Downscaling 15-day Ensemble Weather Forecasts and Extension to Short-term Climate Outlooks

James A. Renwick*, A. Brett Mullan, Craig S. Thompson, and Alan Porteous
NIWA Wellington, New Zealand

Abstract
A diagnostic downscaling system was applied to daily output from the NCEP GFS Ensemble Prediction System, to provide an ensemble of estimates of daily climate variables at 114 New Zealand climate stations for a 15-day forecast interval. Downscaled forecasts were validated on a day-by-day basis and for time averages through the 15-day forecast interval. Fifteen-day average forecast anomalies of temperature and precipitation were used to estimate probabilities of monthly tercile outcomes for six climate regions of New Zealand, using a conditional climatology approach.

Daily forecasts exhibited useful (and statistically significant) levels of skill through the first week of the forecast interval, while five-day averaged forecasts generally exhibited useful skill out to the day 6-10 average. Forecast probabilities of wet and dry spells were found to have a small amount of skill through week two of the forecast interval, in the sense of predicting whether the week will be generally wet or dry. Based on forecast average 15-day anomalies, one-month forecasts of mean temperature exhibited positive skill, while precipitation forecasts exhibited no skill. The downscaling system appears to be a useful tool for downscaling medium-range forecast output from a global NWP model.

1. Introduction

Statistical downscaling of numerical weather prediction (NWP) model output has long provided guidance for operational weather prediction activities, using the Model Output Statistics (MOS) and/or Perfect Prognosis (PP) frameworks (Glahn and Lowry

* Corresponding author address: Dr J. A. Renwick NIWA Private Bag 14-901 Wellington NEW ZEALAND. e-mail: j.renwick@niwa.co.nz
The simplest approach is PP, which takes diagnostic relationships between analysed meteorological fields (temperature, winds, etc) and local surface weather observations, and applies them to fields predicted by an NWP model to predict local surface weather. The assumption is that predicted fields exhibit the same statistical characteristics as the analysed (“observed”) fields, hence are considered “perfect”. Any biases in the NWP model fields are carried through into the downscaled predictions. The advantage of a PP approach is that it is portable and can be applied to any NWP model. Forecast skill will improve as the forecast model improves, without the need to change or update the PP downscaling system.

The more complex MOS approach uses relationships between observed surface weather and predicted meteorological fields, where the predictions come from a particular configuration of a particular NWP model, with separate statistical relationships defined for each forecast interval of interest. In this way, MOS takes account of bias and random errors inherent in NWP model output. However, a MOS system is not transferable between models or even between different versions of a single NWP model, but must be updated to take account of any changes to the model for which it is defined.

In the context of climate modelling and prediction, statistical downscaling schemes are usually of the PP variety, using relationships between observed large-scale climate parameters (sea-surface temperatures, atmospheric circulation indices, etc) and monthly or seasonal averages of local climate anomalies. Such relationships may be applied to a set of climate change projections from global models to estimate local-scale climate changes implied by modelled changes in the larger-scale climate system (e.g., Mullan et al. 2001).

This paper explores the application of a PP downscaling system to a set of ensemble NWP 1-15 day forecasts. The downscaling is described in detail in a companion paper (Renwick et al. 2009). It uses a regression approach which relates synoptic-scale variability to station-level daily climate parameters. The motivation was to assess the information content for New Zealand of extended-range ensemble weather forecasts (Buizza et al. 2003; Hamill et al. 2004; Pelly and Hoskins 2003), and to explore whether the information inherent in the 15-day forecasts could be used to make
skilful one-month short-term climate predictions. The latter step is achieved by use of a conditional climatology of one-month anomalies (conditional on anomalies in the first 15 days of the month), to provide probabilistic one-month predictions dependent only upon initial conditions.

The intra-seasonal time scale (two weeks to one month) is potentially of great interest to user communities (Maunder 1986) as it is the time scale of much operational planning for weather-affected industries (e.g., construction, renewable energy generation, agriculture). Generally however, climate prediction focuses on the seasonal (three-month) average scale or longer (George and Sutton 2006; Rowell 1998; Shukla 1998; Shukla et al. 2000). In contrast, weather predictions cover only the coming week or two, since chaotic effects come to dominate signals in the initial conditions (Kalnay 2003; Weickmann and Berry 2007). The monthly time scale presents a difficult forecast problem, where the weather “noise” often over-shadows the climate “signal”, i.e. the one month scale is too long for a weather forecast based on initial conditions, and too short for a climate forecast based on boundary conditions. The approach to one-month prediction outlined here represents a hybrid approach, making use of information inherent in initial conditions.

The paper is set out as follows. First, data sets are described, and the statistical and data analysis approaches are outlined. Results are presented on the skill of medium-range (two-week) forecasts from the ensemble NWP system, and on the skill of one-month forecasts derived from the ensemble output. The final section summarises the method and results, and discusses the utility of the approach for operational forecasting.

2. Data and Methodology

a. Development data

Two kinds of data have been used in the development of the prediction system discussed here. The first is the observational data series used in the development of a
Figure 1: Map of New Zealand, indicating the location of climate stations used in the downscaling scheme (dots) and the grid spacing of the NCEP ensemble forecast fields (crosses, also grid point locations for the NCEP/NCAR reanalysis fields). The climate forecast regions are indicated by the thick solid lines and the region numbers.

statistical downscaling system designed for New Zealand daily climate observations. The development of that system and the data sets employed are discussed in detail in a companion paper (Renwick et al. 2009). A 15-year time series (1989-2003) of gridded 1000 hPa height (H1000) and 850 hPa temperatures (T850) over the New Zealand
region (Fig. 1), from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis (Kistler et al. 2001), were related in a regression analysis to daily climate variables at a number of New Zealand sites. The climate data were extracted for a total of 114 locations, from the National Climate Database maintained by NIWA. Variables of interest were precipitation, temperature (maximum, minimum, and 10 cm earth), solar radiation, and wind run. Station locations are illustrated in Fig. 1.

b. NWP model output

For the prediction experiments, H1000 and T850 fields were extracted from the NCEP GFS ensemble prediction system. The NCEP forecast model is presently run four times daily out to a 384 h lead time. Using the “bred vectors” approach (Toth and Kalnay 1997) up to May 2006, and more recently the ensemble transform method (Wei et al. 2008), a set of initial perturbations is produced. When the prediction experiments began, the ensemble had 10 members (5 “positive” and 5 “negative” perturbations) plus the control forecast, making a total of 11 members. In May 2006, the ensemble size was increased to 14 (plus control), and since March 2007 has contained 20 members (plus control). The integrations of the individual ensemble members are generally carried out at lower spatial resolution than the deterministic control run.

The NCEP model used in the reanalyses is an older and lower-resolution version of the operational NCEP global forecast model, but shares essentially the same dynamical core. The operational modelling system is regularly updated, through an ongoing series of upgrade cycles, as detailed at www.emc.ncep.noaa.gov/modelinfo/. The most significant upgrades in the period analysed here include an increase in horizontal resolution (May 2005), and increased ensemble size and changes to the method for calculating initial perturbations (May 2006). Both the reanalyses and the ensemble output are made available at 2.5° latitude-longitude resolution, though both models are integrated at higher resolution(s).
Routine downloading of NCEP ensemble NWP output began in August 2004. Twice daily fields of 1000 hPa geopotential height and 850 hPa temperature, from initial times at 00 and 12 UTC, were downloaded from the National Oceanographic and Atmospheric Administration (NOAA) Climate Diagnostics Center (CDC) web site, (www.cdc.noaa.gov/index.html), where for a given run from a specified initial time, all members of the ensemble and all forecast intervals in 12 h steps out to 384 h are packaged into single files, one for each of a range of model variables. Model values are extracted over the New Zealand region (30°-50°S and 160°E-170°W). This paper reports on trials over the two years between August 2004 and July 2006. Given the number of on-going model changes occurring through the trial period, no attempt has been made to compare the sub-periods with different numbers of ensemble members.

c. Downscaling

The downscaling approach is discussed in detail in a companion paper (Renwick et al. 2009). It consists of a set of regression equations defining diagnostic relationships between quantities derived from reanalysis H1000 and T850 fields (Kistler et al. 2001) and daily climate variables (anomalies from station climatologies) at 114 New Zealand sites. Regression equations were defined seasonally for six climate parameters: precipitation amount, maximum and minimum temperature, 10cm Earth temperature, solar radiation, and wind run; and for 8 event probabilities: 24h precipitation greater than 0, 1, 5, 10, 20 mm, 24h precipitation greater than median amount for the time of year, maximum temperature greater than 25°C, and minimum temperature less than 0°C. Candidate predictors were principal components of the first five empirical orthogonal functions (EOFs, Wilks 1995) of the H1000 and T850 fields, plus gradient winds and geostrophic vorticity calculated at each of the station locations, from the H1000 field. The system was shown to perform well on an independent set of reanalysis fields.

The use of such a system with predicted large-scale fields is a perfect prognosis (PP) approach to statistical forecasting, as discussed in the Introduction. Diagnostic relationships developed between reanalysis fields and contemporary station
observations are applied to an ensemble of forecast fields, resulting in an ensemble of estimated station observations, each a representation of station-level conditions if that particular ensemble member verified correctly. As discussed above, such an approach makes no allowance for forecast model error, either systematic or random. The ensemble provides an estimate of the random error (but not systematic error) inherent in the forecast, conditional on model limitations and the spread of initial conditions in the ensemble, while the statistical post-processing merely provides the downscaling step.

Validation of GFS forecast fields against reanalyses over the New Zealand region showed a small negative bias in T850, on the order of 0.5K averaged over the grid points in the New Zealand region, but variable with forecast interval and in time (related to operational model upgrades). This was taken account of in crude terms by applying a small bias correction to the T850 field, adding the same value at each grid point used. For the two year period reported on here, the correction was 0.25K out to 168h, and 0.8K beyond that. The correction is presently 0.6K out to the 168h forecast and 0.2K beyond that, reflecting changes in model characteristics with increased resolution and ensemble size. No bias adjustment is made to the H1000 fields (average biases were less than 5 geopotential metres at all forecast intervals, from mid-2005).

d. Conditional climatology of monthly climate anomalies

For development of the one-month conditional climatology used in conjunction with the 15-d downscaling, high-resolution gridded data sets of daily rainfall, and maximum and minimum temperature (Tait et al. 2006) were used, over the 30-year period 1974-2003. Values were calculated on a grid with approximately 5km horizontal spacing, using a spline approach that makes use of all available daily observations, plus background geographical and climatological information. The paper by Tait et al. (2006) describes the method as applied to rainfall data.

The monthly climate was defined as tercile outcomes. Terciles were defined at each location as equi-probable thirds of the full distribution, labelled “below normal” (coolest or driest one third of the distribution), “near normal” (central one third of the
distribution), and “above normal” (warmest or wettest one third of the distribution). Results were then inspected to determine sensitivity to location, season, orographic effects, and so on. Starting from daily values, monthly average anomalies were calculated at each grid point, and a tercile assignment made for each month and grid point, over the 30 year period. Again from the daily data, 15-day anomalies (days 1-15 of each month) were calculated, normalised to zero mean and unit standard deviation with respect to the long-term 15-day mean and standard deviation for the calendar month. The day 1-15 anomalies were then assigned to 21 bins of varying width such that there would be an equal number of observations in each bin if the distribution of anomalies was normal. From the binned data, the probability was calculated, over the 30-year period used, of the occurrence of each 1-month tercile T given a 15-day anomaly A (i.e., the conditional probability P(T|A)).

In operational trials, forecast 15-day anomalies were substituted for observed anomalies, retaining the perfect prognosis approach used in the downscaling. Tercile probabilities were calculated for each of six regions of New Zealand, as used in routine seasonal forecasting through the NIWA National Climate Centre (www.niwa.co.nz/ncc), by averaging forecast temperature or precipitation anomalies over all stations in each region. The region definitions are shown in Fig. 1.

3. Results

a. Conditional climatology specification

Figure 2 illustrates the result of the conditional climatology calculations, for the whole country and year, and for different latitude bands and seasons, calculated from the high resolution gridded data sets. The maximum and minimum temperature curves are quite similar and symmetrical, as might be expected. The precipitation curves are asymmetrical – e.g., a very dry first half-month can still lead to a tercile 3 full-month, but a very wet first half-month can never end up in tercile 1. The results show that, as a first approximation, the same tercile probabilities apply anywhere in New Zealand and in any season, but different probabilities apply according to the climate element (rainfall or
temperature). The tercile probabilities shown graphically in Fig. 2 are also listed in Table 1.

Figure 2: The probability of monthly tercile 1, 2 and 3 as a function of the anomaly for the first 15 days of the month. Separate panels given for maximum temperature (left), minimum temperature (centre) and rainfall (right). Top row uses all New Zealand grid points and all months; 2nd row splits results out for two regions (north of 37°S and south of 45°S, approximately); 3rd row splits results out by season (summer and winter).

While the tercile conditional probabilities listed in Table 1 apply quite generally around the country, there are some systematic regional variations. Figure 3 shows the 1-month tercile probabilities (for maximum (tmax) and minimum (tmin) temperature, separately) when the first 15-day average lies in the near-zero anomaly bin (within 0.060 standard deviations of the mean). When the month starts near normal, tercile 2 is the most likely outcome for the month as a whole (as expected), in all locations. However, the probability of tercile 2 decreases from north to south. The tercile that ‘gains’ is tercile 3 in the case of tmax, and tercile 1 in the case of tmin. When tmax is normal in the first half-month, the full month anomaly is equally likely to be in either tercile 1 or tercile 3 in the north of the country, but in the southern South Island tercile 3 is more likely than tercile 1 (while tercile 2 is still the most likely outcome). Conversely, when tmin is normal in the first half-month, the full month anomaly is equally likely to be in either tercile 1 or
Table 1: One-month tercile probabilities (averaged over all New Zealand grid points) when day 1-15 has the normalised anomaly as shown in first column. Probabilities shown for maximum temperature (tmax), minimum temperature (tmin) and rainfall (rain).

<table>
<thead>
<tr>
<th>Day 1-15 anomaly (in SD units)</th>
<th>tmax</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>&gt; +1.669</td>
<td>0.1</td>
<td>4.6</td>
<td>95.3</td>
<td>0.1</td>
<td>4.1</td>
<td>95.8</td>
<td>0.0</td>
<td>2.9</td>
<td>97.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.309, 1.669</td>
<td>0.7</td>
<td>11.4</td>
<td>87.9</td>
<td>0.7</td>
<td>10.0</td>
<td>89.3</td>
<td>0.1</td>
<td>14.2</td>
<td>85.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.067, 1.309</td>
<td>2.1</td>
<td>19.8</td>
<td>78.1</td>
<td>1.6</td>
<td>20.3</td>
<td>78.1</td>
<td>0.6</td>
<td>22.6</td>
<td>76.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.876, 1.067</td>
<td>3.3</td>
<td>27.2</td>
<td>69.5</td>
<td>2.9</td>
<td>28.7</td>
<td>68.4</td>
<td>1.6</td>
<td>30.7</td>
<td>67.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.712, 0.876</td>
<td>4.7</td>
<td>33.7</td>
<td>61.6</td>
<td>5.3</td>
<td>35.9</td>
<td>58.8</td>
<td>3.0</td>
<td>37.8</td>
<td>59.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.566, 0.712</td>
<td>6.3</td>
<td>38.9</td>
<td>54.8</td>
<td>7.7</td>
<td>40.2</td>
<td>52.1</td>
<td>5.1</td>
<td>42.6</td>
<td>52.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.431, 0.566</td>
<td>8.7</td>
<td>42.7</td>
<td>48.6</td>
<td>9.3</td>
<td>44.6</td>
<td>46.1</td>
<td>6.9</td>
<td>47.0</td>
<td>46.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.303, 0.431</td>
<td>11.0</td>
<td>46.2</td>
<td>42.8</td>
<td>11.7</td>
<td>47.3</td>
<td>41.0</td>
<td>9.4</td>
<td>49.5</td>
<td>41.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.180, 0.303</td>
<td>13.9</td>
<td>49.4</td>
<td>36.7</td>
<td>15.7</td>
<td>49.3</td>
<td>35.0</td>
<td>12.6</td>
<td>50.9</td>
<td>36.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.060, 0.180</td>
<td>17.8</td>
<td>51.9</td>
<td>30.3</td>
<td>19.4</td>
<td>50.6</td>
<td>30.0</td>
<td>16.5</td>
<td>51.2</td>
<td>32.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.060, 0.060</td>
<td>22.0</td>
<td>52.3</td>
<td>25.7</td>
<td>25.3</td>
<td>49.2</td>
<td>25.5</td>
<td>21.8</td>
<td>50.1</td>
<td>28.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.180, -0.060</td>
<td>27.7</td>
<td>51.8</td>
<td>20.5</td>
<td>30.6</td>
<td>48.1</td>
<td>21.3</td>
<td>28.0</td>
<td>47.5</td>
<td>24.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.303, -0.180</td>
<td>33.2</td>
<td>50.8</td>
<td>16.0</td>
<td>35.1</td>
<td>47.7</td>
<td>17.2</td>
<td>34.5</td>
<td>44.1</td>
<td>21.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.431, -0.303</td>
<td>39.5</td>
<td>48.2</td>
<td>12.3</td>
<td>41.3</td>
<td>45.2</td>
<td>13.5</td>
<td>40.7</td>
<td>40.7</td>
<td>18.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.566, -0.431</td>
<td>46.4</td>
<td>44.1</td>
<td>9.5</td>
<td>48.4</td>
<td>40.6</td>
<td>11.0</td>
<td>46.7</td>
<td>37.7</td>
<td>15.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.712, -0.566</td>
<td>54.6</td>
<td>38.8</td>
<td>6.6</td>
<td>56.2</td>
<td>35.1</td>
<td>8.7</td>
<td>54.3</td>
<td>32.7</td>
<td>13.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.876, -0.712</td>
<td>64.0</td>
<td>32.4</td>
<td>3.6</td>
<td>63.2</td>
<td>30.1</td>
<td>6.7</td>
<td>61.7</td>
<td>27.3</td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1.067, -0.876</td>
<td>73.7</td>
<td>24.5</td>
<td>1.8</td>
<td>69.2</td>
<td>26.7</td>
<td>4.1</td>
<td>67.8</td>
<td>23.0</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1.309, -1.067</td>
<td>81.9</td>
<td>17.4</td>
<td>0.7</td>
<td>74.8</td>
<td>23.1</td>
<td>2.1</td>
<td>71.9</td>
<td>20.7</td>
<td>7.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1.669, -1.309</td>
<td>90.1</td>
<td>9.5</td>
<td>0.4</td>
<td>83.0</td>
<td>16.2</td>
<td>0.8</td>
<td>77.3</td>
<td>17.2</td>
<td>5.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; -1.669</td>
<td>95.4</td>
<td>4.4</td>
<td>0.2</td>
<td>95.9</td>
<td>4.0</td>
<td>0.1</td>
<td>82.1</td>
<td>14.2</td>
<td>3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

tercile 3 in the south of the country, but in the northern North Island tercile 3 is more likely than tercile 1. The reasons for such latitudinal variations are unclear, and are being investigated.

b. Downscaling: two-week forecasts

An example downscaled forecast, of maximum temperature in Wanganui (see Fig. 1) is illustrated in Fig. 4. The median forecast tmax is shown as the black line, the interquartile range (IQR, estimated from the ensemble spread) is shown as the shaded region, and the climatological mean tmax for Wanganui is shown as the red line. Observed values are shown in blue. This example illustrates a number of typical features of the forecast system:
Figure 3: One-month tercile probabilities for maximum and minimum temperature, when the first 15-day average lies in the near-zero anomaly bin (within 0.060 standard deviations of the mean). The x-axis labels show the latitude bands used to bin the data: e.g., “[36,38)” means all latitudes in the gridded data set greater than or equal to 36°S but less than 38°S, etc.

- The IQR increases with forecast interval, as the ensemble spread increases,
- About 50% of the observed values fall within the IQR envelope, as expected, although the IQR appears to be underestimated over the first few days of the forecast period,
- The median forecast is quite accurate over the first 5 or so days of the forecast interval,
- The trend in the forecast (e.g. above climatology on days 3-5 but tending below climatology during week two) is representative of the observations, although errors on individual days may be large, beyond day 5.
Analogous comments may be made for forecasts of the other daily climate variables, allowing for differences in forecast skill between the climate variables.

![Example plot of a 15 d downscaled maximum temperature forecast from mid-November 2004, for Wanganui.](image)

**Figure 4**: Example plot of a 15 d downscaled maximum temperature forecast from mid-November 2004, for Wanganui. The heavy black line indicates the median of the forecast ensemble, the grey shaded area encompasses the inter-quartile range of the ensemble distribution, the red line is the daily climatological maximum temperature for Wanganui, and the blue line is the observed maximum temperature for the period indicated.

The geographical distribution of the skill of ensemble-mean forecasts (calculated over all days in the two-year validation period) is illustrated in Figs. 5 and 6, which show percentage explained variance (compared to climatology) for day-3 and day-8 forecasts respectively, for the six climate variables. At day-3, forecast skill was lower than the skill found in independent trials of the downscaling system, using reanalyses rather than forecast fields (Renwick et al. 2009). The reduction in explained variance was around 5% for temperature variables and solar radiation, but closer to 10% on average for precipitation and wind run. By day-8, forecast skill was significantly reduced again, but
was still positive even for precipitation amount, suggesting that there is deterministic information in the ensemble mean into the second week of the forecast interval, though it may be too low to be practically useful.

Figure 5: Percent explained variance for ensemble-mean day-3 forecasts of (a) wet-day precipitation, (b) maximum temperature, (c) minimum temperature, (d) solar radiation, (e) earth temperature, and (f) wind run. The period of record used was August 2004 through May 2005. The contour interval is 5%, 40% contour black, blue contours less than 40%, and red contours greater than 40%.
The two–year validation period was considered too short to give reliable statistics on the seasonality of forecast skill. However, results from independent trials using a ten year time series of NCEP reanalyses (Renwick et al. 2009) suggest that downscaling skill is highest in summer for temperature variables, and highest in winter for other variables (precipitation, solar radiation, wind run).

The statistics shown in Figs. 5 and 6 are calculated relative to a base of zero explained variance for “climatology”, i.e. using the long-term mean for the time of year.
as the forecast. Skill statistics were also calculated for persistence forecasts, using current observations as the forecast for each day over the coming fifteen. The PP forecasts exhibited greater explained variance than persistence forecasts in almost all cases out to day 8. At day 3, the average improvement in explained variance, compared to persistence forecasts, was around 27%. At day-8, the average improvement was around 6%.

To gauge the statistical significance of forecast skill, a set of 400 Monte Carlo experiments was carried out, to estimate the distribution of scores from a series of random forecasts drawn from the observational record. Forecast skill was greater than the 95th percentile of the distribution of random forecasts in almost all cases out to day 8. Beyond day 10, all measures of forecast skill were less than the 95th percentile of the distribution of random forecasts at more than half of the sites used. At day 3, the explained variance of the forecasts was less than the 95th percentile of the random distribution at three sites for predictions of the probability of precipitation greater than 10mm, and at three sites for the probability of maximum temperature greater than 25°C. By the day-8 forecast interval, the explained variance of the forecasts was less than the 95th percentile of the random distribution at around one third of sites for the above two parameters. For the non-probabilistic forecasts, day-8 explained variance exceeded the 95th percentile from the random trials at all sites for all temperature variables (max, min, 10cm Earth), at all but two for wind run, at all but seven for solar radiation, and at 90 sites (all but 24) for precipitation amount.
Figure 7: Forecast skill, averaged over all stations, for daily forecasts, as a function of forecast interval in days. Variables are (a), probability of measurable precipitation, (b) probability of at least 1mm precipitation, (c) probability of at least 10mm precipitation, (d), probability of at least median precipitation, (e) wet-day precipitation, (f) maximum temperature, (g) minimum temperature, (h) solar radiation, (i) earth temperature, and (j) wind run. In panels (a)-(d), the solid lines indicate mean ROC for yes/no probability forecasts, and the dashed lines indicate mean RPS skill. In panels (e)-(j), the solid lines indicate the average explained variance for daily forecasts, and the crosses indicate average fraction of explained variance for five-day means of forecasts centred on the day indicated.
Spatially averaged skill as a function of forecast interval is shown in Fig. 7, for the climate variables themselves, and for probabilistic forecasts of various precipitation events (measurable rain, at least 1mm, etc). Skill scores were calculated at each site and averaged in space. The probability forecasts are verified in terms of the Ranked Probability Skill Score (RPSS, Wilks 1995), and the Relative Operating Characteristic (ROC, Mason and Graham 1999; Wilks 1995). The RPSS is a measure of the relative RMS error of the forecast cumulative probability distribution (as a fraction of the RMS error of a climatological forecast), and is zero for a forecast as skilful as a forecasting climatological probabilities, negative for forecasts less skilful than climatology, and up to +1 for perfect forecasts. The ROC is a measure of the ability of the forecast probability to discriminate event occurrence. The ROC score used here is scaled to be analogous to the RPSS: the ROC is zero for a forecast as skilful as a forecasting climatological probabilities, negative for forecasts less skilful than climatology, and up to +1 for perfect forecasts.

All quantities show essentially zero skill beyond day 12, for individual days. As noted above, forecast skill was generally not statistically significantly different from zero beyond day 10. Five-day mean forecasts show somewhat more skill (shown as crosses in Fig. 7) than forecasts for individual days, but the skill of the 5-day mean forecast was often similar to the mean of the skill of the forecasts for each of the five days. The exception is for precipitation, where the 5-day mean forecasts are noticeably more skilful than the individual day forecasts that go into the mean. This may be due to the more “noisy” nature of daily precipitation, compared to temperature. Multi-day averages smooth out some of the random variability and leave a more obvious low-frequency signal. In other words, the results suggest that while the deterministic skill of precipitation forecasts may be low, there is skill in predicting periods of rain, even though the day to day timing and amounts may not be skilful. The ROC curves for precipitation probability forecasts also suggests that there is discriminatory information in the probability forecasts right through the forecast interval.
Figure 8: Hit rate (percent) for predictions of wet spells (more than half the period being rain days); (a) of days 3-7 having more than 2 wet days, and (b) of days 8-14 having more than 3 wet days. The contour interval is 5%, with values less than 60% in blue, values greater than 60% in red, and the 60% contour in black.

To investigate the above idea further, a series of trials were carried out with the precipitation probability forecasts. For each forecast day, the prediction was classed as “wet” if the ensemble median probability of more than 1 mm of precipitation was greater than the climatological probability for the time of year. Otherwise, the prediction for that day was classed as “dry”. Forecasts were then grouped into 5- or 7-day periods predicted to be “wet” if more than half the days (i.e. at least 3 or 5, respectively) were predicted to be “wet”. The observations were similarly grouped and labelled wet or dry. The hit rate (percentage of forecasts correctly predicting a “wet” or “dry” period) for the days 3-7 5-day period, and the days 8-14 7-day period (“week two”) is illustrated in Fig. 8, for the two year trial.
Consistent with the apparent extra skill of 5-day mean precipitation forecasts seen in Fig. 7, the hit rates shown in Fig. 8 are around 70% for much of the country, even for week two of the forecast interval. Although the forecast and observed frequencies were strongly weighted towards “dry” periods (less than half the days being rain days), by a factor of 4:1 for the 5-day period and 9:1 for the 7-day period, overall Hanssen scores\(^1\) (Wilks 1995) were positive in both cases, being 0.34 for the days 3-7 5-day period and 0.14 for the days 8-14 7-day period.

However, the hit rates illustrated in Fig. 8, and the associated skill scores, are only a small average improvement on those for persistence forecasts (using the frequency of wet days in the past week as the forecast for days 8-14). Forecast skill was less than that for persistence forecasts, and below the 95th percentile from the set of Monte Carlo random forecasts, at around one third of stations. The stations where the forecasts performed poorly (compared to persistence and random predictions) were almost all in the east of the country, where conditions are driest on average. Such results suggests that the information content of week-two forecasts of precipitation occurrence is marginal, especially in eastern regions.

c. One-month forecasts

As described above, a conditional climatology technique was used to estimate probabilities of one-month tercile outcomes for precipitation and mean temperature, from the downscaled ensemble forecasts. Mean temperature predictions were calculated as the average of maximum and minimum temperature forecasts. Predictions were averaged across the ensemble, across forecast intervals (for days 1-15 of the month) and within geographical regions 1-6 (Fig. 1). Probabilities for each temperature and precipitation tercile were taken from look-up tables developed by comparing observed 15-day anomalies against full-month outcomes (section 2c, Table 1). Here, the averaged forecasts were again used in a perfect prognosis sense, as if they were observed departures for the first half of the month.

\(^1\) The Hanssen score measures the ability to correctly predict a categorical outcome (e.g. wet or dry). It is zero for a forecast of climatological conditions, negative for forecasts less skilful than climatology, and up to +1 for perfect forecasts.
Forecasts were assessed for the 24 months between August 2004 and July 2006, in terms of their Hanssen score and ranked probability skill score (RPSS, Wilks 1995). Contingency tables were calculated using the probabilistic tercile forecasts directly, rather than by making the forecast binary by labelling it according to the most likely tercile. For example, if the forecast probabilities were 0.2, 0.7, and 0.1 for terciles 1, 2 and 3 respectively, and the outcome was tercile 1 (say), column 1 of the contingency table would be incremented by 0.2 in row 1, 0.7 in row 2, and 0.1 in row 3. Scoring this way takes some account of the “sharpness” of the forecasts (closeness to a categorical forecast), as well as their accuracy. It is also a more conservative scoring system. Assigning forecast states to the most likely forecast tercile before calculating a contingency table (i.e. by converting the above example forecast to be 0, 1, 0) generally results in higher hit rates and Hanssen scores.

The contingency table for mean temperature is shown in Table 2, with summary scores. Temperature forecasts show a small amount of skill compared to a climatological forecast, with a Hanssen score of 11.6% and RPSS of 19.8%. They are markedly more skilful than persistence forecasts for the same period (27% hit rate, Hanssen score -10%), and forecast skill is close to the 99th percentile of skill statistics from random Monte Carlo trials.

Table 2: Contingency table (percent occurrence) for two years’ of monthly tercile forecasts of mean temperature. The Hanssen score and RPSS are indicated below the table.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11.4</td>
<td>13.8</td>
<td>6.6</td>
<td>31.8</td>
</tr>
<tr>
<td>Tercile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11.3</td>
<td>15.9</td>
<td>13.4</td>
<td>40.6</td>
</tr>
<tr>
<td>3</td>
<td>4.7</td>
<td>8.7</td>
<td>14.1</td>
<td>27.5</td>
</tr>
<tr>
<td>Total</td>
<td>27.5</td>
<td>38.4</td>
<td>34.0</td>
<td></td>
</tr>
</tbody>
</table>

Hit rate 41.4%, Hanssen score 11.6%, RPSS 19.8%
During the validation period, tercile 1 was somewhat under-represented in the observations with 27.5% occurrence (compared to the long-term mean of 33.3%), with terciles 2 and 3 slightly over-represented. In contrast, tercile 3 was under-represented in the forecasts, and tercile 2 over-represented, suggesting a cold bias and a tendency to “predict” tercile 1. The tendency towards tercile 1 was expected, given that the 15-day anomalies used to estimate tercile probabilities are heavily averaged and display lower variance than the observed 15-day anomalies. Yet, the forecasts exhibit some positive skill, for a forecast interval (one month) where predictive skill is known to be low (e.g., Rowell 1998).

One-month precipitation tercile forecasts display no skill (Table 3). The hit rate of 36% is only slightly above the climatological hit rate of 33%, and the hit rate from persistence forecasts (35%). The hit rate, Hanssen score and RPSS were all less than the 90th percentile of scores from random trials. In the observational sample used, drier and near-normal conditions (terciles 1 and 2) were over-represented and tercile 3 was relatively rare. The distribution of forecasts precipitation probabilities followed the observed distribution, but again exhibited the tendency towards tercile 2 noted in the discussion of temperature forecasts, with the average forecast probability of tercile 2 being nearly 46%.

Table 3: As for Table 2, but for precipitation.

<table>
<thead>
<tr>
<th>Observed Tercile</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td>1</td>
<td>11.8</td>
<td>12.8</td>
<td>4.5</td>
</tr>
<tr>
<td>Tercile</td>
<td>2</td>
<td>18.2</td>
<td>19.0</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.8</td>
<td>10.2</td>
<td>5.1</td>
</tr>
<tr>
<td>Total</td>
<td>39.8</td>
<td>42.0</td>
<td>18.1</td>
<td></td>
</tr>
</tbody>
</table>

Hit rate 36.0%, Hanssen score 0.9%, RPSS 8.6%

To arrive at the one-month forecasts discussed here, global ensemble forecast information was statistically downscaled to a number of point locations, then re-
averaged across regions and in time. Is it possible to obtain more skill by relating ensemble-mean model output to regional climate anomalies directly? Tests were carried out (not shown) where ensemble mean model output was statistically related to regional average temperature and precipitation anomalies, using a combination of regression and EOF analysis as discussed above and in Renwick et al. (2009). The levels of skill obtained over the same two-year validation period (Aug 2004-Jul 2006) were almost identical to those reported in Tables 2 and 3, underlining the linearity of the approaches used here.

On the basis of the scores presented here, one-month temperature forecasts using this method appears to be of some practical use, but one-month precipitation probability forecasts appear to be of marginal utility at best, despite the apparent skill of the 1-2 week precipitation probability forecasts from the downscaling system. One-month forecasts generated as described here compare favourably to other statistically-based one-month forecast information used routinely at NIWA (e.g. Francis and Renwick 1998; Mullan and Thompson 2006; Zheng and Renwick 2003) in trials during 2007 (not shown).

5. Discussion and Summary

A perfect prognosis system, for downscaling large-scale atmospheric circulation over New Zealand to daily climate anomalies at land stations, has been applied to daily NWP output from the NCEP GFS ensemble prediction system. The system was used to directly illustrate the likely range of outcomes at stations for each of the realisations of the large scale circulation in the ensemble. Downscaled forecasts were validated on a day-by-day basis and for time averages through the 15-day forecast interval. Fifteen-day average forecast anomalies of temperature and precipitation were used to estimate probabilities of monthly tercile outcomes for six climate regions of New Zealand.

Daily forecasts from the downscaling system exhibited useful levels of skill through approximately the first week of the forecast interval, as expected from general considerations of atmospheric predictability. Five-day averaged forecasts generally exhibited some skill out to the day 6-10 average. Forecast probabilities of a rain-day (at least 1mm of precipitation in 24h) were found to be marginally skilful through week two
of the forecast, in the sense of predicting whether the week will be generally wet (more than half the days being rain-days) or dry.

Downscaled forecasts (as illustrated in Fig. 4) were made available to a small group of potential users from late 2004. There was general agreement that the visual display used, clearly illustrating the forecast trend in comparison to climatology, would be useful in a qualitative sense as input to decision-making in the agriculture and other climate-sensitive sectors. For example, decisions around the timing of fertilizer application, or about the deployment of a work force for outdoor maintenance work, could be informed by qualitative indications of the likelihood of rain, or of a period of high or low temperatures over the coming ten days or so. Some users are presently trialling the use of 15-day forecasts for operational decision-making.

One-month probabilities of terciles of precipitation and mean temperature for regions of New Zealand were calculated using a conditional climatology technique, using the ensemble-mean forecast anomalies for the first 15 days of the month. One-month forecasts of mean temperature exhibited positive skill, while precipitation forecasts exhibited essentially no skill.

The extension to one month predictions was somewhat successful for temperature but of little use for precipitation. The relatively low predictability of precipitation processes (beyond a short-term forecast) implies that there is little correlation between forecasts of the first two weeks' precipitation and precipitation totals for the whole month. Yet, the conditional climatology calculations discussed in section 4a suggest there is a relatively strong relationship between observed 15-day and one-month precipitation totals. Hence, the lack of skill of the one-month forecasts discussed here is related to errors in the forecast amounts for the first 15 days of the month.

For a number of users in the agriculture and other climate-sensitive sectors of the economy, the greatest utility to be gained from the forecasts discussed here is likely to come from combining them with information from other sources. Future work will investigate the integration and optimal combination of downscaled ensemble forecasts with observations, short-term deterministic NWP predictions, and with longer-scale
seasonal climate forecasts, to develop a picture of the evolution of local climate anomalies from the recent past into the medium-term future.

Acknowledgements

The authors thank staff at NCEP and at NOAA CDC for advice, information, and assistance with data sets and data transfers. Comments of two anonymous reviewers helped considerably with the structure of the paper and the interpretation of the results. The work described here was funded by the New Zealand Foundation for Research Science and Technology under contracts C01X0202 and C01X0701, and associated capability funding.

References


*Submitted to Weather and Climate September 2008 Revised manuscript submitted May 2009*