INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **26**: 485–504 (2006)
Published online 21 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.1261

ANALOGUE FORECASTING OF NEW ZEALAND CLIMATE ANOMALIES

A. BRETT MULLAN* and CRAIG S. THOMPSON
National Institute of Water and Atmospheric Research, Wellington, New Zealand

Received 9 July 2004
Revised 25 July 2005
Accepted 25 July 2005

ABSTRACT

An analogue forecast scheme is described for multifield prediction of monthly and seasonal New Zealand climate anomalies on the basis of the methodology of Livezey and Barnston (1988) for US seasonal temperatures. The method is applied to predicting terciles of temperature and precipitation for six regions of New Zealand. Empirical orthogonal function analysis is used to reduce sea surface temperature and sea-level pressure predictors down to a set of five independent indices, which incorporate variations due to El Niño-Southern Oscillation, Indian Ocean sea temperatures and a wave 3 pattern in the Southern Hemisphere westerlies. A full bootstrap cross-validation procedure is carried out, along with Monte Carlo tests, to assess the skill of the method on independent data and to determine the significance of the results. Significant skill is found for seasonal temperature forecasts for the summer and winter seasons; there is less success in predicting monthly temperatures or rainfall at either timescale. Considerable care is required to constrain the climate state vector, from which analogues are defined, and to constrain the search procedure itself, in order to produce results that are stable with respect to small parameter changes in the model. For the New Zealand region, 5 to 7 is found to be the optimum number of 'closest analogues', and the inclusion of anti-analogues improves the predictions, at least in the seasonal case. Skill in predicting regional temperature and rainfall is shown to be related to a combination of skill in predicting sea-level pressure patterns and to how strongly these patterns project onto temperature and rainfall anomalies. Copyright © 2005 Royal Meteorological Society.

KEY WORDS: New Zealand; long-range forecasting; analogues; bootstrap cross-validation; ENSO; EOF analysis; rainfall; temperature

1. INTRODUCTION

The use of analogues to generate climate forecasts is an attractive idea, not the least because of its conceptual simplicity, and many meteorological institutions around the world either still use analogue forecasts or have done until recently (WMO, 2002). The analogues may be used to forecast temperature and precipitation anomalies directly, or through the intermediary of an anomaly flow pattern. In some cases, the analogues are used more indirectly: e.g. the Australian Bureau of Meteorology determines analogue years of the Southern Oscillation Index (SOI) as a first step to subsequent analyses (Drosdowsky, 1994; Voice *et al*., 1996).

The technique involves searching the historical data, identifying previous periods that resembled the immediate past period and predicting the following month's or season's climate anomalies on the basis of what happened on those previous occasions. Analogues have been used in climate forecasting for a long time. Namias (1968) reviews the early history, and Nicholls (1980) presents a somewhat more recent view. There was a resurgence of interest in analogue techniques at the end of the 1980s, particularly in the United States, with the publication of Livezey and Barnston (1988) and subsequent papers (Barnston and Livezey, 1989; Livezey *et al*., 1990; Chapman and Walsh, 1991; Barnston and van den Dool, 1993; Livezey *et al*., 1994). The ongoing interest in analogue techniques is evidenced by a recent paper (Wetterhall *et al*., 2004) that uses analogues to downscale general circulation model precipitation predictions over Europe.

*Correspondence to: A. Brett Mullan, National Institute of Water and Atmospheric Research, Kilbirnie, Wellington, New Zealand; e-mail: b.mullan@niwa.co.nz

Analogues have even been tested as a forecast tool on the daily timescale (Gutzler and Shukla, 1984; Van den Dool, 1989, 1994). Fraedrich *et al.* (2003) describe an analogue model that shows skill in forecasting tropical cyclone tracks in the Australian region. Van den Dool (1994) considered how close a match could be expected between daily flow patterns. According to his estimates, a library of 10^{30} years would be needed in order for two observed flows to match within observational error over the entire Northern Hemisphere. Obviously, a weaker requirement could be placed on a 'match', or the geographic area under consideration could be reduced. What this calculation does show is the importance of the number of degrees of freedom when identifying suitable analogues.

Analogy can be defined in a number of ways. Gutzler and Shukla (1984) define three different measures, applied in their case to daily winter flow patterns at 500 hPa. Analogues were defined in terms of (1) the root-mean-square difference between pairs of maps; (2) the spatial covariance between two maps over the region of interest and (3) the spatial correlation. The first measure is most sensitive to anomaly amplitudes, the second to matching up pronounced highs and lows, whereas the third is more sensitive to the phase of the patterns with less emphasis on amplitudes. Each measure will tend to generate a somewhat different set of analogues, although there will certainly be commonly occurring analogues, and it is not clear whether there is an optimum way in which analogue similarity should be identified. Barnett and Preisendorfer (1978) called this type of analogy definition the *classical* approach and described two other methods that took account of the evolution of the forecast and observed states. In their comparative study, the use of state evolution in defining an analogue sometimes gave a superior prediction and at other seasons and lead times gave a worse result.

For monthly and seasonal forecasting, the emphasis is no longer entirely on initial atmospheric conditions, and any anomalies at the lower boundary become important. For the Southern Hemisphere, this means primarily sea surface temperature (SST) anomalies, although the state of seasonal sea ice around Antarctica may also play a role. Numerical model simulations have demonstrated that both initial conditions *and* boundary conditions must be considered (Palmer and Anderson, 1994). The predictors developed in the New Zealand context are therefore drawn from indices describing both the mean sea-level pressure (MSLP) anomalies and the SST anomalies at the start of the forecast period.

There is also a link between analogue forecasts and ensemble forecasts, which are at the forefront of long-range forecasting research. Ensemble integrations involve the use of numerical weather prediction models where the simulations are extended well beyond the limit of predictability of individual synoptic weather systems. Each integration starts from a slightly different initial atmospheric state: these may be perturbations about an analysis at the same starting time, or forecasts initialised at staggered times. Whatever the method of generating the ensemble, one ends up with a set of forecasts from which a probability distribution could be calculated. The analogue method also comes up with a set of forecast states, and from this point on both analogue and ensemble forecasting have the same problem: what is the best way to generate a forecast from a group of possible future states, given that some disagreement within the group is inevitable? The simplest answer is to take the ensemble average, although even here it is necessary to assess how many ensemble or analogue members should be included. Anderson and Stern (1996) argue that ensemble (or analogue) forecasts are most useful when the forecast distribution is significantly different from an appropriate climatological distribution, and simply using the ensemble mean discards valuable information.

Section 2 details the data sets used. In developing the predictors, previous research guides the selection of the most appropriate variables. A number of observational studies have identified the SOI as a useful predictor of New Zealand rainfall and temperature anomalies (Gordon, 1986; Mullan, 1995, 1996). In general, during El Niño events New Zealand experiences lower temperatures, with more precipitation in the south and west and less precipitation in the north and east of the country. Mullan (1998) has found Indian Ocean sea temperatures to have an important influence on New Zealand too. Higher SSTs in the Indian Ocean result in stronger anticyclones over and north of the North Island and more persistent northwesterly airflow onto the remainder of the country. This results in higher South Island temperatures, wetter conditions in the west of the South Island and drier conditions in the north and east of the country. The same teleconnection pattern has been identified in Australian research, where it produces drier winters in southern Australia (Drosdowsky, 1993; Smith, 1994).

The Indian Ocean SST influence is confined to the austral autumn and winter seasons. The El Niño-Southern Oscillation (ENSO) effect can occur at any time of year, but is strongest in the spring and summer half-year. (Note that Southern Hemisphere seasons are used throughout, so summer is taken as December, January and February; autumn as March, April and May; and so on.) Thus, when predicting New Zealand climate variability, it is important to allow for seasonal stratification of the data. This is catered for within the analogue model by retaining the same predictors year round, but allowing these predictors to be weighted differently from season to season according to a hindcast optimisation.

The analogue methodology is described in Section 3, and the model is now run operationally at the National Institute of Water and Atmospheric Research (NIWA). As part of NIWA's programme on climate variability, monthly climate outlook meetings are held to discuss future circulation and climate anomalies over New Zealand (Salinger, 1996). A range of guidance material is used, and the analogue model that identifies monthly and 3-monthly analogues is a very useful tool in developing a mental picture of climate developments.

2. DATA

Two rather distinct data sets were developed for this study, and they required considerable preprocessing. The *predictand* data set comprised time series of the monthly or seasonal anomalies of temperature and precipitation over New Zealand. A decision was taken early on to generate forecasts for regions within New Zealand, rather than for individual sites, which meant amalgamating station data into coherent regions. The *predictor* data set was made up of circulation anomalies local to New Zealand (MSLP) and SST anomalies over a wide extent of the Southern Hemisphere. Because of the interest in predicting flow anomaly, the MSLP was used as a predictand as well as a predictor. The analogue model was developed using data for the period 1957–1994, and additional validations were made on independent data up to 2003.

2.1. Predictands

The New Zealand station data used are monthly average temperature and monthly total precipitation, taken from the NIWA Climate Database for a wide range of locations around New Zealand. The station data set and initial processing were the same as described in Mullan (1998). A total of 51 temperature stations and 74 precipitation stations were used. The annual cycle was removed from the temperatures by subtraction of the monthly mean for each calendar month, at each station. Rainfalls were normalised by conversion to percentage deviation from 1957 to 1994. Three-monthly seasonal anomalies were generated by taking 3-month averages of the individual monthly normalisations.

In order to amalgamate the stations into regions, empirical orthogonal function (EOF) analysis was applied, followed by varimax rotation. EOF rotation, as recommended by Richman (1986), acts to isolate subgroups of stations that co-vary similarly, and thus is useful for regionalising variations. Six coherent regions were identified from the rotated EOF analysis of the rainfall data and are shown schematically in Figure 1, three regions in the North Island and three in the South Island (see Mullan, 1998, for full details of the procedure and the actual eigenvector patterns). Variations in temperature across the country can be explained by fewer rotated EOFs (Mullan, 1998, retained only three EOFs). Two of the temperature regions correspond quite closely to Regions 5 and 6 of Figure 1, with the third temperature district encompassing the entire North Island. However, in order to keep the analogue forecasts as straightforward as possible, temperature is forecast for the same six regions as for rainfall. Regional averages of temperature and rainfall anomalies comprised the basic predictand input to the analogue search model.

2.2. Predictors

The predictor data comprise MSLP and SST anomalies. The MSLP fields are taken from the NCEP/NCAR reanalysis data set for the period July 1957 through December 1994. The annual cycle was removed by subtraction of the long-term mean pressure for each month at each grid point.

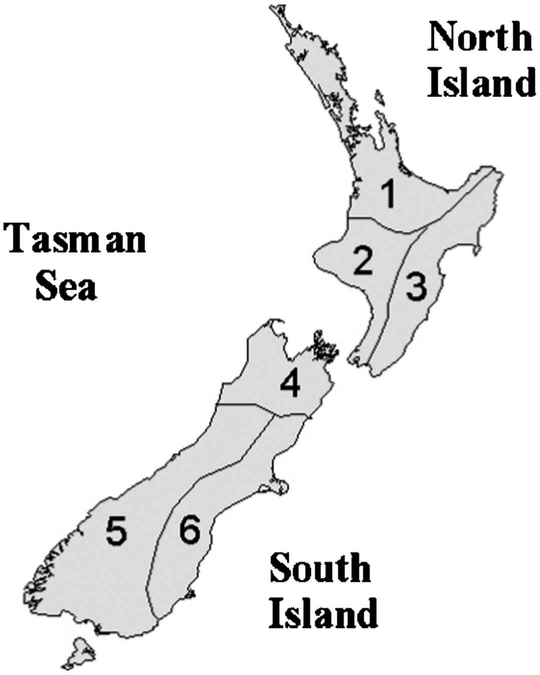

Figure 1. Location map showing the six regions of New Zealand for which temperature and rainfall hindcasts were generated

EOF analysis using a correlation matrix was performed, followed by varimax rotation, for the domain 15°–65 °S by 120 °E–160 °W, which covers the Tasman Sea–New Zealand region. The first four rotated EOFs were retained, and these are shown in Figure 2, which also records the explained variance in the monthly anomalies for the 450-month period July 1957 to December 1994. Virtually identical patterns, and associated time series, are found in seasonal data. The first four EOFs, denoted MSL1 through MSL4, account for 73% of the monthly variance over this region. MSL1 shows a pattern of northeast (or southwest) flow over New Zealand, with a high (low) pressure centre to the southeast of the country. MSL2 shows weaker (stronger) westerly flow south of about 50 °S. MSL3 shows an anti-cyclonic (cyclonic) anomaly in the south Tasman Sea, centred near 45 °S, 160 °E. MSL4 shows an anti-cyclonic (cyclonic) anomaly south of Australia with southerly (northerly) flow in the southwest Tasman.

The SST data used was the UK Meteorological Office SST data set, known as *HadISST* (Rayner *et al.*, 2003). The SST grid-point data were averaged over eight 'key areas' identified by Mullan (1998) as having significant correlations with New Zealand rainfall and temperature variations (Figure 3). These areas, SST1 through SST8, can be described as follows: New Zealand region (SST1), Australian Bight–Tasmania (SST2), NINO3 region in the eastern tropical Pacific (SST3), NINO4 region in the central tropical Pacific (SST4), New Caledonia region north of New Zealand (SST5), central Pacific east–northeast of New Zealand (SST6), central Indian Ocean (SST7) and equatorial Indian Ocean (SST8).

Since these 12 predictors are not all independent of one another, a second EOF analysis was then applied, this time without rotation, following the approach of Barnston and Livezey (1989). The resulting final set of five predictors is discussed in Section 4.

3. ANALOGUE METHODOLOGY

The analogue technique involves searching the historical data, identifying previous months (or seasons as appropriate) that resembled the immediate past period and predicting the following month's climate anomalies on the basis of what happened on those previous occasions. While this is straightforward in principle, a number

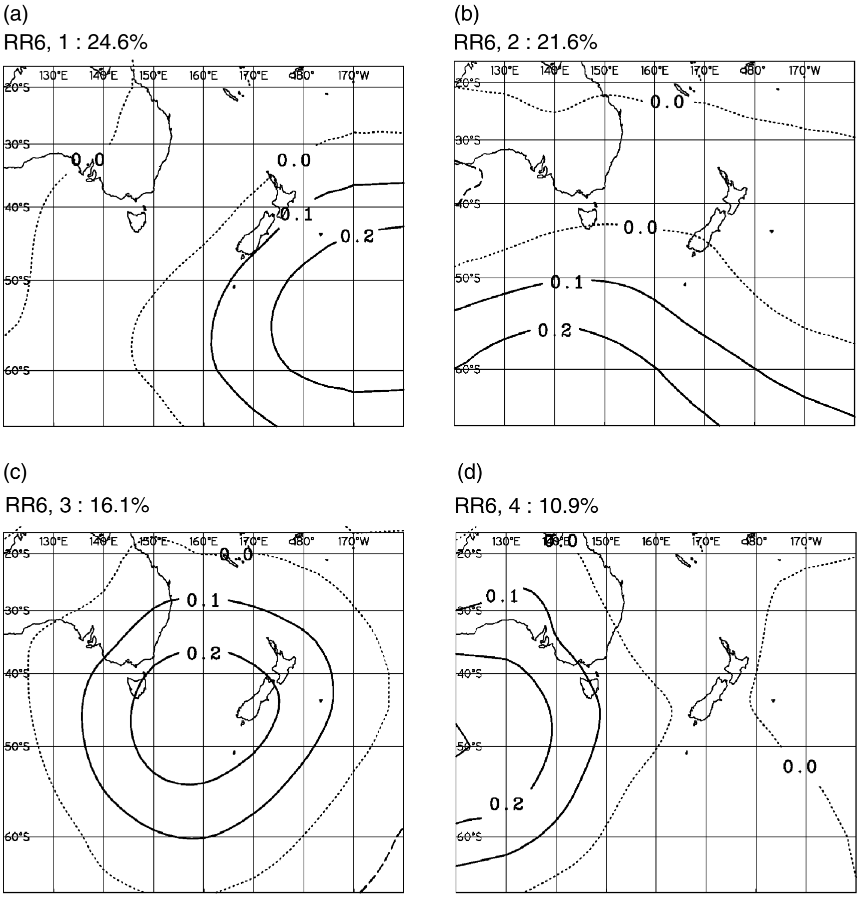

Figure 2. The leading four rotated EOFs of seasonal mean MSLP anomalies over the New Zealand region, all seasons of the year combined. Contours of the pattern weightings are plotted every 0.1, with negative contours dashed and the zero contour dotted. The percentage variance accounted for by the REOF, and the REOF number, is given above each panel

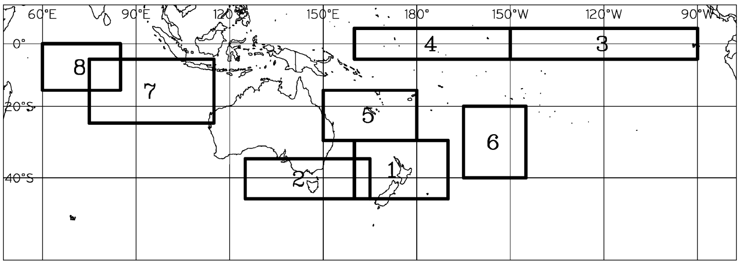

Figure 3. SST index 'key areas', delineating regions that have significant lag correlations with New Zealand temperature and rainfall (from Mullan, 1998)

of decisions have to be made about the experimental design. Firstly, a quantitative method for assessing 'similarity' between periods in the historical data is required. Secondly, a decision must be made on how to score the forecasts. This is intrinsic to the whole procedure, because the similarity formula used to select the analogues needs to be optimised, through a hindcast bootstrap validation, in order to maximise the success rate. An additional parameter that the hindcast validation exercise can determine is the optimal number of analogue months to use in the forecast. It has been known since the early work of Bergen and Harnack (1982) that forecasts based on only the single best analogue are subject to a large sampling error and do not verify well.

3.1. Analogue/anti-analogue similarity

The purpose is to find months in the historical record (the *analogue* month) when *the state of the climate is closely similar* to the analysis (or *base*) month, for which a forecast is to be made. In looking for potential analogues to the base month in other years of record, analogues are searched for in the same calendar month and also in adjacent months. By this means, the number of potential analogues is increased threefold.

There are a number of ways in which a 'climate state vector' (CSV) and analogue 'similarity' may be defined. However, in terms of the experimental set-up, the CSV (see e.g. Barnett and Preisendorfer, 1978; Livezey and Barnston, 1988) is simply the time series of the five predictor indices, INDEX_1 to INDEX 5. The analogue similarity metric, S, is measured in terms of the weighted root-mean-squared difference between the two states as

$$\mathrm{S} = \sqrt{\sum_{1}^{N} wi(I_{\mathrm{a}}^{i} - I_{\mathrm{b}}^{i})^2}$$

where I_{a}^{i} and I_{b}^{i} are the ith indices for the candidate analogue and base months respectively and wi is the weighting given to the ith index. The smaller the similarity value, the closer the agreement between the analogue and base months. The weights wi are determined by the bootstrap cross-validation exercise. This metric is the so-called 'classical' approach (Barnett and Preisendorfer, 1978) based on an instantaneous comparison of analogue and base months.

As first suggested by Van den Dool (1987) and developed by Livezey and Barnston (1988), 'anti-analogue' situations are also sought. For anti-analogues, the indices describing the CSV are reversed, and previous months are searched for situations most similar to the 'reverse' climate. Mathematically, if the similarity metric S is a function of the base month indices, $\mathrm{S} = \mathrm{S}(I_{\mathrm{b}})$, then the closest anti-analogues have the smallest $\mathrm{S}(-I_{\mathrm{b}})$. Thus, two models are tested on the data set: an analogue-only model and a 'mixed' model where both analogues and anti-analogues are allowed for simultaneously.

3.2. Scoring of forecasts

The measure of forecast success chosen is the Hanssen Skill Score (HSS) of 3 × 3 contingency tables (i.e. predicted *vs* observed terciles for rainfall and temperature over the six regions of New Zealand). Other measures of skill could of course be used, and will likely have some small impact on the optimised predictor weights. However, contingency tables are a good way of explicitly presenting uncertainty in the forecasts, a desirable attribute to emphasise to prospective users.

The data for regional rainfall and temperature anomalies were therefore discretised and placed into terciles for validation purposes. Since the analogue model is also to be applied to forecasting future MSLP anomalies, terciles of the rotated EOF scores (the time series associated with the MSLP patterns of Figure 2) were also computed.

When an anti-analogue is selected, the forecast tercile is 'reversed' (i.e. tercile 1 becomes tercile 3, etc) and entered into the contingency table. Initial tests with the model where the 3 × 3 contingency table was accumulated for all six regions combined produced a very low skill score overall. This result of grouped stations scoring poorly has been noted in previous work (e.g. Bergen and Harnack, 1982). The US operational system also uses different predictors for each station (Livezey and Barnston, 1988).

3.3. Analogue search procedure

The principle behind the analogue search procedure is to perform a base-line bootstrap or 'cross-validation' on N years of data, where each datum in the series is used for both the determination and verification of skill in the optimised index weights. Cross-validation is relatively intensive on computing resources; each year of the data is withheld in turn (i.e. N times) for verification purposes and the $N - 1$ remaining years are used as the developmental data set.

The analogue search is performed for each region and for rainfall or temperature separately with optimised index weights wi computed for each season separately. Taking (austral) spring, e.g. analogues are sought for September (and October, then November) by searching for closest similarity in other years in months July/August/September (and August/September/October, then September/October/November). The search procedure, which is detailed below, is performed for each of the six regions in turn. The sequence of steps is described for monthly forecasts, but the same procedure is applied to 3-month forecasts where the seasonal stratification is according to the centre month of the period. The flowchart in Figure 4 summarises the steps described in the following text, and a specific example (Table II) is discussed later in Section 4 to help the reader understand the procedure details.

3.3.1. Tuning of index weights. The bootstrap procedure involves looping over the $N(= 38)$ years of data. Tuning index weights consists of two main steps:

(1) Looping over the remaining $N - 1$ years to identify the best set of index weights. Three weights (values of 0, 1 and 4) were selected on the basis of pre-testing of the number and values of weights and from computational-time considerations. A total of approximately 3^5 weighting combinations are possible (actually slightly less because of effective duplicates) from which to select the closest potential analogues

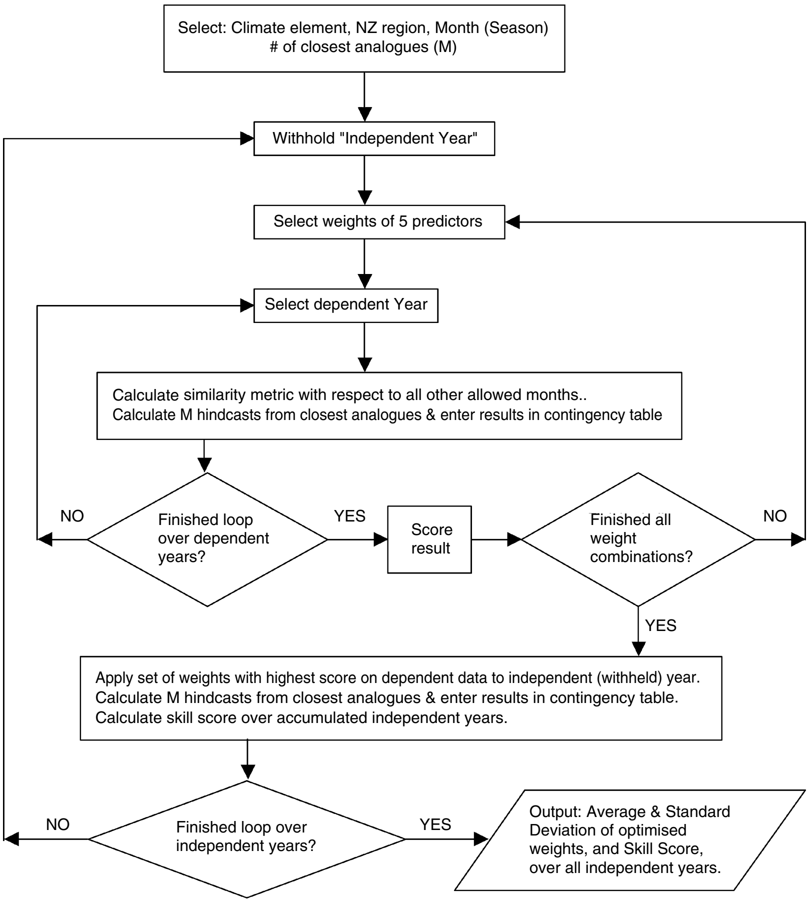
Select: Climate element, NZ region, Month (Season)
# of closest analogues (M)
Withhold "Independent Year"
Select weights of 5 predictors
Select dependent Year
Calculate similarity metric with respect to all other allowed months..
Calculate M hindcasts from closest analogues & enter results in contingency table
NO
Finished loop over dependent years?
YES
Score result
Finished all weight combinations?
NO
YES
Apply set of weights with highest score on dependent data to independent (withheld) year.
Calculate M hindcasts from closest analogues & enter results in contingency table.
Calculate skill score over accumulated independent years.
NO
Finished loop over independent years?
YES
Output: Average & Standard Deviation of optimised weights, and Skill Score, over all independent years.

Figure 4. Flowchart showing step-by-step process for tuning weights of analogue model

to the analysis (or base) month. Using the 0,1,4 subset of weights is essentially a computational short cut to efficiently span the subspace. Results are relatively insensitive to the precise value of the weights, although it is crucial to allow the possibility of a zero weight (i.e. elimination of predictors), as discussed in Section 4.2.

(2) Evaluation of skill. For a given combination of index weights, over all base months in the season (a total of $3 \times [N-1]$) should be looped. For each base month, the similarity metric for potential analogue months ($3 \times [N-2]$ for analogue-only models, twice this number for a mixed analogue–anti-analogue model) should be evaluated. The 'M' best analogues, with the lowest similarity metric, are selected and used to determine the rainfall or temperature forecasts for the following 1-month (i.e. the verifying month) period. A 3×3 contingency table is accumulated ($3 \times M \times [N-1]$ entries when all base months have been considered), matching the observed tercile categories with the forecast tercile categories. (Anti-analogue terciles are reversed.) The accumulated contingency table for the given set of index weights with the Hanssen skill-score test statistic, (HSS) should be calculated (Hanssen and Kuipers, 1965). Then the next weight combination is selected until the approximately 3^5 combinations are tested.

3.3.2. Evaluation of skill on independent data. The best set of weights is then used to select analogues for the 'independent' (withheld) data year. Forecast and observed tercile information is then entered in another contingency table, which will have $3 \times M$ entries for each independent year. HSSs can be computed for each year separately. However, these scores are unstable for small samples, and indeed the score is indeterminate if, e.g. the observed tercile was the same for all 3 months in the seasonal sample. Thus, in practice, the HSS is accumulated over the N independent years.

The next year is then selected as the independent year and the above search procedure is repeated until all N years of data have been analysed and the best sets of index weights computed. The final result is N sets of 'best weights', i.e. for each independent year, there is, at least potentially, a different weight combination that maximises the skill on dependent data. In practice, the weights often do not vary much with the independent year. Section 4 provides an example that helps explain how the search procedure works.

3.4. Stability of results

The search procedure described above involves trying all possible weight combinations across all five predictors and taking the single best combination. In practice, the objective is to find a volume in the solution phase space where the dependent skill score maximises. The exhaustive search samples this phase space at discrete points. Near neighbours to the selected best point will also have similar skill scores. Also, as the search procedure loops through the 38 independent years, noisiness in the data can make the selected best point jump around. A single very bad verification for an 'outlier' set of best weights can destroy what appears to be a skilful forecast model in all other years.

After an analysis of the causes of this instability, it was concluded that just as it is unwise to choose the single closest analogue ($M = 1$), it is also unwise to take the single best set of weights. The stability of results was greatly enhanced by taking an average over the closest 'K' sets of top weights, where K was a small number. Examples and implications are described in Section 4. The flowchart (Figure 4) omits this step for ease of presentation.

3.5. Significance testing

The analogue model involves a number of parameters, some of which are tuned explicitly in the hindcast cross-validation (i.e. the index weights). An important preselected parameter is the number of closest analogues M that are made use of in the forecast. It is also well known that with analogue models, the skill score decreases as the total number of analogues (M) increases. Obviously, the skill score will approach zero (climatology) if *all* other months in the data set are included as analogues. The best value for M is unclear, except that it should be greater than 1. Livezey and Barnston (1988) found that 8 to 11 closest analogues were optimal for an analogue model of US winter temperatures. Because of the number of choices available in specifying model parameters, it is crucial that significance testing be done very carefully.

Gordon (1982) presented a generalised skill score for evaluating the skill of categorical forecasts, and, in particular, also showed how to assess the significance of the resulting score. Gordon's method involved computing the 'no-skill' standard deviation and applying the Student's t-test to check significance. The standard deviation depends, in general, on the entries in the contingency table, the probabilities of occurrence of the forecast categories and the mark or penalty assigned to each forecast/observed combination. In the idealised case of equal probabilities of occurrence (close to what is expected with a tercile data set), the standard deviation becomes independent of these quantities and simplifies to $\sigma = 1/\sqrt{2 \times N_{\text{eff}}}$, where N_{eff} is the effective number of degrees of freedom.

Applying this formula suggests that the null hypothesis of 'no skill at the 95% level' would be rejected in favour of a hypothesis of positive skill (therefore a one-sided test is appropriate) for an HSS that exceeded approximately 0.050 for number of closest analogues $M = 5$ in the case of 1-month forecasts and 0.085 in the case of 3-month forecasts. Corresponding 99% significance levels would be 0.070 and 0.120 respectively, again for $M = 5$. The appropriate significance level reduces as M increases.

Gordon's approach would suggest that the significance levels depend only on the sample size and are independent of season and climatic element. However, it does depend on simplifying assumptions for a large sample (i.e. central limit theorem). Also, the observed tercile frequencies will not be exactly equal, since the number of years in the data set (38) is not divisible by three, and the terciles are computed separately for each month or centred 3-month period. There have been a number of papers that have pointed out problems with assessing significance of skill scores (Barnston and van den Dool, 1993; Elsner and Schmertmann, 1994). These problems arise from degeneracy in the data and become particularly severe when predictor–predictand relationships are weak (Barnston and van den Dool, 1993), which is a fairly standard occurrence in long-range forecasting.

To minimise these instabilities in cross-validated skill, the following approach has been taken:

(1) The 'K' top scoring set of weights was averaged, as noted above. This greatly improved stability of the resulting skill scores.
(2) A Monte Carlo simulation approach for estimating the significance levels was adopted instead of using Gordon's (1982) theoretical formula. To test the significance of a particular model, the exact same bootstrap cross-validation procedure described in Section 3.3 is used, but the 'observed' anomaly data is randomly reordered and terciles are recalculated. Having determined the optimised index weights and associated skill score, the data are randomised again and the cross-validation are repeated. Tests were made for 500 randomised samples, for a range of cases: varying M, K, type of analogue model, climate variable, region and season.
(3) The skill scores were computed for a range of M, typically from 3 to 9, and the skill had to be significant over most of this range for the particular forecast model to be accepted.

The simple direct calculation of significance level from the t-test described earlier was found to be a reasonable guide, although the Monte Carlo results did show variations with the data set (e.g. temperature *vs* rainfall) and with the season.

4. RESULTS

4.1. Final set of predictors

The first important result is the determination of a parsimonious set of predictors, which are discussed in this subsection. The remainder of Section 4 considers the application of these predictors in the analogue model. Section 2 described a group of 12 predictors derived from an EOF analysis of MSLP and from previous work on SST indices at key locations. After applying a second EOF analysis, the leading five EOFs explained 73.6% of the monthly variance of the original 12-predictor set. Table I shows the variance contribution of each eigenvector and the contributions to that EOF made by the four MSLP and eight SST variables.

Table I. Principal component scores of four MSLP REOFs and eight SST regions as they contribute to the final five predictor indices used in analogue search. Scores larger than 0.30 in magnitude are given in bold. The overall variance of each EOF is also shown

Variable Variance (%)	INDEX_1 **27.3**	INDEX_2 **15.0**	INDEX_3 **14.3**	INDEX_4 **9.9**	INDEX_5 **7.1**
MSL1	−0.23	0.03	−0.30	**−0.60**	0.17
MSL2	−0.22	−0.19	−0.14	**−0.46**	**−0.60**
MSL3	0.03	**0.47**	**−0.46**	0.03	0.02
MSL4	0.06	**0.39**	**−0.43**	0.09	−0.18
SST1	**−0.36**	0.28	0.14	−0.24	0.19
SST2	−0.15	**0.49**	0.07	0.18	**−0.35**
SST3	**0.43**	−0.02	0.04	−0.20	**−0.35**
SST4	**0.45**	−0.08	−0.00	−0.24	−0.17
SST5	−0.29	0.17	**0.43**	−0.28	0.03
SST6	−0.30	−0.01	0.25	0.29	**−0.51**
SST7	0.23	**0.38**	**0.35**	−0.16	0.08
SST8	**0.35**	0.29	0.30	−0.19	−0.02

These final five predictors, simply called INDEX_1 to INDEX_5, are used to define the analogues. Although these indices were derived from a more fundamental data set that relates specifically to New Zealand local climate, it is of interest to examine the broader-scale teleconnections associated with these indices. From Table I, positive INDEX_1 is associated (because of high principal component scores) with stronger southwesterlies across New Zealand (negative MSL1, MSL2), colder sea conditions surrounding the country (negative SST1), high SST anomalies in the tropical Pacific (positive SST3, SST4) and warmer Indian Ocean seas (positive SST7, SST8). All these features together comprise what is known about an El Niño 'warm' event in the tropical Pacific and its teleconnection patterns to New Zealand (Gordon, 1986). The time series of INDEX_1 is indeed significantly correlated with the SOI (−0.66 on monthly data and −0.70 on 3-month running means over the model training period). Figure 5(a) shows the time series that has been normalised by its mean and standard deviation, with its sign reversed to highlight the similarity to the SOI. The time series shows the large La Niña events of the 1970s, the 1982/1983 and 1987 El Niños and the long-running El Niño in the early 1990s.

Figure 5(b) shows the time series of the normalised INDEX_5. This is of interest because indices 4 and 5 are the indices that show the greatest trend: in this case, there seems to be an abrupt change at the end of the 1970s. Some abrupt changes in rainfall in New Zealand have been identified as occurring at this time (Salinger and Mullan, 1999), which may be related to regime changes of the Interdecadal Pacific Oscillation (Salinger *et al.*, 2001). It is possible that INDEX_5, which is most highly correlated to increased westerlies south of New Zealand (Figure 2 and Table I), is picking up on this regime change. Figure 6 shows the seasonal correlation, all seasons combined over 1958–1994, between indices INDEX_2 through INDEX_5 and the Southern Hemisphere MSLP anomalies. The SOI–MSLP teleconnection pattern, the Tahiti–Darwin seesaw, is of course so well known (e.g. see Trenberth and Shea, 1987) that it is unnecessary to show the INDEX_1 correlation.

Positive INDEX_2 is associated with higher pressures in the Tasman Sea and south of Australia (positive MSL3, MSL4 in Table I), in warmer seas south of Australia (positive SST2) and in the Indian Ocean (positive SST7, SST8). The MSLP–INDEX_2 correlation (Figure 6(a)) indicates a very strong and significant variation in zonal pressure gradient across about 60 °S, which is clearly hemispheric in extent. A similar pattern is present in all seasons individually (not shown): in summer the correlations show the greatest zonal uniformity, and the weakening of the correlation gradient that is apparent in the South America–Atlantic sector is mainly a feature of winter.

The striking zonal uniformity in the MSLP correlation with INDEX_2 suggested that this time series may be related to the 'high latitude mode' (HLM) that is a notable and stable feature of the Southern Hemisphere

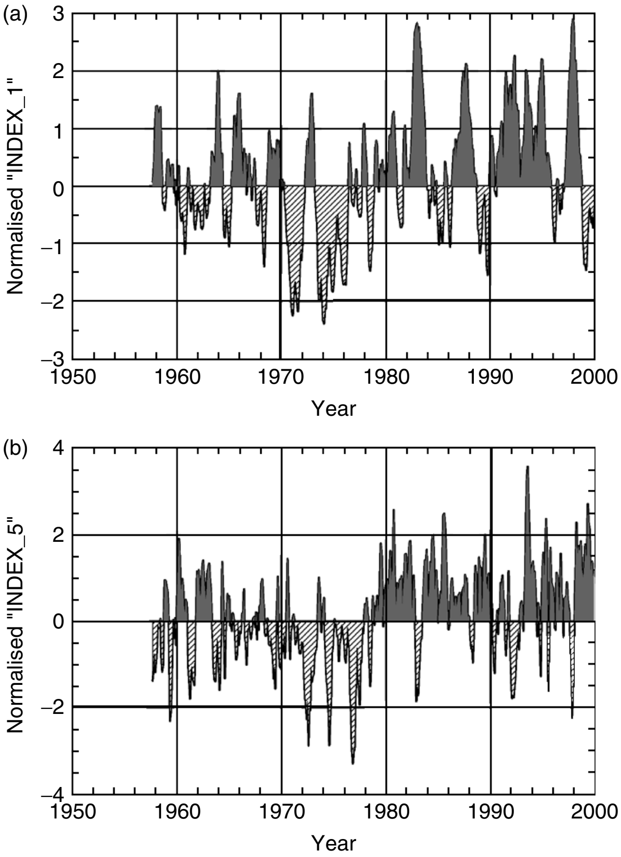

Figure 5. Time series of INDEX_1 and INDEX_5 derived by EOF reduction of four MSLP and eight SST indices. Plots show 3-month running means

circulation (Kidson, 1986, 1999), and is also known as the *Antarctic Oscillation*. A time series of the HLM was derived from NCEP/NCAR reanalysis detrended MSLP and 500 hPa heights and from correlations calculated with the analogue indices. The HLM Index, defined here as the first unrotated EOF over 20 °S–80 °S, is positive when the polar vortex is enhanced, with negative height anomalies over the South Pole and stronger westerlies at about 55 °S. The HLM is more strongly correlated to INDEX_2 than to any of the other indices: +0.60 for monthly data (+0.39 for 3-monthly data) using 500 hPa data. These correlations are virtually the same for MSLP data instead, and are only slightly weakened if the grid-point fields were not detrended first. INDEX_5 also has a highly significant, if somewhat weaker, correlation with the HLM (+0.48 for 1-month data on 500 hPa data).

INDEX_4 and INDEX_5 Figure 6(c) and (d)) both show a marked wave 3 pattern, with an out-of-phase relationship between the two. There are many studies of Southern Hemisphere circulation that comment on the importance of zonal wavenumber 3 at middle and high latitudes. A large increase in the amplitude of wavenumber 3, at 500 hPa and 60 °S, was noted by van Loon *et al.* (1993) as occurring between 1977 and 1981, which corresponds well with the time series in Figure 5. The late 1970s also corresponded with the marked weakening of the semi-annual oscillation (van Loon *et al.*,1993).

Lastly, INDEX_3 appears to be more localised to Australasia in its influence (Figure 6(b)). Positive INDEX_3 is associated with cyclonic anomalies south of Australia and New Zealand and with higher SSTs north of New Zealand (positive SST5) and in the Indian Ocean (positive SST7, SST8). INDEX_3 also shows the greatest seasonal variation in its correlation pattern with MSLP.

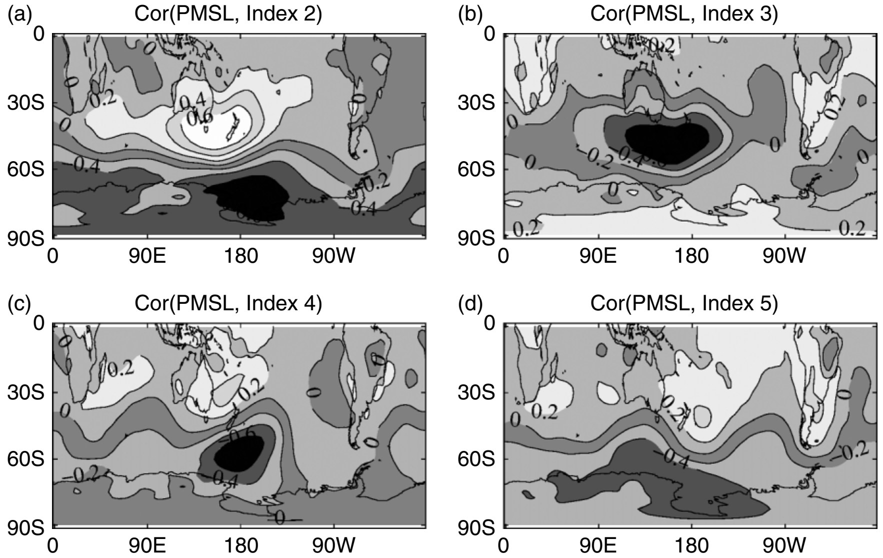

Figure 6. Contemporary seasonal correlations, all seasons combined, between NCEP MSLP anomalies over the Southern Hemisphere and times series of analogue indices INDEX_2 to INDEX_5 for the period 1958–1994

4.2. *Application of search procedure*

Table II shows an example of 1-month rainfall forecasts for Region 5 (west of South Island) for the winter season. In this case, the five closest analogues ($M = 5$) are selected from an analogue-only model. Predictors IND3 and IND4 are selected every year, with maximum weighting on almost all occasions. Predictors IND1 and IND2 are never selected in the best weight set, except on just one occasion: when year 27 (1983) is withheld as independent data, the tuning on the remaining 37 years selects IND1 with low weighting. Two columns of skill scores are shown in Table II. The dependent skill against each year is the HSS on the $3 \times M \times [N - 1]$ forecasts for the best weight set. The final column is the skill score on the independent data that accumulates year by year over the 38 years of data. This independent skill score is highly unstable for the first few years, and the number of interest is the final value after all 38 years have been assessed. In this example, the HSS on independent data is 0.091.

As the final step in the analogue search procedure, the best sets of index weights are averaged to get an overall set of 'optimised index weights' for the month and region in question. It is this set of optimised index weights that would be used in any operational version of the analogue model, provided the model was deemed to be producing statistically significant forecasts. Applying the optimised weights across the whole data set results in another skill score (0.137 in Table II) that can no longer be considered independent, of course; this score will almost always lie between the true independent skill (0.091) and the average of the N dependent skills (0.141). In an extreme case where the same set of best weights are chosen every year (this never actually happens), all three skill scores would be the same.

The variation of the index weights over these N independent years is computed by the root-mean-squared deviation (average of the standard deviations for each index separately). This provides an assessment of the 'consistency' and stability of the chosen weights; a small RMS variation indicates a highly consistent set of weights, as in the chosen example of Table II.

The 'exhaustive' search option through all 0,1,4 weight combinations across five predictor indices was not the only option considered in the preliminary test of the analogue model. One option tested was to apply a 'simplex downhill search' algorithm (Press *et al.*, 1986). While this method was much faster in finding the best set of index weights wi, it did not always converge to the same solution. Particularly, in cases where the solution surface was 'flat' with no well-defined minimum (actually maximum in the skill score), the simplex search converged to different points depending on the initial guess set provided. In most of these situations, the

Table II. Sample output from analogue bootstrap test. For each withheld independent year (not all years shown), the columns below show the best weights for each of the five predictor indices, the same weights normalised out of 1000 and Hanssen skill scores (×1000) on dependent and independent data. See accompanying text for further explanation Case: 1-Month forecast, winter rainfall, Region 5 Options: Analogues only, # closest analogues (M) = 5

Year	Best weights					Normalised best weights					Skill Score	
	1	2	3	4	5	IND1	IND2	IND3	IND4	IND5	Dep	Ind
1	0	0	4	4	1	0	0	444	444	111	151	0
2	0	0	4	4	1	0	0	444	444	111	152	−33
3	0	0	4	4	1	0	0	444	444	111	135	140
4	0	0	4	4	1	0	0	444	444	111	118	220
5	0	0	4	4	1	0	0	444	444	111	141	142
6	0	0	4	4	1	0	0	444	444	111	117	177
7	0	0	4	4	1	0	0	444	444	111	139	171
8	0	0	4	4	4	0	0	333	333	333	146	125
9	0	0	4	4	1	0	0	444	444	111	135	133
10	0	0	4	4	1	0	0	444	444	111	154	122
11	0	0	4	4	1	0	0	444	444	111	151	91
12	0	0	4	1	1	0	0	667	167	167	138	74
....												
24	0	0	4	4	1	0	0	444	444	111	138	73
25	0	0	4	4	1	0	0	444	444	111	149	66
26	0	0	4	4	1	0	0	444	444	111	151	56
27	1	0	4	4	4	77	0	308	308	308	129	54
28	0	0	4	4	1	0	0	444	444	111	137	56
29	0	0	4	4	1	0	0	444	444	111	134	61
30	0	0	4	1	1	0	0	667	167	167	158	56
31	0	0	4	4	1	0	0	444	444	111	146	65
32	0	0	4	4	4	0	0	333	333	333	138	68
33	0	0	4	4	1	0	0	444	444	111	133	81
34	0	0	4	4	1	0	0	444	444	111	145	86
35	0	0	4	4	1	0	0	444	444	111	132	92
36	0	0	4	4	1	0	0	444	444	111	149	86
37	0	0	4	4	1	0	0	444	444	111	132	97
38	0	0	4	4	1	0	0	444	444	111	160	91
	Average over all years					2	0	452	400	145	141	91

Skill over all years using average weights 137
Index weight 'consistency' 64.6

forecast skill would not be significant, so the analogue model would have been discarded anyway. However, any unnecessary instability in the method was to be avoided.

With the exhaustive search option, it was important to allow for predictors to be omitted. Thus, tests with no zero-weight option (e.g. weights 1, 2, 4) showed that forcing the inclusion of all predictors was definitely detrimental in some cases.

4.3. *Stability of predictor weights*

Looping through the set of weight combinations of the five predictors, an HSS was calculated for the dependent 37 years and then applied to the omitted year. Figure 7 shows two examples, where the HSS is averaged over the 38 independent years (each year potentially having a different set of predictor weights as in Table II). An important modification here was that instead of choosing the single weight combination that scored highest on the dependent data (as in Table II), an average over the set of K best weights was used. This was found to be necessary to minimise the effect of outliers and produce a more stable solution. Figure 7(a) compares the skill for $K = 1$ (the single best set of weights) *versus* $K = 5$, as the number of closest analogues used is allowed to vary. The skill score maximises near 2–3 closest analogues, and then again at 9–10. There is no obvious reason why an intermediate number of analogues (like 7) should be so much worse. Using $K = 5$ produces a result that is much more stable. Figure 7(b) shows another comparison: here it is apparent that the forecast models for 1-month winter rainfall in Regions 1 and 2 are unstable since the significance is reduced drastically when the predictor weights are ‘smoothed’ over a larger region of phase space. The model for Region 5, however, remains significant. Subsequently, $K = 5$ was used for all model selection.

4.4. Skilful forecast models

Tables III and IV list all forecast models found to exhibit significant skill (95% level) on independent data, where the skill level was estimated according to the Monte Carlo simulations described above. A number of

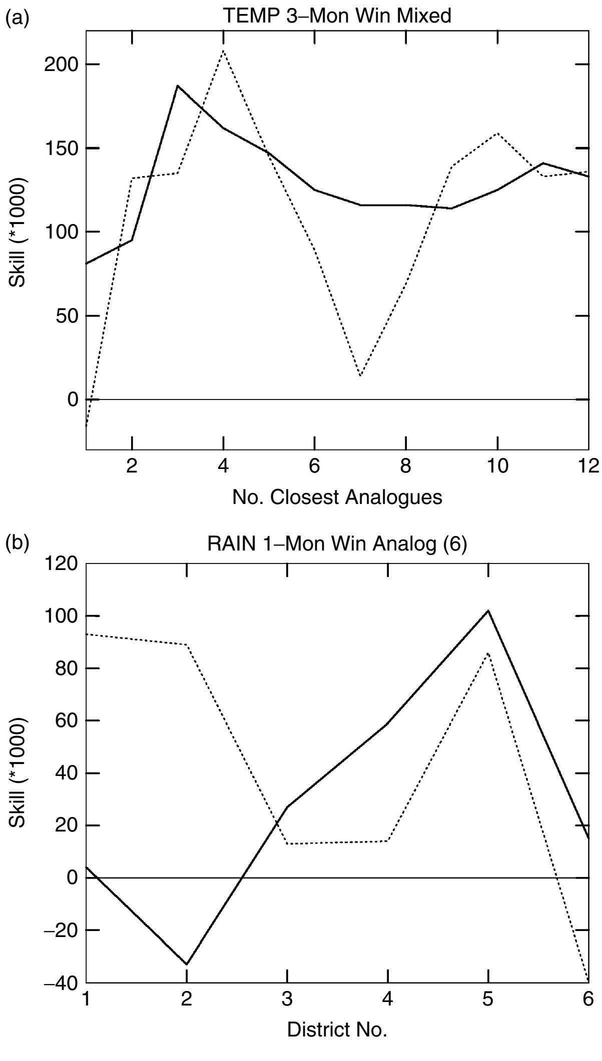

Figure 7. Hanssen skill scores (×1000) on independent data, using average of K best predictor weights $K = 1$ (dotted) and $K = 5$ (solid): (a) 3-monthly winter temperature forecasts for Region 1, using mixed analogue–anti-analogue model, showing variation with the number of closest analogues used; (b) 1-monthly winter rainfall forecasts, an analogue-only model with six closest analogues, showing variation with region

Table III. Hanssen skill scores (×1000) for all analogue (A) and mixed (M) models showing skill on independent data for temperature and rainfall over 1-month and 3-month periods. Table entries are ordered horizontally by season (spring, summer, autumn, winter) and vertically by region (as in Figure 1). Blank entries indicate no skilful model

		Temp 1-month				Temp 3-month				Rain 1-month				Rain 3-month			
		Sp	Su	Au	Wi	Sp	Su	Au	Wi	Sp	Su	Au	Wi	Sp	Su	Au	Wi
Region 1	A											70				83	70
	M					89		79	142								
Region 2	A	74				107						70	124		80		
	M	75			67	96		117	84						101		
Region 3	A		74	58	68	189			129	86	80				105	101	
	M	68			116	173			101	110			91			87	77
Region 4	A		108		95	110			145				79				72
	M	74			71	119			163				74		88		81
Region 5	A		83			130			75	113							
	M					220			193								
Region 6	A		102		70				139				80				
	M		119		84	144			135				78	89			

Table IV. As in Table III, but for MSLP forecasts, ordered vertically by EOF pattern (as in Figure 2)

		MSLP 1-month				MSLP 3-month			
		Sp	Su	Au	Wi	Sp	Su	Au	Wi
EOF 1	analogue					72		74	151
	mixed					101	73		105
EOF 2	analogue	101	79		92	204	91	72	183
	mixed		107			256	87	90	148
EOF 3	analogue			68					
	mixed	86		135				89	63
EOF 4	analogue		84		94	80			
	mixed					116		104	

general comments can be made about the success of the forecast models in terms of model type (analogue-only or mixed analogue–anti-analogue), forecast element (temperature, rainfall, MSLP), forecast period (1 *vs* 3 months), season and region or EOF pattern.

It is clearly useful to have both analogue and mixed models. There are about an equal number of occasions when only an analogue model is available as when only a mixed model is available. However, for those cases where both model types verify significantly, the mixed model has a higher skill score about twice as often as the analogue model. Comparing the forecast periods, there are more 3-month than 1-month models, with substantially higher skill, for both temperature and pressure. However, the skill of the rainfall models shows little difference between the two forecast periods. A seasonal breakdown of the successful models suggests little influence on forecasts of pressure. However, it appears to be 'easier' to forecast rainfall in the winter season, and temperature in either winter or spring. The inability to forecast over summer and autumn is most noticeable for 3-month temperatures.

In terms of geographic variation, Region 1 (north of North Island, Figure 1) and Region 5 (west of South Island) are the most difficult parts to forecast for. Of the four pressure patterns, EOF2, a measure of the strength of the westerlies over and south of South Island has the most skilful models, and, in particular, the 3-month winter and spring models have some of the highest skill scores of all the analogue models tested.

4.5. Validation on independent period

The analysis of model performance so far has focussed on the 1958–1994 period over which the bootstrapping approach was applied. Skill was also assessed over the completely independent period from September 1994 to June 2003. This substantially shorter recent period leads to noisier results, so skill is aggregated over all seasons and over combined regions, one each for the North Island and South Island. The HSS is also compared with other measures of skill frequently found in the literature.

A wide variety of validation procedures exist (see, e.g. Chapter 7 of Wilks, 1995) for quantifying the strength of relationships between matched pairs of model forecasts and the corresponding observed outcomes. Since the analogue model provides regional probabilities of occurrence of precipitation and temperature for each tercile of the distribution, only performance measures that take account of probabilistic information are examined. Examples are shown for seasonal forecasts, where greater skill was demonstrated in the bootstrap tests. Table V shows the HSS as before and also the Ranked Probability Skill Score (RPSS, see Wilks, 1995). Of these two skill scores, the RPSS is the more stringent statistic that penalises severe predictions in the incorrect tercile category with high probabilities. Figure 8 shows the skill of temperature and rainfall 3-month forecasts in terms of the relative operating characteristic (ROC) curves (Mason and Graham, 1999), which characterise the relationship between the hit rate and false-alarm rate. These quantities, derived from contingency tables for each tercile category, are conditional probabilities of a ‘yes’ prediction given either occurrence or non-occurrence of the event (Wilks, 2001). For predictions to show skill, the hit rates must be larger than the false-alarm rates, i.e. the curves should lie in the upper left quadrant above the diagonal ‘no-skill’ line. ROC scores (Table V) can be calculated for each tercile separately as the area between the diagonal and the ROC curves for the tercile in question.

Seasonal rainfall predictions (Figure 8) for the North Island are quite poor, with the HSS and the RPSS being 0.004 and −0.03 respectively. The diagram also shows that the middle tercile (near normal rainfall) has negative forecasting skill, with the (dotted) line in the lower right half of the plot field. The ROC score for tercile 2 is −0.16. South Island seasonal rainfall is influenced to some extent by the strength of the westerlies over and to the south of New Zealand, and the westerlies correspond to the EOF2 pattern that is best predicted by the analogue models. The two scalar skill scores show a very small and insignificant improvement over their climatological values. The ROC scores are positive for all terciles and particularly so for the above normal tercile, suggesting that the analogue models are most successful in the ‘heavy’ rain situations.

The lack of apparent skill in the seasonal precipitation predictions for the North Island has also been noted by others (Francis and Renwick, 1998), who found for New Zealand no significant forecast relationships for rainfall anomalies at both monthly and seasonal timescales.

Seasonal temperature predictions over the North Island indicate an encouraging level of skill in the model. For forecast probabilities that are typically within 20% of their climatological values, the RPSS are often in the range of 5–20% (Goddard *et al*., 2003). Seasonal temperature predictions over the South Island are not as skilful as those for the North Island. When compared to the North Island, the scalar skill scores are lower, and the deviations from the no-skill line in the ROC curves are smaller. Note, again, that forecasts of tercile 2 (near normal temperature) show no skill in the ROC scores. This tendency for lower skill in middle tercile forecasts has been noted by others (Van den Dool and Toth, 1991).

5. SUMMARY AND DISCUSSION

The methodology and bootstrap validation of an analogue model for forecasting New Zealand regional rainfall and temperature anomalies has been described. The five predictors used have a straightforward physical interpretation, are associated with large-scale fluctuations in Southern Hemisphere flow and are therefore expected to be stable. The first predictor is essentially the SOI, and three of the next four indices show teleconnection patterns of hemispheric extent, which provides a lot of confidence that the data set reduction procedure has produced a sensible result. The analogue indices INDEX_2, INDEX_4 and INDEX_5 may well be useful as predictors of climate variations in other parts of the Southern Hemisphere besides New Zealand.

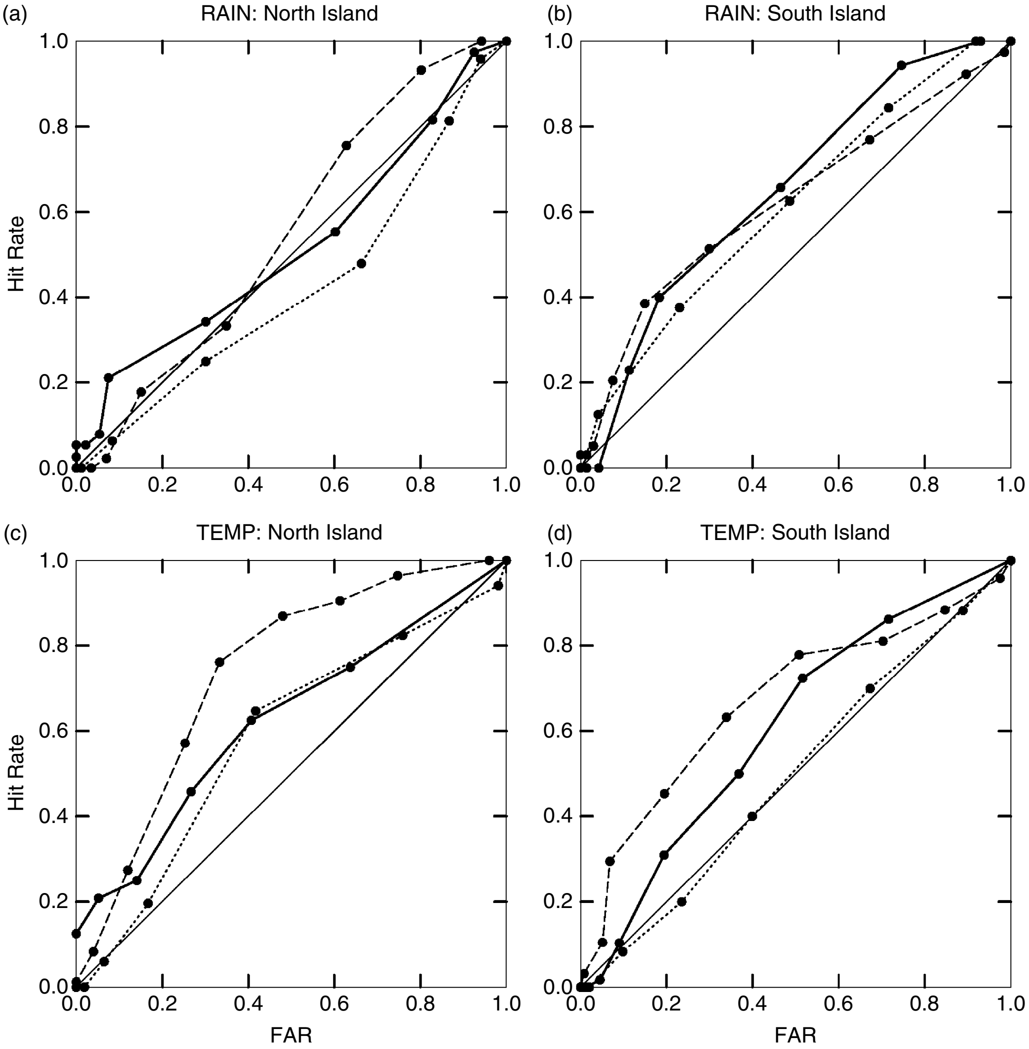

Figure 8. Relative operating characteristics curves for 3-month forecasts over 1994–2003 period of: (a) North Island rainfall, (b) South Island rainfall, (c) North Island temperature and (d) South Island temperature. The hit rate *versus* false-alarm rate is plotted separately for terciles 1 (heavy solid line), 2 (dotted) and 3 (dashed). Skilful forecasts lie in the upper left half of the plot, above the thin solid diagonal line

Table V. Skill scores (×1000) on independent period September 1994–June 2003, for 3-month predictions of temperature and rainfall terciles. Results are aggregated over all seasons and over the three North Island and three South Island regions

Skill score	Rainfall		Temperature	
	North Island	South Island	North Island	South Island
Hanssen	4	68	105	65
Ranked probability	−30	20	230	20
ROC (Terc 1, 2, 3)	40 −160 100	300 220 240	260 180 480	220 0 340

There were many options available in the tuning of index weights, and great care was taken to minimise the effect of outliers, which could otherwise lead to a choice of weights that did not validate well on independent data. Test results where the 3 × 3 contingency table was accumulated for all six regions combined produced a very low skill score overall. Hence, the decision was taken to optimise for each region separately. This approach probably requires some philosophical justification. Not requiring a common set of indices and weights for all regions means accepting that there is no unique set of large-scale anomalies (in SST and circulation) that lead to a coherent New Zealand–wide pattern of MSLP, rainfall and temperature anomalies. Instead, it is argued that the appropriate best analogue depends on the region (and climatic element) being predicted, e.g. some elements of the CSV may be important predictors for next month's rainfall on the east coast of the South Island, but predictors for temperature forecasts in the west of the North Island require different weighting.

Another consistency test that increases confidence in the results is to compare how the skilful models for rainfall or temperature (by season and region) match up to skilful models of MSLP (by season and EOF). This can be understood in terms of Tables III and IV, which show the significant analogue models, and from Table VI, which notes when there are significant correlations between the EOF time series and regional anomalies. The example is discussed for 3-month correlations and models, but similar comments apply to the 1-month timescale. The 3-month analogue temperature models in Table III suggest there is little skill in the summer and autumn seasons in almost all regions of the country. In five of the six regions, EOF3 is the pressure pattern most strongly correlated to summer temperature (Table VI), but no skill is shown in predicting this EOF (Table IV). While EOF3 also cannot be predicted for spring either, this pattern is not significantly associated with spring temperatures except for Region 2. For autumn temperatures, predicting EOF1 would seem to be crucial, and although a validated MSLP model for this season is available, the skill is not high (Hanssen score of 0.074 in Table IV).

For seasonal rainfall predictions, a noticeable feature is that more of the EOFs tend to project on to the North Island regions, where three or even all four of the EOFs 'need' to be predicted. This is a difficult task and possibly the reason why North Island rainfall validated so poorly on the recent decade. Region 5, the west of South Island, has no skilful rainfall models in any season. This can be explained qualitatively in terms of the importance of EOF3 for this region and the absence of good MSLP models for EOF3. Thus, although the presence or absence of skilful analogue models cannot be justified in all cases, there is generally a clear association with how significantly a particular pressure EOF projects onto the data and whether the analogue approach is able to predict that MSLP pattern.

Research has shown that it is generally more difficult to predict variations within a season than variations of the season as a whole (Rowell *et al*., 1995). Thus, it is not surprising that the 1-month forecasts are much less successful than the 3-month ones. Having the option of using anti-analogues in addition to analogues increases the usefulness of the forecasts, particularly to the seasonal case. Where both analogue and mixed models are available, more often than not the mixed model demonstrates higher skill.

Table VI. Seasonal correlations, by season, between EOF time series and regional rainfall and temperature anomalies. The EOF number (1–4) is shown only where correlation is significant at the 95% level, or (in bold) at the 99% level

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
RAIN						
Spring		**1** 2 **3** **4**	**1** **2** **3** **4**	**1** **2**	**3** **4**	**3** **4**
Summer	1 **3** 4	**2** **3** **4**	**1**	**2** 4	**3**	3
Autumn	1 2 3 4	**1** **3**	**1** **2** **3**	**1** **2**	**3**	
Winter	**2** **3**	**1** **2** **3** 4	**1** **2** **3**	**1** **2**	**2** **3**	**3**
TEMP						
Spring	**2** **4**	**1** **3**	**1**	**4** **1** **1**		
Summer	**1** **3**	**1** **3**	**1** **3**	**1**	**1** **3**	**1** **3**
Autumn	**1** **4**	**1** 4	1 2	**1** 3 **4**	**1** 4	**1**
Winter	**2**	**1** **2** **3**	**1** 3 **4**	**1** 3 **4**	**1** **2** **4**	**1** **2**

A number of issues associated with analogue models have not been addressed in this paper. For example, Barnett and Preisendorfer (1978) suggest that analogues be chosen using the recent history of the climate system instead of just the latest or current state. Thus, rather than a similarity metric derived from the latest month or season, a weighted metric from two consecutive months or seasons could be used. Another relevant issue is whether some *a priori* indicator of confidence in the forecast can be estimated. The analogue similarity metric is obviously an indicator one might use – do forecasts based on closer past analogues score higher?

One question that has been considered is whether the forecast can be improved by 'sharpening' the tercile probability distribution. Sharpness refers to the forecast having large deviations from mean values fairly frequently, which is accepted as being more informative than forecasts near the climatological distribution (Wilks, 2001). Since there is a sample of closest analogues to select from, a simple sharpening algorithm can be applied and the consequences assessed. Tests to maximise the sharpness of the forecast tercile distribution were carried out by ranking the closest analogues, successively dropping off the poorest match and recalculating the tercile forecasts. The number of analogues retained maximises the difference between the highest probability tercile and the second highest. For example, if ten analogues led to a tercile forecast of 50 : 40 : 10 and the poorest two analogues were in the middle tercile, then dropping these off would lead to the sharper probability distribution of approximately 62 : 25 : 13. Applying this algorithm, with a lower limit of at least three analogues retained, produced the interesting result that the HSS was almost always improved but that the RPSS was almost always worsened. Thus, the resulting skill of the sharpening procedure depends on the skill score one chooses to use.

In implementing the analogue forecast system operationally at NIWA, a few modifications to the procedures described above have been made. One is to use a default set of 10 closest analogues in the models: this is a convenient number when specifying tercile probabilities and avoids giving the impression of spurious precision in the forecasts. Another modification is that after seeking analogues using the hindcast optimised predictor weights, the weights are 'dithered' and the search repeated. The dithering is implemented by increasing and decreasing each weight in turn by 5% and renormalising. There are occasions when this leads to the discovery of very close (low similarity metric) analogues not previously identified.

Since a number of models (i.e. for different region or climate element) can have closest analogues in common, it is sometimes helpful to aggregate all the chosen analogues, order them in terms of frequency of selection and inverse similarity score and apply these analogues to regions or climate elements where there is *no* direct model available. This additional information, assigned a lower confidence than direct model predictions, can be included with the wide range of other predictive information in the monthly climate outlook discussions.

ACKNOWLEDGEMENTS

The authors are grateful to the U.K. Meteorological Office for making available the sea surface temperature data set. The NCAR/NCEP reanalysis data were provided through the NOAA Climate Diagnostics Center (http://www.cdc.noaa.gov/). This research was supported by the New Zealand Foundation for Research, Science and Technology under contract C01X0202.

REFERENCES

Anderson JL, Stern WF. 1996. Evaluating the potential predictive utility of ensemble forecasts. *Journal of Climate* **9**: 260–269.

Barnett TP, Preisendorfer RW. 1978. Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of the Atmospheric Sciences* **35**: 1771–1787.

Barnston AG, Livezey RE. 1989. An operational multifield analog/antianalog prediction system for United States seasonal temperatures. Part II: spring, summer, fall, and intermediate 3-month period experiments. *Journal of Climate* **2**: 513–541.

Barnston AG, van den Dool HM. 1993. A degeneracy in cross-validated skill in regression-based forecasts. *Journal of Climate* **6**: 963–977.

Bergen RE, Harnack RP. 1982. Long-range temperature prediction using a simple analog approach. *Monthly Weather Review* **110**: 1083–1099.

Chapman WL, Walsh JE. 1991. Long-range prediction of regional sea ice anomalies in the arctic. *Weather and Forecasting* **6**: 271–288.

Drosdowsky W. 1993. Potential predictability of winter rainfall over southern and eastern Australia using Indian Ocean sea-surface temperature anomalies. *Australian Meteorological Magazine* **42**: 1–6.

Drosdowsky W. 1994. Analogue (non-linear) forecasts of the Southern Oscillation Index time series. *Weather and Forecasting* **9**: 78–84.

Elsner JB, Schmertmann CP. 1994. Assessing forecast skill through cross validation. *Weather and Forecasting* **9**: 619–624.

Fraedrich K, Raible CC, Sielmann F. 2003. Analog ensemble forecasts of tropical cyclone tracks in the Australian region. *Weather and Forecasting* **18**: 3–11.

Francis RICC, Renwick JA. 1998. A regression-based assessment of the predictability of New Zealand climate anomalies. *Theoretical and Applied Climatology* **60**: 21–36.

Goddard L, Barnston AG, Mason SJ. 2003. Evaluation of the IRI's "Net Assessment" seasonal climate forecasts, 1997-2001. *Bulletin of the American Meteorological Society* **84**: 1761–1781.

Gordon ND. 1982. Evaluating the skill of categorical forecasts. *Monthly Weather Review* **110**: 657–661.

Gordon ND. 1986. The Southern Oscillation and New Zealand weather. *Monthly Weather Review* **114**: 371–387.

Gutzler DS, Shukla J. 1984. Analogs in the wintertime 500 mb height field. *Journal of the Atmospheric Sciences* **41**: 177–189.

Hanssen AW, Kuipers WJA. 1965. On the relationship between the frequency of rain and various meteorological parameters. *KNMI Mededelingen en Verhandelingen* **81**: 2–15.

Kidson JW. 1986. Index cycles in the Southern Hemisphere during the global weather experiment. *Monthly Weather Review* **114**: 1654–1663.

Kidson JW. 1999. Principal modes of Southern Hemisphere low-frequency variability obtained from NCEP-NCAR reanalyses. *Journal of Climate* **12**: 2808–2830.

Livezey RE, Barnston AG. 1988. An operational multifield analog/antianalog prediction system for United States seasonal temperatures. I. System design and winter experiments. *Journal of Geophysical Research* **93**: 10 953–10 974.

Livezey RE, Barnston AG, Neumeister BK. 1990. Mixed analog/persistence prediction of United States seasonal mean temperatures. *International Journal of Climatology* **10**: 329–340.

Livezey RE, Barnston AG, Gruza GV, Ran'kova EY. 1994. Comparative skill of two analog seasonal temperature prediction systems: objective selection of predictors. *Journal of Climate* **7**: 608–615.

Mason SJ, Graham NE. 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting* **14**: 714–725.

Mullan AB. 1995. On the linearity and stability of Southern Oscillation-climate relationships for New Zealand. *International Journal of Climatology* **15**: 1365–1386.

Mullan AB. 1996. Nonlinear effects of the Southern Oscillation in the New Zealand region. *Australian Meteorological Magazine* **45**: 83–99.

Mullan AB. 1998. Southern Hemisphere sea surface temperatures and their contemporary and lag association with New Zealand temperature and precipitation. *International Journal of Climatology* **18**: 817–840.

Namias J. 1968. Long range weather forecasting – History, current status, and outlook. *Bulletin of the American Meteorological Society* **49**: 438–470.

Nicholls N. 1980. Long-range weather forecasting – value, status and prospects. *Reviews of Geophysics and Space Physics* **18**: 771–788.

Palmer TN, Anderson DLT. 1994. The prospects for seasonal forecasting: a review paper. *Quarterly Journal of the Royal Meteorological Society* **120**: 755–793.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1986. *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research* **108**(D14): 4407. doi:10.1029/2002/JD002670.

Richman MB. 1986. Rotation of principal components. *Journal of Climatology* **6**: 293–335.

Rowell DP, Folland CK, Maskell K, Ward N. 1995. Variability of summer rainfall over tropical north Africa (1906-92): observations and modeling. *Quarterly Journal of the Royal Meteorological Society* **121**: 669–704.

Salinger MJ. 1996. Painting images: the art and science of monthly climate prediction. *Weather and Climate* **16**: 13–22.

Salinger MJ, Mullan AB. 1999. New Zealand climate: Temperature and precipitation variations and their links with atmospheric circulation 1930-1994. *International Journal of Climatology* **19**: 1049–1071.

Salinger MJ, Renwick JA, Mullan AB. 2001. Interdecadal Pacific Oscillation and South Pacific climate. *International Journal of Climatology* **21**: 1705–1721.

Smith I. 1994. Indian ocean sea-surface temperature patterns and Australian winter rainfall. *International Journal of Climatology* **14**: 287–305.

Trenberth KE, Shea DJ. 1987. On the evolution of the Southern Oscillation. *Monthly Weather Review* **115**: 3078–3096.

Van den Dool HM. 1987. A bias in skill in forecasts based on analogues and antilogues. *Journal of Climate and Applied Meteorology* **26**: 1278–1281.

Van den Dool HM. 1989. A new look at weather forecasting through analogues. *Monthly Weather Review* **117**: 2230–2247.

Van den Dool HM. 1994. Searching for analogues, how long must we wait? *Tellus* **46A**: 314–324.

Van den Dool HM, Toth Z. 1991. Why do forecasts for "near normal" often fail? *Weather and Forecasting* **6**: 76–85.

van Loon H, Kidson JW, Mullan AB. 1993. Decadal variation of the annual cycle in the Australian dataset. *Journal of Climate* **6**: 1227–1231.

Voice M, Beard G, Mullen C. 1996. May the odds be with you – seasonal and multiseasonal prediction. *Weather and Climate* **16**: 29–40.

Wetterhall F, Halldin S, Xu C-Y. 2004. Statistical precipitation downscaling in central Sweden with the analogue method. *Journal of Hydrology* **306**: 174–190.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences*, Academic Press: San Diego, California 467.

Wilks DS. 2001. A skill score based on economic value for probability forecasts. *Meteorological Applications* **8**: 209–219.

WMO. 2002. Long-range forecasting progress report for 2001. WMO/TD–No. 1150, LRFP report series No. 8, World Meteorological Organisation: 89.