Towards the Development of Tailored Seasonal Forecasts for the Hawke's Bay Region

Prepared for Hawke's Bay Regional Council

July 2016

Prepared by:

Nava Fedaeff

Nicolas Fauchereau

For any information regarding this report please contact:

Nava Fedaeff

Climate Scientist – Auckland

Climate Application

+64-9-375 6337

Nava.Fedaeff@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd

Private Bag 99940

Viaduct Harbour

Auckland 1010

Phone +64 9 375 2050

Quality Assurance Statement		
signature	Reviewed by:	Andrew Tait
signature	Formatting checked by:	Andrew Tait
signature	Approved for release by:	Ken Becker

Contents

Executive Summary

This report, prepared for the Hawke's Bay Regional Council, describes a tailored statistical seasonal forecasting scheme developed for the Hawke's Bay region. This forecasting scheme is based upon tercile probabilities i.e. what is the chance that rainfall/temperature over the next 3 months will be below normal, normal or above normal.

The dataset used in this study is an extension of the dataset used in a previous report - *Relationship between Climate Modes and Hawke's Bay Seasonal Rainfall and Temperature* (Fedaeff & Fauchereau, 2015). This follow up study builds upon knowledge already established and explores different methodologies for producing seasonal forecasts tailored to the region.

To develop the experimental forecasting scheme, methods borrowed from the field of Machine Learning (ML) were utilised (Simon, 2013). The idea was to derive a general formal relationship (a 'model' in the largest sense of the term), usually in the form of a mathematical transformation or set of rules, linking the predictors (climate drivers) to the predictands (seasonal rainfall and temperature forecasts), such as when presented with new observations of the predictors, the model would be able to successfully predict the value of the predictand.

Regression approaches were briefly tested in the initial stages of the study. As the ultimate goal was to produce categorical probabilistic forecasts, classification algorithms were found to be the most suited method.

After carrying out various analyses, it was established that the best classification pipeline involved using Sea Surface Temperature anomalies over a large Pacific domain (60°S, 40°N, 120°E, 70°W) as inputs. It was found that the accuracy of a forecast using this method for temperature was about 61% for the region as a whole and varied between 62% and 67% for individual stations. For rainfall the forecast accuracy was about 60% for the region as a whole and varied between 56% and 63% for individual stations. It should be noted that as this forecasting scheme is based on tercile probabilities, the theoretical accuracy of a random forecast (i.e. if you were to guess at random) is 33%.

The results of this study show that there is significant potential to operationally implement a tailored seasonal climate forecast for the Hawke's Bay region. Tercile probabilities can be calculated for the whole region (based on the VCSN regional index) and/or for 12 rainfall stations and 3 temperature stations. Such a forecast could be evaluated alongside global dynamical models and the Seasonal Climate Outlook issued by NIWA (https://www.niwa.co.nz/climate/sco) to improve the management of climate sensitive activities in the region. The forecasting template developed in this study has the potential to be applied to other regions once it has been operationally validated.

1   Introduction

NIWA recently completed an Envirolink-funded study which identified a number of climate oscillations that influence seasonal rainfall and temperature in Hawke's Bay (Fedaeff & Fauchereau, 2015). The South Pacific Subtropical Dipole (SPSD), the El Niño /Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD) were found to have statistically significant relationships with Hawke's Bay's climate. Having established that there is a relationship between these climate cycles and temperature and rainfall in the Hawke's Bay, a follow up study was commissioned to explore the efficacy of a tailored seasonal forecasting scheme for the region.

While short-term weather forecasts now have a high degree of reliability, probabilities associated with climate forecasts (long-range estimates of what will happen over the coming three months) are generally not as high. Despite this, seasonal climate forecasting is a rapidly developing area of science, and forecasts are becoming increasingly more accurate. In particular, when there is a strong driver present such El Niño, there is generally higher seasonal forecast accuracy. Seasonal climate forecasting has the potential to enable managers in climate affected sectors to reduce unwanted impacts or take advantage of favourable conditions. In New Zealand long-range forecasts are valued by many sectors, such as farmers, emergency services, regional planners and policymakers. The Hawke's Bay Regional Council (HBRC) has identified weather and climate as having a far reaching effect on the Council's operations in areas such as civil defence, flood protection work, land management and water allocation.

There are several approaches to seasonal forecasting. Statistical predictions are based on regional historical relationships between physical variables such as temperature and precipitation with statistical models of varying degrees of sophistication. Dynamical prediction methods are a relatively recent endeavour and use complex dynamical numerical models of the main Earth system components (Troccoli, 2010). As dynamical models are very sophisticated and currently operate on coarse spatial scales, this report explores a range of statistical methodologies to produce seasonal forecasts for the Hawke's Bay region and evaluates the performance of the best performing solution. The forecasting scheme presented in this study could be evaluated alongside global dynamical models and the Seasonal Climate Outlook issued by NIWA (https://www.niwa.co.nz/climate/sco) to improve the management of climate sensitive activities in the region.
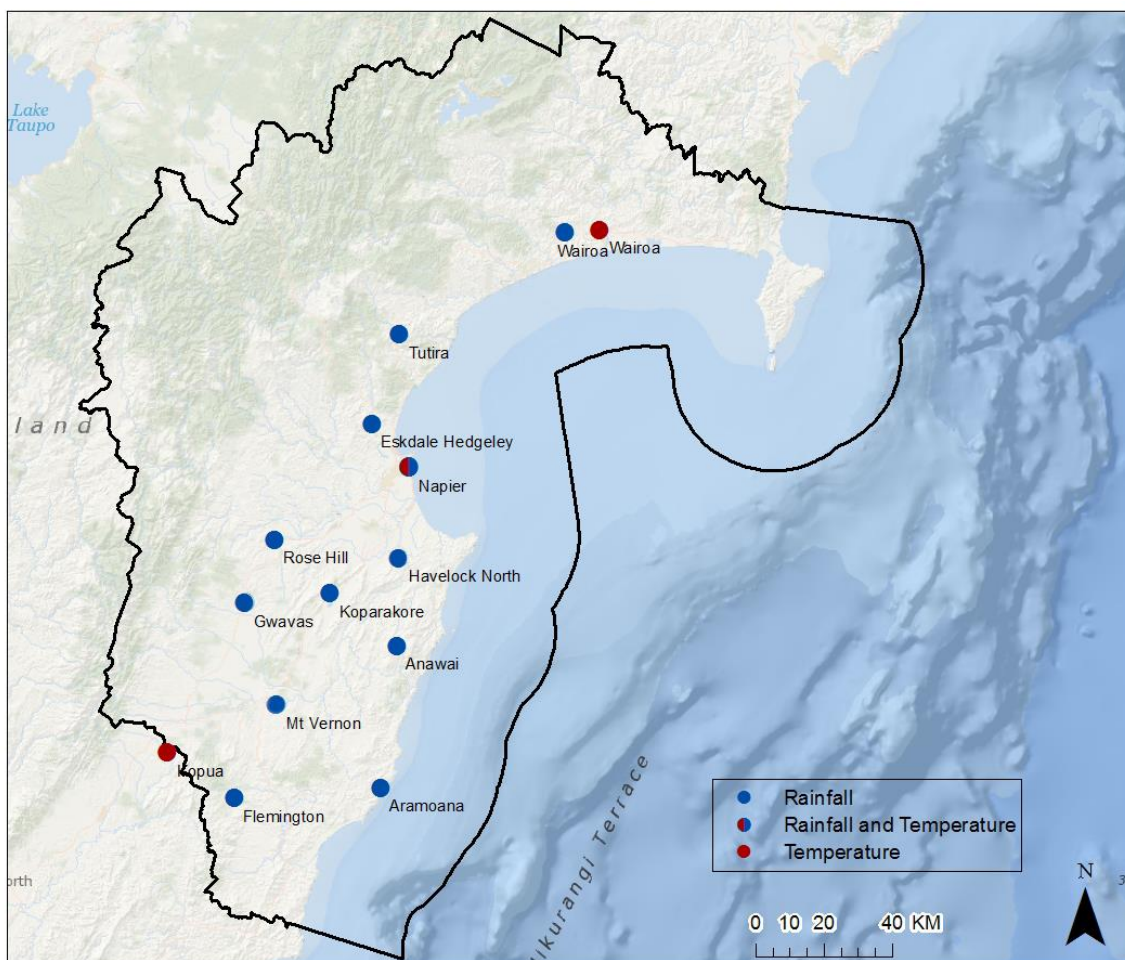
This report describes a tailored seasonal forecasting scheme for the Hawke's Bay, including the components contributing to the forecasts and the methodologies which were explored to produce them. Additionally, the operational system requirements for forecast generation and delivery are described.

2   Methodology

2.1   Data availability and homogenisation

The dataset used in this study is an extension of the dataset used in the initial report - *Relationship between Climate Modes and Hawke's Bay Seasonal Rainfall and Temperature* (Fedaeff & Fauchereau, 2015). Full details of data availability, quality control and homogenisation applied is detailed in the earlier report.

Homogeneity tests carried out for the initial report identified 15 rainfall and 7 temperature sites which had records suitable for further analysis. For the purposes of the follow-up research, some of the homogenised records did not meet requirements for analysis. For rainfall, Hasting was excluded as 32% of data was missing. The Whanawhana, Te Wairere and the Tuai composite record was also not used as the data did not cover the 1981-2010 climatology period. For temperature, only Napier, Kopua and Wairoa had records with sufficient data. The sites chosen for analysis are displayed spatially in Figure 1 and listed in Tables 1 and 2. For several sites, composite climate records were created to extend the record length and fill missing gaps. Further information on data availability is included in Appendix A.

Figure 1: Rainfall and temperature sites used for analysis.

Table 1:   Rainfall records selected for analysis (shown in grey shading).

		Records		Years of record	Status	Altitude (m)	Notes
Station	Location	Begin	End				
D87731	Erepeti	1/05/1928	31/03/2007	78	closed	405	
D87712	Onepoto	1/09/1890	31/05/1935	45	closed	-	
D87812	Tuai	1/09/1923	28/02/2005	81	closed	274	
	Tuai composite	1/09/1923	31/03/2007	83			
D97031	Wairoa, Waiputaputa Station	1/04/1911	31/10/2015	104	open	29	Base period 1981-2010
D96081	Te Rangi Maungaharuru	1/01/1919	31/12/1983	65	closed	335	
D96251	Te Wairere	1/01/1932	31/12/2007	76	closed	564	Daily record starts in April 1932.
D96281	Tutira	1/11/1894	31/08/2004	110	closed	201	
D96282	Tareha	3/11/1949	31/03/2016	65	open	430	
	Tutira composite	1/11/1894	31/03/2016	122			Base period 1981-2010
D96541	Whanawhana	1/01/1906	30/04/1984	78	closed	293	
D96653	Rose Hill	1/01/1954	31/03/2016	62		171	Daily rainfall record starts in 1954. Base period 1981-2010
D96591	Napier Nelson Park	1/01/1870	29/02/2016	146	open	2	
D96484	Napier Aero Aws	1/10/1990	31/03/2016	26	open	3	
	Napier composite	1/01/1870	29/02/2016	146			Base period 1981-2010.
D96691	Havelock Nth, Te Mata	1/09/1889	28/02/1985	95	closed	140	Daily rainfall record starts in 1917
-	Kopanga Homestead	1/02/1938	29/02/2016	78	open	-	
D96681	Hastings	1/01/1928	31/08/1966	38	closed	14	
	Havelock North composite	1/09/1889	29/02/2016	127			Daily rainfall record starts in 1917. Base period 1981-2010.
D96882	Anawai	1/07/1925	30/04/2014	88	open	442	Base period 1981-2010.
D96741	Gwavas	1/09/1889	31/03/2016	124	open	244	
D96731	Smedley	1/11/1964	31/03/2015	50	open	457	
	Gwavas composite	1/09/1889	31/03/2016	127			Base period 1981-2010.
D96951	Mt Vernon 2	1/02/1886	31/03/2016	130	open	155	
D06051	Waipukurau Aero	1/01/1945	31/07/1994	49	closed	137	
D96962	Waipawa EWS	28/07/2007	30/04/2016	7	open	130	
	Mt Vernon composite	1/02/1886	31/03/2016	130			Base period 1981-2010.

Station	Location	Begin	End	Years	Status	Altitude	Notes
D06181	Aramoana	1/01/1907	29/02/2016	109	open	27	Daily record begins in March 1907. Base period 1981-2010.
D96891	Waimarama	1/08/1888	30/11/1989	100	closed	3	
D06452	Tawadale, Wimbledon	1/01/1946	30/09/1999	53	closed	107	
B96051	Tarawera	1/07/1908	31/08/1975	67	closed	497	
D96483	Eskdale Hedgeley	1/08/1894	31/03/2016	122	open	34	Base period 1981-2010.
D96771	Koparakore	1/01/1902	30/04/2011	109	closed	48	
D96861	Pukehou, Te Aute Station	1/10/1892	31/07/1997	103	closed	107	
	Koparakore composite	1/01/1902	30/04/2011	109			Base period 1981-2010.
D96925	Makaretu North	1/01/1960	31/12/2014	54	open	396	
D06062	Waiwhero Station	1/03/1951	31/12/2014	63	open	194	
D06142	Flemington	1/12/1958	31/03/2016	57	open	174	
D06151	Motuotaria	1/09/1910	1/06/1962	51	closed	62	
	Flemington composite	1/09/1910	31/03/2016	106			Base period 1981-2010.
D96680	Hastings Aws	1/10/1981	30/04/2016	35	open	16	
D96688	Hastings Fire Station	1/09/1965	31/10/1981	16	closed	12	
	Hastings composite	1/09/1965	30/04/2016	51			

Table 2:   Temperature records selected for analysis (shown in grey shading).

		Temperature records					
Station	Location	Records		Years of record	Status	Altitude (m)	Notes
		Begin	End				
D87811	Waikaremoana Onepoto	1/06/1935	30/06/1990	18	closed	643	Daily data not digitised prior to 1972.
D87812	Tuai	1/01/1982	8/07/1990	8	closed	274	
D97045	Wairoa, North Clyde Ews	31/07/1991	30/04/2016	25	open	15	
D97043	Wairoa Hospital	31/12/1971	31/07/1991	19	closed	20	
D97042	Wairoa, Frasertown	1/01/1964	31/10/1989	25	closed		Daily data not digitised prior to 1972
	Wairoa composite	1/01/1964	30/04/2016	52			Daily data not digitised prior to 1972. Base period 1981-2010.
D96591	Napier Nelson Pk	1/1/1870	29/02/2016	146	open	2	

D96484	Napier Aero Aws	30/09/1990	30/04/2016	24	open	3	
D96481	Napier Aero	31/10/1973	20/03/1990	16	closed	2	
	Napier composite	1/01/1870	29/02/2016	146			Base period 1981-2010.
D06051	Waipukurau Aero	20/12/1944	31/07/1994	49	closed	137	Base period 1961-1990.
D96962	Waipawa Ews	28/06/1907	31/12/2014	7	open	130	
D96743	Gwavas Forest	1/01/1948	30/08/1989	40	closed	335	Daily data not digitised prior to 1972.
D06022	Kopua	1/07/1962	31/03/2016	54	open	311	Daily data not digitised prior to 1972. Base period 1981-2010.
D96680	Hastings Aws	1/10/1981	30/04/2016	35	open	16	
D96688	Hastings Fire Station	1/09/1965	31/10/1981	16	closed	12	Daily data not digitised prior to 1972
	Hastings composite	1/09/1965	30/04/2016	50			Daily data not digitised prior to 1972.

In addition to climate station data, NIWA also has access to the Virtual Climate Station Network (VCSN, see Tait *et al.*, 2006, 2008). Temperature and rainfall indices for the Hawke's Bay region as a whole were derived from the VCSN dataset. This was done by averaging the 533 VCSN agents (virtual stations) falling into the Hawke's Bay region, i.e. including the Hastings, Central Hawke's Bay and Wairoa districts as well as Napier City (see Figure 2).
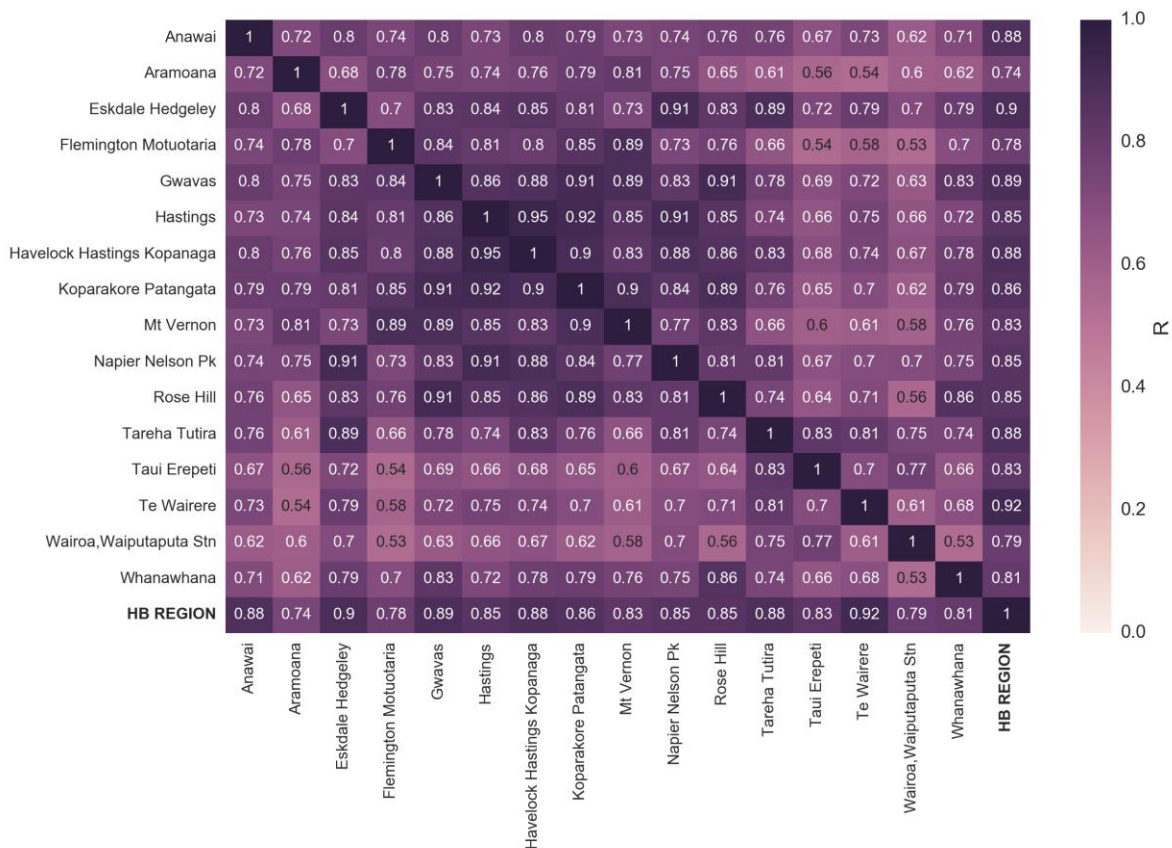
Once the regional VCSN index was obtained, the same processing methodology that was applied to the station data was used. Compared to climate station data, the regional index offers the advantage of presenting a continuous record, and is also updated regularly by NIWA. The disadvantage is that the VCSN dataset only starts in 1972 which reduces the time period available for training statistical models.

Correlations were calculated between time-series of seasonal rainfall anomalies, temperature anomalies and the respective regional index derived from the VCSN dataset over the overlapping time periods (See Figures 3 and 4). The results show that most stations are well correlated to each other as well as to the VCSN-derived regional index, indicating that interannual variability of seasonal rainfall tends to be homogeneous over the whole region. It is however to be noted that there are exceptions to this rule, i.e. some stations (e.g. Wairoa and Flemington) are relatively weakly correlated (R ~ 0.5, i.e. only 25% of common variance), but are still reasonably well-correlated to the Regional Index.
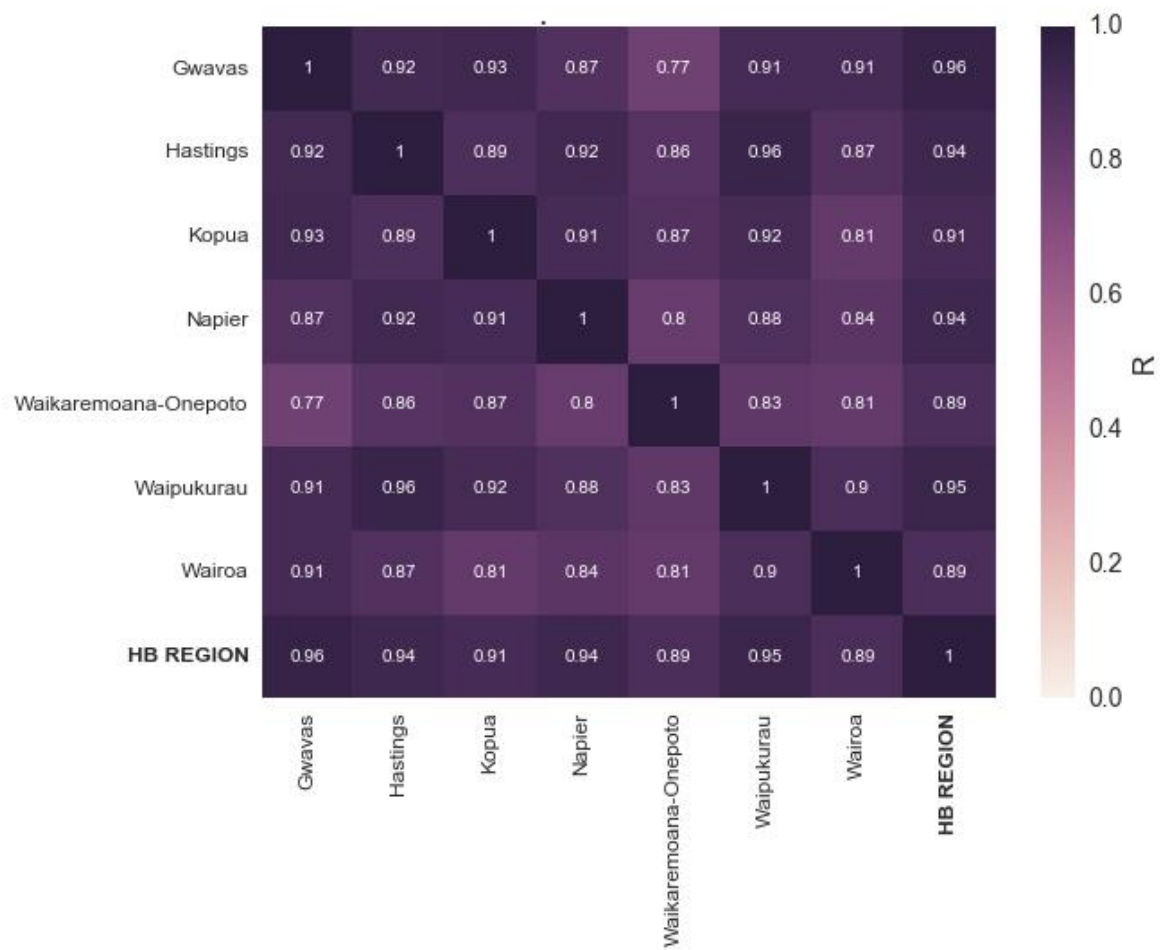
As expected (but with the caveat that most stations present a large proportion of missing values), the spatial homogeneity is greater for temperature than for rainfall, with most correlation values exceeding 0.8. This also means that the Hawke's Bay regional index for seasonal temperature is representative of all stations' interannual variability.

Figure 2: Virtual Climate Stations in the Hawke's Bay Region

Figure 3: Correlation matrix of seasonal rainfall anomlaies between climate stations (virtual and real) within the Hawke's Bay.

Figure 4: Correlation matrix of seasonal temperature anomlaies between climate stations (virtual and real) within the Hawke's Bay.

2.2 Data Processing

2.2.1 Calculation of anomalies

For all stations and regional indices, the 3-month running mean (for temperature) or 3-month running accumulation (for rainfall) was first calculated to obtain overlapping 3-month 'seasons', i.e. JFM (January–March), FMA (February–April), etc. The anomalies were then calculated with respect to the 1981–2010 climatology (or 'normal').

Note that the actual time-series of seasonal anomalies were only used in regression models for initial exploration and for some ancillary diagnostics (e.g. creating correlation matrices between stations time-series and with the respective Hawke's Bay regional index from the VCSN, Figures 3 and 4). Time-series of anomalies were not used in the development of probabilistic categorical forecasts (see section 3.2).

2.2.2   Derivation of tercile categories

Tercile categories (i.e. value of the 33rd and 66th percentile) were calculated over the climatology period of 1981–2010 using the time-series of 3-month means (temperatures) or accumulations (rainfall). The terciles were calculated independently for each 3-month 'season'.
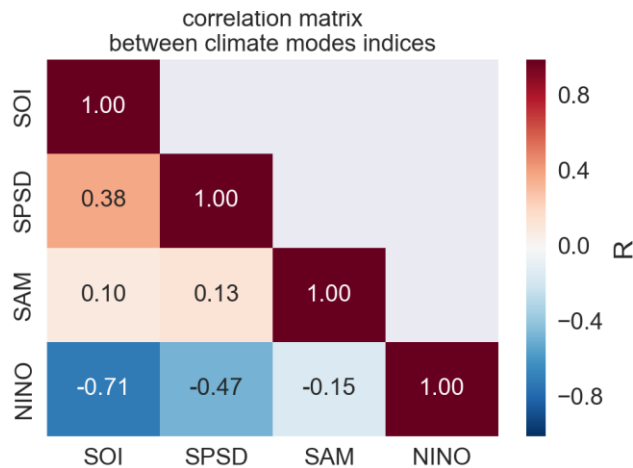
Each season (3-month period) was classified into a category: 'below normal' (below the 33rd percentile), 'normal' (between the 33rd and 66th percentile) or 'above normal' (above the 66th percentile). An integer flag (i.e. -1 for 'below normal', 0 for 'normal', 1 for 'above normal') was also assigned to the season for classification purposes.

2.2.3   Derivation of the potential predictors

In the previous report (*Relationship between Climate Modes and Hawke's Bay Seasonal Rainfall and Temperature*) five large-scale climate oscillations (Interdecadal Pacific Oscillation (IPO), El Niño Southern Oscillation (ENSO), Southern Annular Mode (SAM), Indian Ocean Dipole (IOD) and the South Pacific Tropical Dipole (SPSD)) were reviewed and their relationship to rainfall and temperature in New Zealand and the Hawke's Bay region was summarised. The indices representative of these main climate modes affecting the Southern Hemisphere were considered, and some potential for predictability was inferred from the statistical relationships that were analysed (Fedaeff & Fauchereau, 2015).

While this approach shed light on the potential climate processes and modes affecting Hawkes Bay regional climate, it has some limitations when it comes to developing a statistical forecasting scheme, in particular:

- The climate indices analysed in the initial report are in some cases correlated to each other. This is a problem for many statistical learning algorithms, which require statistical independence of the features (predictor time-series). Figure 5 presents the correlation matrix between the SOI, the NINO3.4 SST index, the SPSD and the SAM monthly anomalies over the 1950-2014 period. Besides the obvious strong negative correlation between the SOI and NINO3.4, note that the SPSD also appears to not be independent from ENSO (correlation of 0.38 with the SOI, and -0.47 with NINO3.4). Based on previous work, it is also known that the SAM is partly correlated to ENSO during the Southern Hemisphere summer.

correlation matrix
between climate modes indices

Figure 5: Correlation matrix between the SOI, the NINO3.4 SST index, the SPSD and the SAM monthly anomalies over the 1950-2014 period.

- ▪ They only account for a limited subset of the potential modes of interannual variability suspected to operate in the climate system.

In other words, the 'ideal' predictors need to be time-series that are both statistically independent and, taken together, represent a large part of the variance of the global or regional coupled ocean-atmosphere system.

Taking these requirements into consideration, **Empirical Orthogonal Function** decomposition (EOF, also known as **Principal Component Analysis** or PCA) was the preferred method of analysis. Hannachi *et al.* (2007) provides further information about the use of Empirical Orthogonal Function decomposition. In short, EOF analysis decomposes (in this particular case) the time-evolution of a climate field (varying along dimensions of time, latitude and longitude) in a set of statistically independent modes, each combining a spatial pattern (the 'EOF') and a time-series (the Principal Component - 'PC') describing the evolution over time of this pattern in sign and amplitude.

The use of this method reduces the dimensionality of the original dataset (which is the size of the number of latitudes times the number of longitudes in this case) to a set of PCs, which can then be viewed as 'synthetic variables'. Usually about 10–20 PCs are sufficient to explain more than 80% of the original field's total variance. In this study we considered the PCs associated with EOF analyses performed on Sea Surface Temperature (SST) alone or SSTs combined with Geopotential height at 850 hPa (i.e. the height of the 850hPa isobar, noted as 'Z850' hereafter) over several spatial domains as potential predictors. In all cases enough PCs were kept to explain > 80% of the original variance, i.e. capturing the essential spatial-temporal signals contained in the SST field or the combined SST–Z850 fields.
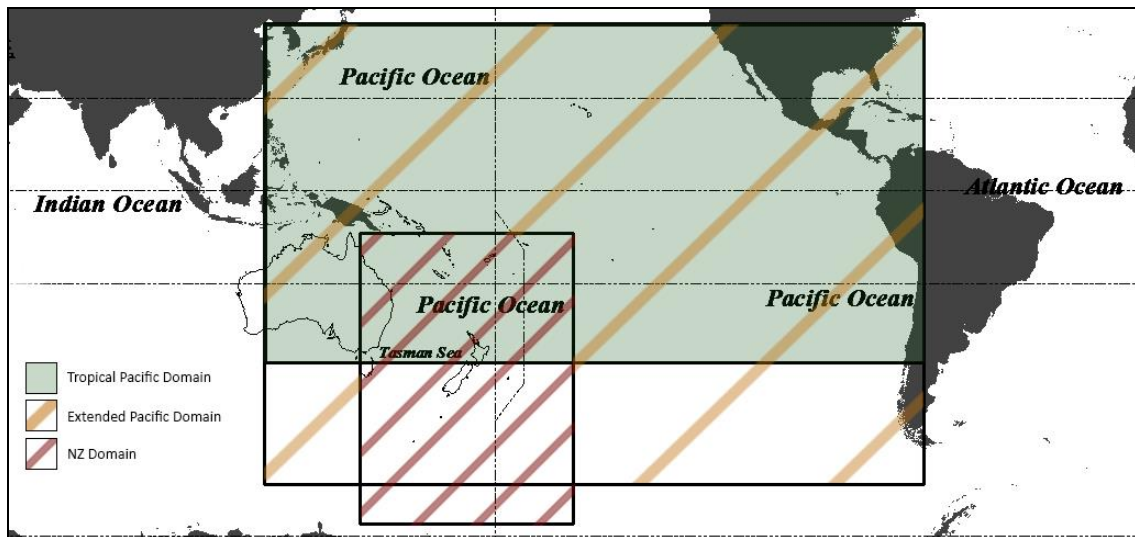
The SST data used in this analysis come from the NOAA ERSST V4 dataset (Huang *et al.,* 2015, 2016). The NOAA Extended Reconstruction Sea Surface Temperature (ERSST) provides global, spatially complete SST data at a monthly time step and a 2 degree spatial resolution for the period 1854–present. In this study, only the post 1950 high quality data period has been considered.

The Z850 data come from the NCEP/NCAR Reanalysis dataset (aka NCEP R1, see e.g. Kalnay *et al*., 1996). A **reanalysis** is a meteorological data assimilation project which aims to assimilate historical

observational data spanning an extended period, using a single consistent assimilation (or "analysis") scheme throughout. It provides physically-consistent gridded atmospheric variables, and – in the case of the NCEP R1 dataset – it is updated regularly, making it suitable to use for operational purposes. Most variables are available on a 2.5 degree grid, from 1948 to present. Again we selected the high quality data period post 1950.

For the SST and SST-Z850 analyses, three spatial domains were considered (Figure 6):

- The Tropical Pacific domain (40°S, 40°N, 120°E, 70°W)

- The New Zealand domain (65°S, 10°S, 145°E, 160°W)

- The large-scale Pacific domain (which includes the NZ domain) (60°S, 40°N, 120°E, 70°W)

Figure 6: Spatial domains used for Principal Component Analysis.

For both SSTs and Z850, the following processing was applied:
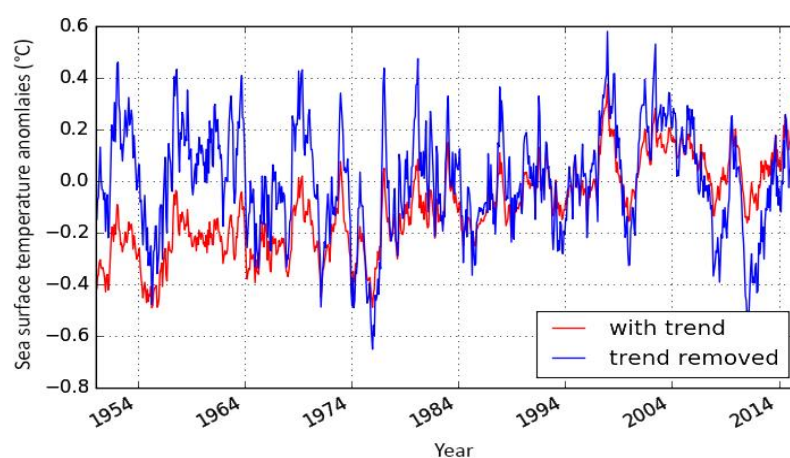
- The high quality observational period January 1950 – March 2016 was selected.

- Monthly anomalies were calculated relative to the 1981–2010 climatology.

- The long-term trend was removed. This step was particularly important in the case of the SSTs, which present a significant long-term warming trend over the period considered. An example of this can be seen in Figure 7, which presents the time-series of SSTs averaged over the larger Pacific domain (60°S, 40°N, 120°E, 70°W) respectively with and without the linear trend removed from the individual grid-points time-series.

- The time-series of monthly detrended anomalies were standardised (mean = 0, standard deviation=1)

- The resulting matrix was weighted by the square root of the cosine of the latitude, to account for the varying area represented by each grid-point.

- The EOF analysis was performed, and the number of PCs sufficient to explain at least 80% of the variance in the analysed field were kept.

A selection of figures showing the EOFs (spatial patterns) and PCs (time-series) coming from the EOF analyses performed respectively on SST monthly anomalies over the larger Pacific domain, the combined SST- Z850 monthly anomalies over the larger Pacific domain, and the combined SST-Z850 monthly anomalies over the regional New Zealand domain can be seen in Appendix B.

In addition, for some of the experiments, the set of PCs was extended to include lagged versions of themselves, i.e. PCs are lagged in time so that all PCs up from 1 to 6 months prior to the month of the observation are included into the predictor set in addition to the original PCs.

Figure 7: Sea surface temperature anomalies (°C) over the larger Pacific domain with and without long-term trend removed.

2.3   Developing a statistical forecasting scheme for Hawke's Bay

2.3.1   General approach

To develop this experimental forecasting scheme, methods borrowed from the field of Machine Learning (ML) were utilised (Simon, 2013). In our case the **predictand** (also called 'target variable' in ML, or 'dependent variable' in statistics) consists of an individual time-series of seasonal precipitation or temperature anomalies or tercile categories ('below normal', 'normal', 'above normal') at an in-situ climate station or derived from a regional index calculated from the VCSN dataset. The **predictors** (usually called 'features' in ML, or 'independent variables' in statistics) are either the time series of climate mode indices as described in the initial Hawke's Bay climate modes report (Fedaeff and Fauchereau, 2015) or the PCs coming from the EOF analyses summarized in Section 2.2.3, lagged in time so that the predictors lead the predictands by 0 or 1 months.

The idea was to derive a general formal relationship (a 'model' in the largest sense of the term), usually in the form of a mathematical transformation or set of rules, linking the predictors to the predictands, such as when presented with new observations of the predictors, the model would be able to successfully predict the value of the predictand.

This very general problem is amenable to two broad approaches: **regression** and **classification**, both fall into the class of **supervised learning** algorithms.

In **regression**, for a predictand y and (a set of) predictors X the goal is to find a function y' = f(X) such as the difference (measured usually in terms of sum of squared differences) between y'and y is minimized. In regression the forecast value is a continuous variable (i.e. in this particular case, temperature anomaly in °C, or rainfall anomaly in mm).

In **classification**, a learning algorithm is presented with a set of features (the predictors) and corresponding labels or classes (the predictand; i.e. here the observed categories: 'below normal', 'normal', 'above normal'). Given a new vector of features, the goal is to predict the new class, or – more interestingly – the probabilities that the new observation belongs to each class. In the context of seasonal climate prediction, the classes are categories, bounded by real values, resulting from the continuous values of the predictand (seasonal values). Such a transformation (from continuous to categorical variable) yielding distinct classes can be performed based on the distribution of the predictand (e.g. percentiles scores) or determined *a priori*. Here the categories are defined according to terciles determined over the climatological period (1981–2010), this definition is adopted by NIWA for the New Zealand Seasonal Climate Outlook (https://www.niwa.co.nz/climate/sco); more details are given in Section 2.2.

Regression approaches were only briefly tested in the initial stages of the study. As the ultimate goal is to produce categorical probabilistic forecasts, classification algorithms was the most suited method. To some extent, the results of regression models can be cast back into the corresponding terciles categories, but obtaining terciles probabilities from regression results is not straightforward and requires a number of assumptions. For these reasons, the results of the regressions approaches are explored only briefly in the study.

2.3.2   'Evolving' the best forecasting pipeline

Many different ML algorithms exist for the classification problem we are faced with (predicting a probability to belong to a class given a set of predictors).

It is beyond the scope of this report to describe at length the specificities of each algorithm that we have tested. In short, both single *classifiers*, as well as some so-called '*ensemble*' methods were tested. In ensemble methods, *several* simple classifiers are trained on different subsets of the data, and the resulting class prediction is obtained via e.g. a majority rule or the average prediction of the individual classifiers. Ensemble methods aim at improving the generalizability and the robustness over single classifiers.

Below is a list of the different ML algorithms that we have considered and tested in this study:

- Single classifiers
  - Support Vector Machine (SVM) – Cortes & Vapnik, (1995)
  - Decision tree – Quinlan, (1987)
  - *k*-nearest neighbor – Altman, (1992)
  - Logistic regression (which despite its name is a classification method) – James *et al.* (2014)
- Ensemble methods – based on randomized decision trees (Breiman, 1998)

- Random Forest – see e.g. Ho, (1998), Breiman, (2001)

- Extremely Randomized Trees (or extra-tree) classifier (Geurts et al, 2006)

Both of these methods use decision trees as their base classifier and combine the individual classifiers by averaging their probabilistic prediction, but differ in the details of how the splits on which the single decision trees are trained are computed.

We invite the reader to refer to the excellent documentation of the Python Machine Learning library 'scikit-learn', which provides high-level interfaces to most Machine Learning algorithms (Pedregosa, 2011). We especially refer the reader to the section on Supervised Learning, available at: http://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

Beyond the choice of the potential predictor set and the choice of the algorithms, other choices need to be made, and therefore tested, notably:

- The choice between different options for **scaling the predictors**: i.e. in ML, for many models (especially linear estimators) it is usually a good idea to scale the predictors, but many different types of scalers are available, i.e. a standard scaler (mean 0, stdev to 1) or a min-max scaler, etc.

- The potential processing of the features (predictors): i.e. features combination via e.g. polynomial functions, pruning via features selection, etc.

- The choice of different tuning **parameters** (or **hyperparameters**), which are the parameters of the model that are *not* learned from the data, but must be set *a priori* by the user, (i.e. the *regularization parameter* in Logistic Regression, the *depth parameter* in a Decision Tree, the *kernel* type (e.g., linear, Radial Basis Functions) in Support Vector Machine).

It then becomes obvious that the space of all available 'pipelines' (i.e. set of combinations of predictor sets, predictors scaling and processing, classification algorithms, multi-dimensional vector of hyper-parameters) becomes extremely large and mostly intractable computationally, precluding a *systematic* evaluation of all the possible combinations to determine the '*optimal*' pipeline.

In this study, we used **genetic programming** methods (Olson *et al.,* 2016) to essentially 'evolve' the best prediction pipelines. A comprehensive treatment of this approach is also beyond the scope of this document, but in a nutshell genetic programming belongs to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover. In this particular case, the possible choices for all the components of a prediction pipeline can be considered as individuals in a population, allowed to 'mutate' and 'reproduce', with the 'fitness' of a solution evaluated according to a criteria, in this case the forecast accuracy.

2.3.3   Evaluation of the prediction pipeline's accuracy

Once the best prediction pipeline was evolved, the performance of the latter was evaluated independently using stratified k-fold cross-validation in order to have a more robust estimation of the true predictive power of the pipeline.

In stratified k-fold cross validation, the dataset, consisting of the features (predictors) and the target (predictand) restricted to January 1950 – December 2014, is divided N times (in this particular instance, we've chosen N=1000) into a training set and a testing set. Here the training set is chosen to represent 80% of the data, and the testing set the remaining 20%. The particular instances (samples, observations) in the training and testing sets are chosen at random with the constraint that the random draw respects as much as possible the class distribution: i.e. in the case of tercile categories as considered here, the proportion of each class (target) in the training and testing sets should be approximately 33%, to ensure that the algorithm is trained using a *balanced* dataset.

For each of the 1000 iterations, the evolved pipeline (potentially including scaling, features selection or combination, algorithm fitting with optimized hyperparameters) is trained using only the observations in the training set, then evaluated independently over the testing set.

The evaluation metric used here is simply the accuracy (i.e. the proportion of samples in the testing set that have been correctly classified). A forecast was deemed correct if the tercile with the highest probability of occurrence eventuated. The accuracy can vary between 0 (i.e. all samples are mis-classified) to 1 (all samples are correctly classified, a perfect forecast) and usually expressed in percentage. We must remember that the accuracy needs to be compared to what would be a 'climatological' forecast, obtained by guessing randomly the class attribution of each sample. Given the class distribution, such a 'random' forecast would give (when numerous samples are considered) a mean accuracy of ~ 0.33 (33%) as we have – by construction – three equiprobable classes ('below normal', 'normal', 'above normal'). Because the climate system is a complex dynamical system, it is theoretically impossible to obtain 100% forecast accuracy, but a reasonable forecast accuracy would need to exceed 33% to be of any practical use for decision-making and risk management.

In order to represent the potential variations in forecast accuracy, the results show not only the average forecast accuracy across the 1000 iterations of the stratified k-fold cross-validation exercise but also its distribution and a non-parametric fit using Gaussian Kernel Density Estimation (KDE).

2.3.4 Independent evaluation of the prediction pipelines over the recent period (January – December 2015)

In addition to the evaluation of the forecast accuracy using stratified k-fold cross validation over the 1950–2014 period (for the climate station data) or the 1972–2014 (for the VCSN-based regional indices) period, we also evaluated the accuracy of the prediction pipelines independently for the January–March 2015 to October–December 2015 seasons for the VCSN-based regional indices.

3 Results

Prior to the results on the classification pipelines, we present very briefly below the methods and results from the initial analyses using regression approaches: i.e. whereby the predictand is the time-series of seasonal temperatures anomalies (in °C) and precipitation (in mm).

3.1 Some results from regression modelling

For the regression analyses, we used both the climate indices that were investigated in the initial Hawke's Bay climate report (ENSO, IOD, SAM, SPSD (The IPO was not considred due to the long time-scale of this cycle); Fedaeff and Fauchereau, 2015) and the PCs coming from the EOF analyses on SST and combined SST-Z850 fields (Section 2.3) as potential predictors.

The indices were considered individually rather than in combination as they are, to some extent, correlated to each other (see Figure 5), a property undesirable for the independent variables of many regression algorithms. In contrast, the PCs coming from the EOFs on SSTs or joint SST-Z850, are by construction orthogonal to each other and therefore uncorrelated (at 0 lag, some complex lead-lag linear relationships between PCs maybe observed, but will be of no concern here).

The results of the regression analyses have been evaluated using Leave One Out cross-validation (i.e the regression model is successively fitted using all but one observation, and predicts the left-out sample) and the metric used is simply the correlation coefficient between the vectors of observed and predicted values.

In all cases presented here, seasonal averages were calculated from the monthly anomalies, and a lag of 0 month was applied, e.g. the prediction of the seasonal rainfall or temperature for the June-July-August period is carried out using the climate mode indices or the PCs up to and including May.

3.1.1   Regression using climate modes indices as predictors

In all cases (i.e. either Hawke's Bay regional VCSN-based indices or climate station rainfall and temperatures), the correlation coefficients between observed and predicted values barely exceeded 0.4. This confirmed our initial assessment that indices of climate modes are not necessarily the best choice in building a forecasting model. While they reveal potential mechanisms at play, they most likely do not represent a large enough part of the total variance of the climate system to be used as predictors operationally.

3.1.2   Regressions using SST or SST-Z850 PCs as predictors

We tested several regression methods using, successively, the available sets of PCs coming from the EOF analyses on SST or joint SST-Z850 fields as potential predictors. We excluded the sets that included the lagged (i.e. from -1 to -6 months) versions of the PCs, as in this case the variables included in the predictor set are not statistically independent (i.e. a PC is significantly correlated to itself at lag of 1 – 6 months).
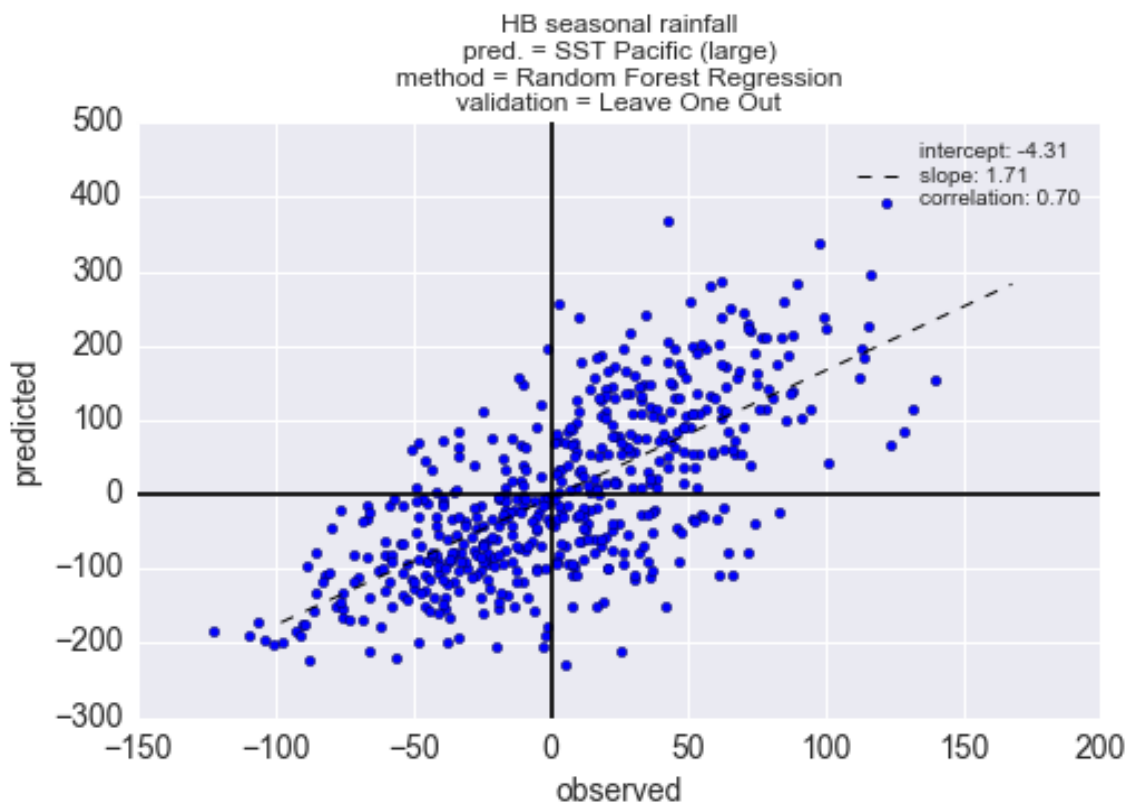
Because the number of features can be large for some sets (generally > 10 PCs), we also tested two methods that penalise for the complexity of the model: the so-called 'Ridge-regression' (or Tikhonov regularization) method tries to reduce the size of the coefficients (Hoerl & Kennard, 1970), while the LASSO (Least Absolute Shrinkage and Selection Operator) method also tries to reduce the size of the coefficients and/or return sparse coefficients (Kukreja *et al.,* 2006).

In addition to these linear methods (Linear Regression and its penalised versions: Ridge Regression and LASSO) we also tested Random Forest Regression; an extension of the Random Forest classification method to regression problems.

Of all the regression methods and the sets of PCs, a Random Forest Regression model using the set of PCs coming from the EOF on monthly SST anomalies over the larger Pacific domain (60°S, 40°N, 120°E, 70°W) performed the best. The mean correlation coefficient (between predicted and observed values, using Leave One Out cross-validation) was 0.78 for seasonal temperatures and 0.7 for seasonal rainfall when the Hawke's Bay regional indices (VCSN based) were considered as predictands. Figures 8 and 9 present respectively the corresponding scatterplots of predicted vs. observed seasonal temperature and rainfall anomalies.

Figure 8: Predicted vs. observed seasonal temperature anomalies (using Leave-One-out cross validation) for the Hawke's Bay regional VCSN-based index.
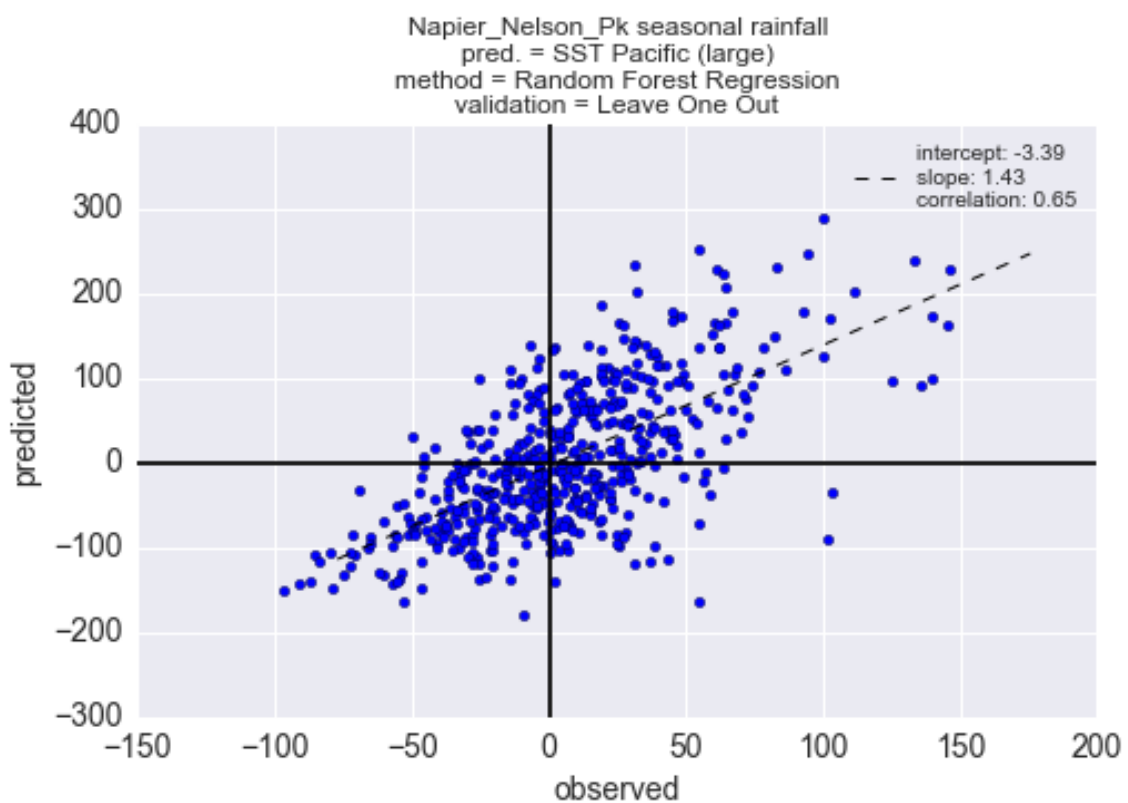

Figure 9: Predicted vs. observed seasonal rainfall anomalies (using Leave-One-out cross validation) for the Hawke's Bay regional VCSN-based index.
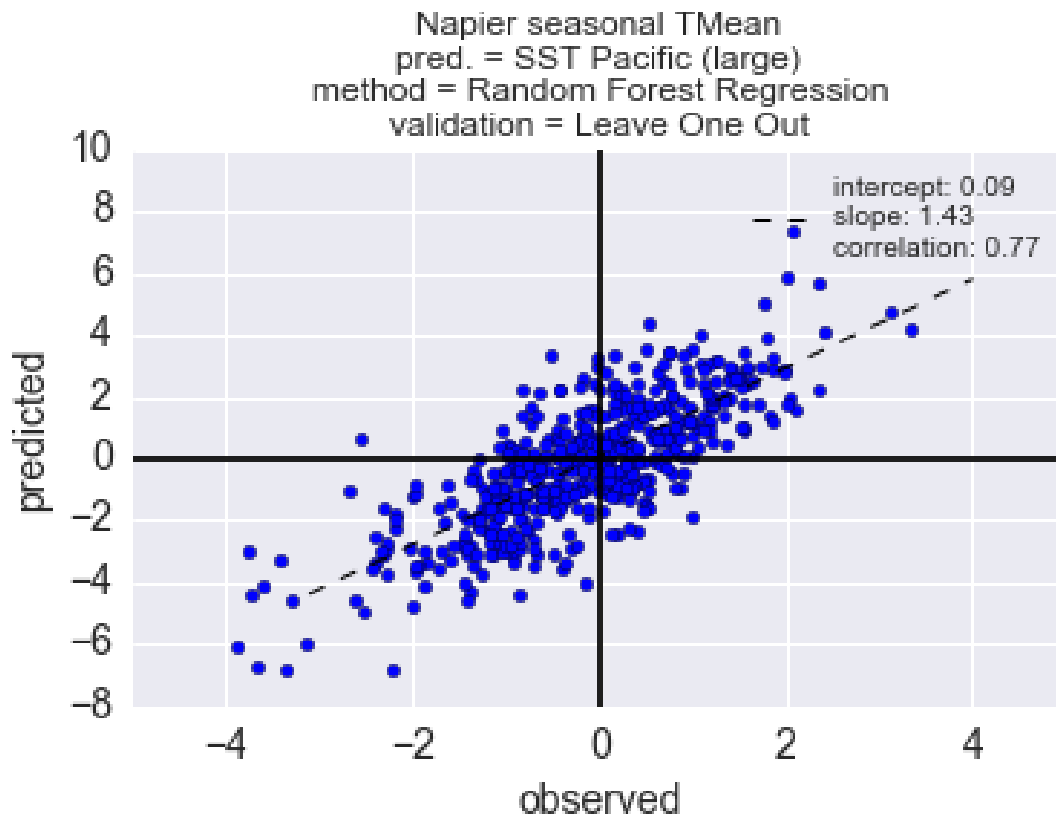
For individual stations, regression models generally yield slightly lower correlations (note that to allow for comparison with the Hawke's Bay VCSN-based regional indexes, the station time-series were truncated to consider only data from 1972).

Correlation coefficients between predicted and observed station seasonal rainfall ranged from 0.64 (Aramoana) to 0.69 (Anawai).

For seasonal temperatures, correlation coefficients varied between 0.75 (Wairoa and Kopua) and 0.77 (Napier). As an illustration, Figures 10 and 11 show the scatterplots of predicted vs. observed seasonal rainfall and temperature anomalies (using Leave-One-out cross validation) for Napier.

Figure 10: Predicted vs. observed seasonal rainfall anomalies (using Leave-One-out cross validation) for Napier.

Figure 11: Predicted vs. observed seasonal temperature anomalies (using Leave-One-out cross validation) for Napier

While these results are encouraging, one must keep in mind that Random Forest Regressors are known for overfitting when used for regression. Moreover, while the transformation from a continuous forecast into terciles categories ('below normal', 'normal', 'above normal') is trivial, estimating the probabilities associated with each category is not straightforward.

3.2 Results using classification pipelines

A detailed exposition of all the results for all the combinations of potential predictors and predictands would be too lengthy, so only the best results are presented here.
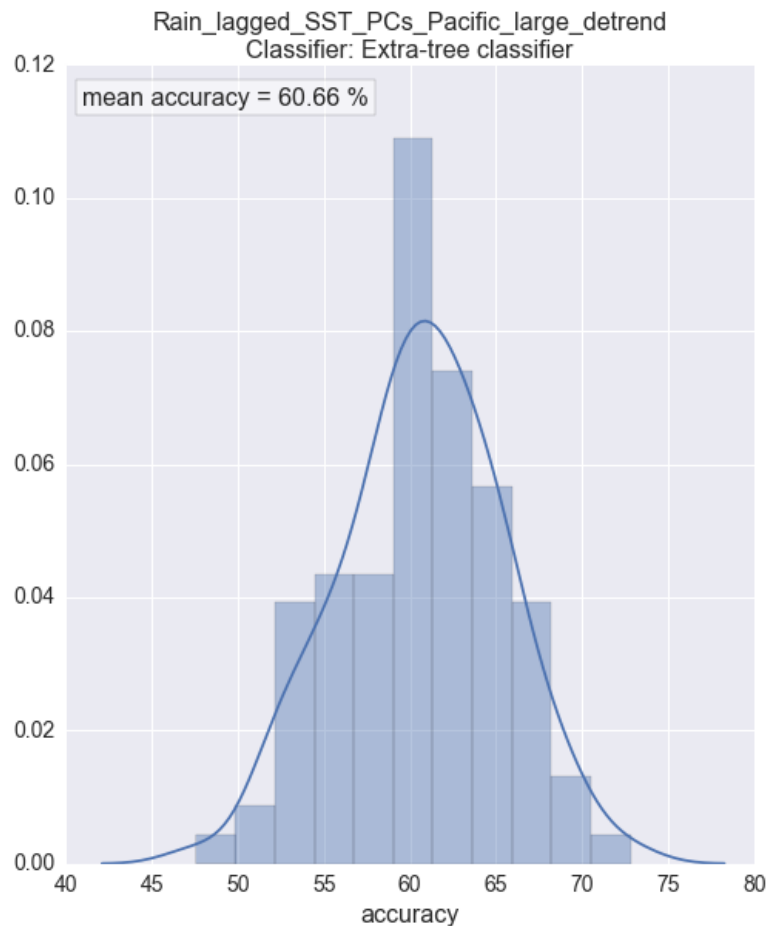
For classification, we first considered the Hawke's Bay regional indexes from the VCSN dataset, and then the climate station data (as was done for regression analysis).

Different pipelines were optimized by use of genetic programming using the PCs coming from the EOF analyses on SST monthly anomalies and joint SST- Z850 monthly anomalies as predictor sets. In all cases presented here, seasonal averages were calculated from the monthly PC values, and a lag of 0 months was applied, e.g. the prediction of the seasonal rainfall or temperature category for the June-July-August period was carried out using the PCs up to and including May.

From an operational perspective, this means that forecasts can be made available within the first 5 days of the first month of the forecast, allowing for the lag to real-time of the SST and/or Z850 datasets (e.g. in the case of SST, the ERSST dataset is usually updated on the 5th day of each month).

3.2.1 Hawke's Bay regional rainfall index results

The best prediction pipeline for the Hawke's Bay region VCSN-derived rainfall index was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain (60°S, 40°N, 120°E, 70°W). The mean accuracy over the 1000 folds was ~61%. Figure 12 presents the full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution.

Figure 12: Full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution for the best prediction pipeline for the Hawke's Bay region VCSN-derived rainfall index. This was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain (60°S, 40°N, 120°E, 70°W).
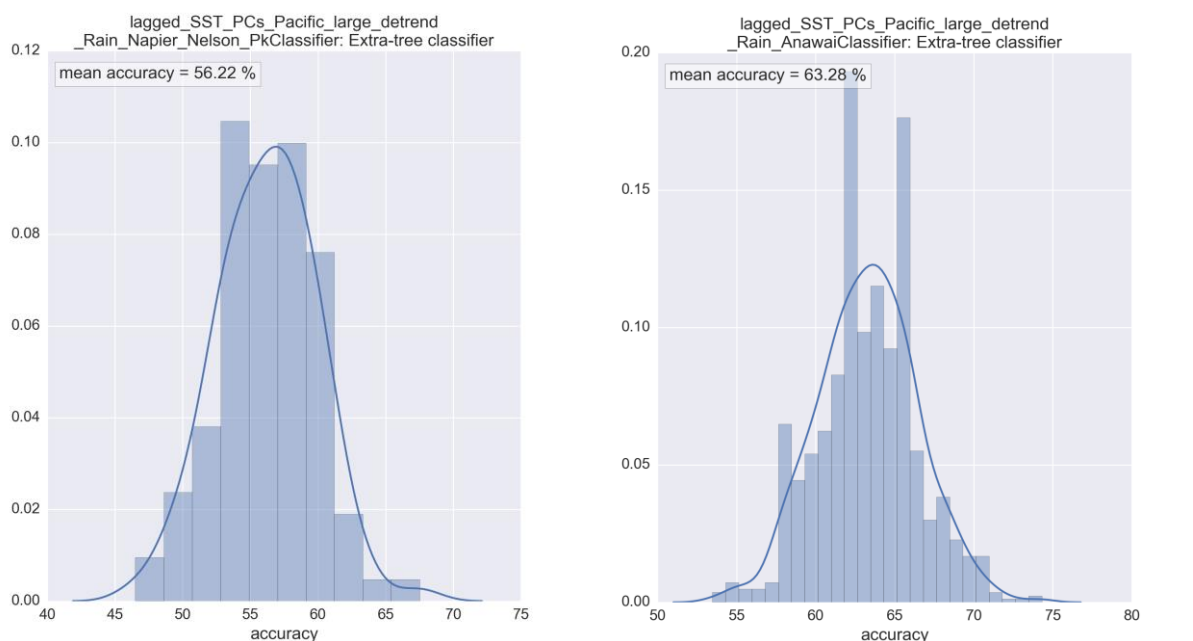
The independent validation of the same classification model over the period January–March 2015 to October–December 2015 however gave an accuracy score of ~50%. This is below the average of the forecast accuracy that was obtained over the development set (1972–2014).

3.2.2 Hawke's Bay station rainfall results

For the 12 selected rainfall stations (Table 1) all seasons available after January–March 1950 to present were considered in building and testing the classification pipelines.

In order to reduce the computational requirements, 100 iterations for the cross-validation procedure were used instead of 1000. Comparisons indicated that this did not significantly affect the validity of the results.

The mean accuracy of seasonal rainfall forecasts varied between ~56% and ~63% depending on the station. Figure 13 shows the distribution of the accuracy scores for Napier (Nelson Park) (lowest accuracy at 56.22%) and Anawai (highest accuracy at 63.28%). Table 3 gives the mean accuracy obtained for each of the rainfall stations.

Figure 13 Full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution for the best prediction pipeline for the Napier and Anawai rainfall stations. This was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain (60°S, 40°N, 120°E, 70°W).
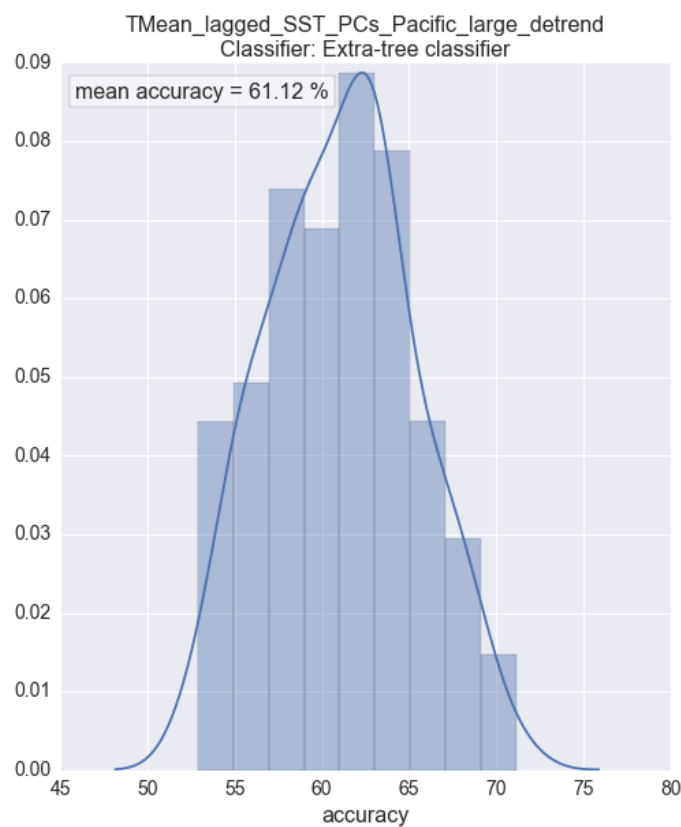
Table 3: Mean accuracy obtained for each of the rainfall stations using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain (60°S, 40°N, 120°E, 70°W).

Station (rainfall)	Mean accuracy (%)
Anawai	63.28
Aramoana	58.77
Eskdale Hedgeley	58.28
Flemington composite	59.14
Gwavas composite	57.68
Havelock North composite	59.09
Koparakore Patangata	58.03
Mt Vernon composite	56.55
Napier composite	56.22
Rose Hill	57.26
Tutira composite	56.86
Wairoa,Waiputaputa Stn	58.20

3.2.3   Hawke's Bay regional temperature index results

As was the case for rainfall, the best prediction pipeline for the Hawke's Bay region VCSN-derived temperature index was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain. The mean accuracy of seasonal temperature forecasts over the 1000 folds was ~61%. Figure 14 presents the full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution. Note that the accuracy scores are less widely distributed than for the Hawke's Bay region VCSN-derived rainfall index (see Figure 12). This indicates that it is reasonable to expect the seasonal mean temperature forecasts to be more consistent (i.e. the accuracy can be expected to present less variability) than the rainfall forecasts when operationalised.
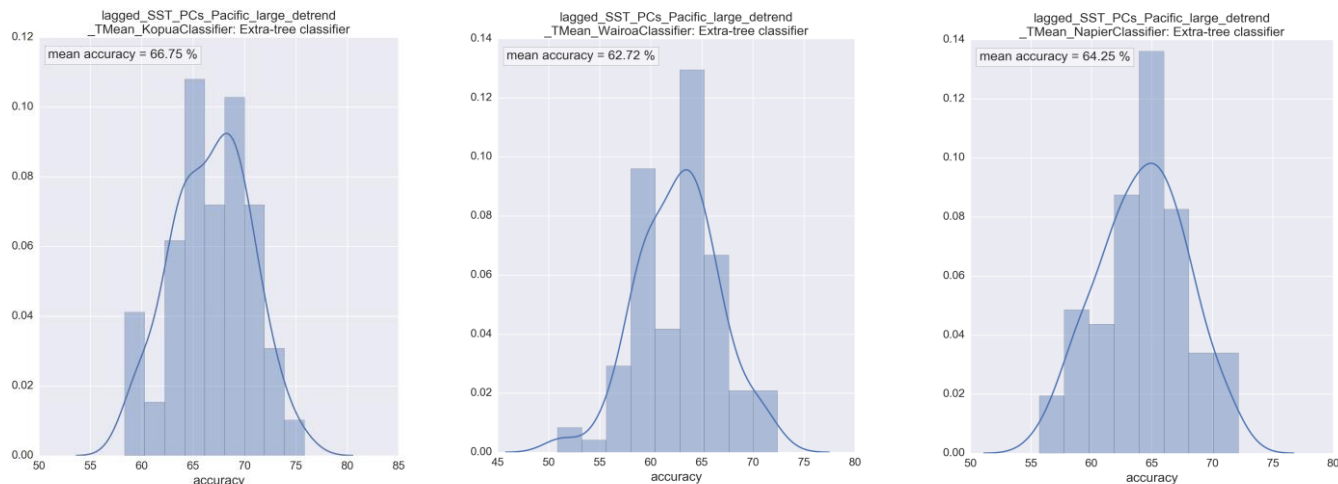

Figure 14: Full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution for the best prediction pipeline for the Hawke's Bay region VCSN-derived temperature index. This was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain (60°S, 40°N, 120°E, 70°W).

The independent validation of the classification model over the period January–March 2015 to October–December 2015 gave an accuracy score of 72%. This is above the average accuracy obtained by cross-validation over the development set (1972–2014).

3.2.4   Hawke's Bay station temperature results

The same methodology was applied to analyse station temperature records as was done for rainfall in Section 3.2.2. Only 3 stations could be considered for seasonal mean temperatures: Kopua, Napier and Wairoa.

Figure 15 presents the distributions of the accuracy scores for the three temperature stations and Table 4 gives the mean accuracy obtained at each site.

Figure 15: Full distribution of the accuracy scores, as well as a Kerned Density Estimation of the true distribution for the best prediction pipeline for the Kopua, Wairoa and Napier temperature stations. This was obtained using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain domain (60°S, 40°N, 120°E, 70°W).

Table 4 Mean accuracy obtained for each of the temperature stations using the lagged PCs from the EOF analysis of monthly SST anomalies in the larger Pacific domain domain (60°S, 40°N, 120°E, 70°W).

Station (TMean)	Mean accuracy (%)
Kopua	66.75
Napier composite	64.25
Wairoa composite	62.72

4   Conclusions

After carrying out various analyses, the results show that there is significant potential to implement an operationally tailored seasonal climate forecast for the Hawke's Bay region.

For seasonal mean temperature, a reasonable accuracy was obtained when the set of lagged PCs associated with an EOF analysis on SST anomalies over the larger Pacific domain was used as a set of predictors, and the extra-tree classifier (extremely randomized Trees) was used as the prediction algorithm. The accuracy was about 61% for the Hawke's Bay region VCSN-derived temperature index, and varied between ~62 and ~67% for individual stations (with the caveat that only three stations were available for temperatures). It should be noted that as this forecasting scheme is based on tercile probabilities, the theoretical accuracy of a random forecast (i.e. if you were to guess at random) is 33%.

For seasonal cumulative rainfall, the best selected pipeline included the same classification algorithm (extra-trees classifier) and set of predictors as for temperature. The accuracy was slightly lower than

for seasonal temperatures (~60 % for the Hawke's Bay regional index, and varying between 56% and 63% for the station rainfall), it also shows a larger variability between the iterations of the cross-validation procedure, and the model did not perform as well for the post-2015 independent validation period.

We carried out a quick comparison of these results and the NIWA Seasonal Climate Outlook (SCO). The NIWA SCO provides a forecast for 6 regions in New Zealand. Hawke's Bay falls into the "East of the North Island" region (which also includes Gisborne and Wairarapa). From May 1999 to present, the forecast provided by NIWA's SCO for the East of the North Island was correct 46% of the time for temperature and 52% of the time for rainfall when applied to the Hawke's Bay region (using the VCSN index for validation). The methodology used to calculate the accuracy of the SCO differs slightly to how the accuracy was calculated for the forecasting scheme in this report. The comparison presented here is provided as a quick guideline and a more thorough validation may be required.

If the scheme presented developed in this study was operationalised, forecasts would be expressed as tercile probabilities (i.e. probability of seasonal cumulative rainfall and seasonal mean temperatures falling in the 'below normal', 'normal' or 'above normal' categories) for the whole region (based on the VCSN index) and/or for 12 rainfall stations and three temperature stations.

These forecasts could be made available on the 5th day of each first month of the 3-month 'season' (i.e. forecast for July-August-September could be made available on the 5th of July). An earlier (e.g. towards the end of June) release could possibly be implemented if deemed necessary by using preliminary Sea Surface Temperature data derived from available daily near-realtime datasets.

A simple operationalisation of this statistical forecasting schemes would involve:

> i) Scheduled download and processing of the SST data, and updating of the PCs values;
> ii) Processing of the updated PC set (i.e. calculation of the last 3-months anomalies, standardization);
> iii) Feeding the updated PC values into the classification models: one model is trained for each variable (Rainfall, Temperature) and each station as well as each Hawke's Bay regional index (one for rainfall and one for temperature), i.e. a total of 12+1 models for rainfall, and 3+1 models for temperature;
> iv) Delivering the tercile probabilistic forecasts for the Hawke's Bay region and for the individual stations: this could take the form of tables, pie-charts, and – in the case of station-based forecasts – could possibly be visualised as an interactive map (an example of interactive map, developed for Samoa for another project, is available at: https://rawgit.com/nicolasfauchereau/drought_risk/master/notebooks/map.html)

Ideally, and for the purpose of continuous 'real-time' validation of the forecasts, the time-series of observed temperatures and rainfall also need to be updated on a regular basis. This could be easily achieved for the VSCN rainfall and temperature index but not for station data as the majority of climate stations in the region are manual. Because of this it takes at least two months for it to be entered in the National Climate Database (http://cliflo.niwa.co.nz/).

We propose that upon reaching a suitable agreement between NIWA and HBRC, an experimental operational scheme could be implemented as early as September 2016, and delivery of the forecasts to the Hawke's Bay regional council to start from October 2016. We would recommend that release of forecasts to the public is delayed until value (i.e. reasonable forecast accuracy) is demonstrated over a period of six months. The accuracy of this forecast could also be evaluated against the performance of the NIWA Seasonal Climate Outlook.

5   References

Altman, N. S. (1992) An Introduction to Kernel and Nearest Neighbor Nonparametric Regression. *The American Statistician, 46*(3): 175-185.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, *26*(3), 801-849.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cortes, C., Vapnik, V. (1995) Support-vector networks. *Machine Learning, 20*(3): 273-297.

Fedaeff, N. & Fauchereau, N. 2015. Relationship between climate modes and Hawke's Bay seasonal rainfall and temperature. NIWA Client Report: AKL2015-016. 50pp

James, G., Witten, D., Hastie, T., Tibshirani, R. (2014): An Introduction to statistical learning: with applications in R. Springer, New York.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3-42.

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, *27*(9), 1119-1152.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Huang, B., Thorne, P.W., Smith, T.M., Liu, W., Lawrimore, J., Banzon, V.F., Zhang, H.M., Peterson, T.C. & Menne, M. (2016). Further exploring and quantifying uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). *Journal of Climate*, *29*(9), 3119-3142.

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

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., ... & Zhu, Y. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, *77*(3), 437-471.

Kukreja, S. L., Löfberg, J., & Brenner, M. J. (2006). A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, *39*(1), 814-819.

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., & Moore, J. H. (2016). Automating biomedical data science through tree-based pipeline optimization. In *European Conference*

on the Applications of Evolutionary Computation (pp. 123-137). Springer International Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

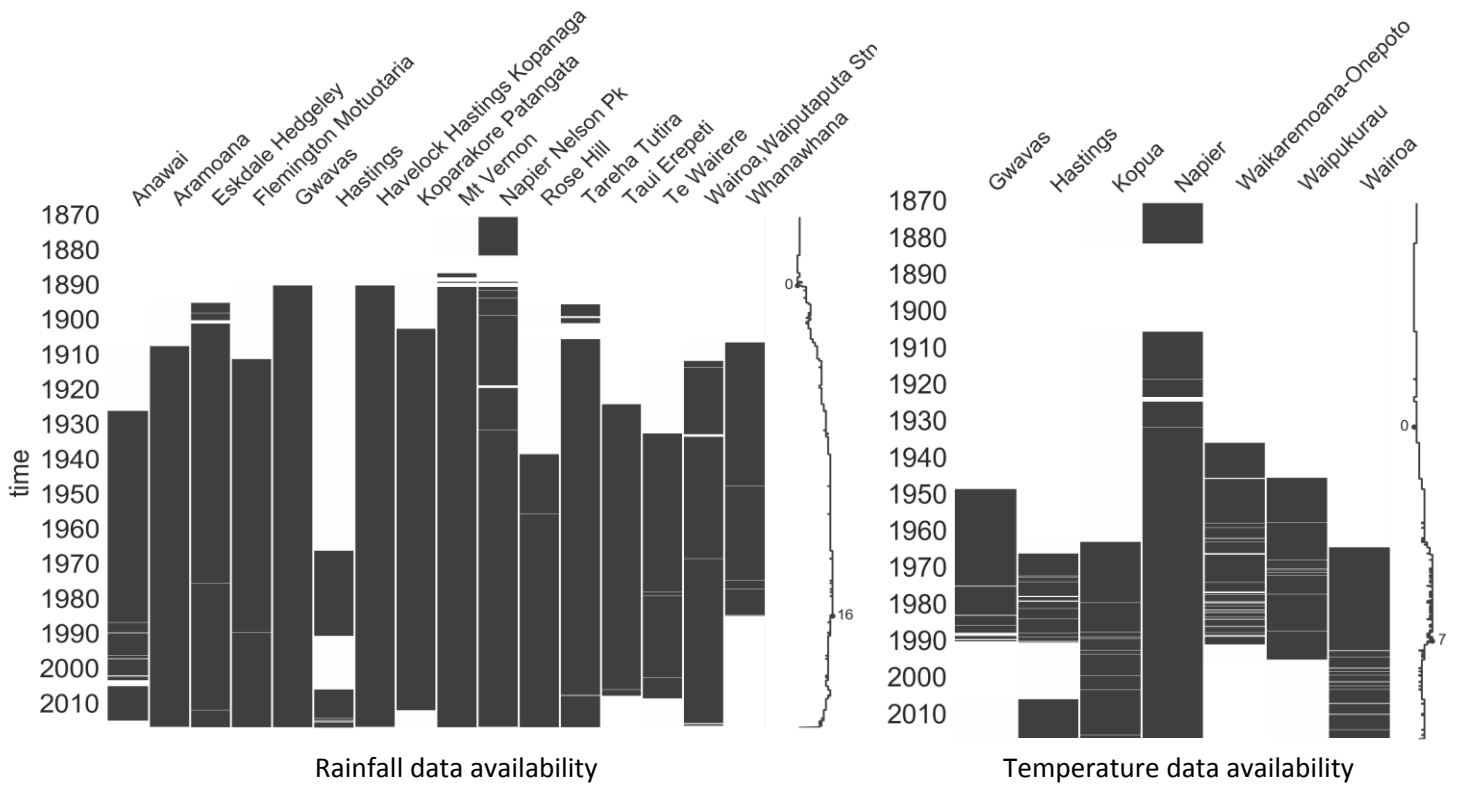Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data* (Vol. 72). John Wiley & Sons.

Tait A., Henderson; R, Turner R, Zheng X (2006) Thin-plate smoothing spline 1034 interpolation of daily rainfall for New Zealand using a climatological rainfall 1035 surface. International Journal of Climatology 26: 2097–2115

Tait, A. (2008). Future projections of growing degree days and frost in New Zealand and some implications for grape growing. *Weather and Climate*, *28*, 17-36.Troccoli, A. (2010). Seasonal climate forecasting. *Meteorological Applications*, *17*(3), 251-268.

Quinlan, J. R. (1987) Simplyfing decision trees. *International Journal of Man-Machine Studies,* 27(3): 221-234.

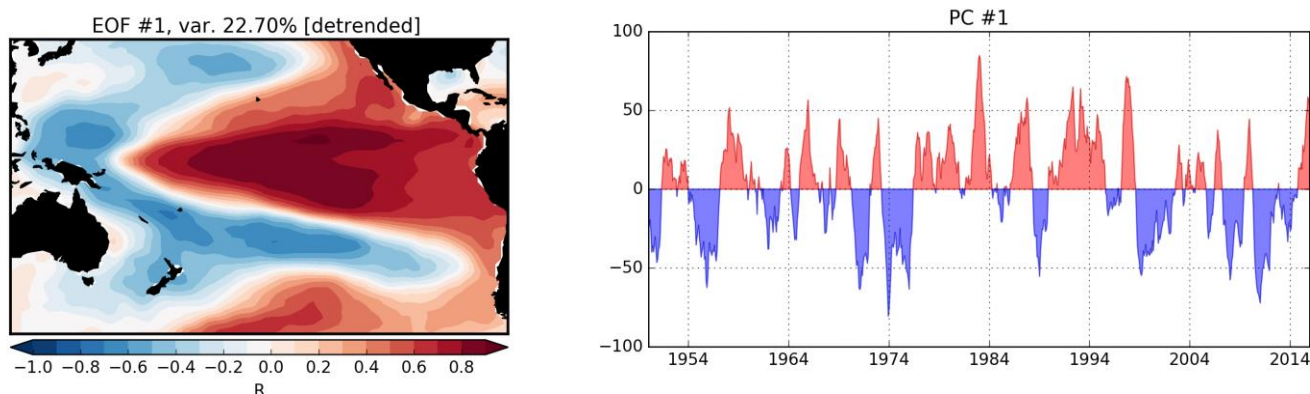# Appendix A   Climate station data availability

For rainfall, the Hastings record was excluded for further analysis as 32% of data was missing. The Whanawhana, Te Wairere and the Taui Erepeti composite records were also not used as data did not cover the 1981–2010 climatology period. For temperature only Napier, Kopua and Wairoa had records with sufficient data.
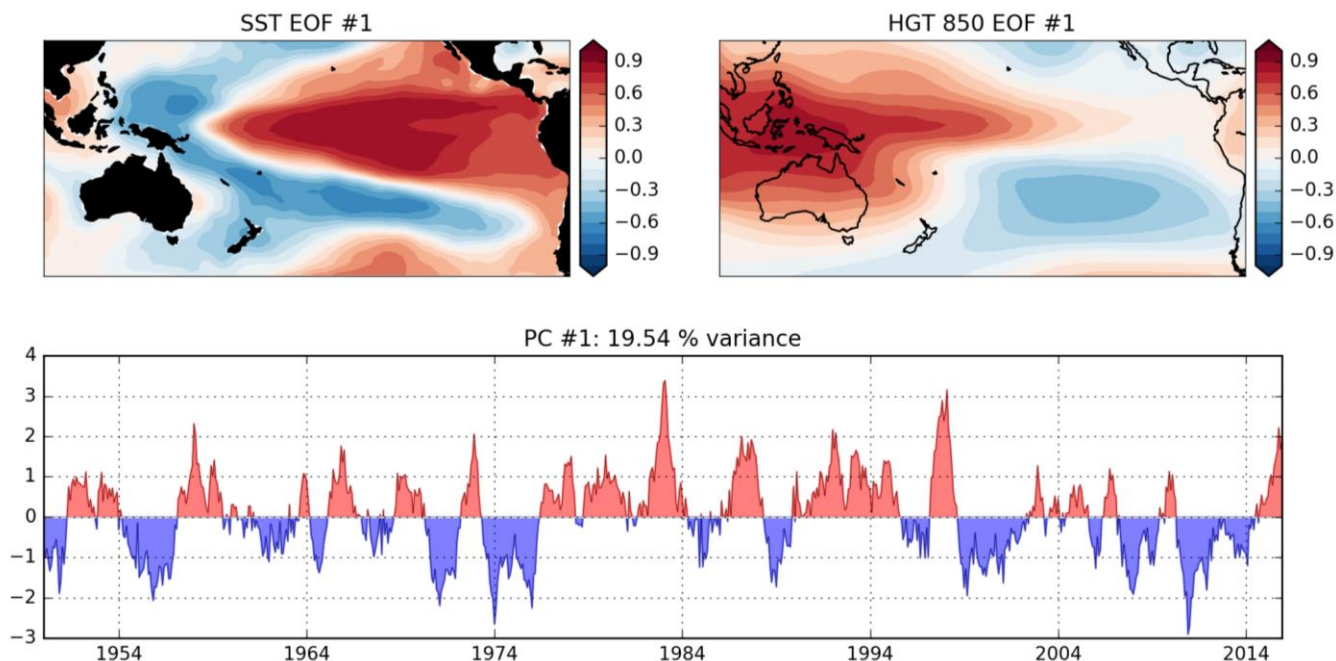

Rainfall data availability

Temperature data availability

Appendix B   Selection of outputs from EOF analyses

Below is a selection of Figures showing the EOFs (spatial patterns) and PCs (time-series) coming from the EOF analyses performed respectively on SST monthly anomalies over the larger Pacific domain, the combined SST-Z850 monthly anomalies over the larger Pacific domain, and the combined SST-Z850 monthly anomalies over the regional New Zealand domain. In all cases the EOF (spatial patterns) are expressed as correlations between the corresponding PC and the original field, and the PCs are expressed in standardized units.
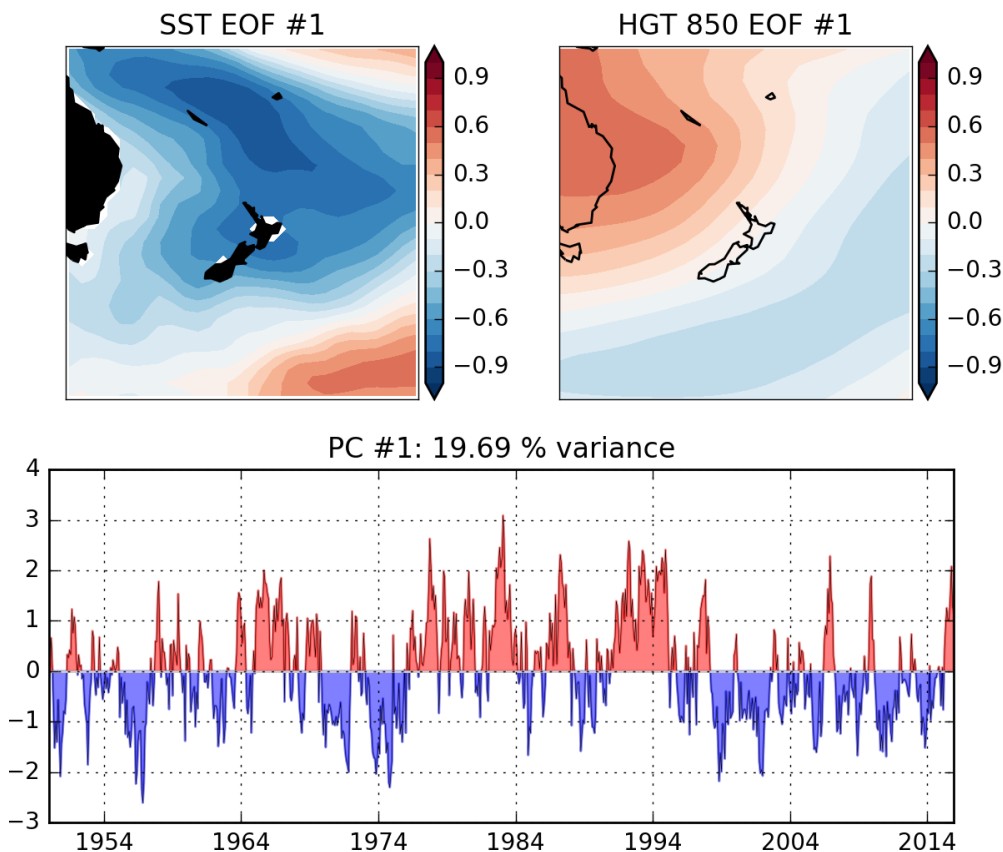

The Figure above presents the 1st EOF / PCs coming from the EOF analysis of monthly SST anomalies over the larger Pacific domain (60°S, 40°N, 120°E, 70°W). The oceanic component of the El Niño Southern Oscillation pattern is clearly recognizable. This mode explains more than 22% of the original variance contained in the original field of detrended monthly SST anomalies.

The Figure above presents the spatial patterns (EOFs, respectively for SST on the left and Z850 on the right) and time-series (PC, bottom) from the EOF analysis performed on the combined SST and Z850 fields for the larger Pacific domain. The SST pattern is indistinguishable from the EOF on SSTs alone (see above), while the spatial

pattern for Z850 represents the Southern Oscillation, i.e. the atmospheric component of the El Niño Southern Oscillation, usually monitored by the Southern Oscillation Index (standardized difference between MSLP in Tahiti and Darwin).

The Figure above presents the spatial patterns (EOFs, respectively for SST on the left and Z850 on the right) and time-series (PC, bottom) from the EOF analysis performed on the combined SST and Z850 fields for the regional New Zealand domain. The EOF pattern for SST shows negative SST anomalies to the north and around NZ and positive SST anomalies to the southeast of the domain. These SST anomalies are associated with higher pressures than normal over Australia and north Tasman and negative pressure anomalies to the SE. Together these patterns mainly consist in the regional expression of the ENSO signal. The corresponding PC is correlated at +0.73 with the 1st PC of the combined SST-Z850 fields in the Pacific region (i.e. the coupled ENSO signal) and +0.58 with the standard NINO 3.4 SST index.