Selected data exploration methods in hydroclimatology

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy

at

The University of Waikato

by

Varvara Vetrova

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2016

Abstract

The volumes of climatological data are rapidly growing due to development of new acquisition platforms and advances in data storage technologies. Such advances provide new challenging problems for data analysis methods. As a result, there is an increasing interest in application and development of new machine learning and data mining methods in climatological data analysis.

This dissertation contributes to the field of data analysis by evaluation of selected methods and developing new techniques with application to hydro-climatological datasets.

It is shown that data pre-processing with the decimated wavelet discrete transform can cause false predictive accuracy in regression machine learning algorithms. A general result is obtained that a decimated wavelet discrete transform based on a pyramidal algorithm requires utilizing some future values of the time series concerned. When the discrete wavelet transform is utilized as a pre-processing step for forecasting the time series, the necessary independence of calibration and validation data is compromised. This in turn translates into over-optimistic forecasting accuracy or even giving the illusion of forecasting skill when there in none. The obtained result is general and has wide implications in any discipline where the discrete wavelet transform is utilised in forecasting frameworks.

In addition, a general framework for creating simple predictive models is presented, based on LASSO regularised regression. The method is illustrated for time series modelling with external event forcing but the approach has general applicability.

As a contribution to association discovery, two tests of bivariate association are developed. The first method is designed for detecting threshold-like associations in a scatter plot with particular reference to testing the significance of the extent of a data-sparse region within a scatter plot. The second method is a more general test of non-random associations. Both methods utilise significance testing based on randomizations.

Finally, LASSO regularized regression is investigated as a tool for discovering informative large scale climatogical predictors of local hydrological processes. A cross-validation scheme is proposed, which is related to practical forecasts of the next time interval to come, while at the same time maximising use of available information. The proposed methodology was applied to a case study predicting next-season river discharges in the Upper Waitaki River in New Zealand. The proposed forecasting methodology and cross-validation frameworks are applicable for similar hydroclimatological forecasting situations. The physical aspect of this part of the study included discovering the influence of the Interdecadal Pacific Oscillation on winter discharges in the upper Waitaki catchment.

Acknowledgments

Behind the scenes of this dissertation lies a trans-Pacific journey from a middle of continent to the middle of ocean through a tropical island.

Whatever the events might have happened along the journey, this study would never be possible without certain inputs.

I would like to sincerely thank my chief supervisor Associate Prof. Earl Bardsley for his interesting productive discussions on my thesis and unconventional chess strategies. There were also fruitful discussions on statistical methodology and pattern recognition. I am very grateful for all his selfless endless patience and support. I am also grateful to his introduction into New Zealand culture and sense of humour and I did learn some rules of cricket from him, specifically about underarm bowling. His depth of knowledge on Russian history is also very impressive.

I would like to express my gratitude to my supervisors, Dr. Robert Durrant, for his support, mathematical/statistical and machine learning discussions, and Dr. Willem de Lange for the support and contributions to my study.

I owe my gratitude to Dr. Jennifer Purdue for all her inspiration and support with my dissertation and for inviting me to New Zealand, especially to McKenzie country, and also to Dr. Kyoko Fukuda for convincing Earl that my poster indeed might be interesting and for inspiring me to explore data mining further.

I am grateful to financial support during my studies, specifically to MBIE PhD Fellowship, Meridian Energy financial support, UoW Doctoral Scholarship for PhD completion, and travel grants and awards obtained during the study – Terry Healy memorial award, SigKDD student travel award, awards from Hydrological and Meteorological societies of New Zealand, Queen Elizabeth II study award.

The great opportunity was provided by Plant and Food Research, and I express thanks to Dr. Andrew McGlone for our productive collaboration during my internship there.

Landcare Research kindly provided with some time support for conferences and PhD work.

Enormous thanks go to my friends and family:

Dr. Nayo, for her invaluable friendship and endless source of inspiration, encouragement, support all along the way, and Indonesian delicacies. Nina for all her friendship, selfless help and great patience despite certain adventures and our mathematical discussions and not only mathematical, and blini.

Ying, for all her support during the beginning of this journey and for introducing me to Chinese culture. June, for our interesting discussions and Korean culture insights.

Malcolm Taylor, for inspiring chats and New Zealand humour and also for sharing his depth of knowledge on New Zealand hydropower industry.

Kyle, for all his questions about statistics, letting me read his math chapters and encouragement that PhD could be indeed completed.

My fellow PhD/Drs friends Alex, Ali, Mariana, Mohhamed, Dan, Elizabeth, Leyala, Tilini for all the help to relax during the studies. Ali especially, for his cheerul company in our office.

Andrew and Anya, for their nice company and samovar tea.

All my Siberian friends, for staying friends all this time despite the distance and for their encouragement to complete this study, Dr. Zina, Katya, Dr. Igor, Oli, Bayrma. Alya and Sima for their understanding.

Special thanks to Dr. Eji for his incredible sense of humour and support which helps me keep going and for tolerating all the mountain walks.

Last, but not least, my Mum and Dad for all their support during all my studies and continuing encouragement and for being happy for me.

Finally, without continuous inspiration of wonderful outdoors of New Zealand this study wouldn’t be possible. The magnificent views from the tops of mountains and volcanoes, coastlines, waterfalls, bush tracks provided so much energy needed to complete this study, apart from coffee and chocolate.

Table of Contents

Chapter 1 Introduction

1.1 Background

We are privileged to witness first-hand a revolution in the digital era. There is a rapid rise of new observation-measurement platforms[1], such as satellites, UAVs, which carry range of sensors capable of capturing many different varieties of data. Advances in technology, in the same time, are bringing increased computational power and storage capacity. Higher resolution capture and sensing of data and rapid dissemination of this complex information-rich data is facilitated by networks such as Internet. The potential benefits of these data for science and society are enormous.

However managing and drawing inferences from this flood of data brings its own special problems. Climatological data especially poses a number of challenges. It spans the multiple dimensions of time, space and numerous input channels of information. In other words, it is characterised by high dimensionality (Figure 1-1). However, time series of continuous reliable direct measurements are relatively short in contrast to the high-dimensional nature of the data.[2] Drawing inferences from such data is conceptually akin to trying to solve a system of $\boldsymbol{n}$ linear equations and $\boldsymbol{p}$ unknown variables, where $\boldsymbol{p}>\boldsymbol{n}$, as in the case here. This is known to be impossible without further constraints. Thus variable selection and new well-motivated models tailored to this setting turn out to be crucial for meaningful inference.

[1] The launch of the first artificial satellite Спутник-1 (Sputnik-1) was a start of new technological era in Earth observations.

[2] Exceptions are long-term time series of proxy point-based indirect measurements, such as coral growth rings, sediment cores, ice cores, tree rings, etc. [26]

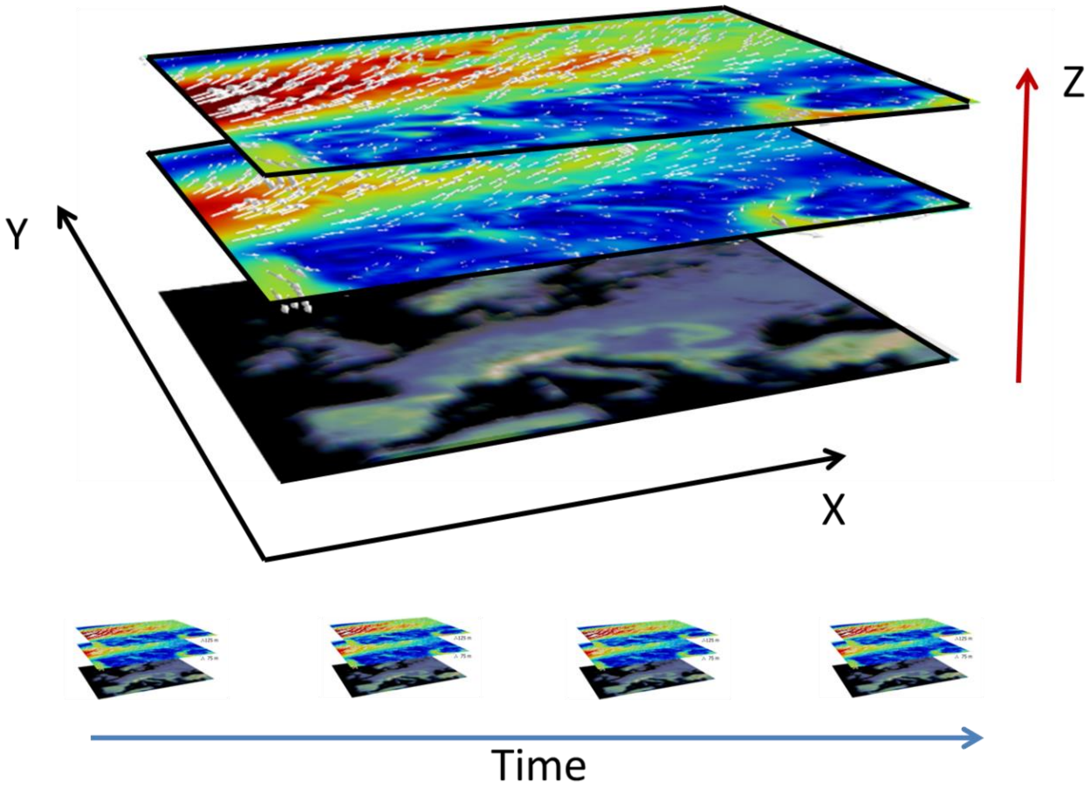

Figure 1-1 High-dimensionality of climatological data. *X*, *Y* corresponds to spatial dimensions, *Z* represents additional dimensions such as elevation, channels of information, etc. *X*, *Y*, *Z* varies with time.

We see that the growing volumes of climatological data, while welcomed, provide challenging problems for data analysis methods and as a result, there is an increasing interest in application and development of new machine learning (ML) and data mining methods in climatological data analysis [1]. Application of ML methods is by no means new in the field of environmental data analysis generally, one of the earliest examples of ML applications to environmental data can be found in [2], where models for wave amplitude prediction were examined. Research in prediction and analysis of climate data started long before ML methods were introduced, see for example [3].

What is new is the resolution of the modern data, speed of acquisition, storage capacity and new tools and techniques for making the most of these. Hsieh [4] provides a general review of ML methods in environmental sciences. Reviews of applications of one of the ML methods, neural networks, in hydrology can be found in [5] and [6].

The terms “machine learning”, “data mining” ,“data-driven methods”, “artificial intelligence” and “statistical learning” are often used interchangeably in the current literature on environmental data analysis [1], [7]–[9]. It is impossible to define a clear separating boundary and we will not attempt to do so. For the interested reader, further information can be found in [10] where one view on concepts of statistical analysis and thinking related to ML is presented.

Taking into account the constant interaction between disciplines, it is probably inevitable that some terminology imported from one scientific field to another is lost in translation. For example the term “cross-validation”, which is a method of estimating prediction error ([11], [12]) has been used in the hydrological literature as a synonym for validation subsample [13]. Such variable use of terminology could be not only confusing but raises concerns of correct estimation of prediction accuracy in the ML models.

Looking further at interaction between ML and hydrology, a certain research trend appears to be emerging, characterised in the hydrological literature by multiple attempts to compare different ML techniques on hydrological datasets [14]–[20]. By no means, this research trend describes whole variability of applications of ML in hydrological literature. Similar developments might also exist in other environmental science subfields not considered here. It worth reference to Hsheh [4], who noted “a word of caution is needed” when different ML methods are compared on a specific dataset, where a bias towards reporting positive results might skew the analysis. Efron, in his commentary to [10] also stated that “New methods always look better than old ones. Neural nets are better than logistic regression, support vector machines are better than neural nets, etc. In fact it is very difficult to run an honest simulation comparison, and easy to inadvertently cheat by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger”.

It appears, however, that the climatological literature expresses different trends with relation to ML methods, for example, clustering analysis, network methods, downscaling [1], [21]–[24].

Going back to hydrology, it seems that an interesting avenue could well be presented for Occam’s principle for application to model simplification. There is an existing preference towards large multi-purpose over-parameterised hydrological models. And although these can provide realistic simulations of nature, it is hard to determine the effect of different parameters on the output. Reduction in model complexity could certainly be explored further and in part of this thesis we analyse several data-driven approaches for this purpose.

An interesting intersection is present between hydrology, climatology and ML. Utilising large scale spatial-temporal climatological information in hydrological forecasts could be viewed as a high-dimensional regression problem [25]. Exploring ML methods for feature selection and dimensionality reduction forms a remaining part of this thesis.

1.2 Research questions

The overall objective of this thesis is to contribute to the field of data analysis by evaluation of selected methods and developing new techniques with application to hydro-climatological datasets.

From the above motivation, the following specific research questions emerged:

- Are machine learning regression methods incorporating decimated wavelet transform really a gain in predictive accuracy?
- How might Occam's razor principle be applied to time series modelling where data is driven by external event forcing?
- How might general non-random association in bivariate data sets be detected more simply?
- How to address the "curse of dimensionality" when high-dimensional climatological data is incorporated into hydro-climatological prediction?

The corresponding research objectives addressed in this thesis are:

1. Determine if data pre-processing with the decimated wavelet discrete transform could give false predictive accuracy in regression ML algorithms.
2. Investigate a specific regularised regression method as a tool for formalised simplification of time series models generally.

3. Develop a new and simple test of general non-random association in bivariate data.
4. Investigate a specific regularised regression method in seasonal streamflow forecasting problem

1.3 Structure of the thesis

The structure of this thesis reflects the structure of climatological data and consequently progresses along the number of dimensions of climatological datasets in consideration (Figure 1-1).

In ML terminology, this thesis considers a specific subfield of ML of regression supervised ML (Chapter 2 and Chapter 4). The thesis also addresses a problem of dimension reduction in high dimensional data through filter and embedded methods (Chapter 3 and Chapter 4).

The thesis consists of five chapters, including this introduction. Chapter two evaluates methods applicable to *univariate* problems – time series analysis, forecasting and modelling; Chapter 3 introduces methods of *bivariate* data analysis and Chapter 4 addresses a problem of *multivariate* high-dimensional analysis. Chapter 5 provides conclusion statements and directions for further research.

Each of the three main chapters consists of two parts which are written in journal publication style, as much of the content has been either published or submitted for publication. An exception is the second part of Chapter 2 which contains a contribution to the second research question.

With respect to chapter content:

Chapter 2 addresses two problems of analysis and prediction of time series. The first part provides a critical analysis on the use of wavelet transform coupled with prediction methods in various fields of study. The common misconception of discrete wavelet transform improving the forecasting ability is addressed. The second part introduces the use of the linear LASSO as a mechanism for creating simple time series models for application where a time series is dominated by external impulses such as individual rainfall events influencing river discharge

Chapter 3 introduces two methods of bivariate data analysis based on randomization methods. The first part describes a significance test of a threshold effect. The second part provides a method of detecting a general bivariate non-random association.

Chapter 4: The first part illustrate the usefulness of the specific instance of LASSO regularized regression as both an exploratory and predictive tool in seasonal streamflow forecasting based on high-dimensional raw climatological data. Its utility is demonstrated with a case study of predicting lake inflows in the Waitaki river catchment, New Zealand. The second part of the chapter provides a detailed study of the influence of a particular climatic index – the Interdecadal Pacific Oscillation on winter inflows in the Waitaki catchment.

Chapter 5 provides conclusions and directions for further research.

References:

[1] V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, Eds., *Machine Learning and Data Mining Approaches to Climate Science*. Cham: Springer International Publishing, 2015.

[2] A. . Ivaknhenko and V. G. Lapa, *Cybernetic Predicting Devices*. Purdue University School of Electrical Engineering, 1965.

[3] W. T. Nakamura, “A study of the variation in annual rainfall of Oahu island (Hawaiian islands) based on the law of probabilities,” *Mon. Weather Rev.*, vol. 61, no. 12, pp. 354–360, 1933.

[4] W. Hsieh, *Machine learning methods in the environmental sciences: Neural networks and kernels*. Cambridge University Press, 2009.

[5] R. J. Abrahart, F. Anctil, P. Coulibaly, C. W. Dawson, N. J. Mount, L. M. See, A. Y. Shamseldin, D. P. Solomatine, E. Toth, and R. L. Wilby, “Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting,” *Prog. Phys. Geogr.*, vol. 36, no. 4, pp. 480–513, 2012.

[6] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, “Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions,” *Environ. Model. Softw.*, vol. 25, no. 8, pp. 891–909, 2010.

[7] G. Cervone, J. Lin, and N. Waters, Eds., *Data Mining for Geoinformatics*. New York, NY: Springer New York, 2014.

[8] B. Sivakumar and R. Berndtsson, *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*. World Scientific, 2010.

[9] M. Kanevski and V. Demyanov, Eds., “Statistical learning in geoscience modelling: Novel algorithms and challenging case studies,” *Comput. Geosci.*, vol. 85, no. Part B, pp. 1–136, 2015.

[10] L. Breiman, “Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author),” *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009.

[12] R. R. Picard and R. D. Cook, “Cross-validation of regression models,” *Technometrics*, vol. 79, no. 387, pp. 575–583, 1984.

[13] D. P. Solomatine and A. Ostfeld, “Data-driven modelling: some past experiences and new approaches,” *J. Hydroinformatics*, vol. 10, no. 1, pp. 3–22, Jan. 2008.

[14] Y. Feng, N. Cui, L. Zhao, X. Hu, and D. Gong, “Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China,” *J. Hydrol.*, vol. 536, pp. 376–383, 2016.

[15] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, “Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir,” *J. Hydrol.*, vol. 476, pp. 433–441, 2013.

[16] M. Behzad, K. Asghari, M. Eazi, and M. Palhang, “Generalization performance of support vector machines and neural networks in runoff modeling,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7624–7629, 2009.

[17] M. Behzad, K. Asghari, and E. A. C. Jr., “Comparative Study of SVMs and ANNs in Aquifer Water Level Prediction,” *J. Comput. Civ. Eng.*, vol. 24,

no. 5, pp. 408–413, 2009.

[18] M. Liu and J. Lu, “Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?,” *Environ. Sci. Pollut. Res.*, vol. 21, no. 18, pp. 11036–11053, 2014.

[19] U. Okkan and Z. A. Serbes, “Rainfall-runoff modeling using least squares support vector machines,” *Environmetrics*, vol. 23, no. 6, pp. 549–564, Sep. 2012.

[20] H. Yoon, S.-C. Jun, Y. Hyun, G.-O. Bae, and K.-K. Lee, “A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer,” *J. Hydrol.*, vol. 396, no. 1, pp. 128–138, 2011.

[21] Q. Y. Feng, R. Vasile, M. Segond, A. Gozolchiani, Y. Wang, M. Abel, S. Havlin, A. Bunde, and H. A. Dijkstra, “ClimateLearn: A machine-learning approach for climate prediction using network measures,” *Geosci. Model Dev. Disucssions*, in review

[22] K. Steinhaeuser, A. R. Ganguly, and N. V. Chawla, “Multivariate and multiscale dependence in the global climate system revealed through complex networks,” *Clim. Dyn.*, vol. 39, no. 3–4, pp. 889–895, Aug. 2012.

[23] Y. Gala, Á. Fernández, J. Díaz, and J. R. Dorronsoro, “Hybrid machine learning forecasting of solar radiation values,” *Neurocomputing*, vol. 176, pp. 48–59, 2016.

[24] Y. Ke, J. Im, S. Park, and H. Gong, “Downscaling of MODIS One Kilometer Evapotranspiration Using Landsat-8 Data and Machine Learning Approaches,” *Remote Sens.*, vol. 8, no. 3, p. 215, Mar. 2016.

[25] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, “Sparse Group Lasso: Consistency and Climate Applications,” in *Proceedings of the 2012 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2012, pp. 47–58.

[26] M. Mudelsee, *Climate time series analysis*. Springer, 2014.

Chapter 2 Critical review of two selected aspects of hydro-climatological time-series forecasting and modelling

2.1 Introduction

The chapter is concerned with two particular approaches to hydro-climatological time-series forecasting and modelling.

The first part of this chapter carries out a critical review of the decimated wavelet transform in time series forecasting, which has gained much recent attention, particularly in hydrology [1]. A number of studies suggest there is a gain in predictive accuracy for machine learning forecasting frameworks utilizing wavelet data-preprocessing, in contrast to the counterparts where no pre-processing has been applied [2]–[10] . This apparent gain in accuracy is investigated here, motivated by the fact that there appears no possible mechanism for such improvements.

The second part of this chapter addresses a more general issue of hydrological time series modelling where data is driven by external event forcing. Specifically, it has long been recognized that many hydrological models are overly complex for a given application to data and many suggestions have been made toward avoiding needless model complexity. The approach adopted here to building simpler models incorporates a formalized method of model simplification. Firstly, an initial model is constructed as a highly over-parameterised linear model subject to linear constraints. This model is then subject to automated forced simplification as part of the calibration process, so the final linear model is as simple as possible while still giving reasonable match to the calibration time series concerned. The goal here is that by avoiding over-fitting in calibration the simpler model will perform better for prediction purposes. The first part of this chapter is presently under review. The second part of this chapter has been published [11].

References:

[1] V. Nourani, A. Hosseini Baghanam, J. Adamowski, and O. Kisi, "Applications of hybrid wavelet–Artificial Intelligence models in

hydrology: A review," *J. Hydrol.*, vol. 514, pp. 358–377, 2014.

[2] Y. Feng, N. Cui, L. Zhao, X. Hu, and D. Gong, "Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China," *J. Hydrol.*, vol. 536, pp. 376–383, 2016.

[3] R. Remesan and J. Mathew, *Hydrological Data Driven Modelling: A case study approach*. Springer International Publishing, 2015.

[4] A. D. Gorgij, O. Kisi, and A. A. Moghaddam, "Groundwater budget forecasting, using hybrid wavelet-ANN-GP modelling: a case study of Azarshahr Plain, East Azerbaijan, Iran," *Hydrol. Res.*, May 2016.

[5] A. A. Najah, A. El-Shafie, O. A. Karim, and O. Jaafar, "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation," *Neural Comput. Appl.*, vol. 21, no. 5, pp. 833–841, Jul. 2012.

[6] R. C. Deo, M. K. Tiwari, J. F. Adamowski, and J. M. Quilty, "Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model," *Stoch. Environ. Res. Risk Assess.*, pp. 1–30, 2016.

[7] V. Nourani and S. Mousavi, "Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method," *J. Hydrol.*, vol. 536, pp. 10–25, 2016.

[8] M. Ravansalar, T. Rajaee, and M. Zounemat-Kermani, "A wavelet–linear genetic programming model for sodium (Na+) concentration forecasting in rivers," *J. Hydrol.*, vol. 537, pp. 398–407, 2016.

[9] M. Shoaib, A. Y. Shamseldin, B. W. Melville, and M. M. Khan, "A comparison between wavelet based static and dynamic neural network approaches for runoff prediction," *J. Hydrol.*, vol. 535, pp. 211–225, 2016.

[10] M. Zounemat-Kermani, Ö. Kişi, J. Adamowski, and A. Ramezani-Charmahineh, "Evaluation of data driven models for river suspended sediment concentration modeling," *J. Hydrol.*, vol. 535, pp. 457–472, 2016.

[11] W. E. Bardsley, V. Vetrova, and S. Liu, "Toward creating simpler hydrological models: A LASSO subset selection approach," *Environ. Model. Softw.*, vol. 72, pp. 33–43, 2015.

2.2 Potential for error in wavelet-based forecasts

Abstract

Time series forecasts which incorporate the decimated wavelet transform are exposed to a subtle error which can have significant impact. The error arises if data pre-processing is applied initially to an entire data set, a situation which may occur frequently. This apparently innocuous step results in validation data indicating higher levels of model forecasting ability than is actually the case. The problem arises from the model calibration phase having the unintended effect of being able to incorporate future data values, thereby obtaining information about the validation set without the user being aware. We suggest that this effect may have resulted in forecasts in many fields being reported with overstated accuracy, even for situations where there may be no forecasting ability. Some selected examples illustrate the exaggerated forecasting accuracy. There would appear to be a case for review of many previous forecasts which incorporate wavelets as part of the forecasting process.

Introduction

Wavelet-based forecasts are now an established element of time series prediction in a number of disciplines, including such disparate subject areas as environment (Wang & Ding, 2004; Liu et al., 2014; Belayneh et al., 2014; Nourani et al., 2014; Kim et al., 2014), finance (Kriechbaumer et al. 2014; Jin & Kim, 2015; Yousefi et al., 2005), electricity load (Conejo et al., 2005 (a,b); Benaouda et al., 2006) and the nuclear industry (Upadhyaya et al., 2014). In essence, the wavelet approach seeks to discover sufficient structure in a time series to enable a degree of self-forecasting in the absence of external information. This would apply, for example, if there were cyclic components in the data.

On the other hand, concerns have been raised over the actual accuracy and general utility of wavelet forecasting, particularly for time series lacking evident structure (Beriro et al.,2012; Zhang et al., 2015). We revisit here the basics of wavelet-based forecasting in the light of some of the concerns raised. We then outline a mechanism by which wavelet forecasts can be inadvertently formulated from an incorrect methodology. This error will yield either false forecast accuracy or give an illusion of time series self-forecasting capability when in fact there is none. We suggest that the error could be sufficiently widespread over many disciplines to the extent that many published wavelet-related forecast methodologies are at the very least in need of clarification.

Forecasting with the discrete wavelet transform

The error arises with application of the discrete wavelet transform (DWT), which is a widely-used pre-processing technique that decomposes a time series into a set of approximation coefficients and detailed coefficients (Misiti et al., 2007; Walker, 1999). The decimated DWT is incorporated in widely-used packages such as the Matlab wavelet toolbox and the Python PyWavelets library.

There are two standard mechanisms described in the literature by which these wavelet components are translated into forecasts of the time series (Murtagh et al., 2004). The first approach extrapolates each wavelet component into the future and the forecast is achieved as the sum of the extrapolations (Kriechbaumer et al. 2014; Yousefi et al., 2005). The alternative method uses the components as independent predictors of a forecasting model such as an artificial neural network or support vector machine (Belayneh et al., 2014; Nourani et al., 2014).

In both approaches there is a necessity for complete independence of calibration (determining parameters) and validation (checking forecast accuracy), as holds for any data-based model evaluation procedure (Hastie et al., 2001, pp. 245-247). Of course, no investigator in the environmental sciences or elsewhere would knowingly allow dependency when evaluating any type of forecasting model. However, dependency can be introduced in subtle ways which may even pass undetected through a journal review process (García-Serrano & Frankignoul, 2014). In the case of wavelet-related forecasting the dependency is introduced through the apparently innocuous step of data pre-processing being undertaken prior to partitioning the time series concerned into the different segments needed for model calibration and validation.

In many publications of wavelet forecasts there is in fact no confirmatory statement that pre-processing did not take place prior to calibration / validation data partitioning. That is, the possibility cannot be excluded that there was incorrect ordering of data pre-processing and partitioning the data into validation and calibration sets. We cannot of course prove that this is the case. However, by illustrating how the error can occur and referencing a few telling examples we make the argument that false wavelet forecasting accuracy has the potential to be widespread over many subject areas.

Source of error

It is perhaps not widely appreciated how the simple exercise of an initial pre-processing of an entire time series will always invalidate any subsequent evaluation of wavelet forecasting accuracy on the validation data. We therefore provide (Supplementary material) a general result that if a decimated DWT is used for pre-processing a time series then this generates the unavoidable requirement of knowing the values of future data when constructing the series of approximation and detailed coefficients. The degree of extension into the future increases with the resolution level of the wavelet transform, regardless of the chosen wavelet filter. Of course, in real-world forecasts the future values are unknown by definition, so future values used in validation give illusionary accuracy.

In this light, it is unfortunate that some published work may have given credibility to false forecasting accuracy. For example, in the context of forecasting metal prices it was noted that wavelet data pre-processing before calibration / validation partitioning supposedly gave much improved forecast accuracy (Kriechbaumer et al. 2014). In fact, this result is most easily explained by the pre-processing creating a lack of independence between the calibration and validation sets.

The issue concerning the evident need to obtain future values has been noted at times in the literature. A specific wavelet forecasting algorithm has been proposed which avoids use of future values (Benaouda et al., 2006, Murtagh et al., 2004). In a hydrological application it was somewhat unrealistically suggested that prior independent estimates of the future values be obtained, which of course would negate the need for forecasts (Zhang et al., 2015). However, the potential impact of future values yielding false validation accuracy appears to have been largely unrecognized, as evidenced by many publications failing to clarify whether data pre-processing was carried out before or after data partitioning into calibration and validation sets.

Examples

We conclude with three examples which demonstrate how incorrect application of the decimated discrete wavelet transform may lead to erroneous forecasts. The first is a very simple illustrative forecasting model applied to synthetic data. The other two examples consider published work from the literature.

For the first example we use the Haar discrete wavelet (Walker, 1999) and make forecasts using wavelet components as predictors. We seek one step ahead forecasting of any arbitrary time series $X = \{x_1, \ldots, x_{100}\}$.

Now apply the decimated discrete wavelet decomposition of the entire time series *X* using the Haar discrete wavelet. Just one level of smoothing is employed in this case, giving decomposition into two time series of wavelet coefficients *a* and *d*, which are respectively approximation coefficients and detailed coefficients. The coefficients corresponding to the odd and even members of the time series are respectively:

$$\begin{aligned} a_{2k-1} &= \left(x_{2k-1} + x_{2k}\right)/2, \quad d_{2k-1} = x_{2k-1} - a_{2k-1} \\ a_{2k} &= \left(x_{2k-1} + x_{2k}\right)/2, \quad d_{2k} = x_{2k} - a_{2k} \end{aligned} \tag{2.1}$$

where k = 1,2,...,50.

It is evident from Eq. (2.1) that the next time step is required in order to calculate the coefficients $a_{2k-1,}\, d_{2k-1}$ for the odd-numbered elements of the time series. It is therefore possible to make a forecast from the expression:

$$x_{i+1} = a_i - d_i \tag{2.2}$$

which uses *d* and *a* as predictors, resulting in half the values in the time series being forecast with perfect accuracy. This would apply even if *X* was a sequence of random numbers.

It follows that if pre-processing via Eq. (2.1) was carried out on any arbitrary data sequence prior to calibration / validation partitioning, then every second validation data point would be perfectly predicted from Eq. (2.2).

This effect is illustrated in Figure 2-1, showing results of forecasting with pre-processing prior to calibration / validation partitioning. The data comprises a sequence of 100 random variables generated from the standard normal distribution. The first 50 values were used for calibration and Eq. (2.2) was used to forecast one step ahead for the other 50 points in turn, giving perfect forecasting of every second point and complete forecast failure for all other points. The net combination of perfect success and failure here gives an illusionary correlation of R = 0.58.

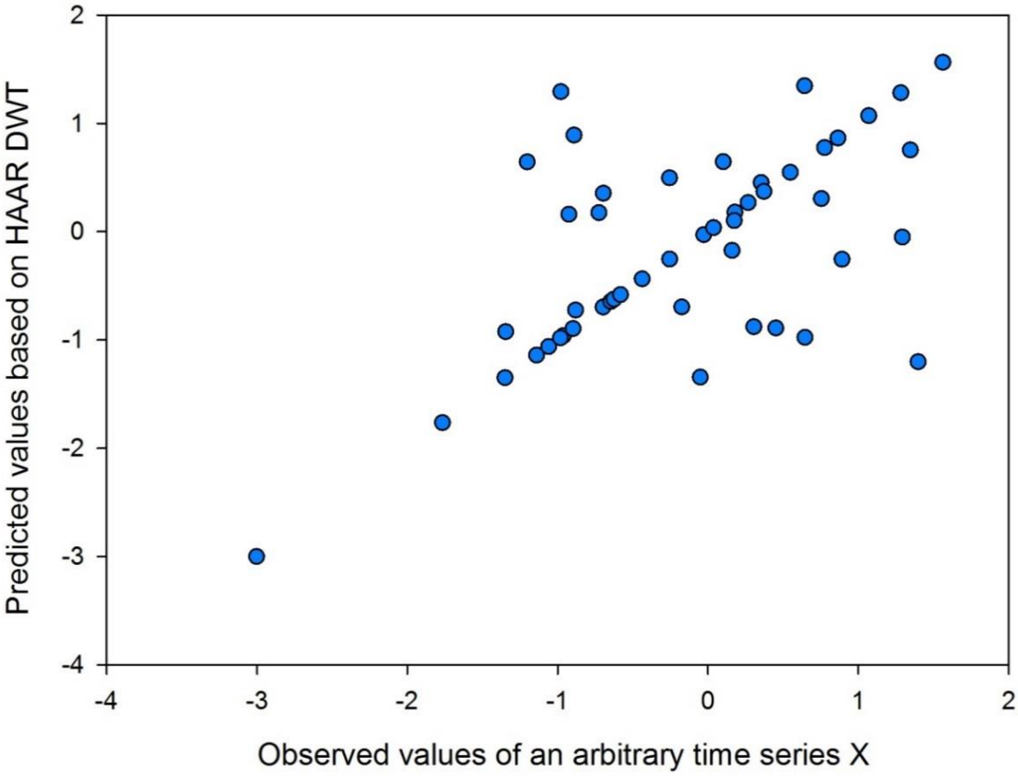

Figure 2-1 Wavelet forecast of a sequence X of 50 random variables from the standard normal distribution (see text for description).

Of course, if plots like Figure 2-1 were to appear in wavelet-related forecasts then the error would be immediately apparent. However, more complex forecasting models are applied in practice so the effect of using future values will not be evident as unusual patterns in the scatter plots.

We next consider a frequently-cited hydrology publication by Wang & Ding (2004), which includes an example of wavelet forecasting of flood discharges of the Yangtze River. Our concern here is that there appears to be a number of flood peaks where the timing of the peak is forecast correctly up to three days in advance

(Figs 4 and 5 of that paper). If this is the case then it strains credibility that a river time series by itself contains sufficient information to forecast the times of its own future discharge peaks.

The authors do not clarify whether pre-processing was undertaken before or after data partitioning, so the possibility is open that the accuracy of forecasting some peak timings is the reflection of the future data effect.

We now illustrate how such evident accuracy might come about, while also fully recognising that this is not necessarily indicative of the analysis of Wang & Ding (2004). Consider the hydrograph-like synthetic time series of Figure 2-2 where the peaks are all even-numbered members of the data sequence. Using the Haar DWT as in the previous example, the peak magnitudes and timings are predicted perfectly. This perfect prediction would not have been evident if the time series had been constructed instead with peaks being odd-numbered elements of the data sequence.

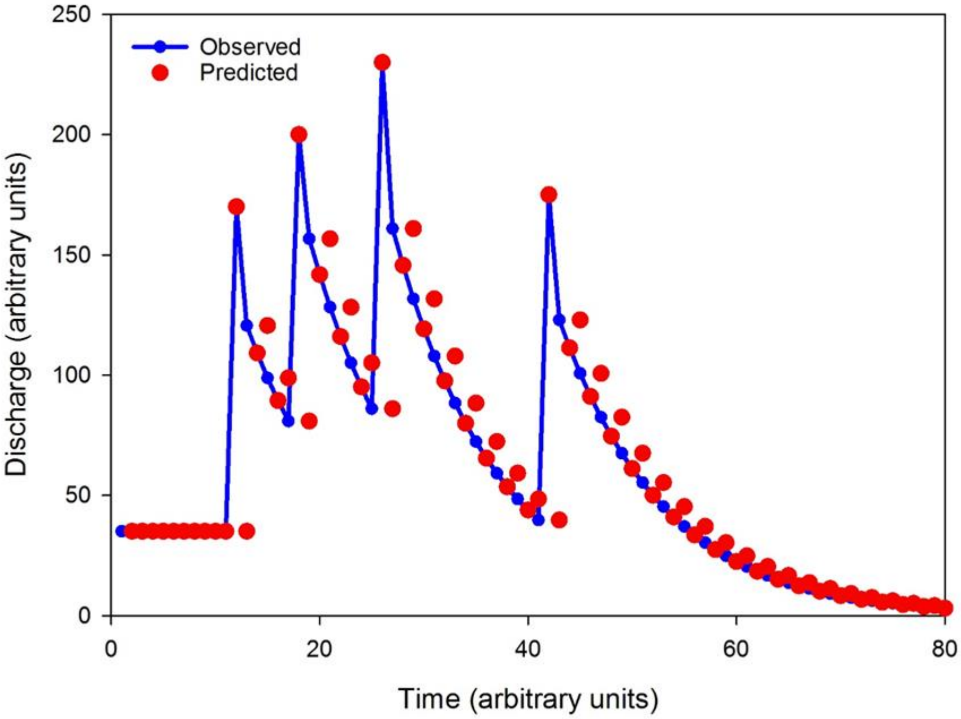

Figure 2-2 Wavelet forecast of a hydrograph-like synthetic time series (see text for description).

Finally, we consider another frequently-cited paper, in this case concerned with wavelet-based forecasting of future crude oil prices (Yousefi et al., 2005).

Given the nature of the data the forecasts appear surprisingly accurate. However, credibility is strained here also, in this case by the authors' own observation that their forecast accuracy does not diminish as the forecasting time horizon is increased.

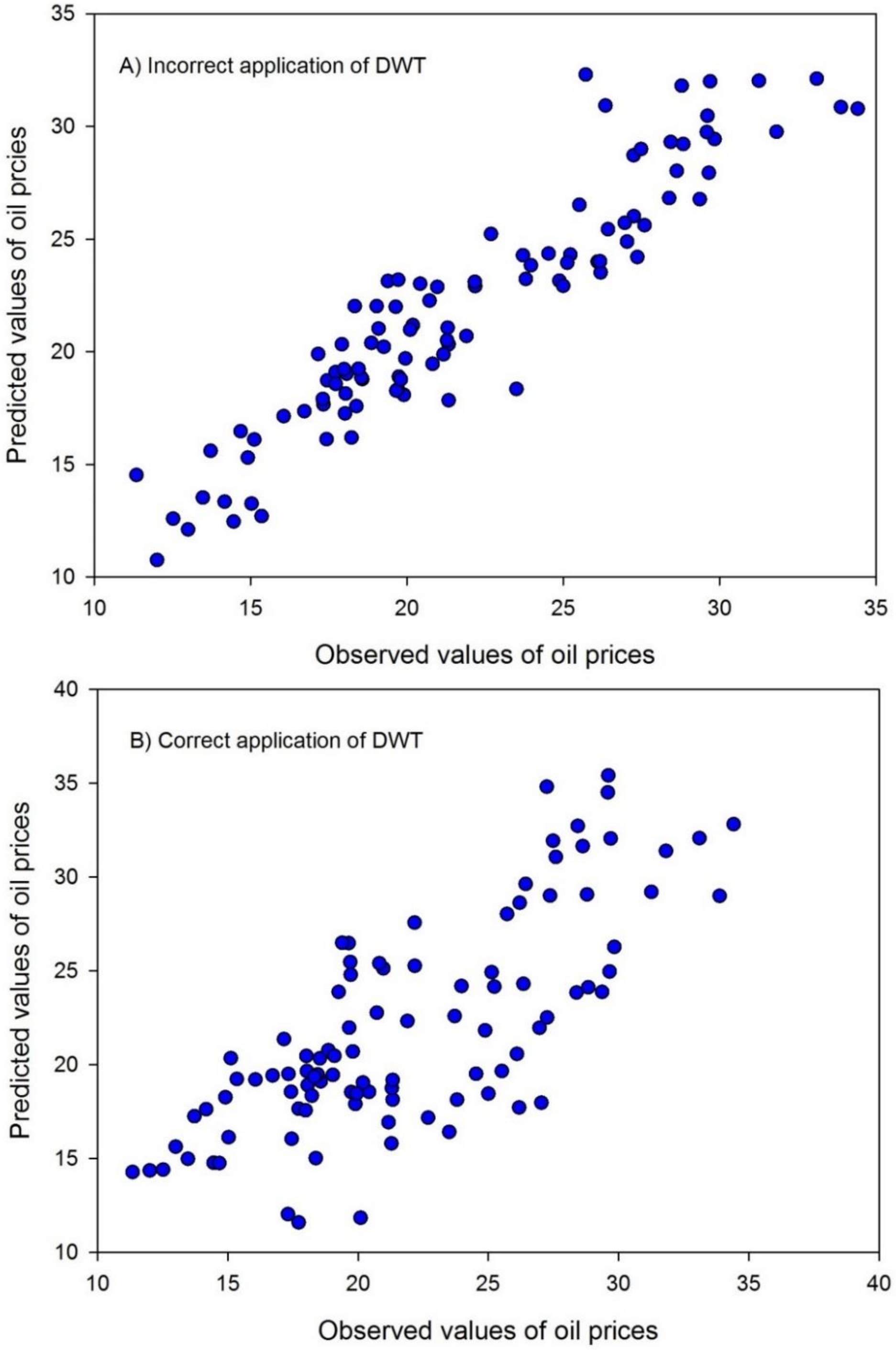

Figure 2-3 Wavelet-based forecasts of oil prices for a 4-month horizon: A) DWT applied to the entire time series prior to splitting into calibration and validation portions; B) DWT applied after splitting into calibration and validation.

As a check, we applied a wavelet transform to the whole data time series of oil prices and used the methodology of Yousefi et al. (2005), which gave forecasts that were of a similar level of high accuracy as reported (Figure 2-3a). However, the accuracy was considerably reduced when the exercise was repeated using wavelet pre-processing subsequent to the data calibration / validation partition (Figure 2-3b). In fact, the forecast accuracy was then no improvement over the reference forecasting methodology used by the authors. Therefore, the use of wavelet methods in this case adds nothing to the forecasts because the apparent accuracy improvement was an artefact of incorporating future values in the methodology.

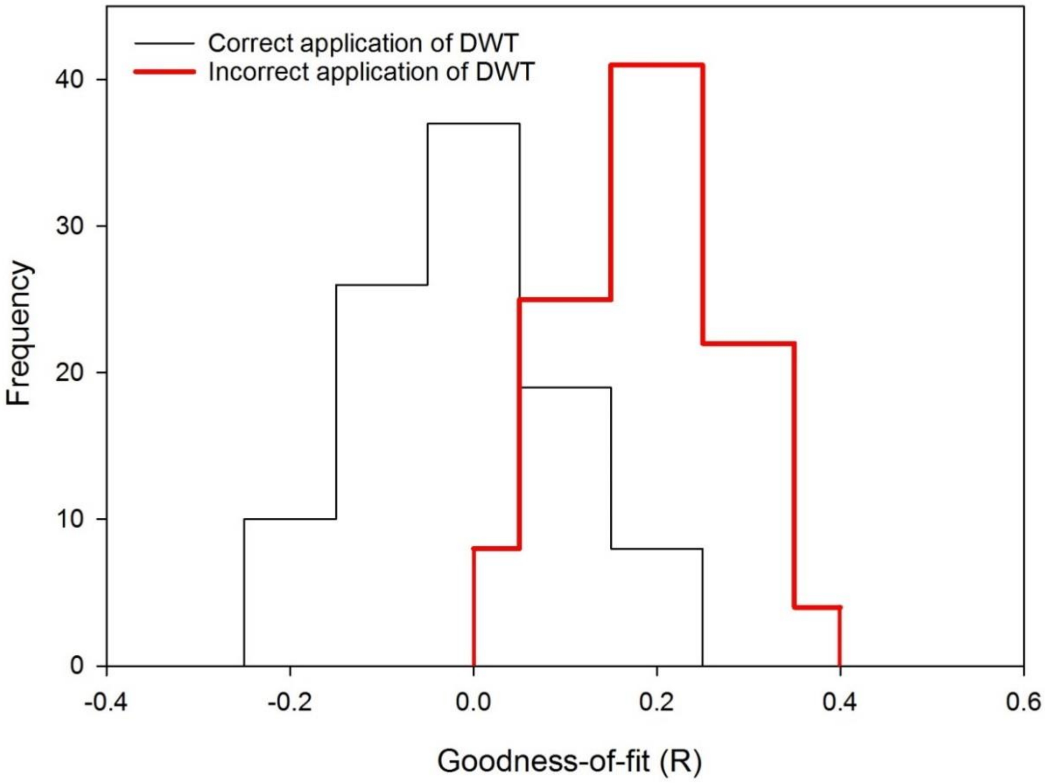

Figure 2-4 Frequency distribution of correlation coefficients R obtained from wavelet forecasting applied to 100 simulated time series oil prices, using both correct and incorrect application of DWT. Note the shift toward positive R for incorrect applications (see text for description).

In order to further demonstrate how DWT can yield false forecasting accuracy in this case, we randomly re-ordered the oil price time series 100 times. For each randomisation we calculated forecast values using both of the two approaches described above. In each case correlation coefficient values were calculated as a measure of the respective forecast accuracies.

Obviously a randomised time series can have no self-predictive property and it would be expected that the frequency distribution of correlation coefficients would centre at zero. This is indeed the case for the correct application of the DWT with data pre-processing carried out before partitioning into calibration and validation sets (Figure 2-4). However, the correlation coefficients are shifted toward positive values if pre-processing is carried out first. This apparent predictive ability is clearly incorrect because there is no predictive possibility in this case.

Conclusion

For wavelet-based time series forecasting we have showed that the simple act of applying the decimated DWT for data pre-processing will always lead to an unintended dependency between the calibration and validation procedures. This then leads to illusionary forecast accuracy as measured against validation data. While we cannot quantify from the literature as to the extent of this error, its potential impact means that published wavelet-related forecasting methodologies may be under something of a cloud. We suggest it should be a requirement that all future publications incorporating wavelet forecasts make a clear statement confirming that any data pre-processing was not carried out before calibration / validation partitioning.

Supplementary material

The discrete wavelet transform pyramidal algorithm in time series forecasting: requirement for future values

We demonstrate here that future values of an input time series must be utilized for the calculation of wavelet components for the decimated discrete wavelet transform. This has particular relevance because it opens the possibility for erroneous forecasting accuracy on a validation data set. This false accuracy will arise whenever wavelet pre-processing is applied to the whole time series prior to splitting the data into calibration and validation subsets.

The decimated discrete wavelet transform is widely applied in wavelet-based time series forecasting and is based on the pyramidal algorithm (Mallat, 1989). The transform is carried out in two parts: decomposition and reconstruction (Figure 2-5).The decomposition component transforms the original time series into sets of *approximation* and *detailed* coefficients, each with several levels of resolution. The reconstruction component serves to re-create the original time series from these coefficient sets.

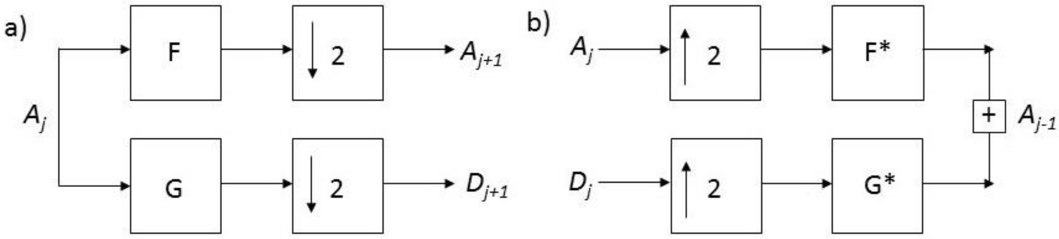

Figure 2-5 The pyramidal algorithm of the discrete wavelet transform: a) decomposition phase; b) reconstruction phase; A_j are approximation coefficients at level *j*; D_j are detailed coefficients at level *j*; F and F* denote convolution with a low-pass filter, G and G* denote convolution with a high-pass filter; ↓2 and ↑2 represents respectively dyadic downsampling and dyadic upsampling, and + indicates summation of the two series.

The decimated discrete wavelet transform based on the pyramidal algorithm leads to a sparse representation of the original time series in such a way that if a time series X has 2^k data points then its approximation and detailed coefficients at resolution level j have 2^{k-j} data points, where $j \in (1,\ldots,k)$. However, in order to use these approximation and detailed coefficients in time series forecasting their interpolated versions are required. These versions contain the same number of data points as the original time series. The interpolation is achieved through the reconstruction component of the pyramidal algorithm.

We now consider in detail how the reconstruction phase of the pyramidal algorithm leads to interpolated time series of approximation and detailed coefficients which require utilization of future data points of *X*.

For the sake of simplicity, consider a process for obtaining an interpolated series of approximation coefficients from a time series $X = \{x_1, x_2, \ldots, x_N\}$, where $N = 2^k$ and there is one level of decomposition *j=1*. Let A_j denote a series of approximation coefficients at level of decomposition *j*, and A_o denotes the input time series *X*.

The following process applies also for obtaining a series of detailed coefficients by utilizing a high-pass filter *G* in place of a low-pass filter *F*. The process of obtaining an interpolated series of approximation coefficients consists of two parts based on decomposition and reconstruction stages of the pyramidal algorithm. We now describe these two stages.

Decomposition stage: The first step of a decomposition stage of a pyramidal algorithm is the convolution of an input time series *X* with a low-pass filter *F*. The convolution operation of a finite sequence $X = \{x_1, x_2, \ldots, x_N\}$ with finite filter $F = \{f_1, f_2, \ldots, f_M\}$ is defined as:

$$X^c = (X * F)[n] = \sum_{m=1}^{M^*} X[n-m+1]F[m] \tag{2.3}$$

where $n = \{1, \ldots, N\}$, m=$\{1, \ldots, M\}$, $M^* = \min(M, n)$.

As can been seen from Eq.(2.3), there are no future data points of *X* required in order to calculate an *n*-th element of a X^c sequence. An *n*-th element of the X^c sequence utilizes at most *n-M+1* past data points of *X*.

The next step of a decomposition stage is a dyadic downsampling, defined as keeping only every second element of the original sequence, yielding a series of approximation coefficients:

$$A_1 = X^d[n] = X^c[2n],\ n = \{1, \ldots, N/2\}\text{, which gives:}$$

$$X^d = \{x_1^d = x_2^c, x_2^d = x_4^c, \ldots, x_{N/2}^d = x_N^c\}$$

The length of the downsampled sequence X^d is evidently half of that of the convoluted sequence X^c and the *n*-th element of the X^d sequence corresponds to the *2*n*-th element of the sequence X^c.

Reconstruction stage: In order to obtain a series of approximation coefficients with the same number of elements *N* as the original time series, the reconstruction procedure is utilized with dyadic upsampling of X^d and then convolution with a reconstruction low-pass filter *F**. This procedure is similar to the reconstruction phase of the pyramidal algorithm (Figure 2-5) where a series of detailed coefficients *D* D_j are assumed to be a sequence of zeros.

The dyadic upsampling of a sequence consists of inserting zeros at every even index, so dyadic upsampling of the sequence $X^d = \left\{x_2^c, x_4^c, \ldots, x_N^c\right\}$ will lead to $X^u = \left\{x_2^c, 0, x_4^c, 0, \ldots, x_N^c, 0\right\}$. Then the convolution is applied to X^u according the Eq. (2.3), which gives a series of approximation coefficients:

$$R = \left(X^u * G\right) = \left\{r_1, r_2, \ldots, r_N\right\} \tag{2.4}$$

where *R* is an interpolated version of sequence A_1 and has the same number of data points *N* as the original time series *X*.

We now show how future data points of *X* must be used in order to calculate data points of the sequence *R*.

From the convolution definition in Eq. (2.3) it can be seen that an element x_{k+1}^c is required in order to calculate the value of an odd data point r_k, *k*={1,3,..,*N*-1}. However, an element x_{k+1}^c in turn requires a data point x_{k+1} of the input time series X. Therefore, for every odd data point of the interpolated sequence *R* there are always future data values of *X* required from one time step ahead. This effect is independent of the length or values of the utilized wavelet filter.

Thus far, only one level of decomposition of wavelet transform has been considered. We now show that the same necessity for future data points of *X* holds for every level of decomposition.

Let there be some value of $j > 1$ levels of decomposition applied to the time series $X = \left\{x_1, x_2, \ldots, x_N\right\}$. The data points of the resulting downsampled time series are obtained as: $X^d = \left\{x_1^d = t\left(x_{2^j}\right), x_2^d = t\left(x_{2^j*2}\right), \ldots, x_{N/2^j}^d = t\left(x_N\right)\right\}$

where $x_k^d = t(x_{k*2^j})$ indicates that the data point x_{k*2^j} of the input time series *X* is required in order to calculate the element x_k^d.

In order to obtain the same number of elements *N* as in the input time series *X*, the same number of levels *j* of the reconstruction phase are required. Applying upsampling and deconvolution *j* times will lead to $R = \{r_1, r_2, \ldots, r_N\}$, where $r_{(k-1)*2^j+1} = t\left(x_{k*2^j}\right), k = \{1, 2, \ldots, N/2^j\}$, where *t(*)* denotes that the future data point x_{k*2^j} is required in order to calculate the element $r_{(k-1)*2^j+1}$.

In summary, the reconstruction phase of the pyramidal algorithm is applied in order to obtain the detailed or approximation coefficient series *R* on *j* levels of decomposition with the same number of data points as in the original time series *X*. This leads to the first element of the block of 2^j data points in *R* being calculated using the last element of the corresponding block of data points of *X* (Figure 2-6). As with the case of *j*=1 considered previously, this holds regardless of data values or size of the wavelet filters applied.

Therefore, applying a decimated discrete wavelet transform based on the pyramidal algorithm prior to splitting the approximation and detailed coefficients into calibration and validation parts must lead to an overoptimistic goodness-of-fit for a wavelet-based forecasting method as measured on validation data.

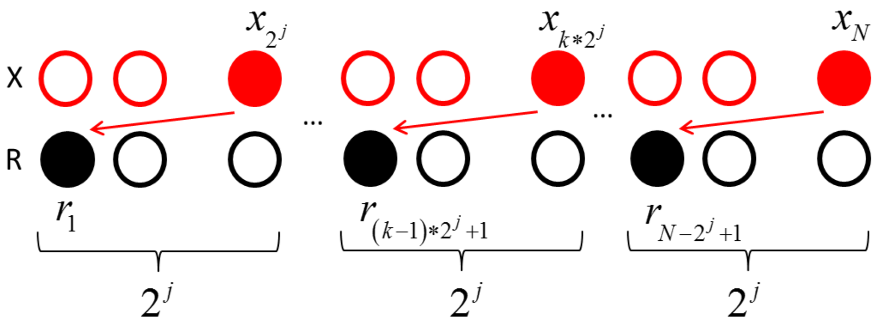

Figure 2-6 Propagation of future values by the discrete wavelet transform based on the pyramidal algorithm with *j* levels of decomposition/reconstruction. Every first element of the blocks of 2^j data points of the reconstructed time series *R* is calculated using the last element of the corresponding block of 2^j data points of the input time series *X*.

References:

Belayneh, A., Adamowski, J., Khalil, B., & Ozga-Zieleinnski, B. (2014). Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression methods, J. Hydrology, 508, 418-429

Benaouda, D., Murtagh, F., Starck, J.L., & Renaud, O. (2006). Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting, Neurocomputing, 70, 139-154

Beriro, D.J., Abrahart, R.J., Mount, N.J., & Nathanail, P. (2012). Letter to the Editor on "Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models" by O. Kisi and J. Shiri, Wat.Res.Man., 26(12),3653-3662

Conejo, A., Plazas, M.A., Espinola, R., & Molina, A.B. (2005a). Day-ahead electricity price forecasting using the wavelet transform and ARIMA models, IEEE transactions on power systems, 20(2), 1035-1042

Conejo, A., Contreras, J., Espinola, R., & Plazas, M.A. (2005b). Forecasting electricity prices for a day-ahead pool-based electric energy market, International Journal of Forecasting, 21(3), 435-462

García-Serrano, J., & Frankignoul, C. (2014). Retraction: High predictability of the winter Euro–Atlantic climate from cryospheric variability, Nature Geoscience, 7, E2

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer (New York), pp. 245-247

Jin, J., & Kim, J. (2015). Forecasting natural gas prices using wavelets, time series, and artificial neural networks, Plos One, 10(11), e0142064
Kim, Y., Suk Shin., H., & Plummer, J.D. (2014). A wavelet-based autoregressive fuzzy model for forecasting algal blooms, Environmental Modelling & Software, 62, 1-10

Kriechbaumer, T., Angus, A., Parsons, D., & Casado M.R .(2014). An improved wavelet-ARIMA approach for forecasting metal prices, Res. Policy, 39, 32-41

Liu, D., Niu, D., Wang, H., & Fan, L. (2014). Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm, Renewable energy, 62, 592-597

Mallat S.G (1989) A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 11(7), 674-693

Misiti, M., Misiti, Y., Oppenheim, G., & Poggi, J.M. (2007). Wavelets and Their Application (ISTE, 2007)

Murtagh, F., Starck, J.L., Renaud, O. (2004). On neuro-wavelet modeling, Dec. Sup. and Systems, 37,475-484

Nourani, V., Baghnam, A.H., Adamowski, J., & Kisi, O. (2014). Applications of hybdrid wavelet-Artificial Intelligence models in hydrology: A review, J. of Hydrology, 514, 358-377

Upadhyaya, BR., Mehta, C., & Bayram, D. (2014). Integration of time series modeling and wavelet transform for monitoring nuclear plant sensors. IEEE Transactions on Nuclear Science, 61(5), 2628-2635

Walker J. S (1999). A Primer on Wavelets for Their Scientific Applications (Studies in Advanced Mathematics, CRC Press)

Wang, W., & Ding, J. (2004). Wavelet network model and its application to the prediction of hydrology, Nature and Science, 1(1), 67-71

Yousefi, S., Weinreich, I., & Reinarz, D. (2005). Wavelet-based prediction of oil prices, Chaos, Solutions and Fractals, 25, 265-275

Zhang, X., Peng, Y., Zhang, C., & Wang, B. (2015). Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences, J.Hydrology, 530, 137-152

2.3 Toward creating simpler hydrological models: A LASSO subset selection approach

The remainder of this chapter is concerned with proposing a framework for constructing no-more-complex-than-necessary event-forced time series models with particular reference to hydrology. Algorithmic simplification as part of this process is achieved through use of the LASSO (Least Absolute Shrinkage and Selection Operator). The approach first constructs a flexible but deliberately over-parameterised linear model, defined by a large number of linear expressions subject to linear constraints. The model fitting process and model simplification are then carried out concurrently, with fitting using any suitable linear method such as least absolute deviations. During the calibration fitting the original over-parameterised model is formally forced to a simpler model no more complex than required for application to the calibration data concerned. Unlike the usual model calibration process, the model after calibration is different and simpler.

There is of course an apparent contradiction in carrying out linear modelling of nonlinear hydrological processes. However, a case is made that constrained linear modelling of the type considered here can be formulated to be as "nonlinear" as necessary, through the use of linear combinations of nonlinear basis functions. A basis function can be defined as an element of a particular basis for a function space. For example, a quadratic polynomial comprises the basis functions 1, x, and x^2 and the expression $a*1+b*x+c*x^2$ is a linear combination of basis functions.

A requirement before the simplification process is the creation of the initial linear model for the nonlinear situation under study. That is, the entire model must be specified as a sequence of linear constraints, with fitting to data being a linear programming (LP) minimisation of absolute deviations or similar linear measure.

Achieving accurate linear approximation of a nonlinear process is not necessarily a trivial task and a full review of all mechanisms by which it might be achieved is beyond the scope of this study. However, a sense of the type of formulation required is illustrated in this section with respect to creating a linear approximation for a general nonlinear time series model.

The nonlinear conceptual time series model is familiar and not necessarily specific to hydrology: events occur at points in time and each event marks the initiation of a continuous non-negative nonlinear response which may be of arbitrary form but eventually declines to zero with increasing time. The responses may change from one event to the next depending on event magnitude and the current state of the system. The event responses sum together, producing the model time series output for some recording point. Variations of this approach have long been used in the context of rainfall-runoff modelling where the current state of a catchment influences the nature of runoff responses from rainfall events, with the individual event responses summing to give model discharge, possibly superimposed on a constant baseflow.

A conversion of this conceptual model to a linear approximation model is demonstrated by first considering a single event and its response. Defining this event to occur at time $t = \tau$, the magnitude of the response at any subsequent time t is expressed as a weighted finite mixture of L pre-selected non-negative nonlinear functions *g(t),* all with origin at time τ:

$$f(t, Z_\tau) = \sum_{i=1}^{L} \omega_i(Z_\tau)\, g_i(t), \quad t \geq \tau \tag{2.5}$$

The $\omega_i(Z_\tau)$ terms in Eq. (2.5) are non-negative weighting expressions which give greater or lesser emphasis to individual $g_i(t)$ functions. The particular set of *g(t)* functions chosen by the modeller would be representative of a range of possible event responses for the physical process under consideration. For example, in the case of a rainfall-runoff model this could be a number of different hydrograph forms characteristic of the catchment type and size. The choice of *g(t)* functions will inevitably include some which will not in fact be helpful for a given application to data. However, these redundant functions will be eliminated later in the simplification process. A greater number of *g(t)* functions would be chosen for a model intended for more general use. This will result in a greater number of *g(t)* eliminations during the subsequent simplification when applied to data.

The $\omega_i(Z_\tau)$ terms in Eq. (2.5) are linear combinations of a set of *M* independent variables *Zτ* whose magnitude may have influence on the system at event time τ:

$$\omega_i(Z_\tau) = \sum_{j=1}^{M} a_{ij} Z_{\tau j}, \qquad a_{ij} \geq 0,\ Z_{\tau j} \geq 0 \tag{2.6}$$

The avoidance of negative $\omega_i(Z_\tau)$ terms ensures Eq. (2.5) cannot yield a negative prediction for the positive-valued response process concerned.

Eq. (2.6) defines the *i*th of the *L* weighting expressions in Eq. (2.5) and is a linear combination of the same set of independent variables *Z*. However, the weighting coefficients a_{ij} differ in value from one weighting expression to the next. The initial choice of the *M* independent causal variables represents a physical working hypothesis and it may happen that most are later eliminated in the linear LASSO simplification process.

In summary, the nonlinear response following a single event is modelled as a positive-valued weighted mixture of pre-chosen nonlinear *g(t)* functions, with their associated weights being linear combinations of *M* independent causal variables. As noted earlier, pre-chosen functions like *g(t)* are referred to as basis functions (Bishop, 2006, p.138) and when used as weighted mixtures can approximate many different nonlinear functions when *L* is sufficiently large.

With respect now to multiple events, the events are defined to occur at respective times τ[1], τ[2], τ[3] …, with the same set of *g(t)* functions and *Z* variables operative for each τ[i]. However, the respective weights *ω(Z)* for each of the *g(t)* will differ from one event to the next because the magnitudes of the *Z* variables change with time.

The model time series output at any given time *t* is the sum of the responses from all previous events up to that time. Defining *t* = 0 as the start of the time series, at some subsequent time *t* there will have been *K(t)* previous events which occurred at times τ[1], τ[2] ...τ[K] . Therefore, at time *t* the model-generated value *h(t,Z)* can be written:

$$h(t,Z) = \omega_0 + \sum_{n=1}^{K(t)} f\left(t, Z_{\tau[n]}\right) \tag{2.7}$$

where $Z_{\tau[n]}$ denotes the magnitudes of the independent variables at the time of the *n*-th event. The constant ω_0 may be set to zero depending on the context. For example, a non-zero value might represent some constant baseflow in a rainfall-runoff model.

As an aside, if an event can be thought of as the input of a set of particles into a store and the response is the time-varying rate of exit of those particles from the store, then at any time *t* the model-defined mean residence time *T(t,Z)* of particles (derived from all the prior events) exiting the store is given by the weighted average of the previous event times:

$$T(t,Z) = \sum_{n=1}^{K(t)} \tau[n] f\left(t, Z_{\tau[n]}\right) / \sum_{n=1}^{K(t)} f\left(t, Z_{\tau[n]}\right) \tag{2.8}$$

Eq. (2.8) could have application, for example, in considering the age of water exiting from a catchment.

Having expressed the conceptual model as a linear approximation, it remains to set out the coefficients as an LP minimisation matrix. The coefficients are all constrained to be non-negative to avoid negative *h(t,Z)* values and the minimisation is with respect to least absolute deviations.

Following from Eq.(2.7), the matrix will have *U* rows (with each row corresponding here to a unit of time) and *L* × *M* +2*U* columns, with one additional column with all values set to 1.0 if ω_0 is permitted to have an unknown nonzero value. The 2*U* columns here are the utility fitting variables required for least absolute deviations regression, two per data observation, and are not part of the model (Bloomfield and Steiger, 1983, ch. 6). As far as the model parameters are concerned, $\omega_0$would be an unknown to be solved for, along with the *L* × *M* unknown *a* coefficients. All the unknowns are constrained to be non-negative in the LP solution, as required by the specification of Eq. (2.5) and Eq. (2.6).

The number of matrix rows *U* will generally be less than the original number of rows in the time series concerned. This is because the user must define the first row in the matrix corresponding to a *t* large enough to avoid any response effects which may be still present from events prior to the start of the time series at *t* = 0.

The LP matrix does not define a linear model with *M* independent *Z* variables corresponding to the independent *X* variables of a standard linear regression. The *Z* variables do not influence the model in a direct linear way, but indirectly through the intermediary of determining the weight magnitudes via the time-varying weighted linear combinations of the *L* different nonlinear *g(t)* functions. It may happen that the *Z* variables are themselves outputs from pre-chosen nonlinear expressions. The $L \times M$ variables here might be better termed pseudo variables because they combine the effect of different *g(t)* functions as opposed to physical variables.

This type of initial model with numerous *g(t)* functions will inevitably result in many superfluous parameters and there is no suggestion that such models should be applied directly in practice. Instead, they are only a means to an end and serve as the necessary preliminary stage before initiating the subsequent linear LASSO simplification to produce models for application.

The LASSO (Least Absolute Shrinkage and Selection Operator) was introduced by Tibshirani, 1996 as a means of eliminating less informative variables in least squares multiple linear regression. It has been applied in many fields but has only relatively recently been introduced into the hydroclimatic literature (Hammami et al., 2012). To our knowledge, the present study is the first application of the LASSO in the more general context of model simplification rather than simply selection of a subset of informative independent variables in linear regression.

Briefly, the LASSO concept maintains the linear regression approach of seeking to match a linear function to a data set, but with the additional aspect of some degree of forcing of the parameters toward zero. Scaling is required prior to avoid preferential elimination of parameters because of units of measurement. A user-specified positive parameter λ defines the relative partitioning between optimising the parameter values toward data matching or forcing the parameters to zero. A large value of λ will cause all parameters to be set to zero while a small λ will not have any simplifying effect. See also Hammami et al., (2012), Wheeler (2009), and Tibshirani (2011).

The deleted variables will be those whose elimination has least effect on fitting the linear regression model to data while all parameters are being forced toward zero. Deletion might arise, for example, if a variable has weak explanatory power. Alternatively, some variables may be highly correlated so that when one is eliminated to zero another can take its place.

The nonzero parameters remaining after a LASSO process will have values biased toward zero. This can be offset with a subsequent standard linear regression with the independent variables now being just that surviving subset. The parameter values from the second regression will usually have larger absolute values and the model will better fit the data because the biasing effect will be at least partly removed.

However, the least-squares LASSO has a disadvantage for model simplification purposes. Specifically, linear constraints cannot be included without transforming the fit procedure into a quadratic optimisation exercise, which may be slow to run for large problems and will not necessarily yield a global minimum.

There is particular advantage in being able to incorporate linear constraints into models, both as part of the model description and because the constraints may result in many parameters never becoming part of the model. Such model improvement from inclusion of constraints has previously was noted, for example, by Gharari, et al. (2014).

The constraints here might be as simple as avoiding negative discharge in a hydrological model or could be a more complex constraint set to approximate some physical process.

We therefore utilise here the linear LASSO (Wang, et al., 2006) as the LASSO version most suited to model simplification. This permits linear constraints while still giving a single optimal global solution for calibration fitting with specified λ. In addition, the linear LASSO can be applied when there are more parameters than data points.

The LASSO simplification is achieved by way of the positive LASSO simplification parameter λ. The role of λ is discussed in more detail in Chapter 4 but for the purposes of model simplification it is sufficient to note that as λ increases more of the model parameters are forced to zero, resulting in a simpler model. It is then of interest to check whether the simpler model avoids over-fitting issues and leads to improved prediction on independent data.

By the way of illustration, the results of calibration with simplification of a linear rainfall-runoff model are given here. Further details can be found in (Bardsley et al, 2015). The initial linear model in this case comprises a multi-component weighted finite mixture distribution, with each distribution being a pre-calculated nonlinear form, thus avoiding the use of nonlinear parameters.

This model was used with calibration / simplification with respect to the short time series shown in Figure 2-7. The partition of the time series into calibration and evaluation data sets was such that the calibration data set had less data range than the evaluation set, providing a more difficult challenge for the simplified version of the model.

As it happened, just the constraint of non-negative discharge (no force of simplification) is itself sufficient to reduce the initial 163-parameter linear model to just 13 parameters while still maintaining $\lambda = 0$. However, even 13 parameters evidently is still too many, as seen by poor matching to the evaluation data set to the right of Fig 2-7. In particular there is considerable error in predicting the magnitude of the large flood peak in the evaluation set.

Figure 2-8 gives the corresponding time series for the best evaluation data match where the forced simplification reduced the model to 6 parameters and the model remained constant over the λ range $0.9 \leq \lambda \leq 2.0$. It is of interest that this simpler model apparently gives a reasonable estimate of the 12.8 m^3s^{-1} peak discharge around hour 500, despite the peak discharge of the calibration data being much lower (6 m^3s^{-1}). This simplified model comprises three *g(t)* functions and six parameters, with one of the six being a non-zero baseflow constant.

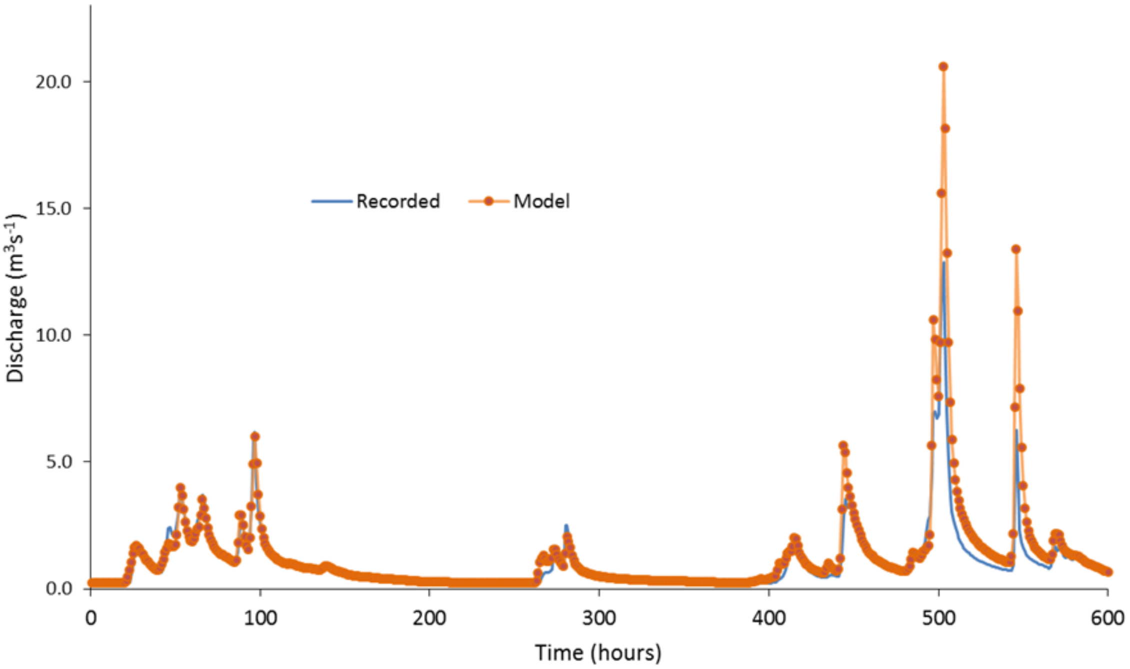

Figure 2-7 Model calibration fitting (λ = 0, 13 nonzero parameters) prior to simplification. The calibration data is for the first 400 hours, with the last 200 hours being used for evaluation.

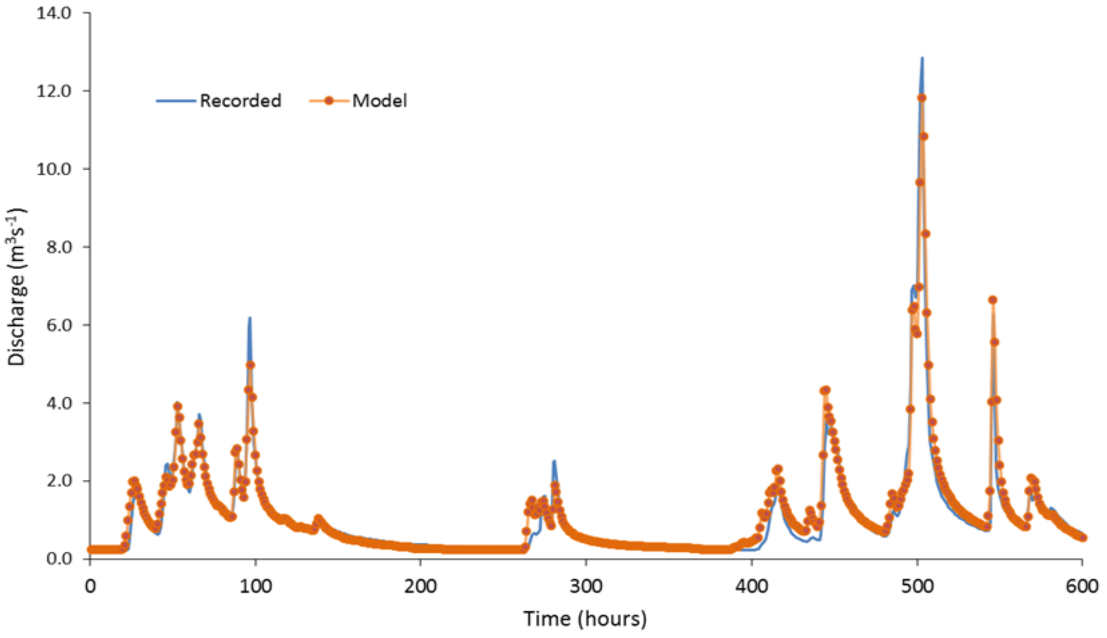

Figure 2-8 Model fitting (6 parameters) to calibration data after previous linear LASSO simplification with ($0.9 \leq \lambda \leq 2.0$), giving the best match to evaluation data.

References

Bardsley, W.E., Vetrova V., and S. Liu, 2015. Toward creating simpler hydrological models: A LASSO subset selection approach,Environ. Model. Softw., 72, 33–43

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Singapore.

Bloomfield, P., Steiger, W., 1983. Least Absolute Deviations: Theory, Applications and Algorithms. Birkhäuser.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., Savenije, H.H.G., 2014. Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. Hydrology and Earth System Sciences 18, 4839-4859.

Hammami, D., Lee, T.S., Ouarda, T.B.M.J., Lee, J., 2012. Predictor selection for downscaling GCM data with LASSO. Journal of Geophysical Research 117, D17116, doi:10.1029/2012JD017864.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58, 267–288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society, 73, 273–282.

Wang, L., Gordon, M. D., Zhu, J., 2006. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. IEEE Sixth International Conference on Data Mining, Proceedings, 690–700.

Wheeler, D.C., 2009. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environment and Planning A 41, 722 –742.

2.4 Conclusions

The first part of this chapter demonstrated the general result of the requirement for an impossible forecasting situation such that future values of an input time series must be known for all forecast methodology which uses calculation of wavelet components for the decimated discrete wavelet transform based on the pyramidal algorithm. The need for future data values in fact increases with the resolution level of the wavelet transform, regardless of the chosen wavelet filter. It was also shown that if pre-processing of time series with the decimated wavelet transform is performed prior to data separation into calibration and validation portions then there will always be dependency between the calibration and validation parts, preventing the necessary independence of the two data subsets. Illusionary and over-optimistic forecasting accuracy is the inevitable consequence of such dependency, regardless of the forecasting machine learning method of choice. This also has implications on the veracity of published forecasts in economics and other disciplines.

The second part of the chapter provided a general time series modelling simplification framework, being in principle applicable over many fields of nonlinear hydrology and in other subject areas of environmental science as well. This development is still in its early stages, but the most likely first extensions of the method might be toward creating simpler distributed hydrological models where the most sensitive spatial locations governing the output process will be preferentially selected from the simplification. This could result in considerable reduction in the spatial input information required by such models while also avoiding overfitting.

Chapter 3 Associations in bivariate data: a randomization approach

3.1 Introduction

Testing of association between two variables is a useful step of data analysis and also as screening method to reduce dimensionality when there is a large number of possible predictor variables. In the field of machine learning such dimension reduction techniques are referred to as filter methods. In the context of environmental data analysis, filter methods are particularly useful in identifying informative predictors in forecasting models, for example in determining which atmospheric-oceanic circulation modes may influence local hydrological processes.

This chapter introduces two methods of significance testing for the associations in bivariate data, utilising environmental data examples. The first part of the chapter introduces a significance test for the specific case where there is an apparent data-sparse zone in a scatterplot. Such data sparsity might represent some environmental threshold effect which operates to reduce the frequency of values over that portion of the scatterplot. The second part test constitutes a generalisation of the approach presented in the first part and introduces a simple non-parametric test of significance of general associations between two variables.

The first part has been published in *Hydrology and Earth Systems Sciences* [1] and the second part is currently under review.

Reference:

[1] V. V Vetrova and W. E. Bardsley, "Technical note: A significance test for data-sparse zones in scatter plots," *Hydrol. Earth Syst. Sci*, vol. 16, pp. 1255–1257, 2012.

3.2 A significance test for data-sparse zones in scatter plots

Abstract

Data-sparse zones in scatter plots of hydrological variables can be of interest in various contexts. For example, a well-defined data-sparse zone may indicate inhibition of one variable by another. It is of interest therefore to determine whether data-sparse regions in scatter plots are of sufficient extent to be beyond random chance. We consider the specific situation of data-sparse regions defined by a linear internal boundary within a scatter plot defined over a rectangular region. An Excel VBA macro is provided for carrying out a randomisation-based significance test of the data-sparse region, taking into account both the within-region number of data points and the extent of the region. Example applications are given with respect to a rainfall time series from Israel and also to validation scatter plots from a seasonal forecasting model for lake inflows in New Zealand.

Introduction

A visual examination of hydrological scatter plots is a useful first step toward considering possible relationships between variables, or for evaluation of the worth of hydrological forecasting models via validation plots of observed and predicted values. It is intuitive that we tend to focus on regions in scatter plots with greatest data density as this suggests the highest degree of association and worth the most effort in further refinements – see, for example, Green and Finlay (2008). However, a sufficiently extensive data-sparse zone in a scatter plot can be of interest also as this may suggest that for a specific magnitude range one variable might restrict the other.

For hydrological variables, the transition between data-sparse and data-dense fields in scatter plots will most likely be a poorly-defined boundary which can be thought of as a stochastic frontier, for which a range of estimation techniques are available (Hall and Simar, 2002; Florens and Simar, 2005; Delaigle and Gijbels, 2006; Kumbhakar et al., 2007). Our focus here is not on boundary estimation as such, but rather on providing a significance test against the null hypothesis that a data-sparse zone in a scatter plot has arisen by random chance. Specifically, the purpose of this analysis is to provide a practical significance test for the size of the area of an observed data-sparse region with a linear internal boundary in a scatter plot within the specific rectangular region which just encompasses all the data points. The test requires no assumptions concerning the data. For convenience, the data-sparse area is taken to mean its proportion of the rectangle area. Given that there are m data points within the data-sparse area $\Delta(m)$, the null hypothesis is that a data-sparse region at least as large and containing m data points could have arisen by random chance. Rejection of the null hypothesis does not imply any specific alternative with respect to correlation between the variables, but simply indicates that the data-sparse region is confirmed large enough so as to be unlikely to have arisen by chance. The approach adopted here represents a generalisation of an earlier test described by Bardsley et al. (1999) which was restricted in practical application because it required the data-sparse region to contain no data points at all (m = 0).

The nature of a data-sparse (as opposed to no-data) region is illustrated with respect to the scatter plot in Figure 3-1. The pattern of data points suggests a possible linear rising trend in an upper boundary for October rainfalls at a site in Israel over the period 1951–1987, but with an unusually wet month in October 1986 as an outlier. The $m = 0$ requirement of the original 1999 test required a somewhat unrealistic location of a boundary as being above the outlier (Bardsley et al., 1999, Fig. 3a). A better approach is to deem “data sparse” in this particular case as permitting a single point within the data-sparse region ($m = 1$) which now gives a better linear boundary location just above the other data points (Figure 3-1).

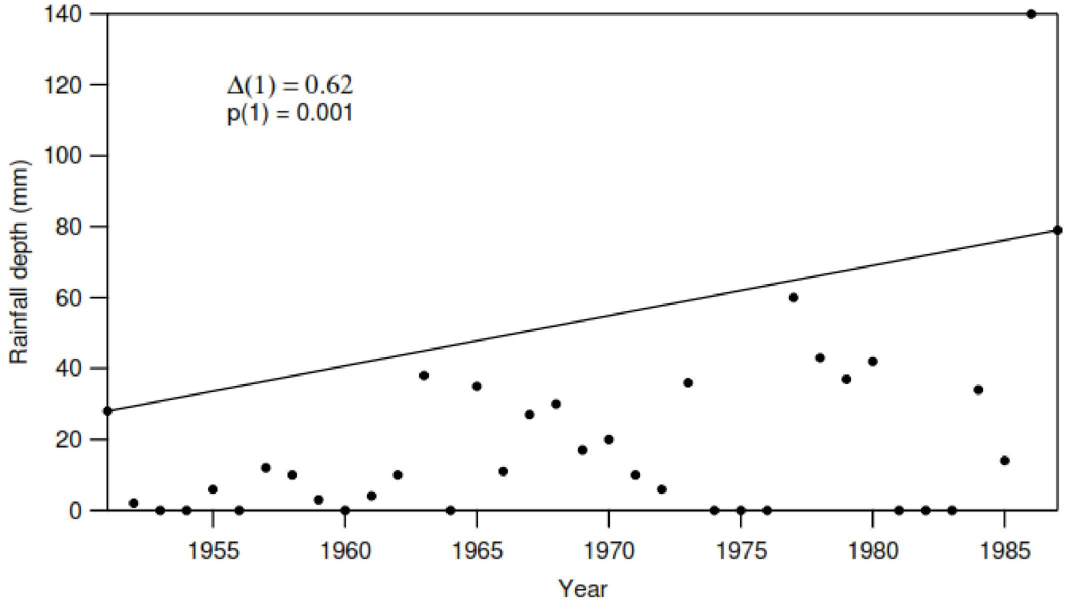

Figure 3-1 Scatter plot of October Rainfall values (1951–1987) at Berurim in southern Israel. Line shows the linear internal boundary of the largest possible data-sparse region for m = 1 (see text for definitions).

The test

Following Bardsley et al. (1999), the significance test of the present paper is based on a standard randomisation approach. That is, the x-coordinates of the data points are randomly reassigned, giving rise to a different pattern of points in the scatter plot in the rectangular region. For example, if the x-axis represented yearly values, then this would amount to a random reordering of years. After a given random reordering of x-coordinates, a check is made in the algorithm whether the largest upper-left region (with linear internal boundary) containing *m* data points is larger than the upper-left data-sparse area in the original scatter plot.

This random reordering of the x-coordinates is repeated many times, and the proportion of times *p* that the original data-sparse area is exceeded is calculated. This *p* value is the probability of obtaining a data-sparse area at least as large as observed in the original scatter plot, given that the null hypothesis is true. Therefore, if *p* is sufficiently small, say less than 0.05, then the size of the original data-sparse region $\Delta(m)$ is deemed statistically significant. The number of random reorderings needed for the required precision of p is determined from the binomial theorem in the usual way (see Appendix in [1]).

A general VBA macro which is unrestricted as to the size of m is described in the Excel spreadsheet supplementary to [1]. The macro appears efficient in trial runs but inevitably will become slower for large numbers of points in the scatter plots coupled with large m. When the macro is applied to the indicated data-sparse region above the line in Figure 3-1 ($m = 1$), the resulting p value is obtained as $p(1) = 0.001$, which is a higher level of statistical significance then the value of $p(0) = 0.02$ listed in Fig. 3a of Bardsley et al. (1999) for the case of $m = 0$. Of course, there is no general guarantee that higher levels of significance will be obtained for the test proposed here, as this is dependent on the data pattern of the scatter plot.

Application to validation scatter plots

Scatter plots most commonly serve as a graphical indication of some degree of association between two variables. In addition, scatter plots are often used in hydrology to give a graphical indication of how well some model fits a set of validation data. The ideal here is to have points scattered close to the 1:1[1] line and Bardsley and Purdie (2007) present an "invalidation test" as one means of testing departure from this situation. However, a validation scatter plot may indicate failure in the sense of poor 1:1 fitting but nonetheless still possess some degree of predictive ability as evident from the pattern of points. For example, the location of a data-sparse region in a validation scatter plot may suggest that low predicted values tend to be associated with low observed values, but increasingly large predicted values result in high or low magnitudes being as likely.

[1] The 1:1 line corresponds to exact matching of observed and predicted values.

An illustration of this situation is given in Figure 3-2, which shows a validation data set with respect to a seasonal lake inflow forecasting model seeking to anticipate total autumn inflow from the standpoint of autumn in the previous year. The lakes concerned (Tekapo and Pukaki) are adjacent New Zealand hydro storage lakes and it is convenient to consider seasonal forecasts of the combined inflow volumes of both lakes. The forecasting model itself is further described in Chapter 4 but our interest here is that the validation scatter plot can be interpreted as the forecasts giving a low probability to high inflows when low lake inflows are forecast. However, at the same time high inflow forecasts may associate with high or low actual inflows. This lends itself to a data-sparse significance test (m = 0) which in fact indicates high significance of the sparse zone above the solid line with p(0) = 0.0004.

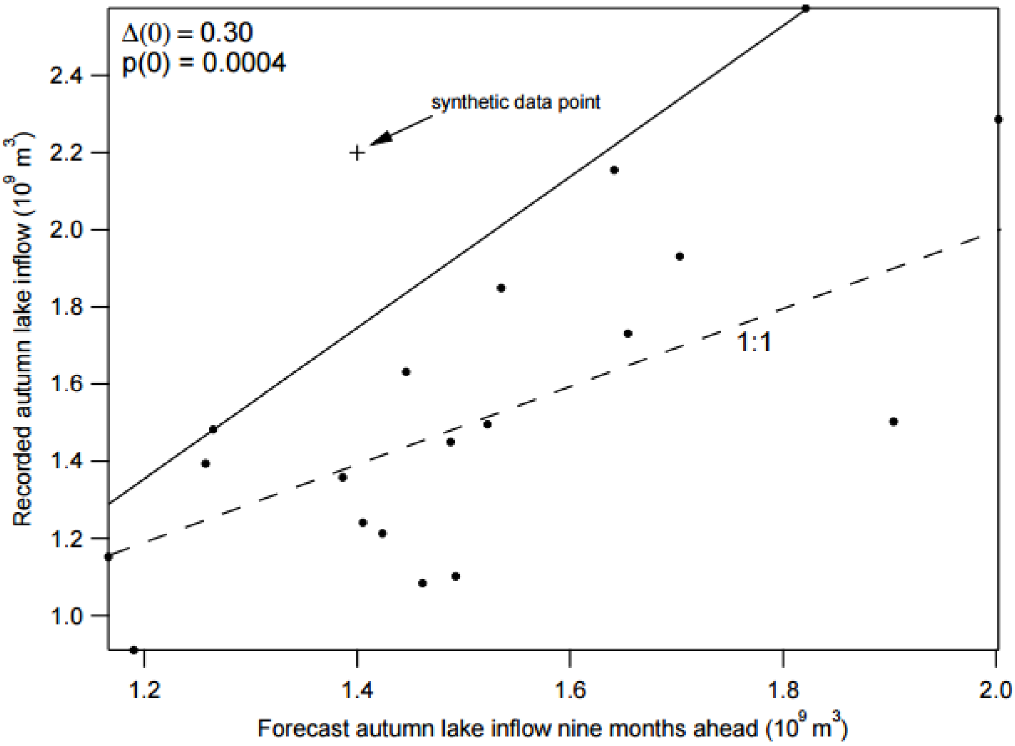

Figure 3-2 Validation plot for a model forecasting combined autumn river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

Although an m = 0 test may appear sufficient here, there could be concern over robustness of the conclusion because of the small number of data points involved. The m > 0 test gives an empirical means of robustness checking because artificial data points can be inserted into the data-sparse zone and a check made to see if statistical significance is maintained. For example, inserting the single synthetic data point indicated in Figure 3-2 yields p(1) = 0.002, which is still highly significant. The forecasting model in this case should probably be robust therefore against a future real data point appearing in the sparse zone. Further synthetic data points could be inserted if required. The autumn forecasting model here is restrictive in that low forecast flows will tend to be below the solid line. However, forecast high flows in reality could be anywhere within the magnitude range. The forecasting value is with respect to a high probability that a forecast flow will not be in the data-sparse zone, as opposed to being near or far from the 1:1 line.

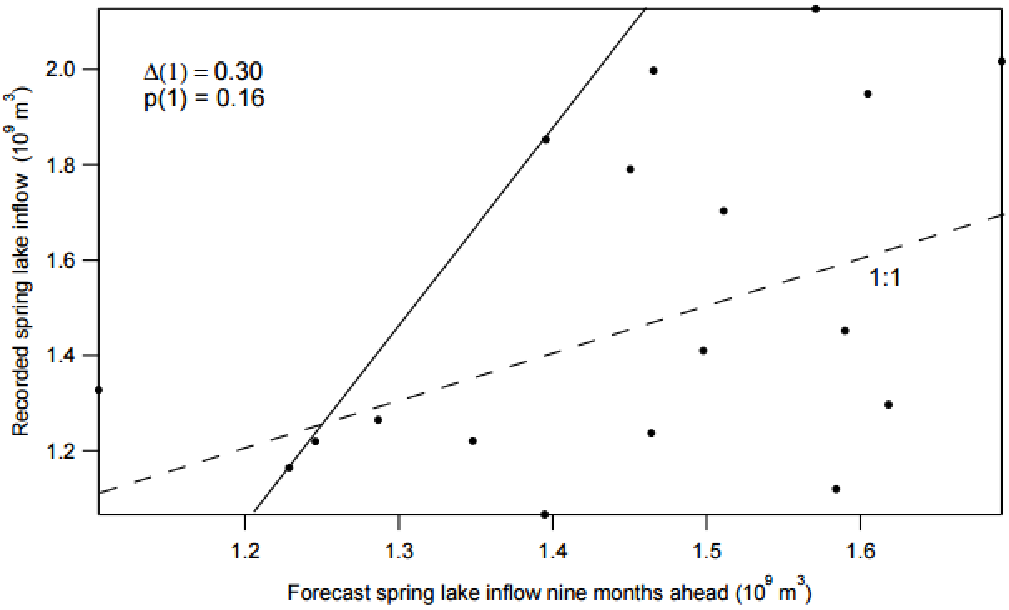

Figure 3-3 Validation plot for a model forecasting combined spring river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

This view of forecasting value is also illustrated in Figure 3-3 , showing in this case the validation results of a model for forecasting spring inflows into the two lakes, where the model is forecasting from the previous spring. The predictive model clearly fails in the sense of any 1:1 matching, but the hope might be that the indicated solid line approximates an upper bound to actual inflows when forecasts are in the range 1.20– 1.45 × 109 m^3 . As with the autumn model, the validity of this upper bound is tested for significance via the macro with respect to the relative size of the upper left (*m = 1*) data-sparse empty corner. It happens in fact that the macro-derived p(1) value of 0.16 indicates the sparse zone size is no larger than expected from chance. The predictive model therefore fails not only in the 1:1

sense, but also in the sense of establishing the existence of an upper boundary which might permit an estimated upper bound to some forecast inflows.

Discussion and conclusion

There is an element of subjectivity introduced for the test considered here with m > 0, in that sometimes it will not be evident which value of *m* best defines a data-sparse region. Some trial and error process will most likely be required in such instances. With respect to further development, the test approach considered here should be amenable to generalisation such as allowing for curved inner boundaries and incorporating multiple dimensions. However, the randomisation algorithms may become complex and slow. As noted in Bardsley et al. (1999), there will be data situations where linear regression is the most appropriate analysis technique. In other situations where data-dense and data-sparse fields are separated by an approximate linear boundary, the test given here should find practical applications for both associations between variables and also for checking validation scatter plots under situations of restricted forecasting ability.

References

Bardsley, W. E. and Purdie, J.: An invalidation test for predictive models, J. Hydrol., 338, 57–62, 2007.

Bardsley, W. E., Jorgensen, M. A., Alpert, P., and Ben-Gai, T.: A significance test for empty corners in scatter diagrams, J. Hydrol., 219, 1–6, 1999.

Delaigle, A. and Gijbels, I.: Data-driven boundary estimation in deconvolution problems, Comput. Stat. Data An., 50, 1965–1994, 2006.

Florens, J.-P. and Simar, L.: Parametric approximations of nonparametric frontiers, J. Econometrics, 124, 91–116, 2005.

Green, M. B. and Finlay, J. C.: Detecting characteristic hydrological and biogeochemical signals through nonparametric scatter plot analysis of normalized data, Water Resour. Res., 44, W08455, doi:10.1029/2007WR006509, 2008.

Hall, P. and Simar, L.: Estimating a changepoint, boundary, or frontier in the presence of observation error, J. Am. Stat. Assoc., 97, 523–534, 2002.

Kumbhakar, S. C., Park, B. U., Simar, L., and Tsionas, E. G.: Nonparametric stochastic frontiers: a local maximum likelihood approach, J. Econometrics, 137, 1–27, 2007

3.3 A simple nonparametric test of bivariate association for environmental data exploration

Abstract

A simple nonparametric test for general bivariate association is proposed on the basis that randomising the data of a scatter plot will tend to result in greater mean nearest neighbour distances if an association is present. The test is carried out by finding the proportion **p** of randomisations giving smaller mean nearest neighbour distances. A well-defined association (not necessarily linear) in the scatter plot will tend to produce values of **p** near zero, while minimal association plot will give larger **p** values. The test is based only on the data, requires no estimation of intermediary entities like joint or marginal distributions, and makes no assumption concerning the origin of the data. The test is not necessarily more powerful or optimal in any sense, but could still provide a useful tool for preliminary investigation of bivariate data sets such as might arise from a study of climatological relationships.

Introduction

In seeking to construct an explanatory model for an environmental variable, it is common to have situations where there are a large number of available candidate variables which may or may not have causal association with that dependent variable. Also, there may be no certainty whether causal associations will always be linear.

One approach for model construction in such situations is to develop procedures for selecting a subset of predictor variables which have some unspecified form of statistical association with the dependent variable. A dimension reduction method is usually applied to reduce redundancy to give minimal correlation between the selected variables. Examples of such studies in environmental applications include Quilty et al. (2016), Tran et al. (2015), Sharma and Mehrotra (2014), Hejazi and Cai (2009), and May et al. (2008). A methodology overview on variable selection in general is included in Huang and Zhu (2016).

There can be many forms of association (not necessarily all causal) and there may be merit in preliminary data exploration to give some first insight into possible causal associations. The simplest approach here is to carry out variable-by-variable checks for any form of bivariate association with the dependent variable. For linear associations this can be achieved by computing correlation coefficients but it is less clear as to choice of an appropriate bivariate index when there is possibility for undefined nonlinear associations.

In the statistical literature there is an extensive history of general bivariate association measures, including both specific bivariate indices and bivariate indices as special cases of multivariate indices. Bivariate association measures to date include two-sample comparisons with respect to empirical distribution functions (Blum et al., 1961), empirical characteristic functions (Kankainen and Ushakov, 1998), information measure (Linfoot, 1957; Kraskov et al., 2004; Sugiyama, 2011; Reshef et al., 2011), general nonlinear functional relations (Delicado and Smrekar, 2009), rank-based methods (Kallenberg and Ledwina, 1999; Heller et al., 2013), bivariate copulas (Schweizer and Wolff, 1981), distance correlation (Szekely et al., 2007), and continuous analysis of variance (Wang et al., 2015). Murrell et al. (2016) introduced a generalized coefficient of variation and Karvanen (2005) gives a resampling test of independence for application to data from two stationary time series.

Many of these tests of bivariate association, though powerful, are not intuitive to the average user and often require nonparametric estimation of intermediary quantities such as joint or marginal probability distributions. We propose here a simple and intuitive nearest-neighbour test for detection of general bivariate associations. The test is truly nonparametric in the sense that only the data values are used, there is no estimation of intermediary entities, and no assumption is made about the mechanism of data generation.

Nearest neighbours have been used previously for association measure in the context of a mutual information estimation (Kraskov et al., 2004). However, we believe our approach to be the first purely data-based use of the nearest neighbour as a test of bivariate association.

Our motivation derives from wishing to formulate a simple test of general bivariate association as a first step toward locating spatial regions which might have influence on variables of interest. For example, rainfall at a given location on land may be partly linked to sea surface temperatures over some nearby oceanic region.

It might happen that the simplicity of the method results in its finding some application for detecting unspecified bivariate associations generally. However, there is no suggestion that the test is more powerful than others or has optimal properties in some sense. In fact, our purpose in this brief communication is simply to introduce the index rather than carry out detailed comparisons with other indices on reference data sets. It could actually be argued that for data exploration purposes high test power may not be desirable because there are so many forms of possible bivariate association that there will inevitably be instances where detected associations are not suited to predictive models.

Section 2 describes the test of bivariate association, Section 3 illustrates the nature of the test using synthetic examples, and Section 4 applies the method for preliminary identification of locations where atmospheric pressure influences westerly wind frequency on the west coast of the South Island of New Zealand.

2. Test of bivariate association

Following from Murrell et al. (2016) we note that randomisation will destroy any existing form of bivariate association and thus gives a useful reference base for no association. That is, for a given bivariate data set $\left\{X,Y\right\}, X \in R,\ Y \in R,$ there can be no association if the *X* values have been rearranged in random order. We next seek a non-specific test statistic of association which can be utilised with randomisation. We adopt the scatter plot mean nearest neighbour distance $\overline{D}$ as an intuitive measure in this regard, and define **p** as a proportion of those scatter plots created from randomisations of *X* which have mean nearest neighbour distances less than $\overline{D}$. If there is some form of association between *X* and *Y* then randomisation of *X* will tend to produce scatter plots with increased mean nearest neighbour values, resulting in smaller values of **p**. If **p** is sufficiently small, say **p** ≤ 0.05, then this can be taken as indicating some form of association of {*X*,*Y*}.

The effect of *X*-randomization is illustrated in Figure 3-4 and Figure 3-5, showing how a single randomization of *X* has varying degrees of disruption depending on the original extent of {*X*,*Y*} association. The well-defined circular pattern of Figure 3-4a is randomised to disorganised spatial scatter in Figure 3-4b, but the minimal {*X*,*Y*} association for the two clusters of Figure 3-4c is largely unchanged from a randomisation of *X* (Figure 3-4d). Similarly, the random scatter of points in Figure 3-5a remains a random scatter after a randomisation (Figure 3-5b) and there should be minimal change in the mean nearest neighbour spacing.

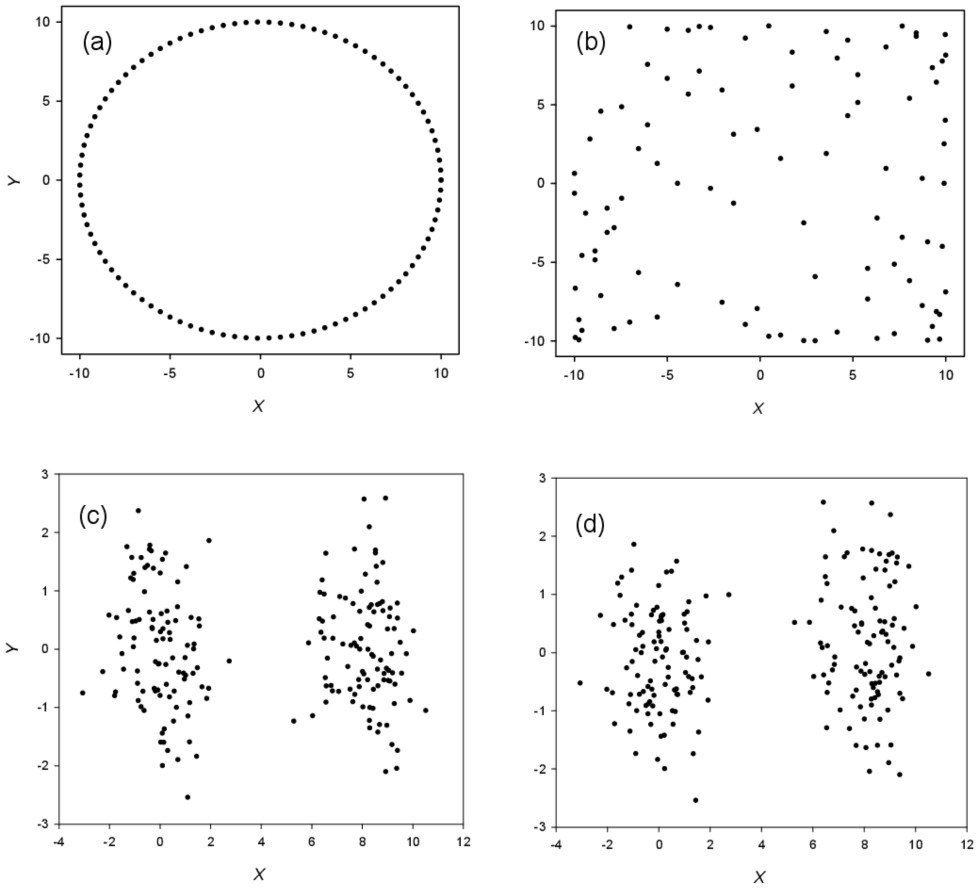

Figure 3-4 A single randomisation applied to a circle pattern (a,b) and to two clusters (c,d).

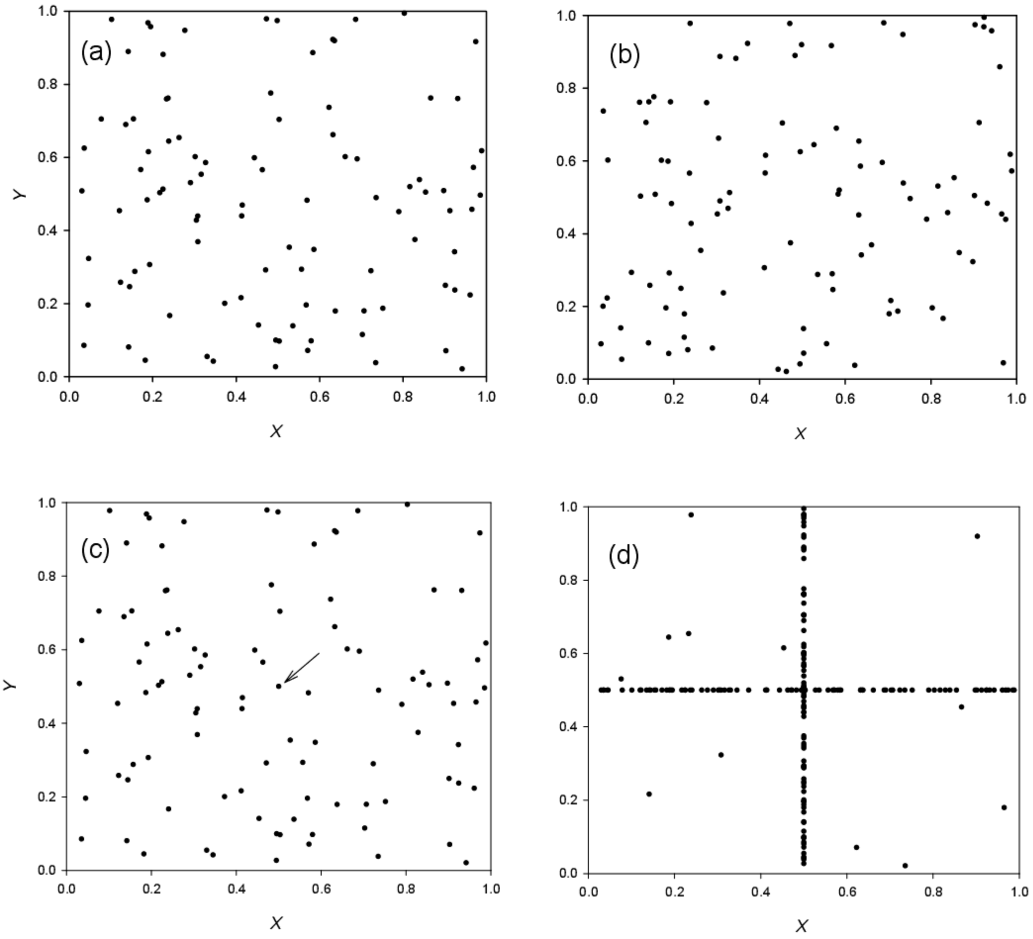

Figure 3-5 A single randomisation applied to a random scatter (a,b) and a random scatter with many points located at the arrowed location (c,d).

Figure 3-5c illustrates how a randomisation may give rise to an apparent spatial pattern. Figure 3-5c is the same *{X,Y}* spatial pattern as Figure 3-5a except that 800 *X,Y* points have been added at one location (arrowed), all with $X = 0.5$, $Y = 0.5$. Randomisation with a single tight cluster of this type gives rise to a cross pattern like that shown in Figure 3-5d. The non-association would still be reflected in high **p** values, however, because outlying data points will tend to come together along the x = 0 or y =0 axes, giving mostly smaller mean nearest neighbour distances than $\overline{D}$.

Synthetic examples

The use of **p** as a bivariate association test is illustrated in Figure 3-6 with four synthetic data sets of 200 values in varying forms of association of $\{X,Y\}$. The

$\overline{D}$ value for each scatter plot was first calculated, followed by randomisations of the respective *X* values to obtain **p** values for the four plots.

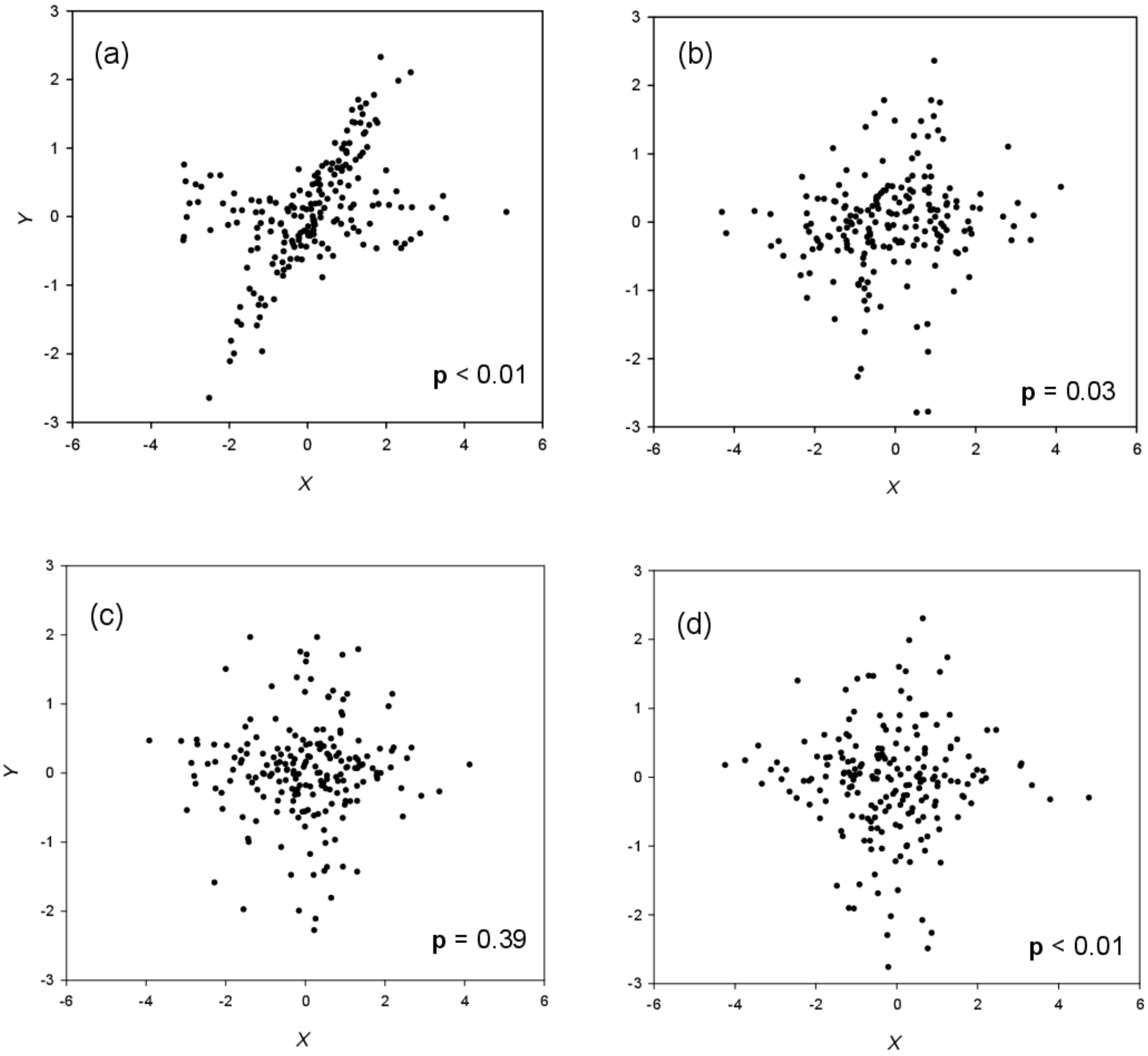

Figure 3-6 The **p** test of bivariate association applied to various scatter patterns.

Figure 3-6a shows an underlying linear association with random noise distributed about y = 0, with the association still being detected as a small value of **p**. Figure 3-6b and Figure 3-6d exhibit a roughly diamond pattern which would be disrupted by randomisation, creating small **p** values. Figure 3-6c suggests no evident association with no pattern of points amenable to randomisation disruption, resulting in the non-significant **p** value of 0.39.

During the randomisations a check was also made on the proportion of instances $\mathbf{p_R}$ that the absolute value of the original Pearson correlation coefficient was exceeded. For Figure 3-6a-d the respective values were $\mathbf{p_R}$ = <0.01, 0.03, 0.75 and 0.88. For Figure 3-6d the difference between $\mathbf{p_R}$ and **p** reflects that the correlation coefficient is specific to detecting linear associations as opposed to general bivariate associations which are detected by the near-zero **p** values.

Figure 3-6 also serves to emphasize that **p** is essentially an exploratory tool and should not be used in a mechanical way to select variables for input to explanatory models. Both Figure 3-6a and Figure 3-6d have small **p** values but if this were real data then Figure 3-6a would be the more interesting because the signal and noise are clearly different in this case and the latter might be amenable to removal after further work.

Similarly, we applied **p** to one of the autocorrelation data-generation models used by Sharma (2000) to test a scheme for optimal predictor variable selection. Specifically, we applied the autocorrelation expression: $X_t = 0.9X_{t-1} + 0.866e_t$ and then tested X_t for association with X_{t-1}, X_{t-2}, .. X_{t-15}. As it happened all **p** values were less than 0.05 except for X_{t-14} and X_{t-15}, again emphasising the exploratory nature of the bivariate **p** value rather than explicit selection of best input variables for prediction purposes.

Data application

The variable *Y* of interest here is the percentage of days per month on the west coast of the South Island of New Zealand when the upper air wind direction is estimated as being between 216°and 365°, coupled with wind speed exceeding 5 ms^{-1}. Such wind conditions at this location are often associated with rain so the monthly frequency of such days is likely to be related to regional monthly precipitation.

The *X* variables in this case are the monthly mean 700 hPa geopotential heights (metres) as referenced to a southern hemisphere 29 × 7 reanalysis data grid of point locations, extending from the Indian Ocean on the west to the Atlantic Ocean to the east (Figure 3-7). The data was not adjusted for seasonal effects, which will contribute a component of noise to any associations.

All grid point locations which produced **p** values less than 0.01 are shown in Figure 3-8. Interestingly, most of the plotted grid points are located along the two lines of latitude 25°S and 55°S, respectively to the north and south of New Zealand.

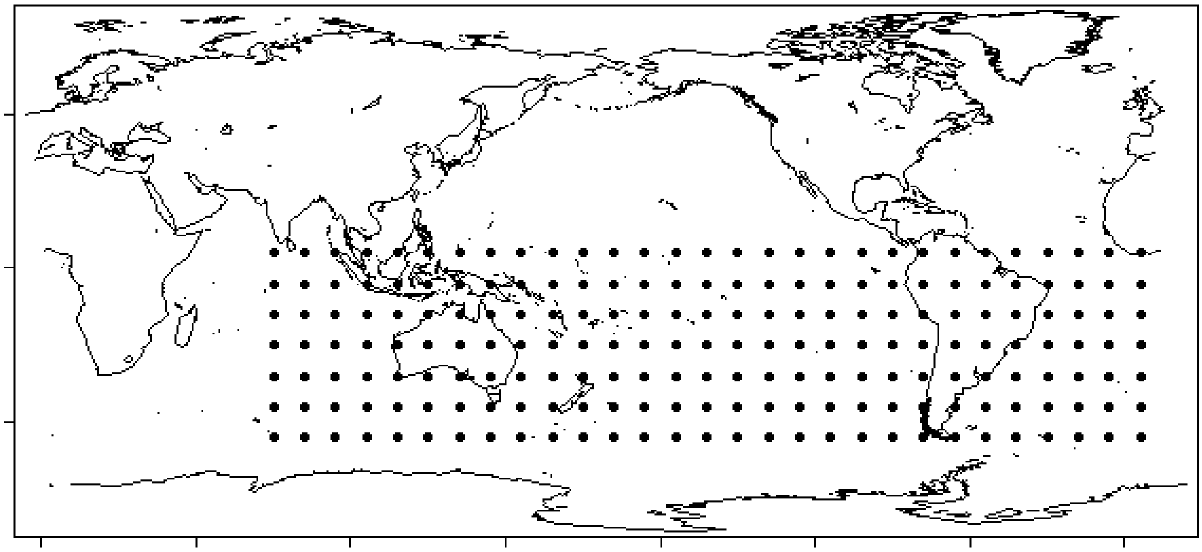

Figure 3-7 Grid point locations for 700 hPa geopotential heights (see text for further description).

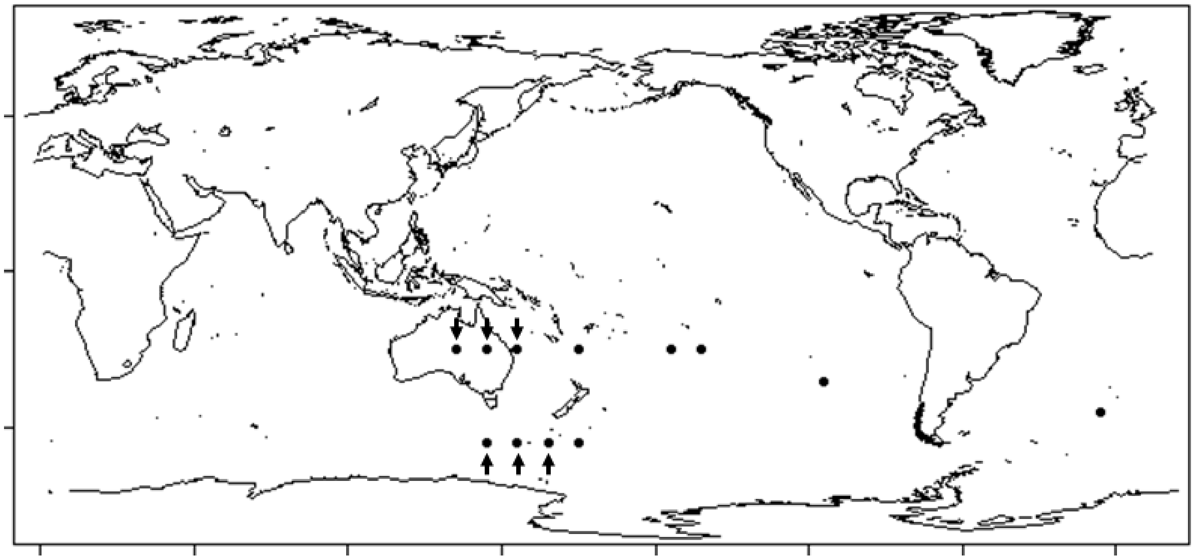

Figure 3-8 Grid point locations for $p < 0.01$. Arrows denote scatter plots shown in Figure 3-9 and Figure 3-10.

The down-arrow symbols of Figure 3-8 show the locations of the three latitude 25°S statistically-significant grid point locations displayed as the three scatter plots of Figure 3-9a-c (from west to east respectively). The association in this case is evidently a tendency for *Y* to increase with *X*. For purposes of comparison Figure

3-9 also includes scatter plots and **p** values for the adjacent three grid points immediately to the north (Figure 3-9d-f). The extent of the association is evidently sensitive to latitude because these grid points do not have significant **p** values and any linear correlation is minimal.

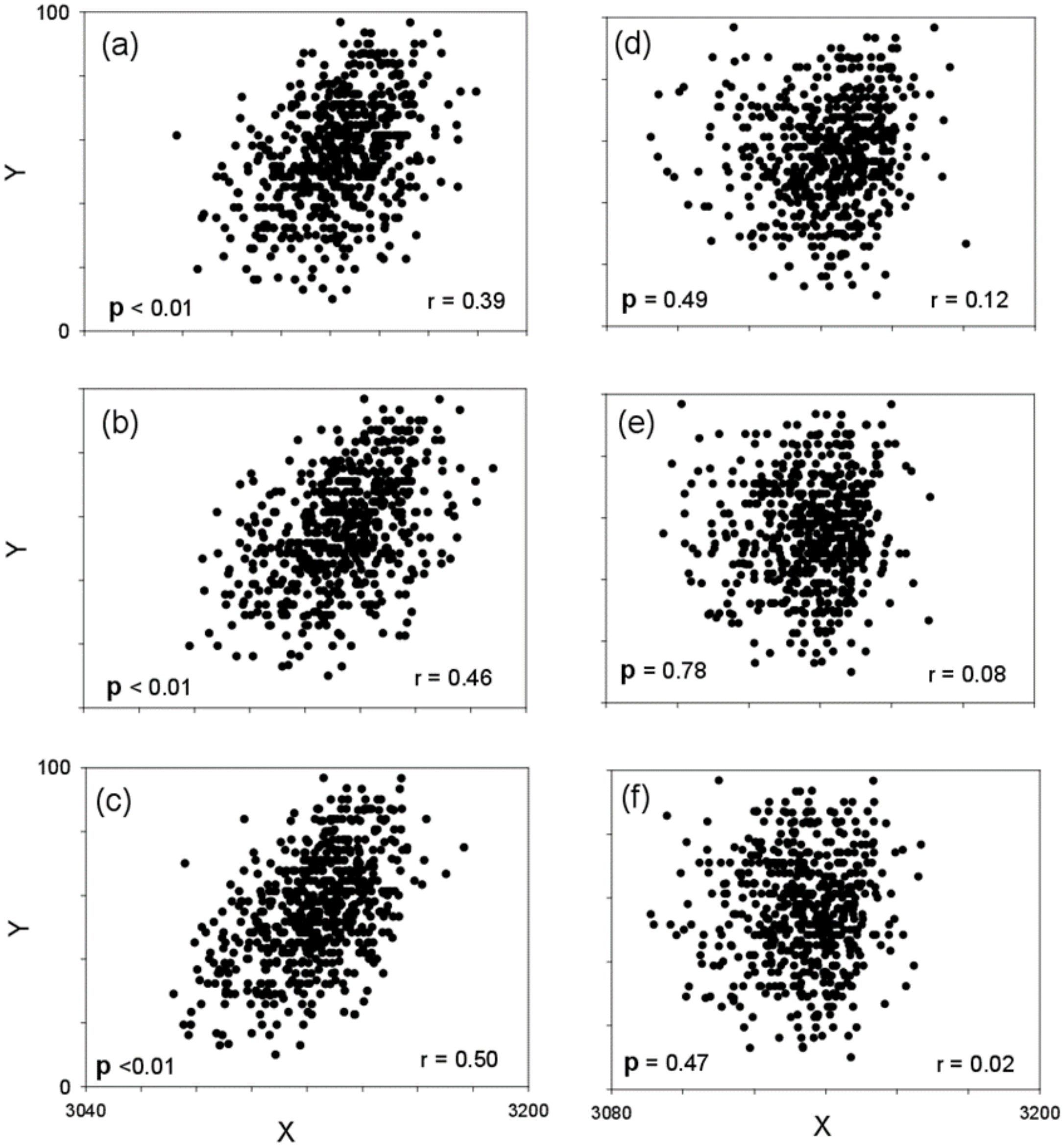

Figure 3-9 Scatter plots for Figure 3-8 down-arrow grid points (a-c) and for the adjacent non-significant grid points immediately to the north (d-f).

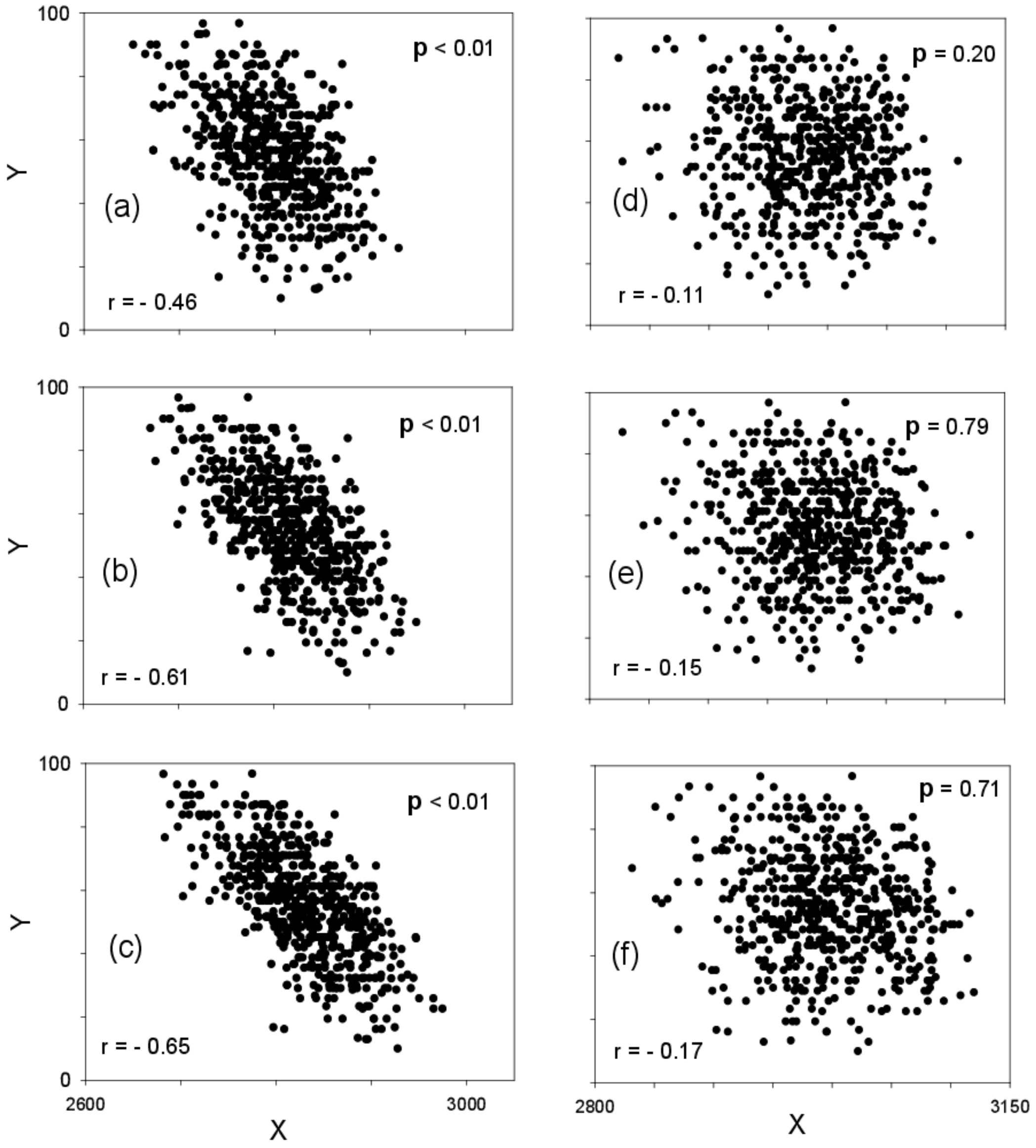

Figure 3-10 Scatter plots for Figure 3-8 up-arrow grid points (a-c) and for the adjacent non-significant grid points immediately to the north (d-f).

Figure 3-10 shows a similar situation for the three indicated significant grid points along latitude 55°S (up arrow symbols on Figure 3-8). In this case, however, the three significant grid points all display negative associations (Figure 3-10a-c). There is again a sensitivity of the association to latitude, with the three adjacent grid points to the north having non-significant **p** values and minimal linear correlation (Figure 3-10d-f).

The detected associations of *X* and *Y* here are to be expected from the definition of *Y*, which incorporates a westerly wind component with respect to the west coast of New Zealand's South Island. Low atmospheric pressure to the north (lower 700 hPa geopotential elevations) would tend favour easterlies rather than westerlies over the South Island. Similarly, South Island easterlies would be favoured by high pressures further to the south, consistent with the negative association. What is perhaps more surprising is the strong latitudinal effect in constraining associations, suggesting that regions of influence in this case may be well defined.

In fact, the associations and non-associations detected here could have been equally well identified just by a table of linear correlation coefficients. That is, the detected bivariate associations were all linear to a first approximation. However, the use of **p** here at least gives some degree of certainty that possible nonlinear associations at other grid locations were not overlooked.

Regions of influence aside, it is inevitable that the use of a large number of grid points will result in spurious detections of general bivariate association in some form. For example, Figure 3-11 shows the scatterplot from the single point identified in the South Atlantic (Figure 3-8). Evidently the **p** significance in this case arises from an elliptical pattern in the scatter, somewhat similar in appearance to Figure 3-6b.

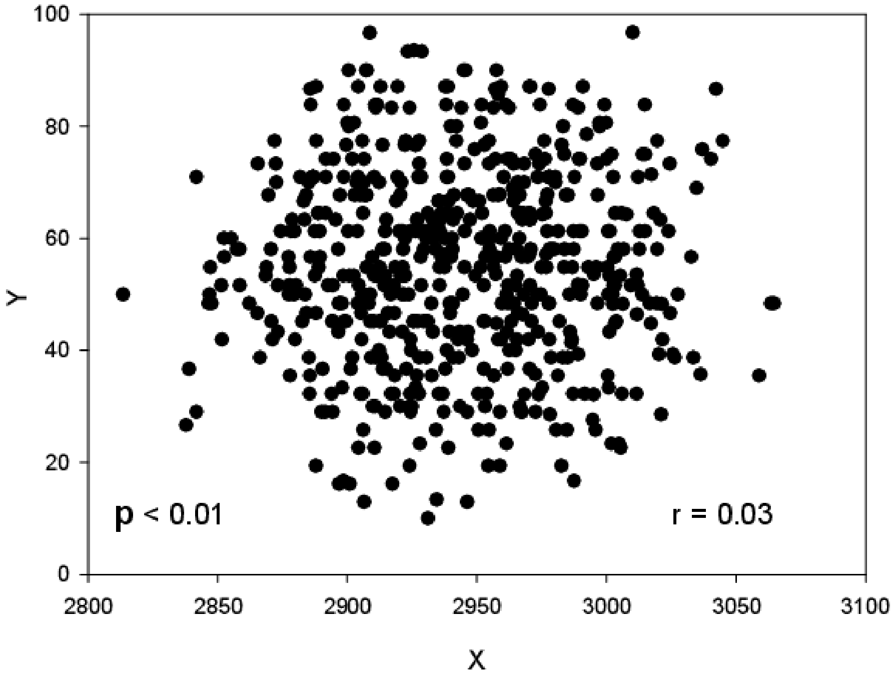

Figure 3-11 Scatter plot for the significant South Atlantic grid point of Figure 3-8.

5. Conclusion

A test of general bivariate association is proposed, based on the degree to which data randomisation induces an increase in scatterplot mean nearest neighbour distance. The simplicity of the test may also result in its finding use in other situations where there is interest in general bivariate associations and test power is not the first priority. However, further work is required for more detailed evaluation of the test properties over the many possibilities of bivariate associations and different sizes of data sets.

The approach is in principle amenable to direct extension to testing *N*-variate general associations with reference to nearest neighbours in *N*-space. This is left an open possibility for now but our feeling is that the test is best suited to preliminary data exploration to exclude non-significant associations prior to the application of more sophisticated methods leading to explanatory model construction.

References

Blum, J.R., Kieffer, J., Rosenblatt, M., 1961. Distribution free tests of independence based on the sample distribution function. Ann. Math. Stat. 32, 485-498.

Delicado, P., Smrekar, M., 2009. Measuring non-linear dependence for two random variables distributed along a curve. Stat. Comput., 19, 255–269.

Hejazi, M.I., Cai, X., 2009. Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. Adv. Water Resour. 32, 582–593.

Heller, R., Heller, Y., Gorfine, M., 2013. A consistent multivariate test of association based on ranks of distances. Biometrika, 100, 503–510.

Huang, Q., Zhu, Y., 2016. Model-free sure screening via maximum correlation. Journal of Multivariate Analysis 148, 89–106.

Kallenberg, W.C.M., Ledwina, T., 1999. Data-driven rank tests for independence. Journal of the American Statistical Association, 94, 285–301.

Kankainen, A., Ushakov, N.G., 1998. A consistent modification of a test for independence based on the empirical characteristic function. Journal of Mathematical Sciences, 89, 1486–1494.

Karvanen, J., 2005. A resampling test for the total independence of stationary time series: Application to the performance evaluation of ICA algorithms. Neural Processing Letters, 22, 311–324.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Physical Review E, 69, 066138.

Linfoot, E.H. 1958. An Informational measure of correlation. Information and control, 1, 85-89.

May, R.J., Holger, R.M., Dandy, G.C., Fernando, T.M.K.G., 2008. Non-linear variable selection for artificial neural networks using partial mutual information. Environ. Model. Softw. 23, 1312–1326.

Murrell, B., Murrell, D., Murrell, H., 2016. Discovering general multidimensional associations. PLoS ONE 11(3): e0151551. doi:10.1371/journal.pone.0151551.

Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap rank-ordered conditional mutual information (broCMI): A nonlinear input variable selection method for water resources modeling. Water Resour. Res. 52, 2299–2326.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C., 2011. Science 334, 1518, DOI: 10.1126/science.1205438.

Schweizer, B., Wolff, E.F., 1981. On nonparametric measures of dependence for random variables. Ann. Statist., 9, 879–885.

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improve water supply management: Part 1 — A strategy for system predictor identification. J. Hydrol. , 239, 232–239.

Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resour. Res. 50, 650–660.

Sugiyama , M., 2011. Least-squares independence test. IEICE Transactions on Information and Systems, E94-D, 1333–1336.

Szekely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances, Ann. Statist., 35, 2769–2794.

Tran, H.D., Muttil, N., Perera, B.J.C., 2015. Selection of significant input variables for time series forecasting. Environ. Model. Softw. 64, 156–163.

Wang, Y., Li, Y., Cao, H., Xiong, M., Shugart, Y.Y., Jin, L., 2015. Efficient test for nonlinear dependence of two continuous variables, BMC Bioinformatics, 16:260. doi: 10.1186/s12859-015-0697-7.

Conclusions

The first part of this chapter provided a general method to test for significance of the size of a data-sparse region in a scatterplot. The second part presents a method for testing for general non-random associations between two variables. Both approaches can be utilised as filter methods to screen for informative bivariate relationships prior to application of a ML method. Both methods may be amenable to extension to the multivariate case, though this is left for further work.

Chapter 4 Incorporating large scale climatic-oceanic data into hydro-climatological analysis and prediction

4.1 Introduction

This chapter examines an issue of incorporating large scale multivariate climatological data into prediction of local hydrological processes. This problem is viewed as a machine learning problem of regression where the number of predictors is larger than the number of observations.

The first part of this chapter illustrates the usefulness of Lasso regularized regression as both an exploratory and predictive tool in seasonal streamflow forecasting based on high-dimensional raw climatological data. Its utility is demonstrated with a case study of predicting lake inflows in the Waitaki river catchment, New Zealand.

The second part of this chapter provides a detailed case study of influence of a particular climatic index – Interdecadal Pacific oscillation on winter inflows in the Waitaki catchment.

The first part of this chapter has been presented as a peer reviewed paper to the workshop on "Data science in food, energy and water" at the 22nd ACM SIGKDD conference on knowledge discovery and data mining. The second part of this chapter was published in the *New Zealand Journal of Hydrology* [1].

Reference

[1] Vetrova, V., and Bardsley, E. 2015. An association between Waitaki River winter headwater flows and the Interdecadal Pacific Oscillation (IPO). Journal of Hydrology, NZ, 54, 2, 103–108

4.2 Finding spatial climate precursors of hydro-lake inflows: Waitaki catchment, New Zealand

Abstract

Predictability of river discharge is a topic of increasing practical importance. In this paper, we address a problem of seasonal forecasting of hydro-lake river inflows using large scale oceanic-climatic predictors. This forecasting problem is high-dimensional with small sample sizes. We highlight the usefulness of the regularized linear regression method Lasso as both a predictive and exploratory tool. This approach allows to explore relationships between large scale climatological processes and local hydrology. A modified cross-validation procedure is proposed to maximize the available information in a forecasting model. We demonstrate the experimental results with a case study of seasonal lake inflow forecasting in the Waitaki catchment in New Zealand.

Keywords

high-dimensionality, regularized regression, cross-validation

Introduction

Seasonal and monthly forecasts of hydro-lake inflows are of significant value for management of power consumption and provision. However seasonal and monthly forecasting poses a challenge, especially in catchments with limited groundwater reservoir influence or relatively small basins which prevent forecasting by delayed flow. Additional climatological information can be utilized in forecasting models, for example when there is a lack of serial correlation between consecutive months and seasons [1-6]. Here we incorporate large-scale climatic information into predictive models to improve forecasting and also provide new insights into physical mechanisms regulating seasonal and monthly catchment precipitation.

Large scale climate data is usually represented as gridded fields such as sea surface temperatures or geopotential heights [7-8]. Utilizing high-dimensional climate data in forecasting models requires some type of predictor selection procedure, because sample sizes are generally small. Typically, climate data has been utilized in streamflow predictive models by way of pre-defined climatological indices, such as the North-Atlantic Oscillation or Nino 3.4 indices [2,4].Correlation analysis has also been used to identify potential regions having the highest correlation with the river flows of a given basin [2,3]. Alternatively, PCA-based methods have also been used to reduce dimensions of climatological predictors [1,3,5].

There has been considerable interest in the KDD[1] community in spatial-temporal mining of climate data [9-11]. Clustering methods for identifying climate regions were developed in [12].

Correlation analysis can help identify linear relationships between individual climate predictors and river flows [2,3]. However correlation analysis cannot capture nonlinear relationships, for example, characterized by the strength of differences between two climatological predictors. Such relationships can arise, for example if the differences between two pressure centers contribute significantly to streamflow rather than values of pressure in two spatial locations individually [13]. These relationships can sometimes be identified when untransformed climatological data is used in predictive models. Untransformed data, however, brings back the issue of the high-dimensionality of predictor space versus the available sample size.

One way in which the issues associated with untransformed data could be resolved is by simultaneously learning the predictor and fitting the model: regularized regression methods are one approach for doing this. Here we will use a Lasso-based regression model for this purpose. In our literature search we could find only one previous example of data-driven seasonal river flow forecasting utilizing regularized regression. However, that investigation used climatic indices rather than the more extensive raw climatological data [6]. Regularized regression methods were applied in [14] for precipitation analysis. However, in that paper only concurrent relationships were examined, whereas here we use time-lagged ocean-climate predictors to predict future streamflow.

[1] KDD – Knowledge discovery and data mining

This section aims to highlight some work in progress to illustrate the usefulness of Lasso regularized regression as both an exploratory and predictive tool in seasonal streamflow forecasting based on high-dimensional raw climatological data. Its utility is demonstrated with a case study of predicting lake inflows in the Waitaki river catchment, New Zealand.

Our study focuses on two main aspects of river flow prediction. Firstly, we propose a modified cross-validation procedure in order to maximise available information in the forecasting model. Secondly, we discuss analysis of the selected spatial predictor variables.

Background and data

The Waitaki catchment in New Zealand was chosen as a case study for this paper. The Waitaki River contains a cascade of eight hydro power stations and produces about 40% of New Zealand electricity (Figure 4-1). River inflows to the Waitaki hydro lakes are subject to strong seasonal variation, with maximum discharge through September-February and minimum flows in March-August. This is in contrast with demand for electricity, which is highest in June-August (the NZ winter). This mismatch makes the problem of seasonal forecasting of hydro-lake inflows significant for hydro-scheme management. Similar problems exist in other countries with significant hydro-electric power component.

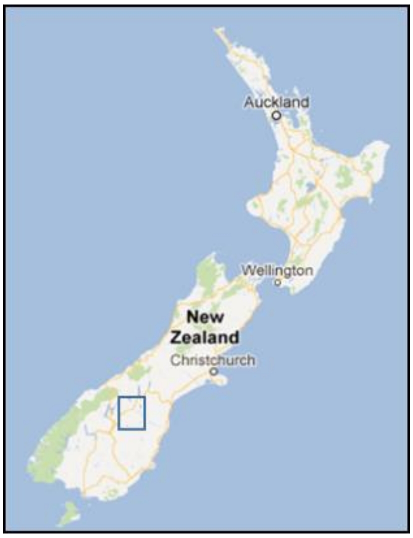

Figure 4-1 Location of the Upper Waitaki catchment (rectangle), New Zealand

Previous studies have analysed global atmospheric circulation influences on hydrology of the Waitaki catchment [15,16]. There has been previous study on seasonal prediction of streamflow in the catchment where PCA was used to reduce the dimensionality of climate predictors [5].

For this study, seasonal lake inflow volumes were provided by Meridian Energy Ltd as derived from water balancing using lake level differencing coupled with recorded outflows.

Sea surface temperature data and 700hpa geopotential heights were used in this study as potential predictors [7,8]. The spatial extent was selected as (54°N,56°E)-(86°S,172°W) in order to allow for potential teleconnection influence of the North Pacific and Indian Oceans.

Seasonal predictive models were analysed for the contrasting seasons of autumn (March, April and May) and spring (September, October, November). 700hpa geopotential heights and sea surface temperatures of previous three months were used as predictors.

Seasonal streamflow prediction using LASSO

There are 1400 potential oceanic-climate predictors for each type of predictor variable and only 67 observations streamflow values, so regularization in the predictive regression models is essential.

The regularized linear regression LASSO was chosen as a predictive model in this study:

$$\min_{(\beta_0,\beta)\in R^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda||\beta||_1$$

where N is the number of observations, x_i is a $P\times 1$ vector of predictors, y is the $N\times 1$ vector of dependent variables, β_0 is the intercept term, β is the $P\times 1$ vector of regression coefficients, $\lambda\geq 0$ is a regularization parameter. The parameter λ plays a trade-off role between model complexity and calibration accuracy. Increasing λ forces some of the regression coefficients to zero and thus eliminates the corresponding predictors from the model.

Cross-validation framework

In order to determine the optimal value of the regularization parameter λ, a cross-validation procedure is normally used [17]. There has been previous research on modification of cross-validation methods in time series and in regularized regression models [18, 19]. However, cross-validation in the context of time series data in regularized regression poses additional challenges.

In the case of time series prediction, a typical k-fold cross-validation procedure might cause validation subsets to precede training subsets in time. However, in a practical setting of forecasting inflows, a predictive model is going to be used only to predict future values of time series. Also, evaluation of a model on a subset of past values might be affected by a change in large scale climatic phenomena[2]. This happened for example with circulation changes at the end of 1970s related to Pacific decadal oscillation. We propose here a modified cross-validation scheme (Algorithm 1) where cross-validation folds consists of one element immediately after the training fold.

The proposed cross-validation procedure in Algorithm 1 (below) starts with reserving a first block of data as a training set. A LASSO model then is fitted on the training set. Some coefficients β_j will be forced to zero in the fitted model, therefore corresponding predictors will be eliminated from the model. In order to avoid bias towards small values of coefficients β, a regression model is refitted on renewed set of predictors with λ set to 0. Finally, a prediction is made on the data point which follows the training set in order to select λ.

At a next step of a cross-validation, the training set size is increased by one data point, namely the next point in the time series. Therefore the training set is constantly growing and the validation fold always consist of the data point following the training data.

The process described above is repeated for incremented values of the regularization parameter λ.

[2] In the area of data stream mining such changes are called “concept drift”.

Our proposed cross-validation resembles a likely practical implementation of the forecasting methodology when all information available to date is utilized in constructing the forecasting model.

The R package *glmnet* was used to implement the models [20]. For the purpose of this study only Lasso models were analysed.

Algorithm 1

Input

$Z_0 \in R^{N\times(P+1)}$ – Data matrix with rows (x_i^T, y_i),

K – number of the examples in the initial training set, i.e. the first K examples

SetOfLambda$\in R^L$ – an array of regularization parameter values

Output

$ValSet \in R^{L\times(N-K)}$ – predicted values of the validation example, one row per each λ∈ *SetOfLambda*

```
for each  λi  in SetOfLambda do
    for k from K to N-1 do
        Zt <- Z[1:k,:] # training set, consists of first k observations
        Model <- Fit the Lasso model on a training set $Z_t$, with regularization parameter λi
        $\widetilde{Z}_t$ <- Model with predictors with zero β coefficient eliminated from Zt
        λt <- 0
        $Model_t$<- Fit Lasso model on a training set $\widetilde{Z}_t$ with regularization parameter λt
        ValSet[i, N-k] <-Use $Model_t$ to predict $Y_{k+1}$ with observation $X_{k+1}$
    end for
    #There will be N-K predicted values for each λi
    $F_i$<-Calculated goodness-of-fit per each row of ValSet[:,N-k]
end for
    $\lambda_{min}$ <- select λ with minimum goodness-of-fit
    $F_{min}$=min($F_i$)
Return $\lambda_{min}$, ValSet, $F_{min}$
```

Predictor selection

N-K regression models are generated for a single value of λ as result of algorithm 1. Each of the generated models might be formed by different set of spatial predictors due to the property of LASSO regression which eliminates correlated predictors, leaving only one such predictor in the model [19]. Therefore for different values of λ even if the same number of predictors appears in its model the predictors might be different.

It is of interest to analyse stability of selected spatial predictors in order to compare it with the knowledge of climatological processes in the study area. For this purpose, all predictors with non-zero β coefficients were extracted from each LASSO regression model fit in Algorithm 1. The location of extracted predictors are then plotted on a map. It is interesting to note that all selected predictors were having constant signs of regression coefficients across all the models.

Results

Separate regression models were applied, respectively using 700hpa geopotential heights and sea surface temperatures as predictors.

In the case of 700hpa geopotential heights, maps of selected predictors show different centers of influence in different seasons. It appears that the locality of the Siberian high could have teleconnections linking Northern Hemisphere winter pressures with autumn inflows to the Waitaki lakes (Figure 4-2). This linkage could be modulated by a connection of the Siberian High and the East-Asian monsoon. Interestingly, none of geo-potential heights were selected in the Equatorial Pacific region. The atmospheric pressure locations north and south of New Zealand form a pressure gradient because corresponding regression coefficients are of opposite sign. This pressure gradient represents the westerly wind strength bringing ocean-derived moisture to the Waitaki catchment.

Clearly, if summer pressures are used as predictors it would be unexpected to see the same region of the winter Siberian High being selected from the models. Indeed, Figure 4-3 shows that spring inflows appear to have winter local and equatorial pressure influences. Somewhat unexpectedly, North America atmospheric pressures were consistently selected for both seasons.

The distributions of the selected sea surface temperatures appear to indicate a degree of spatial clustering, suggesting some ocean localities have more influence than others (Figure 4-4).

In order to analyse fit of the regression models, scatterplots of observed versus predicted values of streamflow were constructed for varying values of λ. There is a suggestion in fact that predicted low flows eventuate, but predicted high flows could be associated with flows over the whole discharge range (Figure 4-5 and Figure 4-6).

However, there is need to apply a suitable measure for goodness-of-fit. If the standard MSE measure is utilized, it would not result in such models being preferably selected over others.

As an alternative predictability measure, a suitable approach is a test of the significance of the size of the near-empty space in the lower right of Figures 4-5 and 4-6 [21]. The test is based on randomization procedure with the null hypothesis being that the near-empty region could have arisen by random chance. Applying the test to the respective lower right regions gave p-values of 0.002 and 0.06 for Figure 4-5 and Figure 4-6 respectively. Interestingly, the suggestion is that the lower flows have a higher degree of predictability from the models with a smaller number of parameters (Figure 4-5). The explanation as to why the lower flows should have higher predictability is left to further work.

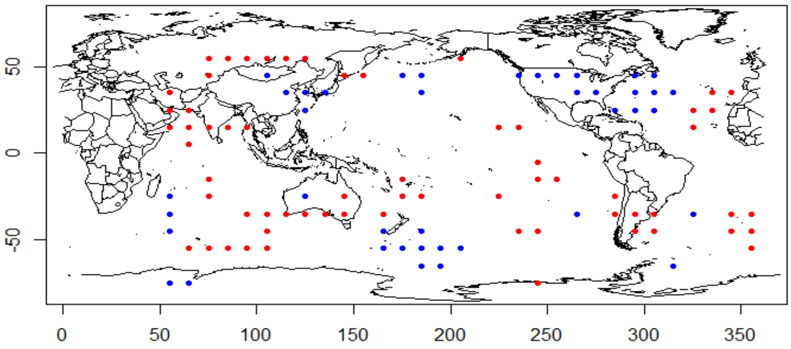

Figure 4-2 The selected 700hpa geopotential height locations for autumn forecasting models (March, April, May). Blue and red points denote negative and positive regression coefficients respectively.

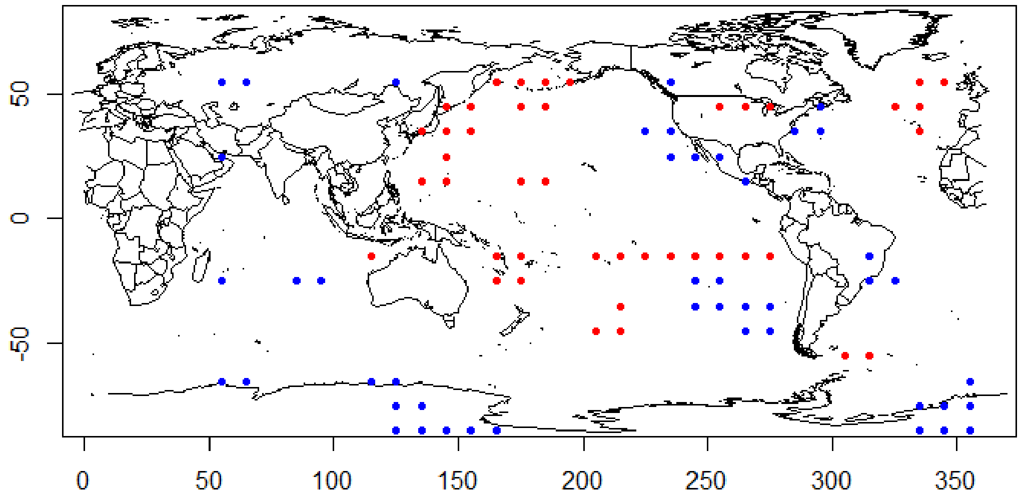

Figure 4-3 The selected 700hpa geopotential height locations for spring forecasting models (September, October, November). Blue and red points denote negative and positive regression coefficients respectively.

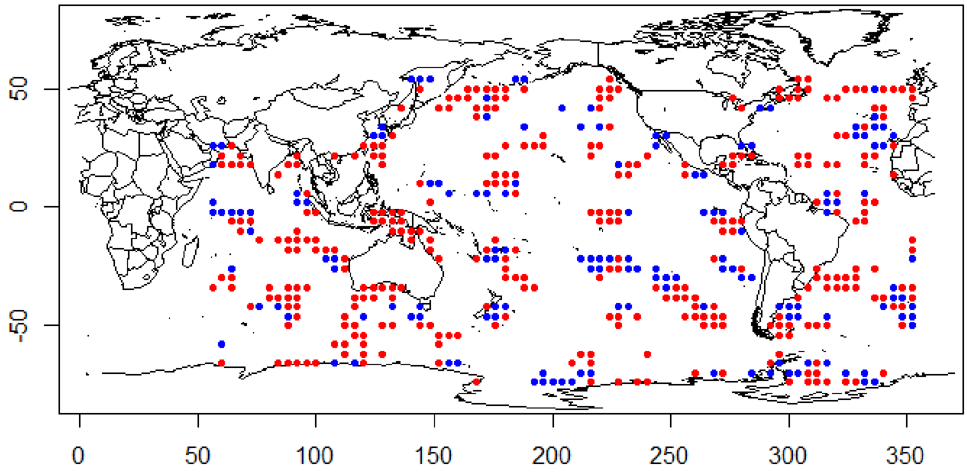

Figure 4-4 The selected February sea surface temperatures locations for autumn models (March, April, May). Blue and red points denote negative and positive regression coefficients respectively.

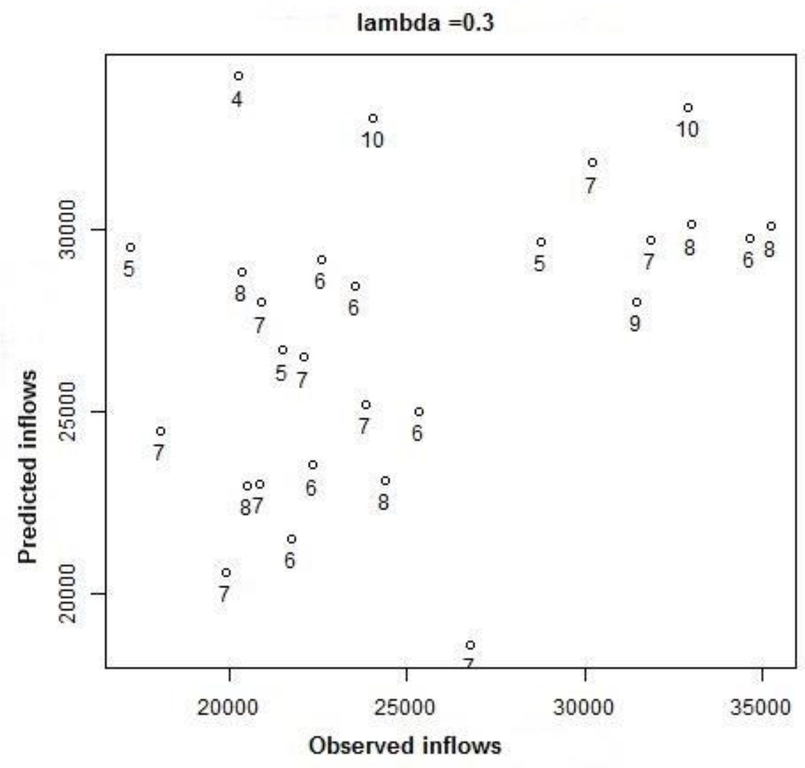

Figure 4-5 Predicted vs Observed inflow values for spring models (September, October, November), 700hpa geopotential heights are used as predictors. Multiplying the axis values by 86400 gives total seasonal inflow volumes in m^3. Numbers next to data points represent number of selected predictors in the particular model.

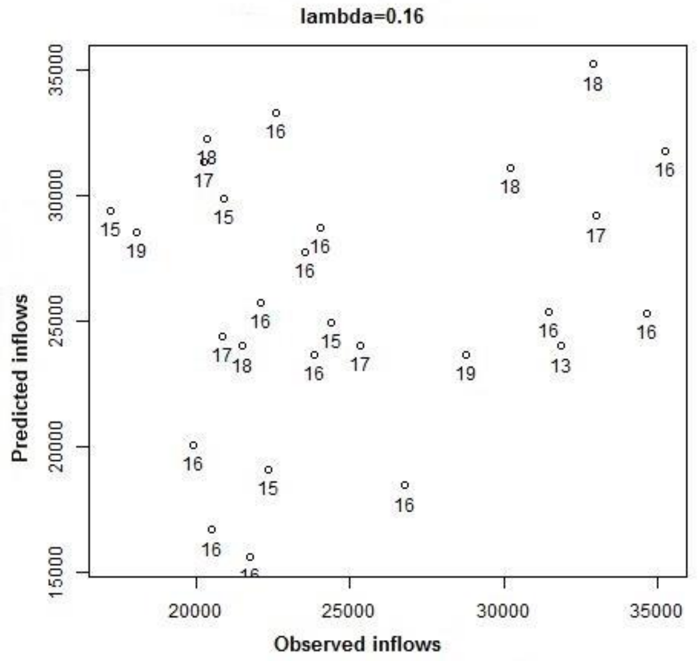

Figure 4-6 Predicted vs Observed inflow values for spring models (September, October, November), 700hpa geopotential heights are used as predictors. Multiplying the axis values by 86400 gives total seasonal inflow volumes in m^3. Numbers next to data points represent number of selected predictors in the particular model.

Conclusions and future work

In this paper, analysis was carried out of predictive models for seasonal river discharge in the headwaters of the Waitaki River, New Zealand. The models utilize regularized linear regression via the LASSO. Predictability is not consistent, however, perhaps due to the operation of threshold effects in the predictor variables. For example, predicted high inflows may be associated with actual inflows which could be high or low, while predicted low inflows translate to an absence of high inflows. Such zone predictability requires quantification with an alternative goodness of fit measure. In future work, we will adapt the approach used in [14] for the predictive task we have considered here. This looks to be a challenging problem: since we work with time series data and predictors with different time lags, it seems that some non-uniform time-dependent weighting may be required here rather than the hard thresholding used in [14]. There is also a need for an evaluation on a per grid point basis of the nature of the associations between inflows and the spatial variable concerned. However, this is beyond the scope of the present thesis study.

Since it is known that climatic phenomena change over time, it would also be of interest to explore changes in relationship between climate variables and local hydroclimatology using a sliding window approach.

References

[1] Talento, S. and Terra, R. 2013. Basis for a streamflow forecasting system to Rincón del Bonete and Salto Grande (Uruguay). Theor Appl Climatol, 114, 73–93.

[2] Hidalgo-Muñoz, J. M., Gámiz-Fortis, S. R., Castro-Díez, Y., Argüeso, D. and Esteban-Parra, M. J. 2015. Long-range seasonal streamflow forecasting over the Iberian Peninsula using large-scale atmospheric and oceanic information. Water Resources Research, 51,5, 3543-3567. doi:10.1002/2014wr016826

[3] Córdoba-Machado, S., Palomino-Lemus, R., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J. 2016. Seasonal streamflow prediction in Colombia using atmospheric and oceanic patterns. Journal of Hydrology, 538, 1-12. doi:10.1016/j.jhydrol.2016.04.003

[4] Wilby, R. L., Wedgbrow, C. S., and Fox, H. R. 2004. Seasonal predictability of the summer hydrometeorology of the River Thames, UK. Journal of Hydrology, 295,1-4, 1-16. doi:10.1016/j.jhydrol.2004.02.015

[5] Purdie, J. M., and Bardsley, W. E. 2009. Seasonal prediction of lake inflows and rainfall in a hydro-electricity catchment, Waitaki river, New Zealand. International Journal of Climatology, 30,3, 372-389. doi:10.1002/joc.1897

[6] Lima, C. H., and Lall, U. 2010. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. Journal of Hydrology, 380,3-4,, 438-449. doi:10.1016/j.jhydrol.2009.11.016

[7] Kalnay et al., 1996. The NCEP/NCAR 40-year reanalysis project, Bull. Amer. Meteor. Soc., 77, 437-470

[8] Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff and H.-M. Zhang. 2015. Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: upgrades and intercomparisons. Journal of Climate. 28,3, 911–930

[9] Faghmous, J.H. and Kumar, V., Data Mining and Knowledge Discovery for Big Data - Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities. Springer Berlin Heidelberg. Vol 1 pages 83-116. (2014).

[10] Lakshmanan, V. (2015). Machine learning and data mining approaches to climate science.

[11] Steinhaeuser, K., Chawla, N.V, and Ganguly, A.R. 2010. An exploration of climate data using complex networks. ACM SIGKDD Explor, 12,25–32

[12] Steinbach, M., Tan, P., Kumar, V., Klooster, S., and Potter, C. 2003. Discovery of climate indices using clustering. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03

[13] Kawale, J., et al., 2013. A graph-based approach to find teleconnections in climate data. Stat. Anal. Data Mining, 6, 158–179

[14] Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., and Ganguly, A. 2012. Sparse Group Lasso: Consistency and Climate Applications. Proceedings of the 2012 SIAM International Conference on Data Mining, 47-58.

[15] Kingston, D. G., Webster, C. S., and Sirguey, P. 2015. Atmospheric circulation drivers of lake inflow for the Waitaki River, New Zealand. International Journal of Climatology Int. J. Climatol., 36,3, 1102-1113. doi:10.1002/joc.4405

[16] Vetrova, V.,and Bardsley, E. 2015. An association between Waitaki River winter headwater flows and the Interdecadal Pacific Oscillation (IPO). Journal of Hydrology, NZ, 54,2, 103–108

[17] Hastie, T., Tibshirani, R., and Friedman, J. H. 2001. The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

[18] Bergmeir, C., and Benítez, J. M. 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

[19] Yu, Y., and Feng, Y. 2014. Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models. Journal of Computational and Graphical Statistics, 23,4, 1009-1027

[20] Friedman, J., Hastie, T., Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33,1, 1-22. URL http://www.jstatsoft.org/v33/i01/

[21] Vetrova, V.V., and Bardsley, W.E. 2012. Technical note: A significance test for data-sparse zones in scatter plots. Hydrology and Earth System Sciences, 16 (4), 1255-1257

4.3 An association between Waitaki River winter headwater flows and the Interdecadal Pacific Oscillation (IPO)

Introduction

Variations and trends in New Zealand temperatures and rainfall have been related to circulation changes in the southwest Pacific, with significant shifts in 1950 and 1975 (Salinger and Mullan, 1999). Salinger *et al*. (2001) showed that climatic shifts over the southwest Pacific region within the periods 1922–1944, 1946–1977 and 1978-1998 are linked to the different phases of the Interdecadal Pacific Oscillation (IPO). Particularly, they suggested that the IPO in the positive phase enhances teleconnections between the El-Nino Southern Oscillation (ENSO) and climate variability over New Zealand. Similarly, Ummenhofer *et al*. (2009) attributed changes in New Zealand summer precipitation during the period 1976-2006 to changes in both ENSO and the Southern Annular Mode (SAM).

We focus here on the specific instance of an apparent similarity between the IPO and winter discharge in the Waitaki headwaters, defined here as the combined inflow volumes into the hydro storage lakes Tekapo and Pukaki. The question considered is why there should be inflow correlation with the IPO in winter but not in the other seasons. Previous work has not had an emphasis on seasonal differences but an IPO influence in the Southern Alps was reported as the change to a positive IPO phase around 1978 resulting in most Southern Alps headwater rivers shifting toward higher discharge magnitudes for both floods and low flows (McKerchar and Henderson, 2003).

Data

The IPO index used here is related to the Pacific Decadal Oscillation (PDO) index. The PDO has been defined as the leading principal component of North Pacific monthly sea surface temperature variability (Mantua et al., 1997). During the positive phase of the PDO the sea surface temperatures (SST) anomalies over the North Pacific are negative while the SST anomalies over tropical Pacific are positive and vice versa. The IPO index describes the similar quasi-symmetrical oscillation over the whole Pacific basin and was shown to be essentially equivalent to the PDO in the North Pacific (Folland et al., 2002).

In the present study, the Waitaki lakes (Tekapo and Pukaki) inflow records were used subsequent to 1940 because earlier records contain large amounts of missing data (Purdie, 2010). The Hermitage station at Mount Cook is used as the source of surface observation of daily values of temperature and precipitation, taken as indicative of the headwater region. Daily temperature mid-ranges are used (mean of the daily maximum and minimum temperature) and we use “rain-day” to mean precipitation (rain or snow) exceeding 0 mm for the day. This will include days of trace precipitation but such data makes up less than 10% of rain-day records. The seasonal lake inflow volumes were provided by Meridian Energy Ltd and were derived from lake level differencing.

Seasonal lake inflows and the IPO

Figure 4-7 shows an approximate similarity in the time variation of the annual smoothed IPO index and Waitaki lake inflows in winter, but not evidently apparent in the other seasons (Figure 4-8). There is always a degree of uncertainty in subjective comparisons of temporal variations derived from heavy smoothing but the winter post-1978 increase in inflows is consistent with the general increase in discharges in the Southern Alps reported by McKerchar and Henderson (2003).

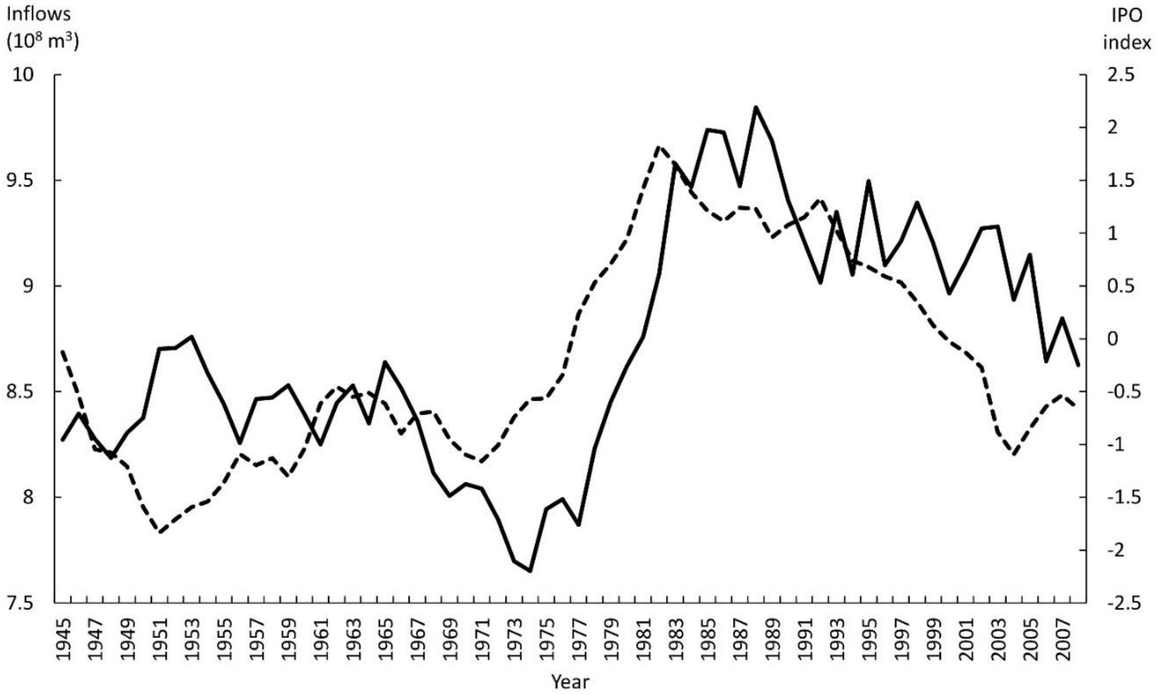

Figure 4-7 Smoothed annual Interdecadal Pacific Oscillation index values (dashed) and smoothed combined winter inflows for lakes Tekapo and Pukaki (solid line). Both plots are 11-year running means.

Proposed climatic mechanism

Taking the winter correlation of Figure 4-7 as real, it remains to establish a causal mechanism which also needs to explain the lack of association in the other seasons.

It is suggested that the different seasonal responses of lake inflows to the IPO derive via the intermediary of different seasonal temperature responses to changes in the phase of the IPO. Taking the Hermitage recording site as broadly representative of local temperature variations, smoothed Hermitage winter temperatures follow roughly similar smoothed trends to the IPO over the period considered (Figure 4-9). However, this correspondence of pattern is not evident in the other seasons (Figure 4-10). In a similar way, there is apparent temporal correlation with winter Hermitage temperatures and winter lake inflows (Figure 4-11), but not in the other seasons (Figure 4-12).

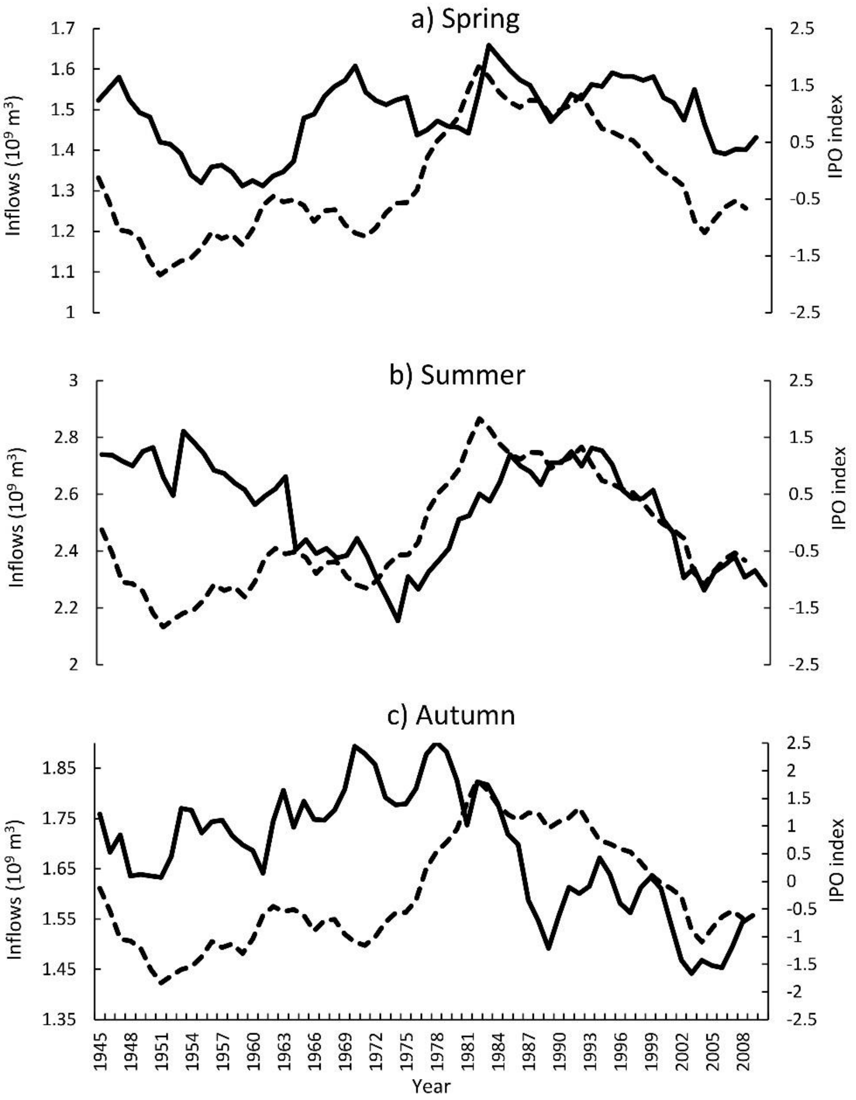

Figure 4-8 Smoothed time series as in the winter plot of Figure 4-7, for the other three seasons: a), b), and c) are spring, summer, and autumn, respectively. The smoothed IPO values (dashed) are replotted for each season.

In seeking a seasonally-variable link between temperature and precipitation leading to inflows, it is interesting to note a correlation between rain-day air temperature and precipitation at the Hermitage station. Taken over all seasons, below 10° C there is a strong association between Hermitage rain-day air temperatures and mean daily precipitation amounts (Figure 4-13). This association weakens at higher temperatures where there is in fact some degree of negative correlation.

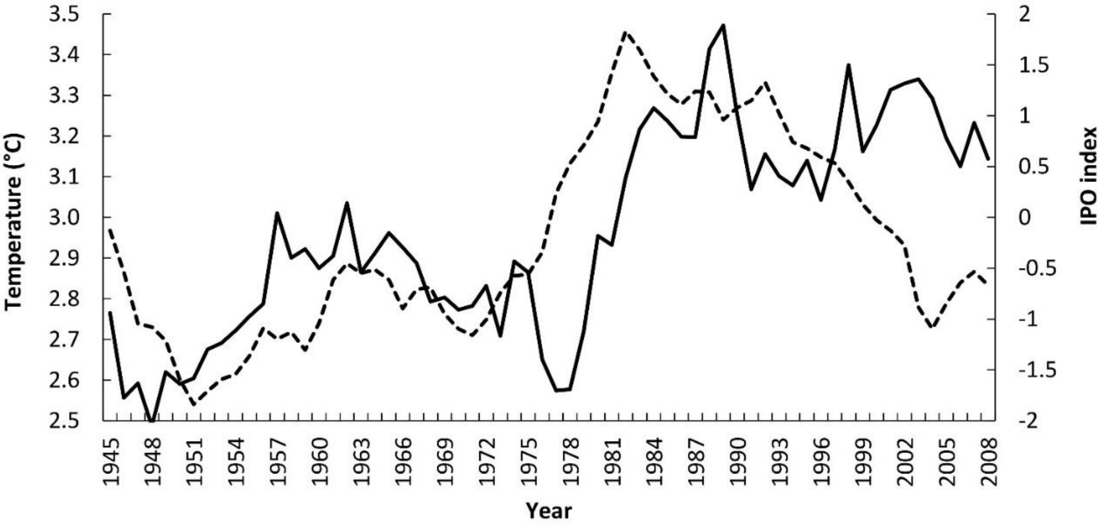

Figure 4-9 Smoothed annual Interdecadal Pacific Oscillation index values (dashed) and smoothed winter averaged temperature at Hermitage station (11-year running means).

Breaking down by seasons, the positive temperature association is particularly well reflected in winter because of the weighting toward greater frequencies of colder rain-day temperatures (Figure 4-14a). A winter rain-day temperature increase of 1° C equates to an increase in mean precipitation amount of about 2 mm. However, there appears no association between winter mean rain-day temperature and rain-day frequency (Figure 4-15). The typically higher rain-day temperatures in the other seasons are in the temperature regions of weaker precipitation-temperature associations of Figure 4-13, so the limited degree of association in the other seasons between temperature and precipitation is to be expected (Figure 4-14, b,c,d). Consequently outside of winter there is minimal association between temperature and Waitaki lake inflows, and hence between the IPO and Waitaki lake inflows.

The actual mechanism linking winter temperature and rain-day precipitation is left open. One possibility is a greater frequency of moist winds from the north-west, which may be the mechanism by which the IPO has an influence on temperature. In this regard Salinger et al. (2001) noted changes in wind patterns associated with the IPO. Kingston et al. (2014) also described associations between north-westerly flow and Waitaki inflows. However, that investigation did not find associations between inflows and larger climatic modes.

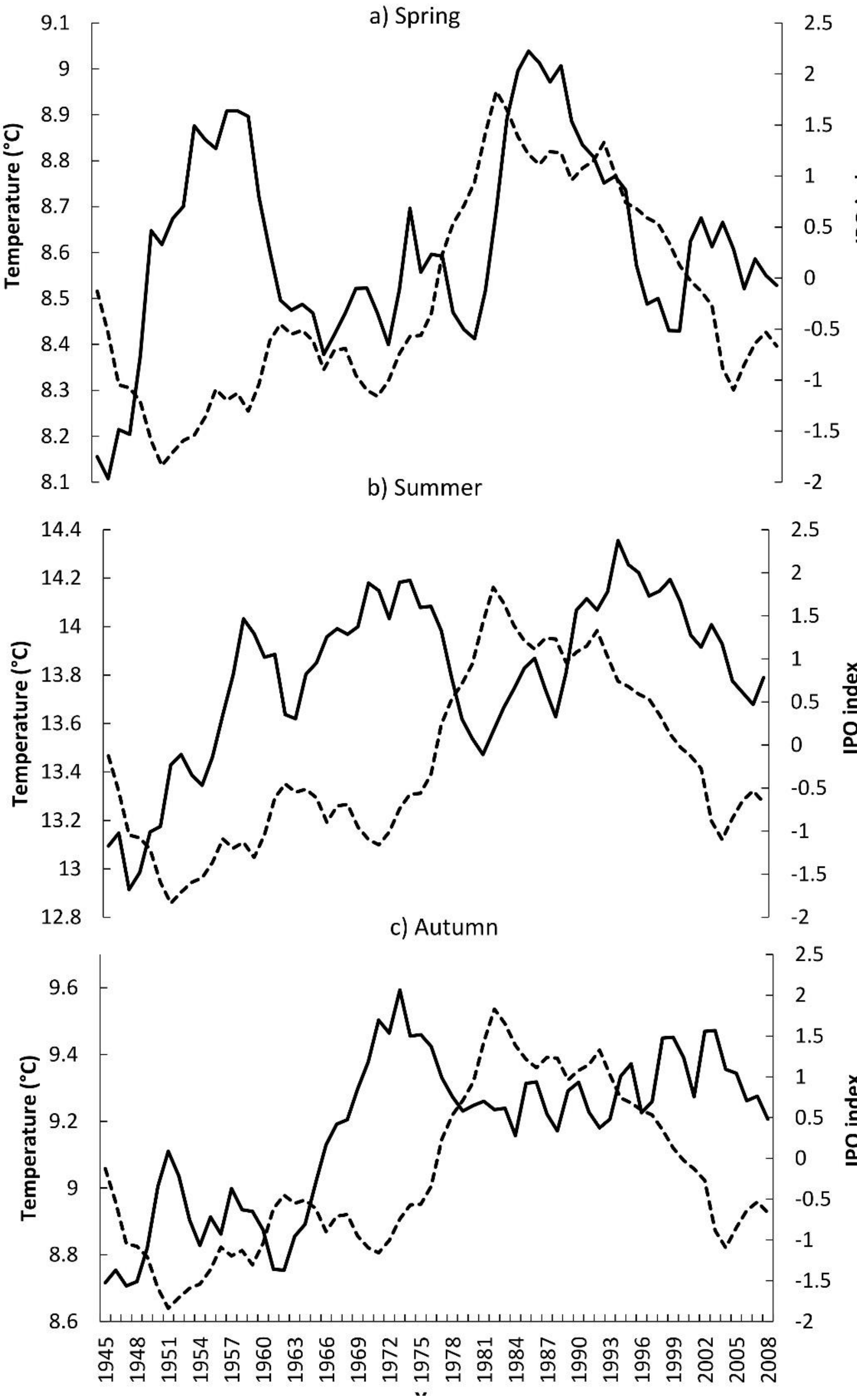

Figure 4-10 Smoothed annual Interdecadal Pacific Oscillation index values (dashed line - repeated) and smoothed mean seasonal temperatures at Hermitage station (solid lines).

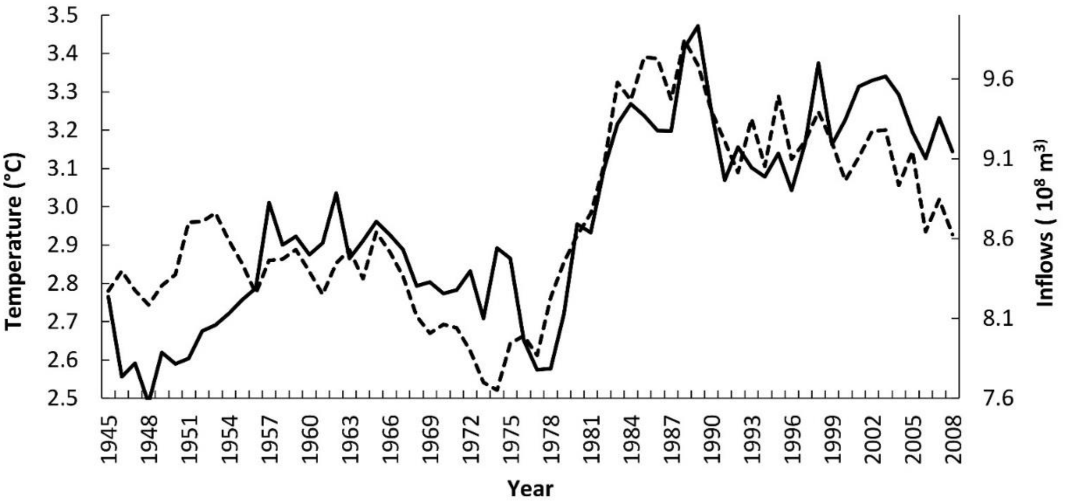

Figure 4-11 Smoothed winter mean temperatures at Hermitage station (solid line) and smoothed winter lake inflows for lakes Tekapo and Pukaki combined (dashed line).

The other open question is whether the apparent IPO correlation with winter temperatures actually translates to sufficient temperature variation to give a causal linkage between temperature and winter runoff in the headwater region (Figure 4-11). One issue here is that there will be within-winter seasonal temperature variation and corresponding within-winter seasonal discharge variation, which will together tend to mask IPO-induced temperature linkages to discharge. Separating these components requires further work so the temperature-related IPO link to winter discharge remains a working hypothesis for now.

The mechanism by which increased winter rain-day magnitudes and temperatures translate to increased winter inflows must involve some degree of interaction with the winter snow pack. Greater precipitation amounts at higher winter temperatures would imply an increased area of the lower elevations receiving precipitation as rain, possibly melting existing snow to add to the increased discharge arising from the greater rainfall magnitudes.

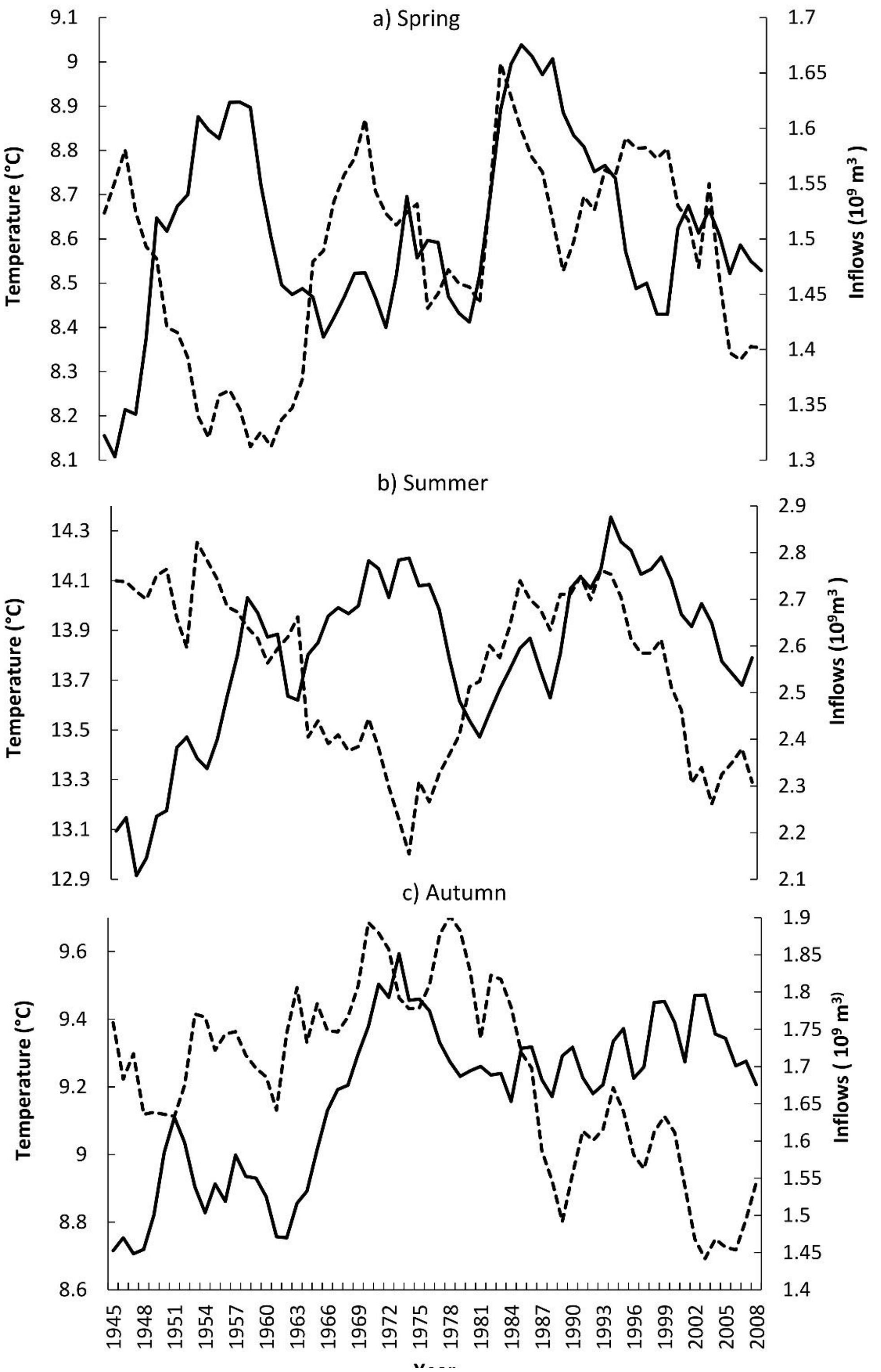

Figure 4-12 Smoothed seasonal mean temperatures at Hermitage station (solid line) and smoothed combined lake inflows, for spring summer and autumn.

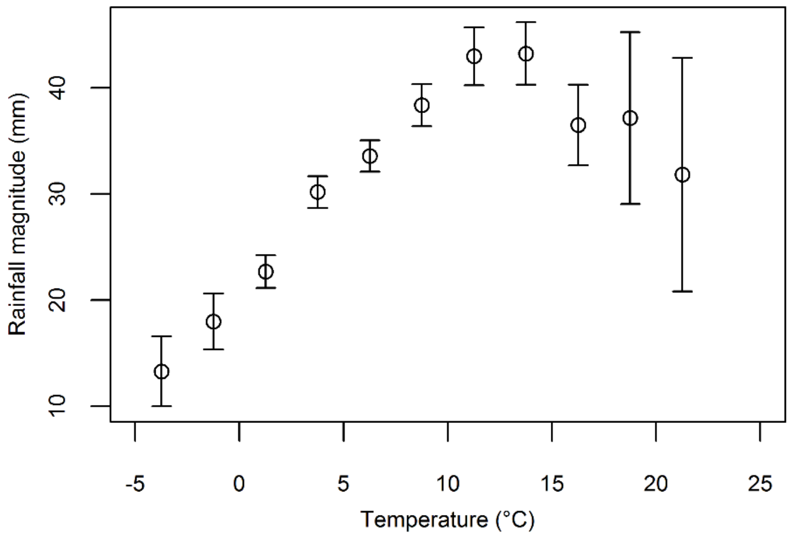

Figure 4-13 Mean daily rainfall vs rain-day temperature at Hermitage station (all seasons). Time period is 1940-2008. Bars denote 95% confidence intervals for the means. The means are all calculated for 2.5° temperature bins.

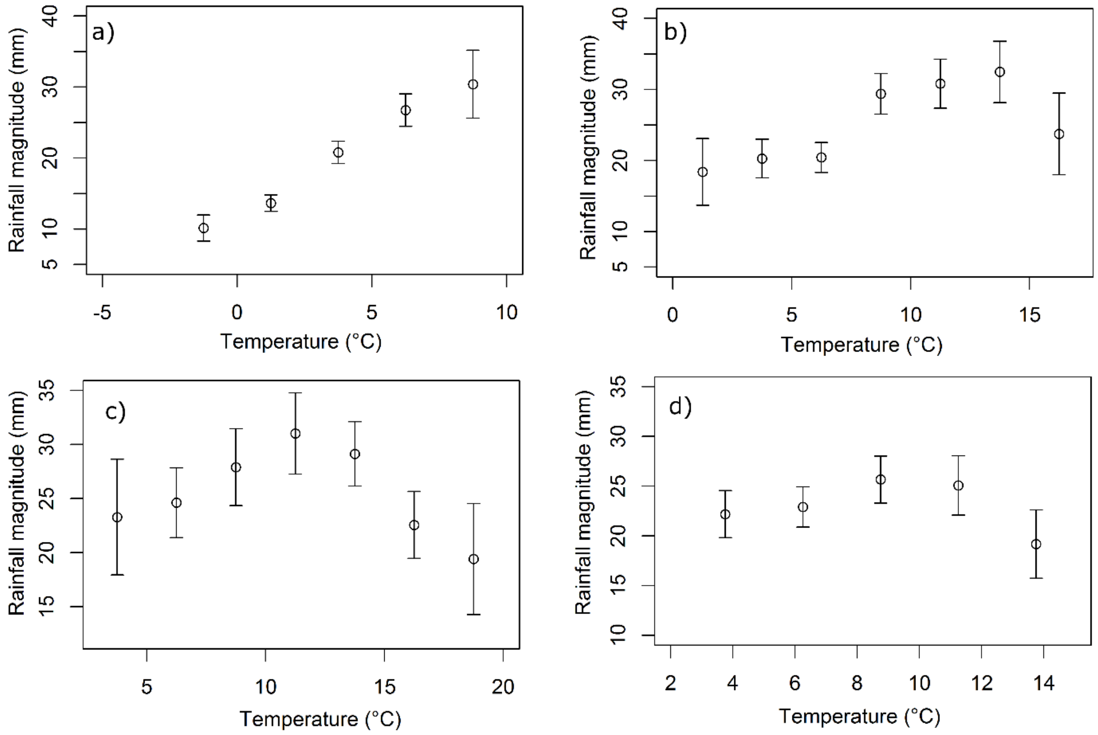

Figure 4-14 Mean seasonal daily rainfall magnitude vs rain-day temperature at Hermitage station: a) Winter; b) Autumn; c) Summer; d) Spring. The greater width of error bars compared to Figure 4-12 is a reflection of the smaller amounts of data.

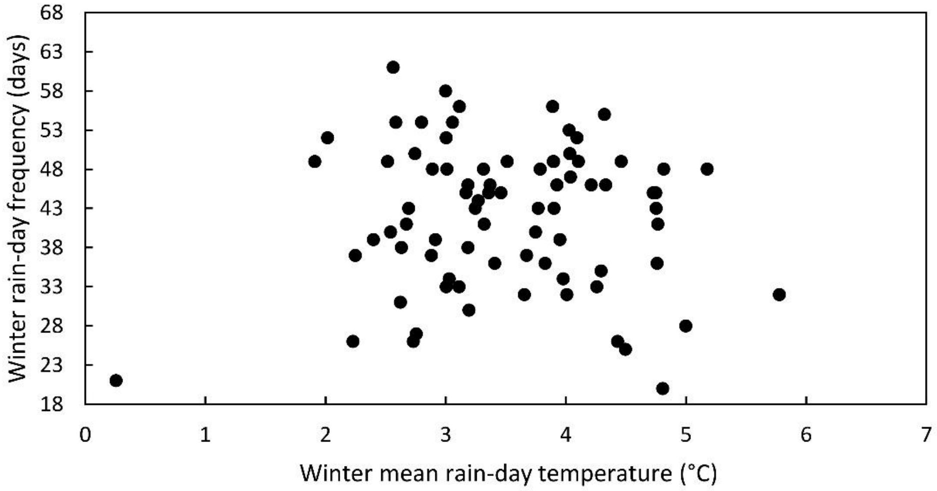

Figure 4-15 Winter rain-day frequencies exhibiting minimal correlation with winter rain-day temperatures (Mt Cook, 1940-2008)

Discussion

There is a climate change implication from the present regional study. Other things being equal, an increase in winter air temperature arising from climate change is likely to result in increased winter precipitation in the Waitaki headwaters. However, for the other seasons there may even be a possibility of decreased precipitation with a temperature increase and thus decreased river discharge.

Further work is needed to investigate in more detail the suggested link, or lack of link, between the IPO and lake inflows for the various seasons. There is also a possibility to repeat this study in nearby regions which may have different hydroclimatic responses. For example, Taylor and Bardsley (2015) found spring IPO values of some use in summer headwater flow forecasting for the Clutha River system. On the other hand, forecasting winter inflows appears most promising for the upper Waitaki because winter temperatures and runoff amounts seem to be more influenced by Pacific atmospheric circulation patterns.

Further work might also be directed toward expanding the somewhat arbitrary "winter" period of June, July and August. Colder conditions might also have a rainfall association in late autumn and early spring, which might be usefully aggregated as part of an expanded "winter" for the purposes of this type of study. There is also a need to seek longer data periods to confirm the conclusions here, as our data incorporated just a single major change of IPO phase.

Finally, we cannot completely rule out at the possibility of imperfect precipitation data playing a role in the results. There is a possibility of some hours delay of snow melt in a gauge until warmer temperatures allowed melting and “precipitation” only then to be recorded. This possibility may require further checking.

Conclusion

It is suggested as a working hypothesis that winter Waitaki headwater discharges have a causal linkage to the IPO via the IPO influencing winter rain-day temperatures. This is by way of a strong correlation between winter air temperatures and mean daily precipitation amount. The typically higher temperatures in the other seasons are in the temperature regions of weaker precipitation-temperature association which in turn contribute to limited inflow correlation. Consequently, the IPO association to Waitaki headwater discharge is confined to the winter season.

References

Folland, C.; Renwick, J.; Salinger, M.; Mullan, A. 2002: Relative influences of the Interdecadal Pacific Oscillation and ENSO on the South Pacific Convergence Zone. *Geophysical Research Letters*, 29 (13): 21-1–21-4.

Kingston, D. G.; Webster, C. S.; Sirguey, P. 2014: "Large-scale climate control on lake inflow in the Waitaki basin, New Zealand", Proceedings of FRIEND-Water 2014, Montpellier, France, IAHS Publ. 363, 138-144.

Mantua, N. J.; Hare, S. R.; Zhang, Y.; Wallace, J. M.; Francis, R. C. 1997. A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, 78 (6): 1069-1079.

McKerchar, A.; Henderson, R. 2003. Shifts in flood and low-flow regimes in New Zealand due to interdecadal climate variations. *Hydrological Sciences Journal*, 48 (4): 637-654.

Purdie, J. M; Bardsley, W. E. 2010: Seasonal prediction of lake inflows and rainfall in a hydro-electricity catchment, Waitaki river, New Zealand. *International Journal of Climatology*, 30 (3): 372-389.

Salinger, M.; Mullan, A. 1999: New Zealand climate: temperature and precipitation variations and their links with atmospheric circulation 1930--1994. *International Journal of Climatology*, 19 (10): 1049-1071.

Salinger, M., Renwick, J.; Mullan, A. 2001. Interdecadal Pacific oscillation and south Pacific climate.*International Journal of Climatology*, 21 (14): 1705-1721.

Taylor, M., & Bardsley, W. E. (2015). The Interdecadal pacific oscillation and the Southern Oscillation Index: relative merits for anticipating inflows to the Upper Clutha lakes. Journal of Hydrology: New Zealand, 54(2), 115–124

Ummenhofer, C.C.; Sen Gupta, A.; England, M.H.: 2009. Causes of late twentieth-century trends in New Zealand precipitation. *Journal of Climate,* 22 (1).

4.4 Conclusions

First, a study of incorporating large scale climatic-oceanic predictors into a forecasting model of a local hydrological process was performed. The forecasting model was considered as a high-dimensional regression, where the number of predictors is larger than the number of observations. The Lasso regularized regression method was investigated as a tool for informative predictor selection in this setting. A case study of season-ahead forecasting of inflows into the hydro-power lakes in Waitaki catchment was investigated. A cross-validation method was proposed which simulates the likely practical situation of forecasting when all available information to date is utilised to construct the current forecasting model.

Second, a study was undertaken of possible physical mechanisms for influence of the Interdecadal Pacific Osciallation on the winter inflows in the hydro-power lakes in the Upper Waitaki catchment.

Chapter 5 Conclusions and further research

In this thesis, results were presented on evaluating selected machine learning methods and developing new methods in the context of analysis of hydroclimatological datasets.

As noted in the Introduction, this thesis aimed to address the following objectives:

1. Determine if data pre-processing with the decimated wavelet discrete transform could give false predictive accuracy in regression ML algorithms.
2. Investigate a specific regularised regression method as a tool for formalised simplification of time series models generally.
3. Develop a new and simple test of general non-random association in bivariate data.
4. Investigate a specific regularised regression method in seasonal streamflow forecasting problem.

Chapter 2 addressed the first two objectives. Firstly, a general result was obtained that a decimated wavelet discrete transform based on a pyramidal algorithm requires future values of time series to be utilized. When the discrete wavelet transform is utilized as a pre-processing step for time series forecasting, the necessary independence of calibration and validation data is compromised. This in turn translates into over-optimistic forecasting accuracy. The obtained result is general and has wide implications in any discipline where discrete wavelet transform is utilised in forecasting frameworks. Secondly, a general framework was presented of time series model construction where there is external event forcing. The framework is based on the LASSO regularized regression method, enabling simplification a deliberately over-parameterised initial model while preserving model capability.

Chapter 3 addresses the third objective by way of two methods of detecting associations in bivariate data. The proposed methods are both based on a randomization approach. The first method aims to detect threshold-like associations while the second is concerned with detecting general non-random associations in bivariate datasets. The methods are both general and applicable in forecasting frameworks when screening of large number of potential predictors is needed.

Chapter 4 addresses the fourth objective by way of using LASSO regularized regression as a tool for discovering informative large scale climatogical predictors of local hydrological processes. A cross-validation scheme was proposed, which is closely related to likely practical forecasts while at the same time maximising use of available information. This chapter utilised a case study predicting seasonal inflows in the Waitaki River catchment. The second part of the chapter investigates the influence of the Interdecadal Pacific Oscillation on winter inflows in the Waitaki catchment. The forecasting methodology and cross-validation frameworks proposed in the first part of the chapter is applicable for similar hydroclimatological forecasting problems.

Further research

Further research directions of interest (by chapter) are:

- Chapter 2 (part 1): Only the decimated wavelet transform as a pre-processing method was investigated as creating potential erroneous forecasting accuracy. However, similar issues arise in applying any other pre-processing method like empirical mode decomposition. It would be of interest to analyse wavelet neural networks as to whether a similar problem is present in that methodology as well.
- Chapter 2 (part 2): The general framework of model simplification presented in this chapter can be extended in a setting with multiple sources of events. Furthermore, its application to groundwater modelling and simplification of existing distributed hydrological models could be investigated.
- Chapter 3: The methods of general association detection presented here might be further extended to the multivariate case. Both methods could be applied for cluster search in climatological data, for example, as replacement of the correlation coefficient as a measure of similarity in climate networks. Further research into computationally optimizing the proposed methods would also be useful.
- Chapter 4: It would be of interest to investigate the sliding window effect on selected predictors in the case study considered in this

chapter. Also, different LASSO methods such as sparse group LASSO and elastic nets could be added for comparison.