ANN in such a way that the ANN portion of the model is used to define
the predictors for the analog model. The end result is a flexible model that maintains the ability of the analog model to preserve inter-variable relationships and to model non-normal and conditional variables like precipitation, while taking advantage of NLPPA’s ability to define an optimal set of analog predictors that results in improved predictive performance.

The remainder of the paper is organized as follows. First, Section 4.2 describes the analog model and the NLPPA model; this is followed by the development of the NLAPPA model. Next, Section 4.3 describes the synthetic benchmark tasks and compares the performance of the NLAPPA model with standard analog models. Section 4.4 applies the NLAPPA model to two real-world downscaling tasks, both focused on multivariate predictand datasets that are of relevance in hydroclimatology. The first involves downscaling of temperature, rainfall, and snowfall at a station near Whistler, British Columbia (B.C.), Canada; the second involves downscaling of precipitation at multiple sites in coastal B.C. Finally, results are discussed in Section 4.5 along with conclusions and recommendations for future research directions.

4.2 Method

4.2.1 Analog model

The analog model is a data-driven nonparametric algorithm that relates a set of analog predictors (or model inputs) to a set of predictands (or model outputs) by assuming that cases that appear “close” to one another in predictor space share similar predictand values.

Given a set of $L$ analog predictors observed at time $t$, $\mathbf{p}(t) = [p_1(t), \ldots, p_L(t)]$, the analog model predicts values of the $K$ predictands at time $t$, $\hat{\mathbf{y}}(t) = [\hat{y}_1(t), \ldots, \hat{y}_K(t)]$ (where the hat denotes the predicted value of a variable), by searching through $N$ historical observations of the analog predictors, $\mathbf{p}_l(h) = [p_l^{(h)}(1), \ldots, p_l^{(h)}(N)]^T$ ($l = 1 \ldots L$), finding the time $u$ with analog predictors $\mathbf{p}_l^{(h)}(u)$ that are closest to $\mathbf{p}(t)$ in terms of weighted Euclidean distance $D$,

$$D = \sqrt{\sum_{l=1}^{L} s_l \left( p_l(t) - p_l^{(h)}(u) \right)^2}, \quad (4.1)$$

and then taking the historical predictand values at time $u$, $\mathbf{y}^{(h)}(u)$, as the predicted values at time $t$, $\hat{\mathbf{y}}(t)$. The scaling factor $s$ is included to allow different weights to be assigned to each analog predictor. Together, this series of operations will be referred to as the analog operator $A$.

If predictions are based on a single historical analog, $A$ can preserve spatial correlations between multiple sites. The analog operator can be extended to use more than one analog or neighbor by taking predictions to be the weighted mean or median of a set of nearest historical cases. This smooths the predictions, which may lead to a better least squares fit with the observed series at the expense of a poorer representation of inter-variable relationships. As the main focus of this study is on multivariate and multi-site precipitation downscaling, all
Chapter 4. Nonlinear analog predictor analysis

subsequent models are based on a single analog, although this is not necessary (or desirable) for more general prediction tasks.

When analog models are used for climate downscaling, predictor variables are typically derived from gridded synoptic-scale fields, e.g., 500-hPa geopotential height or sea-level pressure. To reduce the high dimension of the predictor dataset, PCA is applied to the fields and scores from a smaller number of retained principal components (PCs) are used as the analog predictors (Zorita and von Storch, 1999). Because the PCs are calculated without reference to the predictands, there is no guarantee that the leading PCs will lead to the best predictions. Instead, the use of the leading predictor canonical variates from CCA as analog predictors has been advocated. Each pair of canonical variates (one a linear combination of the predictors and the other a linear combination of the predictands) is found such that the correlation between them is maximized. Analog downscaling based on CCA has been shown to yield improvements in skill over analog models based on PCA (Fernandez and Saenz, 2003).

Another complementary approach involves adjusting the scaling weights \( s \) associated with the predictors to maximize some measure of prediction performance. This approach was taken by Fraedrich and Ruckert (1998), who applied a heuristic optimization algorithm to minimize the least squares cost of analog model predictions, and Mehrhotra and Sharma (2006), who applied an adaptive Metropolis algorithm to maximize the log likelihood of a k-nearest neighbor resampler.

4.2.2 Nonlinear Principal Predictor Analysis


Specifically, NLPPA nonlinearly maps the predictor variables \( \mathbf{x}(t) = [x_1(t), \ldots, x_I(t)] \ (t = 1 \ldots N) \) to a set of principal predictors of lower dimensionality \( \mathbf{p}(t) = [p_1(t), \ldots, p_L(t)] \), and then nonlinearly maps from the principal predictors back to the predictand variables \( \mathbf{y}(t) = [y_1(t), \ldots, y_K(t)] \). The mappings are defined by the ANN architecture shown in Figure 4.1. The encoding mapping is represented by the portion of the ANN passing from the predictors \( \mathbf{x}(t) \) through the \( J \) hidden nodes to the \( L \) bottleneck nodes; this gives the principal predictors,

\[
p_l(t) = \sum_{j=1}^{J} \tanh \left( \sum_{i=1}^{I} x_i(t)w_{ji}^{(1)} + b_j^{(1)} \right) w_{j}^{(2)} + b_j^{(2)},
\]

where \( \tanh(\cdot) \) is the hyperbolic tangent; the adjustable model parameters are each hidden-layer weight \( w_{ji}^{(1)} \), hidden-layer bias \( b_j^{(1)} \), output-layer weight \( w_{j}^{(2)} \), and output-layer bias \( b_j^{(2)} \); and bracketed superscripts keep track of layers. The same naming conventions are used for model
Chapter 4. Nonlinear analog predictor analysis

Figure 4.1: ANN architecture for NLPPA. The multi-layer ANN consists of $I$ input nodes and $K$ output nodes, which correspond to the predictors and predictands respectively. Between the input and output nodes are three sets of internal nodes starting with the $J$ hidden nodes, then $L$ bottleneck nodes, and finally $J$ hidden nodes. Each line represents a parameter in the model.

parameters in subsequent equations. Reduction in dimensionality is achieved by passing the multivariate predictors through the bottleneck nodes.

The decoding mapping is represented by the portion of the ANN passing from $p(t)$ through the layer of $J$ hidden nodes to the output layer; this part of the model maps $p(t)$ to $y(t)$. Predicted values of the $k^{th}$ predictand $\hat{y}_k(t)$ are given by,

$$
\hat{y}_k(t) = \sum_{j=1}^{J} \tanh \left( \sum_{l=1}^{L} p_l(t) w_{lj}^{(3)} + b_{lj}^{(3)} \right) w_{kj}^{(4)} + b_{kj}^{(4)}, \quad (4.3)
$$

Values of the weights and biases in the encoding and decoding portions of the ANN are set by minimizing the mean squared error (MSE) residual between predictions and observed predictands,

$$
C_1 = \frac{1}{K} \sum_{k=1}^{K} \left( y_k - \hat{y}_k \right)^2 + \frac{1}{L} \sum_{l=1}^{L} \left( \langle \langle p_l \rangle \rangle \right)^2 + \frac{1}{L} \sum_{l=1}^{L} \left( \text{var}(p_l) - 1 \right)^2, \quad (4.4)
$$

where $\langle \cdot \rangle$ denotes the temporal mean and $\text{var}(\cdot)$ denotes the temporal variance. The second and third terms are included for convenience to force the mean and variance of each principal predictor to zero and one respectively.

The optimum numbers of hidden nodes $J$ and principal predictors $L$ are chosen to minimize an estimate of generalization error, i.e., the cost function for cases not used to optimize the values of the model parameters, for example through split-sample or cross-validation.
4.2.3 Nonlinear Analog Predictor Analysis

The NLAPA algorithm is based on the analog and NLPPA models described in the previous sections. As in NLPPA, an encoding mapping is used to perform dimensionality reduction, in this case for the purpose of defining analog predictors; the analog predictors are then mapped back to predictand space via a decoding mapping, which, in NLAPA, is performed via the analog operator $A$ instead of the ANN used in NLPPA.

The basic structure of the coupled model is shown in Figure 4.2. The first half of the model is shared with the NLPPA model and consists of an ANN with a single hidden-layer, which nonlinearly transforms the predictors $x(t)$ to a set of new analog predictors $p(t)$. The second half of the model consists of the analog operator $A$, which maps the analog predictors $p(t)$ to the predictands $y(t)$.

Parameters in the ANN portion of the coupled model are estimated by minimizing the cost function,

$$C_2 = \frac{1}{K} \sum_{k=1}^{K} \left( (y_k - \hat{y}_k)^2 \right) + \frac{1}{L} \sum_{l=1}^{L} \left( (p_l)^2 \right) + \left\{ \frac{1}{L} \sum_{l=1}^{L} \text{var}(p_l) - 1 \right\}^2,$$

which is the same as Eq. 4.4, except that the third term constrains the mean of the variances of the analog predictors to one, which allows the analog predictors to be scaled differently from one another. The coupled model structure means that the nonlinear transformation performed by the ANN is optimized based on the performance of the analog model. While the cost function in this study is based on the MSE, it can easily be adapted to fit the needs of the particular...
The discontinuous nature of the analog operator $A$ means that standard nonlinear optimization algorithms that rely on gradient information cannot be used to set the parameters in the NLAP A model. Instead, differential evolution (DE), a global search method that is closely related to genetic algorithms (Price et al., 2005; Storn and Price, 1997), is adopted for its proven neural network training performance (Ilonen et al., 2003; Liu and Lampinen, 2005). As the encoding mapping in the NLPPA model is identical to that in the NLAP A model, weights and biases from the NLPPA mapping can be used to initialize the population members for the DE optimization. As in NLPPA, the optimum numbers of hidden nodes $J$ and analog principal predictors $L$ are chosen to minimize an estimate of generalization error. More details on the model optimization process are given in the subsequent sections.

### 4.3 Synthetic benchmark tasks

The NLAP A model was evaluated on four synthetic benchmark tasks, including the linear, nonlinear, and multicollinear benchmarks used by Mehrhotra and Sharma (2006) to test their analog predictor scaling algorithm, along with a nonlinear, multiplicative benchmark developed for this study. Each synthetic benchmark task consists of a univariate response and a combination of relevant, irrelevant and/or redundant predictors, and noise; model performance on tasks with more than one predictand is considered in Section 4.4.

#### 4.3.1 Linear benchmark

The first synthetic benchmark is constructed as a simple linear combination of two predictors,

$$y(t) = x_1(t) + x_2(t) + \varepsilon(t),$$  \hspace{1cm} (4.6)

where $x_1(t)$ and $x_2(t)$ are drawn randomly from the standard normal distribution $N(0,1)$ and $\varepsilon(t)$ is noise drawn from a normal distribution $N(0,0.7)$. The predictand $y(t)$ is thus a linear function of $x_1(t)$ and $x_2(t)$, with each predictor contributing equally to the response. Two additional predictors, $x_3(t)$ and $x_4(t)$, each drawn from $N(0,1)$, are also included as irrelevant predictors that do not contribute to $y(t)$.

#### 4.3.2 Nonlinear benchmark

The second synthetic benchmark is based on a first order self-exciting threshold autoregressive (SETAR) model,

$$y(t) = 0.4 + 0.8y(t-1) + \varepsilon(t) \text{ if } y(t-1) \leq 0$$

$$y(t) = -1.5 - 0.5y(t-1) + \varepsilon(t) \text{ otherwise}$$  \hspace{1cm} (4.7)
where $\varepsilon(t)$ is drawn from a normal distribution $N(0, 0.6)$. The predictors are the first 6 lagged predictand values $x(t) = [y(t-1), y(t-2), ..., y(t-6)]$. The first predictor contains the signal, with the remaining five variables contributing redundant information. The SETAR model generates a time series that resembles stream flow records from a watershed with multiple streamflow generating processes, e.g., from overland flow and a coupled snowmelt/overland flow mechanism (Lall and Sharma, 1996).

### 4.3.3 Multicollinear benchmark

The third synthetic benchmark consists of the linear combination of three predictors,

$$y(t) = x_1(t) + x_2(t) + x_3(t) + \varepsilon(t) \quad (4.8)$$

where $x_1(t)$ and $x_3(t)$ are drawn randomly from the standard normal distribution $N(0, 1)$, $\varepsilon(t)$ is drawn from $N(0, 0.5)$, and $x_2(t)$ equals $x_1(t)$ multiplied by a noise variable drawn from a uniform distribution $U(0, 2)$. Predictors $x_2(t)$ and $x_1(t)$ are thus highly redundant, exhibiting a coefficient of correlation of 0.87. An additional predictor $x_4(t)$ drawn from $N(0, 1)$ is also included as an irrelevant predictor that does not contribute to $y(t)$.

### 4.3.4 Multiplicative benchmark

The fourth synthetic benchmark is a multiplicative combination of the square of five predictors,

$$y(t) = \left( \prod_{j=1}^{5} x_i^2(t) \right) \varepsilon(t), \quad (4.9)$$

where values of each predictor $x_i(t)$ are drawn from a uniform distribution $U(0, 1)$ and $\varepsilon(t)$ is a source of multiplicative noise that is also drawn from $U(0, 1)$. Another 5 variables, each drawn from $U(0, 1)$, are added as irrelevant predictors that do not contribute to $y(t)$. The statistical properties of the response are similar to those of observed precipitation series, including a highly skewed distribution bounded at zero and a multiplicative, nonlinear relationship with the predictors (Doswell et al., 1996).

### 4.3.5 Synthetic benchmark results

For each of the four synthetic benchmark tasks a total of 1000 cases were generated; 400 cases were used to train the models, and, once models were trained, these cases then formed the historical database from which predictions for the remaining 600 test cases were made. Predictor and predictand data were standardized to unit standard deviation and zero mean prior to modeling. For comparison with the NLAPA model, analog models with scores on the leading PCs selected as analog predictors (analog/PCA), analog models with the leading canonical variates selected as analog predictors (analog/CCA), analog models with predictor
scaling weights $s$ selected by DE optimization (analog/scale), and an analog model with all predictors scaled equally (analog/equal), were also tested.

Initial parameter vectors for population members in the DE algorithm were chosen with equal likelihood from the encoding layers of a trained NLPPA model and randomly from a uniform distribution $U(-5, 5)$. Control parameters for the DE algorithm were set following Ilonen et al. (2003) and Liu and Lampinen (2005). The reader is referred to Storn and Price (1997) and Price et al. (2005) for detailed information on the different control parameters and their effect on the optimization algorithm. Here, the best/1/binomial strategy was selected for use; crossover CR and step size $F$ parameters were both set to 0.9 and the number of population members $NP$ was set equal to the number of weights and biases in the encoding mapping of the NLAPA model. The DE algorithm was terminated following 5000 iterations and runs were repeated 50 times to avoid local minima.

The number of analog predictors and number of hidden nodes in the NLAPA model were selected via 5-fold cross-validation on the training cases. In each cross-validation trial, 80% of the data were reserved to train the model and the remaining 20% to validate the model. Amongst the cross-validation training cases, 80% of the cases were selected for the historical database and the remaining 20% of cases were used to calculate the cost function and set values of the weights and biases. Once the weights and biases were optimized, the cross-validation MSE was calculated on the validation cases. This procedure was repeated four times, each time calculating a new MSE value from the set of validation cases. The mean of the five MSE values was used as an estimate of generalization error; the number of analog predictors and number of hidden nodes were selected to minimize this estimate. Once the optimum numbers of analog predictors and hidden nodes were found, the NLAPA model was retrained on the 600 training cases; 80% of the cases were randomly selected to act as the historical database and 20% of the cases were used to calculate the cost function. The retrained NLAPA model was used to make predictions on the test dataset.

The number of PCs in the analog/PCA model was selected via 5-fold cross-validation. Following Fernandez and Saenz (2003), PC scores were scaled so that the variance of each series was proportional to the amount of variance it explained in the predictor dataset. Because the synthetic test benchmarks all contained a single predictand, a single canonical variate was extracted from each predictor/predictand dataset for use as the analog predictor in the analog/CCA models.

Model performance for the synthetic benchmarks, measured via root MSE (RMSE) on the test cases, is shown in Table 4.1. NLAPA performed best or was tied for best performance with another analog model on all four benchmarks. On the linear benchmark, NLAPA performed as well as the analog/CCA model, which, due to the linear nature of the task, was also expected to perform well. Both the analog/PCA and analog/scale models also accounted for the linear predictor/predictand relationship in reasonable fashion, with the analog/equal model performing the poorest. On the nonlinear dataset, NLAPA and analog/scale models performed the best.
Table 4.1: Root MSE (RMSE) of analog model predictions on test cases of the four synthetic benchmark tasks described in Section 4.3. The bold font indicates the best performing method and underlining indicates the worst performing method.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE Linear</th>
<th>RMSE Nonlinear</th>
<th>RMSE Multicollinear</th>
<th>RMSE Multiplicative</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLAPA</td>
<td>0.62</td>
<td>0.88</td>
<td>0.31</td>
<td>0.78</td>
</tr>
<tr>
<td>Analog/PCA</td>
<td>0.64</td>
<td>1.06</td>
<td>0.37</td>
<td>1.03</td>
</tr>
<tr>
<td>Analog/CCA</td>
<td>0.62</td>
<td>1.43</td>
<td>0.31</td>
<td>1.07</td>
</tr>
<tr>
<td>Analog/scale</td>
<td>0.64</td>
<td>0.88</td>
<td>0.35</td>
<td>0.98</td>
</tr>
<tr>
<td>Analog/equal</td>
<td>0.71</td>
<td>1.09</td>
<td>0.38</td>
<td>1.11</td>
</tr>
</tbody>
</table>

All other models performed poorly, with the analog/CCA model lagging well behind the other models; the nonlinearity of this dataset was not well captured by the linear canonical variate. On the multicollinear dataset, NLAPA and analog/CCA models tied for best performance, followed by the analog/scale, analog/PCA, and analog/equal models. Finally, on the multiplicative dataset, the NLAPA model outperformed all other models by a wide margin. Linear scaling and linear dimensionality reduction algorithms were unable to capture the nonlinear and multiplicative predictor/predictand relationship.

Results for the analog/scale mode on the first three synthetic datasets were consistent with those reported by Mehrhotra and Sharma (2006); in all cases, optimization of the scaling weights lead to improved performance relative to the analog/equal model that weighted all predictors equally. In no case, however, did the analog/scale model outperform NLAPA. Performance of the analog/CCA model was inconsistent. It performed well on datasets with linear predictor/predictand relationships, but performed poorly on datasets with nonlinearity. Again, in no case did the analog/CCA model outperform NLAPA.

4.4 Multivariate downscaling

In the previous section, the performance of the NLAPA model was compared against the performance of other analog models on synthetic test datasets. In this section, NLAPA is applied to real-world climate datasets. Synoptic-scale atmospheric circulation data were matched with two sets of multivariate surface climate data to form downscaling datasets suitable for testing the NLAPA model. The first task involved downscaling daily temperature, rainfall, and snowfall observed at a station near Whistler, B.C. The second task involved downscaling precipitation observed at 10 sites along the coast of B.C. In both cases, synoptic-scale atmospheric circulation data were taken from the National Centers for Environmental Prediction-Department of Energy (NCEP-DOE) AMIP-II Reanalysis (Kanamitsu et al., 2002).
4.4.1 Synoptic-scale atmospheric circulation dataset

Gridded mean sea-level pressure, 700-hPa relative humidity, 700-hPa horizontal wind components, and 500-hPa geopotential height data from the NCEP-DOE AMIP-II Reanalysis were chosen as synoptic-scale predictor fields for the analog downscaling models. Predictor fields are representative of low-level circulation, lower-troposphere moisture availability and circulation, and mid-troposphere circulation conditions that have been used in previous analog-based downscaling studies applied to hydroclimatological variables in B.C. (e.g., Whitfield and Cannon, 2002).

Twenty years of daily data from 1979-1998 (totaling 7305 days) were obtained for a region covering western North America and the North Pacific Ocean at a grid spacing of 2.5×2.5 degrees. The spatial domain is shown in Figure 4.3. The synoptic-scale fields (totaling 550 variables) were combined into a single data matrix, standardized, and then compressed using combined PCA. Scores on the first 15 PCs, accounting for 85% of the variance in the original dataset, were retained as potential predictors in analog models built for the two multivariate downscaling tasks. In all cases, PC scores were scaled according to explained variance on the predictor dataset.

4.4.2 Whistler temperature and precipitation dataset

Analog downscaling models were tested on a single station, multivariate climate dataset observed near Whistler, B.C. (1048898 Whistler, elevation 2158 m); the station location is shown in Figure 4.3. Daily mean temperatures, rainfall amounts, and snowfall amounts from 1979-1998 were obtained from the National Climate Archive of the Meteorological Service of Canada and were used as predictands in the analog downscaling models. Predictors were the retained PC scores from the NCEP-DOE AMIP-II synoptic-scale climate fields.

The retained PCs and the predictand series were split into two datasets. The first 4000 days were selected as the training dataset and the remaining 3305 days as the test dataset. Two benchmark models, one with PC scores as analog predictors (analog/PCA) and the other with canonical variates as analog predictors (analog/CCA), were created to compare against the NLAPA model.

Five-fold cross-validation on the training cases was used to select the number of retained PCs and canonical variates in the analog/PCA and analog/CCA models, as well as the number of hidden nodes and analog predictors in the NLAPA model. Scores on the 8 leading combined PCs were selected as analog predictors in the analog/PCA model. The analog/CCA model employed PCA pre-filtering to reduce the dimensionality of the original grid point predictors prior to application of the CCA model (Barnett and Preisendorfer, 1987). Three leading canonical variates were selected from the 15 retained combined PCs and 3 predictands; the retained canonical variates were used as analog predictors in the analog/CCA model. Two sets of scaling weights were evaluated for predictor variables in the analog/CCA model: all canonical variates scaled equally; and canonical variates scaled according to explained variance on the predictand.
Figure 4.3: Map showing the extent of the synoptic-scale fields from the NCEP-DOE AMIP-II Reanalysis, with the small dots representing centers of the model grid points. The square shows the location of Whistler, B.C. The large dots show locations of the ten daily precipitation observing stations.
Table 4.2: Performance of analog models on test cases for the Whistler downscaling dataset described in Section 4.4. Model statistics are RMSE for temperature (\( T \)), rainfall (\( R \)), and snowfall (\( S \)) predictions; percentage explained variance (EV) for daily temperature, rainfall, and snowfall predictions; ratios of predicted to observed number of temperature events < 0°C (FREEZE); ratios of predicted to observed number of wet days (WET); and ratios of predicted to observed number of snow days (SNOW). Values for the best performing model are shown in bold and values for the worst performing model are underlined.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE ( T ) (°C)</th>
<th>RMSE ( R ) (mm)</th>
<th>RMSE ( S ) (mm)</th>
<th>EV ( T ) (%)</th>
<th>EV ( R ) (%)</th>
<th>EV ( S ) (%)</th>
<th>FREEZE ( T )</th>
<th>WET ( R )</th>
<th>SNOW ( S )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLAPA</td>
<td>4.3</td>
<td>6.2</td>
<td>4.3</td>
<td>72.4</td>
<td>15.9</td>
<td>8.1</td>
<td>1.10</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>Analog/PCA</td>
<td>4.8</td>
<td>6.5</td>
<td>4.7</td>
<td>66.3</td>
<td>6.9</td>
<td>6.9</td>
<td>1.04</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>Analog/CCA</td>
<td>4.4</td>
<td>6.2</td>
<td>5.1</td>
<td>72.8</td>
<td>15.6</td>
<td>3.9</td>
<td>1.17</td>
<td>1.01</td>
<td>1.09</td>
</tr>
</tbody>
</table>

dataset. Little difference in performance was noted between the two, although cross-validation results did favor equal scaling. Reported results are thus for models with canonical variates given equal importance. The NLAPA model was built with three nodes in the hidden layer and three analog variables leading to the analog output layer. Optimization of the weights and biases followed the procedure described in the previous section for the synthetic datasets.

Values of root mean squared error (RMSE) and explained variance (EV) (i.e., squared Pearson product-moment correlation) of predicted temperature, rainfall, and snowfall; the ratio of predicted to observed number of days below freezing; the ratio of predicted to observed number of rainy days; and the ratio of predicted to observed number of snowy days are reported in Table 4.2. NLAPA performed best outright or was tied for best performance on 6 of the 9 measures of model performance and performed worst on just one. Analog/CCA performed best on three measures and worst on three measures, whereas analog/PCA performed best on just one measure and worst on 5 measures. For RMSE and EV, NLAPA and analog/CCA models yielded comparable levels of performance and generally outperformed analog/PCA. With the exception of analog/CCA and the ratio of predicted to observed days below freezing (1.17), predicted numbers of freezing, rain, and snow days were within 10% of observed values for all models.

To gain more insight into the models’ ability to reproduce observed distributions of temperature, rain amounts, and snow amounts, quantile-quantile plots for each model and predictand are shown in Figure 4.4. Visually, all three models represented the temperature distribution well. Under-prediction of rainfall quantiles was evident for all models, with analog/PCA performing worst, particularly for extreme events. Both analog/PCA and analog/CCA models tended to over-predict snowfall quantiles (more so for analog/CCA), whereas the NLAPA model reproduced the observed distribution well at all quantiles.
Figure 4.4: Plots of predicted versus observed quantiles of daily temperature, rain amounts, and snow amounts for NLAPA, analog/PCA, and analog/CCA models.
Table 4.3: Daily precipitation observing stations

<table>
<thead>
<tr>
<th>Station name</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victoria Int’l A</td>
<td>48.65</td>
<td>-123.43</td>
<td>20</td>
</tr>
<tr>
<td>Comox A</td>
<td>49.72</td>
<td>-124.90</td>
<td>24</td>
</tr>
<tr>
<td>Port Hardy A</td>
<td>50.68</td>
<td>-127.37</td>
<td>22</td>
</tr>
<tr>
<td>Estevan Point</td>
<td>49.38</td>
<td>-126.55</td>
<td>7</td>
</tr>
<tr>
<td>Pachena Point</td>
<td>48.72</td>
<td>-125.10</td>
<td>37</td>
</tr>
<tr>
<td>Tofino A</td>
<td>49.08</td>
<td>-125.78</td>
<td>24</td>
</tr>
<tr>
<td>Powell River</td>
<td>49.83</td>
<td>-124.50</td>
<td>52</td>
</tr>
<tr>
<td>Agassiz CDA</td>
<td>49.25</td>
<td>-121.77</td>
<td>15</td>
</tr>
<tr>
<td>Seymour Falls</td>
<td>49.43</td>
<td>-122.97</td>
<td>244</td>
</tr>
<tr>
<td>Vancouver Int’l A</td>
<td>49.20</td>
<td>-123.18</td>
<td>3</td>
</tr>
</tbody>
</table>

4.4.3 Multi-site precipitation dataset

Analog downscaling models were also tested on multi-site precipitation data observed along the south coast of B.C. Daily precipitation data from 10 stations for the period 1979-1998 were obtained from the Adjusted Historical Canadian Climate Data records developed by Mekis and Hogg (1999). Station locations are given in Table 4.3 and are shown in Figure 4.3. Data were used as predictands in the analog downscaling models; predictors were the retained PC scores from the NCEP-DOE AMIP-II synoptic-scale climate fields.

As in the previous downscaling task, the retained PCs and the precipitation variables were split into two datasets. The first 4000 days were selected as the training dataset and the remaining days as the test dataset. Cross-validation was used to select the number of retained PCs, canonical variates, hidden nodes, and analog predictors used in the analog/PCA, analog/CCA, and NLAPA models. Scores on the 8 leading combined PCs were selected as analog predictors in the analog/PCA model. For use in the analog/CCA model, 8 leading canonical variates were selected from the 15 retained combined PCs and 10 precipitation series. Based on cross-validation results, each canonical variate was given equal weight in the analog model. The NLAPA model was built with three nodes in the hidden layer and three analog variables leading to the analog output layer. Optimization of the weights and biases followed the procedure described in the previous section for the synthetic dataset.

Mean values of RMSE, EV, ratio of predicted to observed number of wet days, and ratio of predicted to observed number of days with more than 100 mm of precipitation over the 10 precipitation grid points are reported in Table 4.4. The NLAPA model outperformed both the analog/PCA and analog/CCA models on all statistics except number of wet days, where it and the analog/PCA model performed slightly poorer than analog/CCA (ratios of 0.97 versus 0.99 for analog/CCA). All models gave roughly equal RMSE values, although NLAPA explained more slightly more variance (16.2%) than either analog/PCA (10.9%) or analog/CCA (12.8%). Of the three models, only NLAPA came close to predicting the correct number of days with
Table 4.4: Performance of analog models on test cases for the multi-site daily precipitation downscaling dataset described in Section 4.4. Model statistics are RMSE of predictions; percentage explained variance of predictions; ratios of predicted to observed number of wet days; and ratios of predicted to observed number of days with more than 100 mm of precipitation (EXTREME). Results are mean values over the ten sites. Values for the best performing model are shown in bold and values for the worst performing model are underlined.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (mm)</th>
<th>EV (%)</th>
<th>WET</th>
<th>EXTREME</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLAP A</td>
<td>12.4</td>
<td>16.2</td>
<td>0.97</td>
<td>0.76</td>
</tr>
<tr>
<td>Analog/PCA</td>
<td>12.7</td>
<td>10.9</td>
<td>0.97</td>
<td>0.29</td>
</tr>
<tr>
<td>Analog/CCA</td>
<td>12.7</td>
<td>12.8</td>
<td>0.99</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Figure 4.5: Plots of predicted versus observed inter-site daily precipitation correlations for NLAP A, analog/PCA, and analog/CCA models.

more than 100 mm of precipitation (ratio of 0.76 versus just 0.29 for analog/PCA and 0.42 for analog/CCA).

As a check on the ability of the models to preserve spatial relationships between stations, correlation coefficients between observed precipitation series at the 10 stations (i.e., entries in the upper triangle of the observed correlation matrix) are plotted against correlation coefficients derived from the modeled precipitation series in Figure 4.5. As all predictions were from analog models with one nearest neighbor, all three models realistically reproduced observed spatial variability of the precipitation field. Mean departures from the 1:1 line, as measured by RMSE between observed and predicted inter-site correlations, were greatest for the analog/CCA model (RMSE of 0.045); NLAP A and analog/PCA models both performed slightly better than the analog/CCA model (RMSE of 0.040).

4.5 Discussion and conclusion

This study introduced a new method for synoptic downscaling, called the NLAP A model, which is based on the combination of an ANN with an analog model. The NLAP A model consists of an
ANN encoding layer coupled to a decoding layer that maps the transformed analog predictors to the predictands via an analog model. Parameters in the ANN portion of the NLAPA model are found by minimizing the MSE between observed predictands and predictions from the NLAPA model. In the coupled model structure, the nonlinear transformation performed by the ANN portion of the model is guided by the performance of the analog layer.

Comparisons between the NLAPA model and other analog models on synthetic datasets suggest that it is capable of matching the performance of analog models that rely on linear dimension reduction or predictor scaling algorithms when the underlying predictor-predictand relationships are linear, and beating the performance of these models when the underlying relationships are nonlinear. Tests on real-world datasets suggest that NLAPA can resolve complicated synoptic-to-local-scale relationships, including modeling of non-normal, conditional variables like precipitation, while preserving inter-variable relationships and spatial relationships between sites.

The coupling of ANN and analog models into a single model means that gradient-free optimization methods must be used to set values of the weights and biases in the ANN portion of the model. The DE algorithm (Price et al., 2005; Storn and Price, 1997) was used in this study, with recommended values of its meta-parameters taken from Liu and Lampinen (2005). While results from the test datasets suggest that this set of meta-parameters worked well in practice, alternative strategies should be tested to assess the relative convergence speed and robustness of different optimization procedures.

The NLAPA model, like all analog methods, cannot, due to the sampling step of the analog operator A, predict values outside of the range of the historical record. This means that new record values cannot be set, which may not be realistic when using the model for downscaling from Global Climate Model climate scenarios. In such cases, a post-processing step may be needed to allow extrapolation beyond the range of the historical record, for example using the technique proposed by Imbert and Benestad (2005).

While not explicitly tested on real-world datasets other than those related to downscaling, the NLAPA model could also be used for more general prediction and forecasting tasks in earth systems science. For example, analog models have been used to predict tropical cyclone tracks (Sievers et al., 2000), assess avalanche hazards (Gassner and Brabec, 2002), perform seasonal climate prediction (Barnston and Livezey, 1989), and make short-term weather forecasts (Vandendool, 1989), among other tasks. More work is needed to assess the applicability of the NLAPA model to these subject areas.
Bibliography


Chapter 5

Multivariate ridge regression with negative ridge parameters

5.1 Introduction

Linear regression models for climate downscaling regress local-scale weather elements onto synoptic-scale atmospheric predictors. In general, the downscaling mapping will not account for all of the predictand variance due to unexplained sub-grid scale or other phenomena. To accurately simulate the statistical properties of the predictand, additional steps are therefore required to increase the variance of the predicted series. Three methods have been proposed: (i) inflation, which linearly rescales the predicted series so that its variance matches that of the observed series (Karl et al., 1990); (ii) randomization, which adds random noise to the predicted series (von Storch, 1999); and (iii) expansion, which adds a constraint to the regression cost function that forces observed and predicted variances to be equal (Bürger, 1996, 2002). Expansion, or expanded downscaling, can also be applied to multiple predictand variables, for example weather elements at different observing sites. In this case, the covariance matrix of the predicted series is constrained to match that of the observations.

Each method has advantages and disadvantages. For example, von Storch (1999) suggested that the implicit assumption of inflation and expansion, namely that all the variability in the predictand dataset is related to variability in the predictor dataset, is improper given that predictors in downscaling are derived from synoptic-scale fields with no representation of sub-grid-scale variability. However, Bürger and Chen (2005) found that randomization based on a static noise model failed to represent local changes in atmospheric variability in a climate change simulation, something that was reasonably well accounted for by expanded downscaling. They also found that spatial correlations between sites were misrepresented by inflation, while expansion showed extreme sensitivity to deviations of predictors/predictands from normality.

In terms of their computation, inflation and randomization are applied in post-processing following standard estimation of the regression coefficients, while the constrained cost function of expanded downscaling means that a nonlinear optimization algorithm must be used to solve

---

for the regression coefficients. Expanded downscaling is thus more computationally intensive, but is also more flexible as it is designed to be applied to multivariate downscaling tasks.

As a complement to the three existing techniques for specifying predictand variance and covariance structure, a new method based on the ridge regression model (Brown and Zidek, 1980; Hoerl and Kennard, 1970) with negative ridge parameters (Hua and Gunst, 1983) is introduced in the current study. Whereas positive ridge parameters result in shrinkage of regression coefficients and decreased model variance, negative ridge parameters can effect the opposite change, thus leading to more realistic predicted variance in the case of univariate ridge regression, and covariance, in the case of multivariate ridge regression.

5.2 Data

Models are demonstrated on two simplified downscaling problems. In the first, predictand variables are daily surface mean temperature observations at ten stations in southern British Columbia, Canada (T1 Victoria Int’l A, T2 Comox A, T3 Port Hardy A, T4 Estevan Point, T5 Victoria Gonzales Hts, T6 Bella Coola, T7 Tatlayoko Lake, T8 Agassiz CDA, T9 Princeton A, and T10 Abbotsford A). Data are taken from the Adjusted Historical Canadian Climate Data archive (Mekis and Hogg, 1999; Vincent et al., 2002) for the period 1959-1998. The minimum correlation between stations is 0.88 between T4 and T7, and the maximum is 0.98 between T8 and T10. With the seasonal cycle removed, the minimum falls to 0.62 between T1 and T6 and the maximum to 0.91 between T1 and T10. Downscaling predictors are from the ECMWF ERA-40 Re-analysis 850-hPa temperature field (Brankovic and Molteni, 2004). Data on a 2.5°×2.5° grid are extracted over a domain covering the eastern Pacific Ocean and western North America (35°N-65°N and 145°W-105°W).

The second problem concerns downscaling of daily precipitation at ten stations, again located in southern British Columbia (P1 Victoria Int’l A, P2 Comox A, P3 Port Hardy A, P4 Estevan Point, P5 Pachena Point, P6 Tofino A, P7 Powell River, P8 Agassiz CDA, P9 Seymour Falls, and P10 Vancouver Int’l A). Data are taken from the AHCCD archive for the period 1959-1998. The minimum correlation between stations is 0.29 between P1 and P4, and the maximum is 0.89 between P4 and P6. Downscaling predictors for the precipitation data are ECMWF ERA-40 data (mean sea level pressure, 850-hPa relative humidity, 850-hPa temperature, 700-hPa relative vorticity, and 500-hPa geopotential height fields) on the same grid as described above for the temperature predictors.

5.3 Univariate ridge regression

Given an $n \times p$ predictor matrix $X$, where $n$ is the number of observations and $p$ is the number of predictor variables, and an $n \times 1$ predictand vector $y$, the multiple linear regression model is given by
Chapter 5. Multivariate ridge regression with negative ridge parameters

\[ y = Xb + e \]  

(5.1)

in which \( b \) is the \( p \times 1 \) vector of regression coefficients to be estimated and \( e \) is the \( n \times 1 \) error vector.

In the univariate ridge regression model (Hoerl and Kennard, 1970), \( b \) is estimated by solving a penalized least-squares cost function

\[ C = \frac{1}{n} \sum_{i=1}^{n} e_i^2 + \frac{k}{n} \sum_{j=1}^{p} b_j^2 \]  

(5.2)

where \( k \) is the ridge penalty parameter. The first term is the mean squared error (mse) of the predictions and the second term is a penalty on the square of the regression coefficients. The global minimum of \( C \) is found by the following estimator

\[ b = (X^TX + kI_p)^{-1}X^T y \]  

(5.3)

where \( I_p \) is the \( p \times p \) identity matrix. Values of \( k > 0 \) tend to reduce the magnitude of the estimated regression coefficients, leading to fewer effective model parameters. Ridge regression can be useful for avoiding overfitting (i.e., fitting to noise in the data rather than the signal) and improving numerical stability in the case of an ill-conditioned \( X^TX \) matrix.

5.4 Multivariate ridge regression

Given an \( n \times q \) predictand matrix \( Y \), where \( q \) is the number of predictand variables, the multivariate linear regression model is given by

\[ Y = XB + E \]  

(5.4)

in which \( B \) is the \( p \times q \) matrix of regression coefficients to be estimated and \( E \) is the \( n \times q \) error matrix, where errors for each of the \( q \) predictands are assumed to be independent. The ridge regression estimator

\[ B = (X^TX + kI_p)^{-1}X^T Y \]  

(5.5)

thus yields the same regression coefficients as Eq. 5.3 were it to be applied to each of the \( q \) predictands separately.

Alternatively, Brown and Zidek (1980) proposed a multivariate ridge regression estimator that takes into consideration information across the \( q \) equations, i.e., the between regressions covariance matrix. The penalized least-squares estimator, in the stacked matrix form of Haitovsky (1987), is given by
\[ \mathbf{b}^* = (\mathbf{I}_q \otimes \mathbf{X}^T + \mathbf{K} \otimes \mathbf{I}_p)^{-1}(\mathbf{I}_q \otimes \mathbf{X}^T)\mathbf{y} \]  

(5.6)

where \( \otimes \) denotes the Kronecker matrix product. For example, \( \mathbf{K} \otimes \mathbf{I}_p \) results in the \( qp \times qp \) block matrix

\[
\mathbf{K} \otimes \mathbf{I}_p = \begin{pmatrix}
k_{11}\mathbf{I}_p & \cdots & k_{1q}\mathbf{I}_p \\
\vdots & \ddots & \vdots \\
k_{q1}\mathbf{I}_p & \cdots & k_{qq}\mathbf{I}_p
\end{pmatrix}
\]  

(5.7)

\( \mathbf{K} \) is the \( q \times q \) matrix of ridge parameters, \( \mathbf{y} = \text{vec} (\mathbf{Y}) \) is the stacked \( nq \times 1 \) vector, \( \mathbf{y} = (\mathbf{y}_1^T, ..., \mathbf{y}_q^T)^T \), and \( \mathbf{b}^* = \text{vec} (\mathbf{B}^*) \) is the stacked \( pq \times 1 \) vector of regression coefficients \( \mathbf{b}^* = (\mathbf{b}_1^T, ..., \mathbf{b}_q^T)^T \), where \( \mathbf{b}_q^* \) is the \( p \times 1 \) vector of regression coefficients corresponding to the \( q \)th predictand. With \( \mathbf{K} = 0 \), multivariate ridge regression reduces to standard multivariate regression. If \( \mathbf{K} \) is a diagonal matrix with elements \( \text{diag}(\mathbf{K}) = (k_1, ..., k_q)^T \) then the regression coefficients \( \mathbf{B}^* \) are equal to those found by applying the univariate ridge regression estimator (Eq. 5.3) to each of the \( q \) predictands. Otherwise, control over the modeled covariance is afforded by adjusting the off-diagonal elements of \( \mathbf{K} \).

### 5.5 Univariate ridge regression with negative \( k \)

As ridge regression is primarily used as a means of avoiding overfitting, the ridge parameter \( k \) is normally constrained to be non-negative. However, Hua and Gunst (1983) explored the use of negative ridge parameters, concluding that “the generalized ridge family can be expanded to include estimators which cannot otherwise be modeled as ridge estimators” but that “the efficacy of the use and implementation of [...] negative ridge parameter values is yet to be determined” (p. 43-44). These statements motivated the use of negative ridge parameter values for performing variance inflation.

To illustrate, consider a simplified univariate downscaling task in which surface temperatures at station T1 are regressed onto 850-hPa temperatures at the nearest grid point. For ease of interpretation, both series are standardized to zero mean and unit variance prior to modeling. With no ridge penalty \( (k/n = 0) \), the variance of the downscaled series \( \text{var}(\hat{y}) \) is 0.845 and its mse is 0.154. The series can be inflated following Karl et al. (1990) by multiplying the downscaled series by

\[
\beta = \sqrt{\frac{\text{var}(\mathbf{y})}{\text{var}(\hat{\mathbf{y}})}}
\]  

(5.8)

which gives a series with unit variance – matching the observed – but an increase in mse to 0.161. To investigate the effect of negative ridge parameters on the predictions, ridge regression models with \( k/n \) ranging from -0.2 to 0.2 are estimated via Eq. 5.3. Results are shown in Figure 5.1. While positive ridge parameters lead to series with decreased variance, negative ridge
Figure 5.1: Variances and mean squared errors of surface temperature predictions at station T1 plotted against scaled values of the ridge parameter $k/n$ from univariate ridge regression models. Stars indicate values for the model with no variance inflation ($k/n = 0$); solid dots correspond to the model in which predicted and observed variances are equal ($k/n = -0.08$).
parameters result in inflated variance. From a line search, predicted and observed variances are equal with $k/n = -0.08$. The mse of this model is 0.161, which is the same as the mse of the inflated model with no ridge penalty. In this case, ridge regression with a negative ridge parameter results in model predictions that are the same as those from regression with variance inflation.

In the univariate case, using ridge regression with negative ridge parameters confers no real advantage over variance inflation. Of more interest is the multivariate case, in which there are multiple predictand variables, as variance inflation alone is unable to correctly resolve inter-variable relationships (Bürger and Chen, 2005). The remainder of the paper deals with multivariate downscaling.

## 5.6 Multivariate ridge regression with negative elements in $K$

### 5.6.1 Model optimization

The multivariate ridge regression estimator, given by Eq. 5.6, is used to solve for the regression coefficients $B^*$ in multi-site downscaling. For a given ridge matrix $K$, this leads to the global optimum of the penalized mse cost function, i.e., the multivariate extension of Eq. 5.2. As one of the goals of multi-site downscaling is to accurately simulate the observed covariance matrix, the ridge matrix $K$ is found via a nonlinear optimization algorithm that minimizes the cost function

$$C_{\text{cov}} = \sum_{i=1}^{q} \sum_{j=1}^{q} (\text{cov}(y_i, y_j) - \text{cov}(\hat{y}_i, \hat{y}_j))^2$$

where $\text{cov}(\cdot)$ denotes the covariance between series. Elements of $K$ are free to take positive or negative values.

In contrast, the expanded downscaling model solves for $B$ by minimizing a constrained cost function, where the constraint on the predicted covariance matrix is satisfied by adding a penalty term to the least-squares error

$$C_{\text{ex}} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 + \lambda \sum_{i=1}^{q} \sum_{j=1}^{q} (\text{cov}(y_i, y_j) - \text{cov}(\hat{y}_i, \hat{y}_j))^2$$

in which the coefficient $\lambda$ controls the trade-off between the error term and the constraint term.

### 5.6.2 Temperature downscaling

ECMWF ERA-40 predictor fields and temperature predictand data described above are used to build a multivariate ridge regression model with elements of the ridge matrix $K$ free to take positive or negative values. Prior to modeling, the seasonal cycle is removed from all variables. 850-hPa gridded temperature anomalies are compressed via S-mode principal component anal-
Figure 5.2: Effect of the magnitude of the penalty coefficient $\lambda$ on the mean squared error and the covariance constraint term (i.e., the sum of squared deviations between elements of the observed and predicted covariance matrices) in Eq. 5.10.

As the objective is to accurately model the covariance matrix, $\mathbf{K}$ is assumed to be a symmetric, which results in $q(q+1)/2$ free parameters. A quasi-Newton optimization algorithm (Schnabel et al., 1985) is used to find elements of $\mathbf{K}$ that minimize Eq. 5.9. The optimization is initialized from a diagonal matrix, with $\text{diag}(\mathbf{K})$ determined by applying the line search described in the previous section to each predictand. In other words, the initial $\mathbf{K}$ results in coefficients $\mathbf{B}^*$ that give the same predictions as $q$ variance inflated univariate models.

Three benchmark models are built for comparison. The first two are multivariate regression models ($\mathbf{K} = 0$), one with and one without variance inflation of the $q$ series. The third is an expanded downscaling model in which regression coefficients $\mathbf{B}$ are found by minimizing Eq. 5.10 using a quasi-Newton optimization algorithm. The magnitude of the penalty coefficient
Figure 5.3: Observed versus predicted elements of the predictand covariance matrix for (a) the multivariate regression model, (b) the inflated multivariate regression model, (c) the expanded downscaling model, and (d) the multivariate ridge regression model. Linear fits to the data are shown by dashed lines.

\( \lambda \) is set to 100 for all expanded downscaling models. For reference, Figure 5.2 plots values of the error and constraint terms from expanded downscaling models trained with values of \( \lambda \) set to 0, 0.1, 1, 10, 100, and 1000. A penalty of 100 is sufficient to force the predicted covariance matrix to match the observed covariance matrix closely with only modest impact on the error term.

Elements of the observed and modeled covariance matrices are plotted in Figure 5.3. Without inflation, the multivariate regression model underestimates station variances and covariances between stations (Figure 5.3a). With inflation, variances match observations, but off-diagonal elements of the covariance matrix are now overestimated (Figure 5.3b). Expanded downscaling, with the added covariance constraint, accurately models the covariance matrix (Figure 5.3c), as does the multivariate ridge regression model with optimized \( \mathbf{K} \) (Figure 5.3d). Figure 5.4 shows a graphical representation of the optimized ridge matrix for the 1959-1988 training data. Diagonal elements are negative, which, from the previous section, leads to variance inflation. A mix of smaller magnitude off-diagonal elements, both positive and negative, reduces
Figure 5.4: A graphical representation of the optimized ridge matrix $K$ based on 1959-1988 temperature training data. Shades represent values of $K/n$.

Table 5.1: Performance statistics for model predictions of temperature.

<table>
<thead>
<tr>
<th>Model</th>
<th>rmse ($^\circ$C)</th>
<th>Explained variance (%)</th>
<th>rmse$_{COV}$ ($^\circ$C$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate regression</td>
<td>1.53</td>
<td>94.6</td>
<td>2.62</td>
</tr>
<tr>
<td>Inflated multivariate regression</td>
<td>1.56</td>
<td>94.6</td>
<td>1.96</td>
</tr>
<tr>
<td>Expanded downscaling</td>
<td>1.72</td>
<td>93.4</td>
<td>0.33</td>
</tr>
<tr>
<td>Multivariate ridge regression</td>
<td>1.64</td>
<td>93.9</td>
<td>0.42</td>
</tr>
</tbody>
</table>

the overestimation of off-diagonal elements of the covariance matrix that would result if $K$ were diagonal, i.e., as seen in Figure 5.3b for the inflated multivariate regression model.

Cross-validated performance statistics, including root mean squared error (rmse), percent explained variance (defined here as the square of the Pearson product-moment correlation), and rmse between predicted and observed covariance matrices (rmse$_{COV}$) for the models are summarized in Table 5.1. Values of rmse$_{COV}$ are slightly lower for expanded downscaling than for multivariate ridge regression, which means that the constraint on the covariance matrix is better met by expanded downscaling than by multivariate ridge regression. As expected from Figure 5.2, an improvement in covariance performance by expanded downscaling comes with an attendant increase in rmse and decrease in explained variance relative to multivariate ridge regression. The difference in performance suggests that the quasi-Newton optimization algorithm for $K$ in the multivariate ridge regression model may be getting trapped in a sub-
optimal local minimum of the cost function surface, which would tend to reduce the model’s ability to satisfy the covariance constraint. Alternatively, the difference may be due to the choice of $\lambda$ for the expanded downscaling model. Regardless, performance levels are comparable, suggesting that this may be of little practical concern.

As expected, both expanded downscaling and multivariate ridge regression, with the added constraints on the covariance matrix, have higher rmse and lower explained variance that multivariate regression and inflated multivariate regression. As inflation is a simple linear transformation, both multivariate regression and inflated multivariate regression have the same explained variance, although inflation leads to a model with slightly higher rmse.

### 5.6.3 Precipitation downscaling

The temperature downscaling task is relatively straightforward, as both synoptic-scale and surface temperature fields vary smoothly in space. A more challenging task involves downscaling of precipitation, which is highly variable over small spatial scales. Also, whereas temperature data are close to being normally distributed, precipitation data are strongly non-normal, and this makes modelling more difficult.

To accommodate the non-normality and mixed discrete/continuous nature of the precipitation series, a normalization step, as in Bürger (1996, 2002), is undertaken prior to modelling. Data are first split into 12 groups, 1 for each month. Next, the Bernoulli-gamma distribution (Cawley et al., 2007; Haylock et al., 2006; Williams et al., 1998) is fit, in turn, to precipitation amounts in each group at each station. The Bernoulli-gamma distribution is a mixed discrete/continuous function whose probability density function is given by

$$f(y; p, \alpha, \beta) = \begin{cases} 1 - p & \text{for } y = 0 \\ \frac{p(y/\beta)^{\alpha-1} \exp(-y/\beta)}{\beta \Gamma(\alpha)} & \text{for } y > 0 \end{cases}$$

(5.11)

where $y$ is the precipitation amount, $p$ ($0 \leq p \leq 1$) is the probability of precipitation, $\alpha$ ($\alpha > 0$) is the shape parameter of the gamma distribution, $\beta$ ($\beta > 0$) is the scale parameter of the gamma distribution, and $\Gamma(\cdot)$ is the gamma function. Based on the fitted Bernoulli-gamma parameters, the Bernoulli-gamma cumulative density function is used to express precipitation amounts as cumulative probabilities ranging from 0 to 1. Following Bürger (1996, 2002), cumulative probabilities on dry days are randomly drawn from a uniform distribution ranging from 0 to $1 - p$. (The use of random synthetic values for dry days allows data to be mapped onto the full range of the normal distribution.) Finally, data are normalized by applying the standard normal inverse cumulative density function to the series of cumulative probabilities.

Following normalization of the predictand series, extended PCA is applied to the deseasonalized ECMWF ERA-40 predictor fields. The leading 12 PCs, accounting for 70% of the dataset variance, are retained as predictors. The model training and cross-validation procedure described in the previous section for temperature is then used to complete the precipitation
downscaling.

Elements of the observed and modelled covariance matrices are plotted in Figure 5.5. Multivariate regression without inflation drastically underestimates magnitudes of the variances and covariances (Figure 5.5a), whereas inflation correctly represents variances at stations but results in an overestimation of covariances (Figure 5.5b). Expanded downscaling (Figure 5.5c) and multivariate ridge regression (Figure 5.5d) both provide a more accurate representation of the overall covariance structure, although a small negative bias is evident for the multivariate ridge regression model. However, as shown in Table 5.2, rmseCOV is slightly lower for multivariate ridge regression (20.6 mm²) than for expanded downscaling (22.9 mm²). Satisfying the

<table>
<thead>
<tr>
<th>Model</th>
<th>rmse (°mm)</th>
<th>Explained variance (%)</th>
<th>rmseCOV (°mm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate regression</td>
<td>11.5</td>
<td>23.1</td>
<td>92.7</td>
</tr>
<tr>
<td>Inflated multivariate regression</td>
<td>13.2</td>
<td>23.1</td>
<td>54.2</td>
</tr>
<tr>
<td>Expanded downscaling</td>
<td>13.6</td>
<td>18.1</td>
<td>22.9</td>
</tr>
<tr>
<td>Multivariate ridge regression</td>
<td>12.8</td>
<td>18.9</td>
<td>20.6</td>
</tr>
</tbody>
</table>
covariance constraint leads to a drop in model skill for these two models, although values of rmse are only marginally worse – around 1.5-2 mm – than multivariate regression.

Figure 5.6 shows a graphical representation of the optimized ridge matrix $K$ for the multivariate ridge regression model. Diagonal elements are strongly negative, which implies a large amount of variance inflation, as expected from Figure (Figure 5.5a). Off-diagonal elements are mostly positive, but are of much smaller magnitude, and this reduces the positive bias that would result if variance inflation alone were applied to the model predictions (Figure 5.5b).

### 5.7 Computational cost

Assessing the relative computational cost of multivariate ridge regression versus expanded downscaling is complicated. The number of predictors $p$, the number of predictands $q$, the number of observations $n$, the nature of the nonlinear optimization algorithms, the nature of the matrix inversion algorithm used to solve Eq. 5.6, and the software/programming environment all play important roles in determining the overall computational burden. Consider, for example, the relative amount of time required by the two models to solve the multivariate downscaling task. Timing ratios between multivariate ridge regression and expanded downscaling are presented in Figure 5.7 for different numbers of predictand variables $q$, with the number of predictors $p$, and number of observations $n$ held fixed. Under these conditions, multivariate ridge regression
Figure 5.7: Time required to optimize the multivariate ridge regression cost function (Eq. 5.9), expressed as a ratio with respect to the amount of time required to optimize the corresponding expanded downscaling model. Ratios in excess of one mean that the multivariate ridge regression model take longer to converge than the expanded downscaling model. Values are for different numbers of model predictands $q$. 

![Diagram of time required to optimize the multivariate ridge regression cost function]
converges faster than expanded downscaling for $q = 1, \ldots, 5$ and slower for $q = 6, \ldots, 10$.

5.8 Discussion and conclusions

This paper proposes the use of negative ridge parameters as a means of controlling the covariance structure of a multivariate linear model. Results from two simple multi-site downscaling tasks suggest that proper specification of the ridge matrix $K$ can lead to a model that exhibits realistic levels of variability between sites. Results are comparable to those from the expanded downscaling model of Bürger (1996), although more work is needed to find a more robust optimization algorithm for $K$. Initial results using a global optimization algorithm show promise.

Because optimization of $K$ and the regression coefficients $B^*$ are decoupled in multivariate ridge regression, problem specific cost functions can easily be incorporated into the model building process. For example, consider two of the findings of Bürger and Chen (2005): (i) that the performance of expanded downscaling suffered due to sensitivity of the model to deviations of variables from normality; and (ii) that the correct modeling of the temporal sequencing of events can be as important to the success of a downscaling model as correct specification of covariances. As both multivariate ridge regression and expanded downscaling have the same basic linear model structure, these two problems are shared by multivariate ridge regression. It would, however, be straightforward to use multi-objective optimization to simultaneously minimize Eq. 5.9 along with additional cost functions that gage how well the predictand distributions and autocorrelations are modeled. Because the multi-objective optimization would be wrapped around the matrix equation for $B^*$, identification of optimum predictor transformation parameters could, for example, also be included in the optimization process. Expanded downscaling does not afford this level of flexibility.

The use of negative ridge parameters is not confined to multivariate linear models. Nonlinear models that penalize the magnitude of model parameters, for example the weight penalties in nonlinear canonical correlation analysis (NLCCA) and nonlinear principal predictor analysis (NLPPA) (Cannon, 2006; Cannon and Hsieh, 2008; Hsieh, 2004), might also benefit from taking negative values. However, as the optimization of these models are computationally expensive to start with, performing an additional optimization over multiple penalty parameters may be infeasible. Another option would be to employ kernel methods, which have been proposed as alternatives to the neural networks used in NLCCA and NLPPA, as a means of injecting nonlinearity into a linear downscaling model (Scholkopf and Smola, 2002). Kernel methods have recently started to be applied to downscaling problems, although only in the context of univariate, not multivariate, regression (Tripathi et al., 2006).

In kernel regression, the linear regression problem is reformulated by nonlinearily mapping the predictors from the $p$ dimensional input space to a much higher dimensional feature space. With an appropriate choice of nonlinear mapping function, a regression problem that is non-
linear in input space becomes linear in feature space. In the multivariate ridge estimator (Eq. 5.6), the number of free parameters in the ridge matrix $\mathbf{K}$ is $q(q + 1)/2$, whereas the size of the matrix of regression coefficients $\mathbf{B}$ or $\mathbf{B}^*$ equals the product of the number of predictors and predictands ($p \times q$). In kernel methods, the number of predictors equals the number of cases $n$, which, for daily time series, is usually quite large. This means that the nonlinear constrained optimization required to find $\mathbf{B}$ in expanded downscaling may be impractical with kernel methods, whereas, in multivariate ridge regression, optimization over $\mathbf{K}$ (with $\mathbf{B}^*$ estimated via Eq. 5.6) may be feasible. A study exploring the use of negative ridge parameters in multivariate kernel ridge regression is currently underway.
Bibliography


Bibliography


Chapter 6

Expanded Bernoulli-gamma density network

6.1 Introduction

Statistical downscaling models are used to estimate weather data at a station or stations based on atmospheric circulation data from a numerical weather prediction model or Global Climate Model (GCM). Once a model has been developed, downscaled data can be then be entered into an environmental model, for example a hydrological model for streamflow in a watershed, that requires climate information at a finer-scale (Salathe, 2005; Xu, 1999). Historically, most statistical downscaling models have been developed for observations at a single station. If data are required at multiple stations, then a spatial or multi-site model is required. Maintaining realistic downscaling relationships between sites is particularly important in hydrological models, as streamflow depends strongly on the spatial distribution of precipitation in a watershed. More generally, precipitation is a difficult variable to downscale because of its non-normal distribution and its discontinuous behavior in space and time.

Statistical downscaling methods can be placed into three main categories (Wilby and Wigley, 1997), (i) regression models, (ii) weather classification schemes, and (iii) weather generators, with most multi-site precipitation downscaling algorithms falling into the latter two categories. For example, see work by Wilks (1998), Rajagopalan and Lall (1999), Zorita and von Storch (1999), Charles et al. (1999), Buishand and Brandsma (2001), Yates et al. (2003), and Gangopadhyay et al. (2005), among others. Wilks and Wilby (1999) review stochastic weather generators, including a brief discussion of multi-site methods and approaches that include a weather classification component.

In practice, all three categories of downscaling models have strengths and weaknesses. In the case of regression, models are tightly linked with the the large-scale circulation, and, as a result, their application to future climate scenarios from GCMs is straightforward. The goal of this study then is to address some of the shortcomings of the regression-based approach – namely poor representation of extreme events, poor representation of observed variance/covariance,
and assumption of linearity/normality of data (Wilby et al., 2004) – by introducing a regression model that is fundamentally nonlinear, probabilistic, and targeted on multiple sites. To lay the groundwork for the proposed model, a review of current regression approaches to downscaling is first presented.

Recent studies have adopted multivariate linear regression models such as canonical correlation analysis or maximum covariance analysis (Uvo et al., 2001) for multi-site downscaling tasks. Other methods, like expanded downscaling (Bürger, 1996), extend the multivariate linear regression model by constraining the spatial covariance matrices of the predictions and observations to be the same. As they are linear in their parameters, these methods may fail when nonlinear relationships are present, for instance when trying to predict daily precipitation amounts (Yuval and Hsieh, 2002). Expanded downscaling deals with nonlinearity by normalizing the predictors and predictands prior to estimating the regression parameters. However, Bürger and Chen (2005) found that the resulting model was very sensitive to how the predictors/predictands were normalized. Other forms of nonlinearity, for example higher-order interactions between predictors, may also be difficult to incorporate into the expanded downscaling model.

Flexible nonlinear regression models like artificial neural networks (ANNs), which can represent arbitrary forms of nonlinearity and complicated interactions between predictors, may yield better predictions than classical linear models for a variable like precipitation (Yuval and Hsieh, 2002). Comparisons between ANNs with multiple versus single outputs have demonstrated the potential of the multivariate approach (Caruana, 1997; Mackay, 1998), which suggests that multivariate ANNs may be well suited to multi-site downscaling. However, trials by the author on a multi-site temperature dataset found that model constraints, as in the expanded downscaling model, were needed to ensure that spatial relationships were modeled realistically.

Extension of multivariate ANNs to precipitation, which can be decomposed into separate variables for precipitation occurrence and wet-day amounts, is also not straightforward. Schoof and Pryor (2001) found that ANN downscaling models for the conditional mean of precipitation at a site performed poorly when dry days were modeled at the same time as wet days. Separate models for precipitation occurrence and wet-day amounts are often built for this reason. However, this two step approach may be difficult to apply in the context of multi-site downscaling as, on a given day, precipitation may fall at some locations but not others. It is not immediately obvious how to best combine a multivariate ANN for predicting precipitation occurrence with another for predicting precipitation amounts, especially when constraints on spatial variability are also added to the models.

Regression-based models, whether linear or nonlinear, usually only offer point predictions of the conditional mean, which means that realistic estimates of predictive uncertainty are often not available (Cawley et al., 2007). In addition, the variance of the conditional mean will typically be smaller than the variance of the observed series, due in part to the influence of small scale phenomena that are not represented in the regression model. Expanded downscaling, statistical “inflation”, or the addition of random noise is thus required for the variance of the
Chapter 6. Expanded Bernoulli-gamma density network

predicted series to match that of observations (Bürger and Chen, 2005; von Storch, 1999). Alternatively, the conditional distribution of precipitation amounts can be specified directly, for example by modeling the parameters of an appropriate statistical distribution rather than just the conditional mean. Williams (1998) successfully described seasonal variations in precipitation using an ANN to model parameters of a mixed Bernoulli-gamma distribution. Haylock et al. (2006) used the same model formulation to downscale precipitation at single sites in the UK. Recently, Dunn (2004) proposed the Poisson-gamma distribution for similar purposes. The Bernoulli-gamma and Poisson-gamma distributions can be fit to precipitation series that include both dry and wet days. This means that precipitation occurrence and wet-day precipitation amounts can be specified by the same model. To date, this approach has not been extended to multi-site downscaling.

As an alternative to purely regression-based models, multi-site downscaling models for precipitation have also adopted hybrid approaches whereby a regression model is used as a preprocessing to a nonparametric weather generator, for example the conditional resampling models of Rajagopalan and Lall (1999) and Buishand and Brandsma (2001), among others. Hybrid methods, for example Wilby et al. (2003) and Stahl et al. (2008), are capable of preserving spatial relationships between sites, providing realistic estimates of precipitation occurrence and amounts, including variability, and reproducing the non-normal distribution of precipitation amounts. Unlike pure regression models, hybrid approaches cannot, due to the resampling step, predict daily precipitation amounts in excess of the maximum observed in the historical record. This is true of other nonparametric weather generator or analog downscaling models, although heuristics have been introduced for extrapolating beyond the range of observations in future climate scenarios (Imbert and Benestad, 2005).

The aim of this paper is the development of a regression-based multi-site downscaling algorithm for precipitation, the expanded Bernoulli-gamma density network (EBDN), that combines the strengths of the algorithms noted above. Specifically, EBDN can specify the conditional distribution of precipitation at each site, model occurrence and amount of precipitation simultaneously, reproduce observed spatial relationships between sites, be used as a conditional weather generator to generate synthetic precipitation series, and set new record precipitation amounts. To accomplish these goals, EBDN extends the expanded downscaling model of Bürger (1996) by adopting the ANN-based conditional density estimation approach in conjunction with the Bernoulli-gamma distribution (Williams, 1998).

The remainder of the paper is split into 5 sections. First, data and a benchmark conditional resampling model are described in Section 6.2. Next, expanded downscaling and the Bernoulli-gamma distribution are reviewed in Section 6.3. The EBDN downscaling model is then introduced in Section 6.4. EBDN is applied to multi-site precipitation data and its results are compared against the benchmark model in Section 6.5. Finally, results are discussed in Section 6.6 along with general conclusions and recommendations for future research.
Table 6.1: Daily precipitation observing stations

<table>
<thead>
<tr>
<th>Station identifier</th>
<th>Station name</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Victoria Int’l A</td>
<td>48.65</td>
<td>-123.43</td>
<td>20</td>
</tr>
<tr>
<td>C2</td>
<td>Comox A</td>
<td>49.72</td>
<td>-124.90</td>
<td>24</td>
</tr>
<tr>
<td>C3</td>
<td>Port Hardy A</td>
<td>50.68</td>
<td>-127.37</td>
<td>22</td>
</tr>
<tr>
<td>C4</td>
<td>Estevan Point</td>
<td>49.38</td>
<td>-126.55</td>
<td>7</td>
</tr>
<tr>
<td>C5</td>
<td>Pachena Point</td>
<td>48.72</td>
<td>-125.10</td>
<td>37</td>
</tr>
<tr>
<td>C6</td>
<td>Tofino A</td>
<td>49.08</td>
<td>-125.78</td>
<td>24</td>
</tr>
<tr>
<td>C7</td>
<td>Powell River</td>
<td>49.83</td>
<td>-124.50</td>
<td>52</td>
</tr>
<tr>
<td>C8</td>
<td>Agassiz CDA</td>
<td>49.25</td>
<td>-121.77</td>
<td>15</td>
</tr>
<tr>
<td>C9</td>
<td>Seymour Falls</td>
<td>49.43</td>
<td>-122.97</td>
<td>244</td>
</tr>
<tr>
<td>C10</td>
<td>Vancouver Int’l A</td>
<td>49.20</td>
<td>-123.18</td>
<td>3</td>
</tr>
</tbody>
</table>

6.2 Datasets and benchmarks

6.2.1 Precipitation data

Daily precipitation data from 10 climate observing stations along the south coast of B.C., Canada were used to test the EBDN downscaling algorithm. Data from 1959-1998 were obtained from the Adjusted Historical Canadian Climate Data archive maintained by the Meteorological Service of Canada (Mekis and Hogg, 1999). Station names are given in Table 6.1 and station locations are shown in Figure 6.1.

Southern B.C. is an area of complex terrain that includes coastal influences and a series of north-south oriented mountain/valley systems that extend into the interior of the province. The interaction of complex terrain with synoptic-scale storms in the cool, wet season, along with the influence of convective activity in the warm, dry season, gives rise to large spatial gradients in seasonal and annual mean precipitation and high spatial variability on a daily time scale. Correlations between precipitation series over the period range from 0.29 between Estevan Point and Victoria Int’l A to 0.89 between Estevan Point and Tofino A.

6.2.2 Synoptic-scale predictor data

Synoptic-scale predictors used in the downscaling models were extracted from daily climate fields obtained from the ECMWF ERA-40 re-analysis (Brankovic and Molteni, 2004). Mean sea-level pressure (MSL), 850-hPa relative humidity (RH850), 850-hPa temperature (T850), 700-hPa relative vorticity (V700), and 500-hPa geopotential height (Z500) fields were selected from the archive. Data defined on a 2.5° × 2.5° grid for a spatial domain spanning 145°W-105°W and 35°N-65°N were compressed via an extended principal component analysis. The areal extent of the study area and grid-point locations are shown in Figure 6.1. Sine and cosine of day of year were added as predictors to allow for seasonal changes in predictive relationships.

Initially, models were fitted using principal component scores and the cyclical variables.
Figure 6.1: Map showing the extent of the ECMWF ERA-40 re-analysis domain used in the study. Grid-points are numbered and circles show locations of the daily precipitation observing stations.
Figure 6.2: Multivariate regression tree used for predictor selection in the EBDN model. Splits in the tree are labelled by meteorological field and grid-point number from Figure 6.1. At each node in the tree, bar plots show the relative magnitude of daily precipitation amounts at the 10 stations.

However, the resulting models were found to have sub-optimal prediction performance, which is consistent with results from Cannon et al. (2002). Instead, the final set of predictors was selected from the grid-point, extended principal component, and cyclical variables using a regression tree targeted on the multi-site precipitation series (Cannon et al., 2002; Faucher et al., 1999). The optimum regression tree is shown in Figure 6.2. Of the original set of predictors, only grid-point variables from MSL, T850, and RH850 fields appeared in the decision rules. The first split, on MSL at grid-point 58 (57.5°N and 130°W), was responsible for more than 60% of the regression tree’s reduction in error. This split marks the position and intensity of frontal systems crossing the coast, with splits down the tree further refining the discrimination between wet and dry events.

In summary, the final set of 10 model predictors included MSL at grid-points 45, 58, 114, 127, and 130; RH850 at grid-points 86 and 99; and T850 at grid-points 49, 61, and 86.
6.2.3 Benchmark model

The TreeGen downscaling model from Stahl et al. (2008) is used as a benchmark for comparison with the EBDN. TreeGen is hybrid regression/nonparametric weather generator model that generates multi-site weather series by conditionally resampling historical observations (Buishand and Brandsma, 2001) from synoptic weather types identified via the regression tree weather typing algorithm of Cannon et al. (2002). The algorithm, like the proposed EBDN model, is both nonlinear and probabilistic.

For consistency, predictors from Section 6.2.2 are used as inputs to the TreeGen model. The tree structure shown in Figure 6.2 thus defines the synoptic weather types used in the conditional resampling phase of TreeGen. Once the synoptic map-types have been defined from the historical record, predictor fields are entered into the regression tree and each day is assigned to one of the map-types. Next, precipitation amounts on a given day are predicted using a nonparametric weather generator based on conditional resampling from cases assigned to that day’s map-type (Buishand and Brandsma, 2001). The probability $p_T(i)$ of randomly selecting precipitation amounts observed on day $i$ as the predicted values on day $t$ is taken to be inversely proportional to the Euclidean distance $d(t-1, i-1)$ between the predicted values on the previous day $t-1$ and historical values of the precipitation amounts on day $i-1$,

$$p_T(i) = \frac{(d(t-1, i-1) + 0.1)^{-2}}{\sum_{h \in H} (d(t-1, h-1) + 0.1)^{-2}}$$  \hspace{1cm} (6.1)

where $H$ is the set of historical days assigned to the predicted map-type occurring on day $t$.

6.3 Background

6.3.1 Expanded downscaling

The probabilistic multi-site precipitation downscaling method developed in the current study extends the deterministic expanded downscaling model of Bürger (1996) and Bürger (2002). Expanded downscaling is based on the multivariate linear regression model. Given an $N \times I$ matrix of predictors $\mathbf{X}$, where $x_i(t)$ is the value of the $i$th predictor ($i = 1, ..., I$) at time $t$ ($t = 1, ..., N$), and a corresponding $N \times M$ matrix of predictands $\mathbf{Y}$, where $y_m(t)$ is the value of the $m$th predictand ($m = 1, ..., M$) at time $t$, predictions $\hat{\mathbf{Y}}$ from the standard multivariate linear regression model are given by

$$\hat{\mathbf{Y}} = \mathbf{X W},$$  \hspace{1cm} (6.2)

where $\mathbf{W}$ is the matrix of regression parameters. The least-squares solution to the multivariate linear regression model is given by

$$\mathbf{W} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$  \hspace{1cm} (6.3)
Intercept terms can be incorporated by adding a unit column vector to $X$.

In expanded downscaling, elements of $W$ are instead found by minimizing the least-squares cost function subject to the added constraint

$$S_Y = W^T S_X W = S_Y,$$

(6.4)

where $S_Y$ is the covariance matrix of $\hat{Y}$

$$S_Y = \frac{\hat{Y}_c^T \hat{Y}_c}{N},$$

(6.5)

and $\hat{Y}_c$ is the matrix of predictions with columns centered to zero mean; similarly, $S_X$ is the covariance matrix of the observed predictors $X$ and $S_Y$ is the covariance matrix of the observed predictands $Y$.

### 6.3.2 Bernoulli-gamma distribution

In order to move expanded downscaling from a deterministic modeling framework to a probabilistic modeling framework, an appropriate probability density function (pdf) must be selected to represent the distribution of precipitation at the observing sites. Williams (1998) suggested using a mixed Bernoulli-gamma distribution for describing precipitation series that include both days with no precipitation and days with precipitation. This mixed distribution was used in single site precipitation downscaling models by Haylock et al. (2006) and Cawley et al. (2007).

The Bernoulli-gamma pdf is given by

$$f(y; p, \alpha, \beta) = \begin{cases} 
1 - p & \text{for } y = 0 \\
\frac{p (y/\beta)^{\alpha-1} \exp(-y/\beta)}{\beta \Gamma(\alpha)} & \text{for } y > 0 
\end{cases}$$

(6.6)

where $y$ is the precipitation amount, $p$ ($0 \leq p \leq 1$) is the probability of precipitation, $\alpha$ ($\alpha > 0$) is the shape parameter of the gamma distribution, $\beta$ ($\beta > 0$) is the scale parameter of the gamma distribution, and $\Gamma(\cdot)$ is the gamma function. The mean value of the distribution is $\mu^{(B)} = p\alpha/\beta$ and the variance is $s^{(B)} = p\alpha[1 + (1 - p)\alpha]/\beta^2$.

To illustrate, a Bernoulli-gamma distribution was fit to observations at station C9, the highest elevation site with the largest mean annual precipitation of the 10 stations, and station C10, the lowest elevation site receiving approximately one third the precipitation of station C9. Values of $p$ were estimated directly from the observations, whereas $\alpha$ and $\beta$ were set via the method of maximum likelihood. Histograms of the observed series and the fitted Bernoulli-gamma pdfs, in addition to quantile-quantile plots, are shown in Figure 6.3. The Bernoulli-gamma distribution fit the observed data well at both stations, although the highest quantiles at station C10 were underpredicted slightly.
Figure 6.3: Histograms of observed daily precipitation (grey shading) and fitted Bernoulli-gamma pdfs (dashed line for positive amounts and asterisk for the probability at zero) at (a) station C9 and (b) station C10, along with quantile-quantile plots for (c) station C9 and (d) station C10.
6.4 Method

6.4.1 Bernoulli-gamma density network (BDN)

The EBDN downscaling methodology extends expanded downscaling by (i) using an ANN for the predictor-predictand mapping and (ii) using the Bernoulli-gamma distribution to represent the predictive density of precipitation at multiple observing sites. For reference, Figure 6.4 shows the steps involved in training and evaluating outputs from the model given new data. The first step is to train the ANN.

In most applications, ANNs are trained to minimize a least-squares cost function. The resulting model describes a mapping that approximates the conditional mean of the predictand data. This mapping is optimal if the data are generated from a deterministic function that is corrupted by a normally distributed noise process with constant variance (Bishop, 1995, Sec. 6.1-6.2). When the noise process has non-constant variance or is non-normal then a more general model, one that fully describes the conditional density of the predictand data rather than just the conditional mean, is more appropriate. This can be accomplished by adopting an ANN with outputs for each parameter in the pdf of the assumed noise process. A conditional density...
ANN for a normally distributed noise process with non-constant variance would thus have two outputs: one for the conditional mean and another for the conditional variance (Dorling et al., 2003). For the Bernoulli-gamma density network (BDN), three outputs, one for each parameter in the distribution, are required.

The architecture of the BDN is shown schematically in Figure 6.5. In this example, the model has $K = 6$ outputs, corresponding to the three Bernoulli-gamma parameters ($p$, $\alpha$, and $\beta$) at $M = 2$ sites. The number of model outputs for an arbitrary number of observing sites is $3M$.

Given predictors $x_i$ ($i = 1, ..., I$) for the observed precipitation series $y_m$ ($m = 1, ..., M$), results from the BDN are evaluated in the same manner as a standard ANN. First, output from the $j^{th}$ hidden-layer node $h_j$ is given by applying the hyperbolic tangent function to the inner product between the predictors $x_i$ and the input-hidden layer weights $w_{jk}^{(1)}$ plus the bias $b_j^{(1)}$.
Chapter 6. Expanded Bernoulli-gamma density network

\[ h_j(t) = \tanh \left( \sum_{i=1}^{I} x_i(t) w_{ji}^{(1)} + b_j^{(1)} \right). \]  

(6.7)

The predicted value of the \(k^{th}\) output from the network \(o_k\) is then given by

\[ o_k(t) = g_k \left( \sum_{j=1}^{J} h_j(t) w_{kj}^{(2)} + b_k^{(2)} \right), \]  

(6.8)

where \(w_{kj}^{(2)}\) are the hidden-output layer weights, \(b_k^{(2)}\) are the hidden-output layer biases, and \(g_k(\cdot)\) are functions that constrain the outputs to lie within the allowable ranges of the Bernoulli-gamma distribution parameters. For the first site \((m = 1, k = 1, \ldots, 3)\) this leads to

\[ p_1(t) = g_1(o_1(t)) = \frac{1}{1 + \exp(-o_1(t))} \]  

(6.9)

\[ \alpha_1(t) = g_2(o_2(t)) = \exp(o_2(t)) \]  

(6.10)

\[ \beta_1(t) = g_3(o_3(t)) = \exp(o_3(t)) \]  

(6.11)

and so on for the remaining \(M - 1\) sites. The conditional Bernoulli-gamma pdf at time \(t\) and at site \(m\), \(f_m(y_m(t)|x(t))\), is then given by Eq. 6.6 with the parameters \(p_m(t), \alpha_m(t),\) and \(\beta_m(t)\).

As the BDN gives probabilistic outputs, weights and biases in the model are set following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Cawley et al., 2007; Haylock et al., 2006; Williams, 1998)

\[ L = -\sum_{t=1}^{N} \sum_{m=1}^{M} \log(f_m(y_m(t)|x(t))), \]  

(6.12)

via a quasi-Newton minimization algorithm (Schnabel et al., 1985).

As is the case with a conventional ANN, the overall complexity of the model should be limited to ensure that the model generalizes well to new data. Standard methods for controlling model complexity, for example choosing the number of hidden nodes via cross-validation (Dinopoulos et al., 1999), Bayesian regularization (Cawley et al., 2007; Williams, 1998), ensemble averaging (Cannon and Whitfield, 2002), early stopping (Gardner and Dorling, 1998), or other algorithms, should thus be used to ensure that the BDN does not overfit the training dataset.

### 6.4.2 Conditional simulation of precipitation series

Once a BDN has been trained, it can be used to create synthetic time series of precipitation amounts \(\hat{y}_m\) by randomly generating a series of cumulative probabilities \(z\) from a uniform distribution and then evaluating quantiles from an inverse Bernoulli-gamma cdf, \(F_m^{-1}(z)\), with parameters given by the BDN. This approach was taken by Cawley et al. (2007). However,
as shown by Huth et al. (2001), the time structure of the downscaled series may be poorly simulated if \( z \) is generated by a white noise process.

Following Bürger and Chen (2005), the series \( z \) is instead assumed to be first-order autoregressive AR(1), which means that it can be randomly generated by first creating an autocorrelated series with zero mean and unit variance

\[
    u(t) = u(t-1) \tau + \epsilon(t) \left(1 - \tau^2\right)^{1/2},
\]

(6.13)

where \( \tau \) is the lag-1 autocorrelation of the series to be generated and \( \epsilon \sim \mathcal{N}(0, 1) \). The series \( u \) is then transformed by the standard normal cdf, \( \Phi(\cdot) \), so that it lies in the range \([0, 1]\),

\[
    z(t) = \Phi(u(t)).
\]

(6.14)

The AR(1) model is chosen for its parsimony. In practice, more complicated noise models could be applied instead.

At time \( t \), the value of the synthetic precipitation time series \( \hat{y}_m(t) \) is given by \( F_m^{-1}(z(t)) \) with parameters \( p_m(t) \), \( \alpha_m(t) \), and \( \beta_m(t) \). The value of \( \tau \) is chosen via trial and error such that the lag-1 autocorrelations of the generated series and observed series match as closely as possible. To illustrate, sample results from a BDN model with 5 hidden-layer nodes are shown in Figure 6.6 for station C10 during 1998. Each panel, moving from the conditional Bernoulli-gamma parameters through to estimation of the synthetic series, corresponds to a step in the flowchart shown in Figure 6.4b.

### 6.4.3 Expanded Bernoulli-gamma density network (EBDN)

Following expanded downscaling, the EBDN extends the BDN model by adding a constraint on the simulated covariance matrix to the model cost function. Direct translation of Eq. 6.4, the constrained least-squares cost function from expanded downscaling, to the probabilistic framework of the EBDN leads to the following cost function,

\[
    L_E = -\sum_{t=1}^{N} \sum_{m=1}^{M} \log(f_m(y_m(t)|x(t))) + \gamma \sum_{m=1}^{M} \sum_{q=m}^{M} \left(s_{y_{m,q}} - s_{\hat{y}_{m,q}}\right)^2,
\]

(6.15)

where the first term is the NLPD summed over the \( M \) sites and the second term is the sum of squared differences between elements of the covariance matrix \( S_Y \) for pairs of observed series \( y_m \) and the corresponding elements of the covariance matrix \( S_{\hat{Y}} \) for pairs of predicted series \( \hat{y}_m \). The coefficient \( \gamma \) controls the relative scaling of the NLPD and the constraint on the covariances.

Unfortunately, the evaluation of \( L_E \) is computationally intensive and its minimization is impractical as synthetic series \( \hat{y}_m \) must be generated to estimate the predicted covariance matrix \( S_{\hat{Y}} \). As an alternative, a modified cost function that allows calculations involving \( \hat{y}_m \) to be replaced with ones involving the Bernoulli-gamma distribution parameters \( p_m, \alpha_m, \) and \( \beta_m, \)
Chapter 6. Expanded Bernoulli-gamma density network

Figure 6.6: Time series of (a) conditional Bernoulli-gamma parameters where the probability of precipitation $p(t)$ is shown by the dashed line, the shape $\alpha(t)$ by the solid line, and the scale $\beta(t)$ by the dotted line; (b) AR(1) cumulative probabilities $z(t)$; (c) conditional mean (solid line) and the range from the conditional 5th to 95th percentiles (gray shading); and (d) synthetic daily precipitation (vertical lines). Observed daily precipitation values are shown as dots in (d). Results shown are for station C10 during 1989.
which are available directly and without further computation as outputs from the ANN, is proposed. One simple option would be to calculate the predicted covariance matrix from the conditional means $\mu_m^{(B)}$ instead of the synthetic series $\tilde{y}_m$. This would, however, force the series of conditional means at each site to have the same variance as the observed series, which would degrade the density estimation, essentially resulting in deterministic predictions similar to those from the original expanded downscaling algorithm of Bürger (1996).

Instead, the modified cost function makes use of the fact that the covariance matrix can be expressed in terms of separate contributions from the site variances and the correlations between sites. When estimating elements of the correlation matrix for the predictions, the synthetic series $\tilde{y}_m$ can be replaced with the series of conditional means $\mu_m^{(B)}$. As correlations are insensitive to linear rescaling of variables, this replacement will not, as is the case above for the covariances, tend to degrade the probabilistic performance of the model.

To illustrate, the covariance matrix for the predictions $\tilde{Y}$ can be decomposed as

$$S_{\tilde{Y}} = V_{\tilde{Y}} R_{\tilde{Y}} V_{\tilde{Y}}^T$$

where $V_{\tilde{Y}}$ is a diagonal matrix of standard deviations, i.e., with diagonal elements $s_{\tilde{y}_m}^{1/2}$, and $R_{\tilde{Y}}$ is the correlation matrix

$$R_{\tilde{Y}} = \frac{\tilde{Y}_s^T \tilde{Y}_s}{N},$$

with $\tilde{Y}_s$ the matrix of predictions with columns centered to zero mean and scaled to unit variance. The cost function now becomes

$$\mathcal{L}_E' = -\sum_{t=1}^N \sum_{m=1}^M \log(f_m(y_m(t)|x(t))) + \gamma_1 \sum_{m=1}^M (s_{y_m} - s_{\tilde{y}_m})^2 +$$

$$\gamma_2 \sum_{m=1}^{M-1} \sum_{q=m+1}^M (r_{y_m,q} - r_{\tilde{y}_m,q})^2$$

where the first term is the NLPD, the second term is the sum of squared differences between the observed and predicted site variances, and the third term is the sum of squared differences between elements of the observed and predicted correlation matrices. The coefficients $\gamma_1$ and $\gamma_2$ control the relative scaling of the NLPD and the two constraint terms. In the current study, values of the $\gamma$ coefficients are scaled following the first iteration of the minimization such that all terms in the cost function are weighted approximately equally. From Eq. 6.16 applying separate constraints on the site variances and the inter-site correlations is equivalent to applying a constraint on the covariances.

As mentioned above, evaluation of $\mathcal{L}_E'$ can be sped up by replacing variances/correlations calculated from the synthetic series $\tilde{y}_m$ with values calculated directly from the Bernoulli-gamma
parameters. Predicted variances are given by

\[ s_{\hat{y}_m} = \left\langle s_m^{(B)} \right\rangle + s_{\mu_m^{(B)}}, \] (6.19)

where, from Section 6.3.2, the first term is the mean of the series of conditional variances and the second term is the variance of the series of conditional means; \( \langle \cdot \rangle \) denotes taking the temporal mean of a series. Likewise, values of the correlations are estimated from the series of conditional means \( \mu_m^{(B)} \). After replacement, the final version of the cost function for the EBDN model is given by,

\[ L_E^{(f)} = -\sum_{t=1}^{N} \sum_{m=1}^{M} \log(f_m(y_m(t)|x(t))) + \gamma_1 \sum_{m=1}^{M} \left( s_{y_m} - s_{\mu_m^{(B)}} - \left\langle s_m^{(B)} \right\rangle \right)^2 + \gamma_2 \sum_{m=1}^{M-1} \sum_{q=m+1}^{M} \left( r_{y_m,q} - r_{\mu_{m,q}^{(B)}} \right)^2. \] (6.20)

Following the above substitutions, one problem still remains. The series of conditional means \( \mu_m^{(B)} \) (see Figure 6.6c) will tend to have smaller day-to-day variability than the synthetic series generated by sampling from the conditional Bernoulli-gamma distributions \( \hat{y}_m \) (see Figure 6.6d). As a result, correlation estimates based on \( \mu_m^{(B)} \) exhibit systematic biases with respect to those based on \( \hat{y}_m \). A simple iterative correction algorithm is thus used to nudge the biased estimates towards more realistic values:

(i) The correction algorithm begins by finding ANN weights and biases that minimize the cost function \( L_E^{(f)} \) using values of \( r_{ym,q} \) from the observations and values of the predicted correlations \( r_{\hat{y}_m,q} \) approximated by \( r_{\mu_{m,q}^{(B)}} \).

(ii) Following the minimization, estimates of \( r_{\hat{y}_m,q} \) are made from the trained EBDN based on multiple realizations of the synthetic precipitation series \( \hat{y}_m \), which are generated as described in Section 6.4.2.

(iii) Differences between observed and estimated values of the correlations \( r_{ym,q} - r_{\hat{y}_m,q} \) are then found and are added to the values of \( r_{ym,q} \). This results in adjusted values \( r_{(a)_{ym,q}} \) that are either lower or higher than those observed, depending on whether the bias in modeled correlations is positive or negative.

(iv) Biases in the estimated correlations that result from replacing \( \hat{y}_m \) with \( \mu_m^{(B)} \) are reduced by minimizing \( L_E^{(f)} \) again, this time with the adjusted correlations \( r_{(a)_{ym,q}} \) replacing the observed correlations \( r_{ym,q} \) in the third term of the cost function.

(v) Steps (ii) to (iv) are repeated, each time adding the current estimates of \( r_{ym,q} - r_{\hat{y}_m,q} \) to the previous values of the adjusted correlations \( r_{(a)_{ym,q}} \), and then minimizing \( L_E^{(f)} \).

To illustrate, an EBDN model with 5 hidden-layer nodes was trained on the data described in Section 6.2. The model was initially trained for 100 iterations of the quasi-Newton minimization algorithm; each application of the correction algorithm involved running the minimization
algorithm for another 20 iterations, each time starting from the previous weights and biases. Plots of observed correlations versus predicted correlations based on 30 realizations of $\hat{y}_m$ from the trained model are shown in Figure 6.7. Figure 6.7a shows results following the initialization of the correction algorithm. Without correction, the EBDN model overestimated the degree of spatial correlation between the sites, leading to a positive bias in the predicted correlations. Figure 6.7b shows results following two applications of the correction algorithm. The correction algorithm successfully removed this bias, resulting in a model capable of representing spatial variability between sites in a more realistic manner.

### 6.5 Results

Precipitation data from 10 stations along the south coast of B.C. were regressed, using the EBDN model, onto the 10 synoptic-scale predictors selected in Section 6.2. Models were evaluated using cross-validation. Data were split into four decade-long segments (1959-1968, 1969-1978, 1979-1988, and 1989-1998), models were trained on three of the four segments, and predictions were then made on the held-out segment. Model training followed the procedure outlined in the previous section. This was repeated, each time rotating the held-out segment, until predictions had been made on all data. A secondary 5-fold cross-validation was conducted on the training segments in order to determine the optimal number of hidden nodes in the EBDN, i.e., to avoid model overfitting. Five hidden nodes were selected by consensus.
Chapter 6. Expanded Bernoulli-gamma density network

ing cross-validation, predictions on the test segments were concatenated and validated against observations.

Statistics measuring probabilistic, deterministic, and categorical performance of the EBDN and TreeGen models are given in Table 6.2. In terms of probabilistic performance, the mean NLPD over all stations and days was 2.4 for the EBDN model and \( \infty \) for the TreeGen model. As TreeGen is based on a nonparametric weather generator it cannot generate precipitation amounts in excess of those in the training dataset, i.e., the predicted pdf has a finite upper bound equal to the maximum observed precipitation. Thus, predicted probability densities are zero, with a corresponding NLPD of \( \infty \), whenever precipitation amounts are outside the range of the training data. When out of range values were removed from cross-validation test segments, the NLPD fell to 3.8 for TreeGen and to 2.3 for EBDN. For reference, unconditional Bernoulli-gamma distributions were also fit to observed series at each station. The resulting climatological NLPD was 2.6. Results are consistent with those reported by Cawley et al. (2006) for a single-site precipitation downscaling task wherein a small improvement in NLPD relative to climatology was noted for a Bernoulli-gamma density network.

While both the EBDN and TreeGen models give probabilistic predictions, point forecasts can also be made by estimating the conditional mean of the pdf on each day. Values of the root mean squared error (rmse) and Pearson product-moment correlation based on deterministic predictions are shown in Table 6.2. Values of rmse for the two models were within 0.2 mm of one another and are consistent with those found by Cannon (2007b) using a coupled ANN/analog downscaling model. The EBDN model predicted the spatial mean (i.e., the mean of the 10 stations) more accurately than TreeGen, showing a 10% improvement in rmse, and values of \( r_s \) corresponding to 40% explained variance versus 28% for TreeGen.

Conditional probabilities of precipitation occurrence were calculated from the conditional EBDN and TreeGen pdfs, with cumulative probabilities exceeding 0.5 resulting in a prediction of measurable precipitation. Categorical measures of performance were then calculated based on the 2×2 contingency table (Wilks, 2006). Again, results are given in Table 6.2. Based on the true skill statistic, both models showed positive skill relative to climatology, although EBDN outperformed TreeGen. While TreeGen exhibited a slightly higher probability of detection, this was offset by EBDN having fewer false alarms, a higher hit rate, and a bias ratio closer to one.

To evaluate multi-site performance and the ability of the constrained cost function to replicate the observed covariance matrix, 100 realizations of the synthetic series \( \hat{y}_m \) at the 10 sites were generated following Section 6.4.2. Observed and predicted covariances between sites are plotted in Figure 6.9 for the EBDN and TreeGen models. In all cases, the distribution of the EBDN covariances over the 100 realizations overlapped observed values, whereas the magnitude of TreeGen covariances tended to be lower than observed, particularly for the highest magnitudes. As a further check on the ability of the models to simulate spatially realistic precipitation fields, quantile-quantile plots for observed and predicted daily average precipitation over the 10 stations are shown in Figure 6.8. Results indicate that the distribution of the spa-
Chapter 6. Expanded Bernoulli-gamma density network

Table 6.2: Cross-validated model performance statistics for EBDN and TreeGen downscaling models. NLPD is the mean negative log predictive density; rmse is the root mean squared error of the conditional mean series; rmse_s is the root mse of the spatial mean series; and r_s is the correlation between the observed and predicted spatial mean series. The spatial mean is defined as the mean of the 10 station values. TSS, HIT, POD, FAR, and BIAS are, respectively, the true skill statistic, hit-rate, probability of detection, false alarm rate, and bias ratio for binary predictions of daily precipitation occurrence. Values corresponding to the better performing model are italicized.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>EBDN</th>
<th>TreeGen</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLPD</td>
<td>2.4</td>
<td>∞ (3.8)</td>
</tr>
<tr>
<td>rmse (mm)</td>
<td>11.7</td>
<td>11.9</td>
</tr>
<tr>
<td>rmse_s (mm)</td>
<td>7.4</td>
<td>8.1</td>
</tr>
<tr>
<td>r_s</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>TSS</td>
<td>0.52</td>
<td>0.39</td>
</tr>
<tr>
<td>HIT</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>POD</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>FAR</td>
<td>0.26</td>
<td>0.40</td>
</tr>
<tr>
<td>BIAS</td>
<td>0.93</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Figure 6.8: Quantile-quantile plots of daily station mean precipitation based on 100 synthetic precipitation series from (a) the EBDN model and (b) the TreeGen model.
Chapter 6. Expanded Bernoulli-gamma density network

Figure 6.9: Observed covariances versus boxplots of predicted covariances from 100 synthetic precipitation series from (a) the EBDN model and (b) the TreeGen model. The line inside each box represents the median, boxes extend from the 25\textsuperscript{th} to 75\textsuperscript{th} percentiles, whiskers extend to points within 1.5 times the inter-quartile range, and outliers are shown as circles.

The initial average series from the EBDN model matched observations, whereas quantiles were slightly underpredicted by TreeGen.

To assess the EBDN model’s ability to replicate observed precipitation variability on a seasonal and interannual basis, Figures 6.10 and 6.11 show monthly distributions of four statistics: (i) the number of wet days (NWD), (ii) maximum 5-day precipitation amounts (PX5D), (iii) maximum number of consecutive wet days (WRUN), and (iv) maximum number of consecutive dry days. Two stations, C4, which is located near sea level on the west coast of Vancouver Island, and C9, which is the highest elevation station on the mainland, were selected to represent different climatic regions within the study domain. The degree of correspondence between observed and modeled values over the year are summarized in Table 6.3, which lists rmse values for the median and inter-quartile ranges of the distributions shown in Figures 6.10 and 6.11. In general, the EBDN model performed better than TreeGen, although both models were able to reproduce seasonal variations in precipitation characteristics at the two sites. EBDN had lower rmse values than TreeGen for the median for all four statistics at both stations. Interannual variability, as measured by the inter-quartile range, was better reproduced by EBDN for NWD and DRUN at station C4, but only DRUN at station C9. Eleven of the sixteen statistics were better simulated by EBDN than by TreeGen. It bears noting that these results were from a single synthetic realization from both the EBDN and TreeGen model. As a result, they do not capture the variability due to the stochastic nature of the two algorithms. That being said,
Figure 6.10: Interannual distributions of monthly observed (white), TreeGen modeled (light gray), and EBDN modeled (dark grey) (a) number of wet days (NWD), (b) maximum 5-day precipitation amounts (PX5D), (c) maximum number of consecutive wet days (WRUN), and (d) maximum number of consecutive dry days (DRUN) at station C4.
Figure 6.11: As in Figure 6.10 but for station C9.
Table 6.3: Cross-validated model rmse statistics for median and inter-quartile range (IQR) of monthly number of wet days (NWD), maximum 5-day precipitation (PX5D), maximum number of consecutive wet days (WRUN), and maximum number of consecutive dry days (DRUN) for EBDN and TreeGen downscaling models at stations C4 and C9. Results summarize Figures 6.10 and 6.11. Values corresponding to the better performing model are italicized.

<table>
<thead>
<tr>
<th>Station</th>
<th>Variable</th>
<th>rmse median</th>
<th>rmse IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EBDN</td>
<td>TreeGen</td>
</tr>
<tr>
<td>C4</td>
<td>NWD (days)</td>
<td>0.93</td>
<td>2.58</td>
</tr>
<tr>
<td>C4</td>
<td>PX5D (mm)</td>
<td>12.0</td>
<td>15.4</td>
</tr>
<tr>
<td>C4</td>
<td>WRUN (days)</td>
<td>1.18</td>
<td>2.12</td>
</tr>
<tr>
<td>C4</td>
<td>DRUN (days)</td>
<td>0.59</td>
<td>0.88</td>
</tr>
<tr>
<td>C9</td>
<td>NWD (days)</td>
<td>1.64</td>
<td>3.44</td>
</tr>
<tr>
<td>C9</td>
<td>PX5D (mm)</td>
<td>29.0</td>
<td>31.8</td>
</tr>
<tr>
<td>C9</td>
<td>WRUN (days)</td>
<td>1.00</td>
<td>2.25</td>
</tr>
<tr>
<td>C9</td>
<td>DRUN (days)</td>
<td>1.21</td>
<td>1.27</td>
</tr>
</tbody>
</table>

results from multiple runs were inspected and the general patterns seen in Figures 6.10 and 6.11 and Table 6.3 were also found in the additional model runs.

6.6 Discussion and conclusion

This study introduced a probabilistic synoptic downscaling model for precipitation at multiple sites. The EBDN model modifies the expanded downscaling model of Bürger (1996) and Bürger (2002) by adopting an ANN model that predicts parameters of the Bernoulli-gamma distribution rather than point precipitation values. This extends the density network approach of Williams (1998) and Haylock et al. (2006) from single-site to multi-site datasets. Weights and biases in the ANN model are found by minimizing a constrained NLPD cost function. The constraint on predicted covariances from expanded downscaling is split into separate terms for the site variances and the inter-site correlations in the EBDN model. This reduces the number of computations required to evaluate the cost function by allowing predicted series of Bernoulli-gamma parameters to replace stochastically generated precipitation series in calculations of the constraint terms.

Verification on historical precipitation data from 10 sites on the south coast of B.C. suggests that the EBDN model may be a useful tool for generating multi-site climate scenarios based on synoptic-scale climate data from GCMs. The EBDN model performed better than a benchmark hybrid regression/weather generator model on most verification statistics. While the improvement in NLPD between the EBDN model and the climatological distribution was small, model evaluations suggest that the EBDN model is capable of generating series with realistic spatial and temporal variability. A more rigorous analysis of the model’s ability to reproduce weekly, monthly, seasonal, and annual variability, along with observed trends, is currently underway as
Chapter 6. Expanded Bernoulli-gamma density network

part of a larger project testing a suite of multi-site precipitation downscaling methods, including analog and weather-typing models. It is worth noting that the EBDN model, unlike analog or weather-typing models, does not require any heuristics to extrapolate beyond the range of the training observations (Imbert and Benestad, 2005).

Temporally, constraints on the covariance matrix, along with the value of $\tau$ used in the conditional simulation, are specified using estimates spanning the entire seasonal cycle. Temporal variations could also be incorporated by enforcing seasonal or monthly constraints on the site variances/correlations and by changing $\tau$ throughout the year (Schoof and Robeson, 2003). Minimal changes to the model would be required to accommodate finer-grained control over the temporal variability.

While Bürger (2002) applied the expanded downscaling to multi-site precipitation data, it was originally proposed as a general method for modeling multiple variables (e.g., precipitation, temperature, wind components, etc.) at one or more sites (Bürger, 1996). Extension of the EBDN to variables other than precipitation would require appropriate distributions to be identified and incorporated into the ANN model. Other parametric distributions or mixtures of distributions would, in addition to the pdf and cdf, need expressions for the mean and variance in terms of the distribution parameters to be specified in order to be used within the proposed modeling framework. Work on a more general version of the model is currently underway.

Precipitation in southern B.C. is characterized by high spatial and temporal variability due in part to the area's complex terrain. While results for stations in the wet, maritime climate of the south coast are encouraging, more work is needed to validate the EBDN model in other climatic regions. Preliminary results on data from the Kootenay region of B.C., which is characterized by a more continental climate, suggest comparable levels of skill to those reported here (Cannon, 2007a).

Predictions from the EBDN are probabilistic, which means that any relevant descriptor of the Bernoulli-gamma distribution, including means, variances, quantiles, and posterior probabilities of exceeding a given threshold, can be obtained from the model. As Cawley et al. (2007) demonstrate, estimates of predictive uncertainty from downscaling models can be very useful in impact studies, especially when consequences of extreme events are of key concern. The suitability of the EBDN model for use in generating precipitation data for hydrological models or spatial interpolation algorithms, both of which require spatially realistic input data, should also be assessed. Results suggest that the EBDN model’s ability to reproduce spatial relationships between sites, along with its ability to model precipitation extremes, make it well suited for such tasks.
Bibliography


Chapter 7

Conclusion

7.1 Summary

This dissertation is concerned with the development and application of supervised multivariate statistical models to climate prediction. Five novel models were introduced, two aimed at seasonal climate prediction and three at climate downscaling. A brief summary of each model, focusing on strengths and weaknesses of the proposed approaches, contributions made to the state of the art, potential applications of the research findings, and future research directions, follows.

7.2 Seasonal climate prediction

7.2.1 Robust nonlinear canonical correlation analysis

Chapter 2 introduced a robust variant of NLCCA designed to improve performance on datasets with low signal-to-noise ratios, for example those encountered when making seasonal climate forecasts. The NLCCA neural network model architecture proposed by Hsieh (2000) was kept intact, but the cost functions used to set the model parameters were replaced with more robust variants. The Pearson product-moment correlation in the double-barreled network was replaced by the biweight midcorrelation (Wilcox, 2004). The mean squared error in the inverse mapping networks was replaced by the mean absolute error. Variants of NLCCA with combinations of the four cost functions were demonstrated on a synthetic dataset and were used to forecast sea surface temperatures in the tropical Pacific Ocean based on the sea level pressure field. Adoption of the biweight midcorrelation resulted in improved performance in the face of noise. In particular, the propensity of the double-barreled network to overfit, which was noted by Hsieh (2001a) when using the Pearson product-moment cost function, was removed. Replacing the mean squared error by the mean absolute error resulted in improved performance on the synthetic dataset, but not on the climate dataset except at the longest lead time, which suggests that the appropriate cost function for the inverse mapping networks is more problem dependent.

One drawback of the robust cost functions is that both are not continuously differentiable, which could potentially lead to difficulties when applying standard gradient-based optimization algorithms. Continuously differentiable approximations to the mean absolute error based on the Huber norm (Panayiotis et al., 2006) could be used here. A similar option for the biweight midcorrelation is not obvious. In practice, repeated optimization runs from different initial...
conditions appear to be sufficient to reach a deep local minimum or global minimum of the cost function surface. Alternatively, a gradient-free optimization algorithm, as in Chapter 4, could be adopted.

The propensity of the standard NLCCA model to overfit has hampered efforts at detecting signals that can be exploited for seasonal prediction in the extratropics (Wu and Hsieh, 2004), where potential predictability is lower than in tropical regions (George and Sutton, 2006). Recent work by Hsieh et al. (2006) has, however, identified nonlinear teleconnections between the El Niño-Southern Oscillation and Arctic Oscillation and the Northern Hemisphere atmospheric circulation. The potential exists for robust NLCCA to exploit these signals.

The neural network architecture for NLCCA is but one means of nonlinearly extending CCA. Recent developments in machine learning have led to kernel-based (Lai and Fyfe, 2000) and deflation-based methods for NLCCA (Sharma et al., 2006). Direct comparison between robust NLCCA and these algorithms for seasonal climate prediction is warranted.

### 7.2.2 Nonlinear principal predictor analysis

Chapter 3 introduced a neural network approach for performing NLPPA, the nonlinear extension of PPA (Thacker, 1999). Previously, a truly nonlinear extension of either PPA or RA had yet to be developed. Unlike NLCCA, which extracts correlated patterns from two datasets, with both effectively acting simultaneously as predictors and predictands, NLPPA searches for patterns that explain the maximum amount of variance in a predictand dataset. The asymmetry of NLPPA means that one of the four multi-layer perceptrons required to implement NLCCA was no longer needed, thereby leading to a more parsimonious model. NLPPA was applied to the Lorenz system of equations (Lorenz, 1963) and was compared with the NLCCA model of Hsieh (2000). NLPPA performed better than NLCCA when the synthetic datasets were corrupted with noise.

In its proposed form, the NLPPA model, like robust NLCCA, is restricted to modelling open curves in predictor/predictand space. Periodic and quasi-periodic phenomena are, however, common in the climate system, and might be better represented by closed curves. In the context of unsupervised learning, a periodic variant of nonlinear principal component analysis (NLPCA) was introduced by Kirby and Miranda (1996) and adopted by Hsieh (2001b) for use in climatology. As NLPPA reduces to (open curve) NLPCA, introducing the circular node of Kirby and Miranda (1996) into NLPPA would be straightforward. Development of a circular variant of NLCCA would require choosing a circular measure of correlation (Fisher, 1996), and thus would not be as straightforward as in NLPPA.

An obvious need for future research lies in the application of NLPPA to real-world climate datasets, for example those mentioned above for robust NLCCA. Introducing the mean absolute error cost function from robust NLCCA might also yield further improvements in performance and should be investigated. Regardless, a comparison with robust NLCCA instead of the non-robust variant is needed.
7.3 Climate downscaling

7.3.1 Nonlinear analog predictor analysis

Chapter 4 introduced NLAPA, a hybrid method for climate prediction and downscaling in which an analog model is coupled to a neural network model. NLAPA is loosely based on the model architecture introduced in Chapter 3 for NLPPA. Rather than coupling two multi-layer perceptrons together into a five layer neural network, as is the case in NLPPA, the analog model instead replaces the final two neural network layers. The coupled neural network/analog model maintains the ability of the analog model to preserve inter-variable relationships and model non-normal and conditional variables (such as precipitation), while taking advantage of the neural network’s ability to non-linearly compress the original predictors to a lower-dimensional space in which the performance of the analog model is maximized. NLAPA was tested on a suite of synthetic and real-world hydroclimatological datasets, with results compared against those from three benchmark analog downscaling models (Fernandez and Saenz, 2003; Fraedrich and Ruckert, 1998; Zorita and von Storch, 1999). In terms of root mean squared error, performance of the NLAPA model equaled or exceeded the benchmark models on all synthetic and real-world datasets.

While NLAPA was compared against other analog downscaling models, no attempt was made to compare its performance against other multivariate or multi-site downscaling algorithms. As pointed out by Haylock et al. (2006), an integrated assessment of multi-site downscaling algorithms in terms of their multi-site performance is needed.

One limitation of NLAPA, namely its high computational cost, derives from the discontinuous nature of the analog model. Optimizing the weights and biases in the neural network portion of the model must be done via some form of gradient-free optimization algorithm. While the Differential Evolution optimization algorithm (Price et al., 2005; Storn and Price, 1997) performed well, a more efficient optimization algorithm would be beneficial.

Predictions from NLAPA consist of observations sampled from the historical record. As a result, the model is incapable of extrapolating beyond the range of observations without further processing, for example using algorithms proposed by Imbert and Benestad (2005) and Sharif and Burn (2006).

The idea of using a neural network to non-linearly compress a dataset such that the performance of a second model is maximized, as is the case in NLAPA for the analog model, is one that could be adopted more widely. For example, in hydrology Chokmani and Ouarda (2004) used CCA to define a “physiographic space” in which subsequent interpolation of flood quantiles to ungauged sites was improved. It is conceivable that a coupled neural network/interpolation algorithm might outperform the uncoupled CCA/interpolation approach. More generally, links between NLAPA and unsupervised neural network-based dimensionality reduction algorithms (de Ridder and Duin, 1997), which have been used to improve the performance of classification models (Lerner et al., 1998), are worth exploring.
7.3.2 Multivariate ridge regression with negative ridge parameters

Chapter 5 showed that multivariate ridge regression (Brown and Zidek, 1980) with negative elements in the ridge matrix (Hua and Gunst, 1983) can be used as an alternative to expanded downscaling (Bürger, 1996) for reproducing observed relationships between variables in multivariate linear downscaling. Both methods force the covariance structure of the predictions to match that of observations. Unlike in expanded downscaling, however, an explicit constraint on the covariance matrix is not added to the regression cost function. Instead, regression coefficients are estimated directly via a matrix equation, while ridge parameters, which are free to take positive or negative values, are adjusted iteratively such that the discrepancy between modelled and observed covariance matrices is minimized. Results from multi-site temperature and precipitation datasets suggest that the proposed method is able to constrain the predicted covariance matrix to closely match the observed covariance matrix.

In Chapter 5, it was hypothesized that the multivariate ridge regression approach may be more computationally efficient than expanded downscaling when the number of predictors is large relative to the number of predictands. With the recent adoption of kernel regression, where the potential number of predictors equals the number of samples, in climate downscaling (Tripathi et al., 2006), this hypothesis should be tested.

7.3.3 Expanded Bernoulli-gamma density network

Chapter 6 introduced the EBDN model, a nonlinear, probabilistic, and multi-site precipitation downscaling algorithm. Following Williams (1998), the EBDN model represents the conditional density of precipitation, conditioned on synoptic-scale climate predictors, by a neural network whose outputs are parameters of the Bernoulli-gamma distribution. The single site algorithm of Williams (1998) is extended to multiple sites by adding the expanded downscaling constraint on the predicted covariances (Bürger, 1996), which was explored in Chapter 5, to the neural network cost function. The resulting model can be thought of as nonlinear multivariate regression with a stochastic weather generator component. The EBDN model was applied to a multi-site precipitation dataset from coastal British Columbia, Canada and its results were compared with those from the TreeGen multi-site conditional resampling algorithm (Stahl et al., 2008). The EBDN model was capable of representing the full range of observed variability in precipitation amounts, inter-site correlations, and spatial and temporal intermittency. Results were better than those from TreeGen on eight of nine measures of model performance.

As noted in Chapter 6, the EBDN model is specifically designed as a multi-site precipitation downscaling algorithm. Other weather elements could be incorporated as predictands, but would require specification of appropriate parametric noise distributions (e.g., Gaussian for temperature, Weibull for wind speeds, etc., Wilks, 2006). Alternatively, semi-parametric distributions, using mixtures of simple kernels (Bishop, 1995; Bishop and Nabney, 1996), could also be adopted as a more generic approach.

Temporally, the expanded downscaling constraint that is imposed on the covariance matrix
is specified using estimates encompassing the entire seasonal cycle. As covariances may vary seasonally (Schoof and Robeson, 2003), temporal variations could be incorporated into the model by adding seasonal or monthly constraints to the cost function.

The EBDN model, unlike the NLAPA model from Chapter 4, is capable of extrapolating beyond the range of historical observations. As noted by Hsieh and Cannon (2008), the extrapolation characteristics of nonlinear models have not been explored in detail in the context of climate prediction. Given that the main role of downscaling models is as a tool for post-processing GCM scenarios, which may include significant changes in the large-scale circulation, a more in depth exploration of extrapolation characteristics of statistical downscaling models is warranted.

More generally, the effect of selected predictors on downscaled climate scenarios, including the range in magnitude (and realism) of predicted changes, needs to be assessed. For example, Huth (2004) found that estimates of temperature changes in a doubled CO2 climate from three linear downscaled models differed by over 10°C depending on the set of predictors. Given that the extrapolation characteristics of nonlinear models are more complicated than those of linear models, such an assessment is crucial for a nonlinear downscaling model like EBDN.

Finally, as recommended above for the NLAPA model, a comprehensive evaluation of multi-site downscaling algorithms is needed to better characterize the strengths and weaknesses of the different approaches on a common set of data.

7.4 Overall recommendations

The primary focus of this dissertation was on technique development, that is the creation of multivariate statistical models to address existing gaps in the climate prediction literature. In a broader context, there is need to move beyond technique development to address model validation, particularly from the perspective of end users. In other words, do models actually meet the needs of a given target audience? Rather than basing model intercomparison studies solely in terms of standard meteorological verification statistics, studies should also incorporate domain-specific verification measures. If, for example, outputs from multi-site precipitation downscaling algorithms are being used as inputs to hydrological models, then results from the downscaling algorithms should be evaluated in this context. This point is addressed in recent work by Fowler and Wilby (2007) and Fowler et al. (2007), among others.

As new techniques for climate prediction are introduced, framing research in terms of identifying a single “best” performing model should be avoided, especially if different models have demonstrable skill on a given task. Instead, incorporating information from multiple models to improve prediction performance and estimate predictive uncertainty should be encouraged (Burnham and Anderson, 2004). Multi-model ensembles of seasonal ENSO forecasts have, for example, been shown to provide large improvements in skill over individual models (Tippett and Barnston, 2008). In terms of climate downscaling, Haylock et al. (2006), recommended
that results from multiple downscaling algorithms based on multiple forcing GCMs and emission scenarios be used to better gauge the uncertainty of future climate scenarios at local scales. A Bayesian method for combining multiple downscaling models has, for example, been applied to regional rainfall prediction by Coelho et al. (2006).

Similarly, a greater emphasis on the development of probabilistic rather than deterministic downscaling algorithms would also be profitable (Cawley et al., 2007). Recent examples include nonlinear statistical downscaling methods based on a Bayesian framework, such as the Bayesian neural network (Khan and Coulibaly, 2006) and the relevance vector machine (Ghosh and Mujumdar, 2008).
Bibliography


Bibliography


Bibliography


