Journal of Hydrology (NZ) 44 (2): 91-104, 2005

Skill assessment of a linked precipitation-runoff flood forecasting system

Richard Ibbitt,[1] Craig Thompson[2] and Richard Turner[2]
National Institute of Water and Atmospheric Research
[1] *PO Box 8602, Christchurch, New Zealand*
[2] *Private Bag 14-901, Wellington, New Zealand*

Abstract

Quantitative Precipitation Forecasting (QPF) is increasingly used to extend flood forecasts. In New Zealand a QPF-based flood forecasting procedure based on the Topnet rainfall-to-runoff model has been under trial since 2000. The trial has proved invaluable in identifying system features that need to be made more robust, but it has proved difficult to obtain objective and comparable results on the forecast skill of the system, owing to the variety of conditions under which operational forecasts are made. To examine the skill of the forecast system, a series of forecasts have been made using historical events in a pseudo-real time environment. The tests have been made under controlled conditions, so that the results are comparable across catchments. Forecast skill has been examined with respect to catchment size, inclusion of new flow information at different times through a QPF, and the use of two simple forecast adjustment schemes. Results suggest that forecast skill increases with catchment size, and that shifting and scaling forecast flows to match recently measured flows up to the time of the forecast can improve forecast skill. Overall the forecast skill shown by the trials is not high. One reason for this is the short period of data available for calibrating the rainfall-to-runoff models used for each catchment. The results are likely to define a lower envelop of skill, owing to the necessary omission of features such as upstream information that would be included in an operational system.

Introduction

A modern society requires reliable and accurate flood forecasts to reduce the cost of flooding. Flood forecasting procedures are seldom analysed after events, however, except following major widespread and devastating floods such as those in Southland in 1984 and in the Manawatu in 2004.

Objective comparisons of flood forecasting procedures are difficult, as later forecasts during an event often have more, and better, input information than is available for earlier forecasts. For example, an initial forecast may have information only on rainfall conditions at a single station, whereas later forecasts may also have access to information from stream flow stations. At the time each forecast is made, the objective is to produce the most reliable forecast. To compare the performance of forecast systems, different combinations of forecast circumstances must be reduced to a manageable set. Pearson and Jordan (1991) used a questionnaire-based approach to examine forecast results from a number of operational systems used around New Zealand. They found that in many cases expert assessment as to how the inputs to an event would evolve was an inherent part of the system. They showed

that automatic forecasting procedures were more accurate than manual methods. However, they were not able to reduce the responses to an objective and quantitative set of skill measures, and stated only general conclusions about how well flood forecasting methods work in New Zealand. Because of operational requirements, use of the same forecasting system in a changing operational environment will seldom provide strictly comparable information on system performance. To properly compare the forecasting behaviour of a system, forecasts must be made in a controlled environment, where sources of variation between forecasts are known and can be quantitatively assessed.

The use of Quantitative Precipitation Forecasts (QPFs) to extend flood forecasts is becoming increasingly popular, with the Journal of Hydrology devoting volumes 239 (issues 1-4, 2000) and 288 (issues 1-2, 2004) to QPF and the use of the results for flood forecasting. Ibbitt *et al.* (2000) reported on the use of QPF to produce pseudo-real-time flood forecasts over the Southern Alps of New Zealand. Their results were based on rainfall defined on an approximately 20-km grid, and they noted that calculated rainfalls seemed to underestimate measured amounts by about 50%. Copeland *et al.* (2000) experimented with smaller grids and concluded that a 5-km grid would be adequate to represent the uplift and cooling mechanisms associated with transport of moist air across the Southern Alps. Henderson *et al.* (2002) compared quantitative precipitation estimates with measured runoff and confirmed Copeland's results that modelling precipitation using a 5-km resolution produced estimates that agreed well with measurements.

Table 1 shows a general comparison of the topographic gradients (and hence rates of uplift of moist air) across many of the world's mountain ranges. The Southern Alps stand out because they rise steeply from the coast and the question has been raised as to whether or not the QPF results for the Southern Alps benefit from the rapid uplift

Table 1 – Properties of mountains that significantly obstruct global circulation

Name of mountain range	Annual rainfall m/y	Length km	Height m	Distance from the coast km	Height/Distance from the coast m/km
Andes (South America)	0-2	12,000	7,500	150	50
Coastal Range (USA/Canada)	4+	2,000	4,000	200	20
Drakenbergs (South Africa)	1+	1,000	3,000	100	30
Blue Mountains (Australia)	1-2	1,000	2,000	100	20
Norwegian Alps	2+	1,000	1,000	40	25
Deccan (India)	2	1,000	1,000	50	20
Madagascar	5+	1,000	2,000	100	20
Southern Alps (New Zealand)	5-10	500	3,000	20	150

that is a feature of modelling the Southern Alps. This paper reports on an investigation into how well a flood forecasting system that uses QPFs generated by a meteorological weather model—the Regional Atmospheric Modelling System (Pielke *et al.*, 1992)—works at many locations around New Zealand.

The aim of this study has been to objectively assess the skill of a modern flood forecasting procedure. To eliminate sources of uncontrolled variation in forecast accuracy we have used only QPFs. These forecasts are produced on a 5-km spatial grid using widely accepted and validated meteorological forecasting procedures. We have focused on large floods to avoid biasing the results by forecasting a lot of "non-events". We realise that the forecasting of "non-events" will happen, but we have tried to reduce these by setting objective criteria for initiating flood forecasting procedures. We have used a single rainfall-to-runoff modelling system, albeit with four variants, to adjust the forecast to current river flow conditions and devised an objective way to assess the results from each forecast.

Forecasting analysis assessment scheme

Although meteorologists have standard objective methods for assessing the skill of their forecasts (Wilks, 1995), few of their methods seem to have been applied to assessing the accuracy of flood forecasts. In this paper "skill" concepts used by meteorologists have been adapted for hydrological use. These are used to objectively assess the national performance of flood forecasting around New Zealand using QPFs.

Meteorologists are primarily concerned with forecasting rainfall. They assess their skill by comparing forecast rainfall amounts with those subsequently measured. In all situations there is both a forecast amount and a measurement, although one or other, or both, of these quantities may be zero. Forecast events can be split into discrete categories of event—occurrence or non-occurrence—and a contingency table can be constructed to summarise forecast performance.

The contingency table approach cannot be strictly applied to flood forecasts, since the forecast event, a flood peak, may not occur within the forecast horizon. For example, over the next forecast period, the forecast may be for a continually rising hydrograph. Figure 1 shows a variety of hydrograph forecast situations.

As an alternative to the contingency table approach, it is possible to define a skill score based on an error measurement statistic. A skill score refers to the relative accuracy of a set of forecasts with respect to a set of standard reference forecasts (Wilks, 1995). A formula for skill score based on the mean squared error (MSE) is:

$$SS = 1 - MSE/MSE_{clim} \quad (1)$$

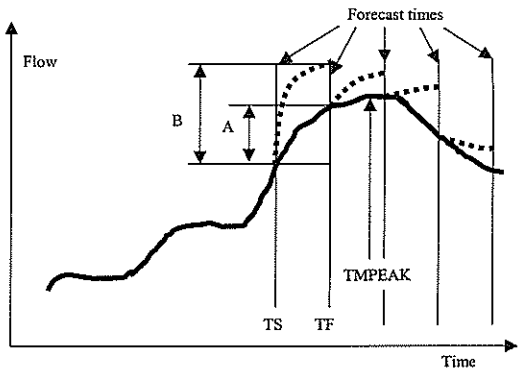

Figure 1 – Different forecast situations during a flood event. The solid line represents the measured flows and the dotted lines represent each forecast. TS is the starting time of a forecast, TF is time to the end of a forecast, TMPEAK is the time of the measured peak, A is the measured increase in flow over the forecast horizon, and B is the forecast increase over the same time period.

where SS is the Skill Score, MSE is the mean squared difference between the forecast variable and the corresponding measured variable, and MSE_{clim} is the mean squared difference between the forecast variable and a "climatological" reference variable. For MSE_{clim} we used the mean, over a sequence of forecasts, of all the measured flows, volumes of flow, or times to peak flow, for a particular river station.

The limits of SS vary from $-\infty$ to 1, with skill scores greater than zero indicating a forecast that is better than the forecast made using just the climatological value, i.e., an improvement over the climatologically-based reference flow. For this presentation we have transformed the SS using the transform:

$$TSS=(SS-1)+1/(1-SS) \quad (2)$$

where TSS is the Transformed Skill Score that varies between $\pm\infty$. Transformed Skill Score values gives a better perspective on the number of results that show skill with respect to those that do not, since the results plot symmetrically about the zero line. Positive values indicate a skillful forecast and can be considered better than forecasts made using just the climatological value.

In the results presented below, Transformed Skill Score values have been calculated for the flow at the end of the forecast interval, the volume of flow during the forecast interval, the peak flow (if both a measured and forecast peak occurred during a forecast interval), and the corresponding time to peak flow.

The forecast process

Two factors were important to draw valid conclusions about the success of the forecasting system: forecasts had to be carried out in a controlled way, and the forecasts needed to cover a range of basin sizes and locations.

To control the environment for assessing the forecast procedures, all forecasts were made using a sequence of QPFs for a succession of 24-hour periods. All forecasts were made at a spatial resolution of 5 km and used the same meso-scale weather modelling system. Forecast sequences were made for seven historical rainfall events that caused high flows in some part of New Zealand. Table 2 lists the storms and the general locations for which forecasts were made. Because of the amount of computing required to produce a sequence of QPFs, forecasts were made for only those regions in which rivers rose significantly during the period of the event.

The first 24-hour forecast of each sequence served as a "warm up" period for the rainfall-to-runoff model, to reduce the impact of initial conditions on subsequent forecasts in the series.

Table 2 – Starting dates, durations, and regions of the storms used for flood forecasting

Start date of storm	Duration of storm (days)	Regions
13 Nov 1999	10	all New Zealand
19 Jan 2000	25	upper South Island
15 Aug 2000	10	North Island, upper South Island
14 Oct 2000	10	lower North Island, South Island
17 Nov 2001	10	lower South Island
2 Dec 2001	10	all New Zealand
18 Jun 2002	21	North Island

The Topnet rainfall-to-runoff model was used. This is a spatially distributed model that uses grid-based rainfall information to derive the different runoffs for each sub-basin within a main basin. Each catchment model that was built was calibrated using 5-km resolution RAMS analysis rainfalls derived for the Southern Alps Experiment (SALPEX) 1996 period and the corresponding measured flows.

We used simulations that approximate operational flood forecast situations, with new forecasts being made each time new data became available. In an operational environment that uses QPFs, a new precipitation forecast becomes available every 24 hours, while flow data can be updated more frequently. To maintain a manageable number of flood forecast runs it was assumed that the flow data were updated every 6 hours, so that new forecasts were also made every 6 hours. Thus the first forecast in each sequence begins with both a new precipitation forecast and measured flow data up to the forecast start time. The next forecast used the same forecast precipitation, but updated flow data. At the fourth forecast a new precipitation forecast becomes available and this is combined with the latest measured flow data to begin a new set of four flood forecasts. We have separately analysed each set of four forecasts.

A flow criterion was set for setting in motion the flood forecasting procedure. Here the flood forecasting system was initiated when either there was heavy rainfall forecast over the catchment (rainfall greater than 100 mm/day) or the flow at the forecast time was significantly greater than normal (flow greater than the flow exceeded 10% of the time during the historical record). When forecast rainfall is in excess of 100 mm/day, a heavy rainfall warning is issued by the meteorological authorities (www.metservice.co.nz/forecasts/email_lists_definitions.asp). A flow criterion was considered necessary, as some forecasting systems automatically initiate an alarm at a pre-determined water level or flow value. The value chosen was considered to be realistic and would also prevent the results being unduly influenced by forecasts of many small events. Thus forecasts were made only when conditions were considered to be potentially flood-producing.

Procedures at the start of a forecast were adjusted to ensure that the initial forecast is consistent with the latest flow measurements. For most flow forecasts produced by rainfall-to-runoff models, at the start of the forecast there is a mismatch between the measured and simulated flows. A number of options are available for handling this mismatch. It can be ignored, but this implies that the latest information in the flow signal is not being used. A factor can be derived, e.g., the ratio of the measured flow to the simulated flow, that when appropriately used, ensures that the measured and simulated flows match at the forecast time. This factor is then used to modify all the subsequent forecast flows. We call this "scaling" and it is one of the options we have used for making use of the latest measured flow information. The forecast flows can be shifted, and scaled to more closely match the measured rate of flow increase just prior to the forecast time. These flows effectively contain information about the rate of response of the catchment to rainfall. So, for example, if the rainfall forecast indicates heavy rainfall will occur before it actually does, the simulated flows are likely to begin rising too early. By comparing the measured and simulated flow it is possible to delay the onset of the hydrograph rise and thereby compensate for some of the error in the forecast rainfall. We have used such a "shifting" scheme to make this adjustment to the flow forecast. For smaller catchments, too long a maximum shift may result in the simulated hydrograph being shifted to match a previous hydrograph rise rather than the

current one, so we set the shift proportional to the square root of the catchment area.

All forecast precipitation is assumed to fall as rain, i.e., no snow modelling has been attempted, although snow amounts are available from the RAMS model.

To assess variability across the country a number of basins were selected, from Mangakahia in Northland to the Mataura in Southland. For realistic forecasts, basins must be large enough to be covered by many rainfall grid points, since QPFs can contain spatial errors where the forecast rainfall field is displaced by one, two, or more grid positions from the actual rainfall field. As the number of rainfall forecast points inside a watershed increases, the homogenising effect of the basin reduces the likelihood of serious errors from a positional error in the entire rainfall field. The basins detailed in Table 3 are generally the largest in their region. However, the requirement that we use measured flows sometimes reduced the size of basin we could use, as the available flow monitoring stations were sometimes well inland. Table 3 lists the final basins used, along with their areas, and an indication of the quality of the calibration that was obtained using the SALPEX96 analysis rainfall data.

Model calibration results

All models were fitted to hourly flow data measured during the SALPEX96 period (9 October to 7 November 1996). These data are continuous, provide national coverage, and contain some storm activity within all the study catchments. They also provide estimates of the hourly rainfall where there are no measurements, and the estimates have been checked extensively against measurements, e.g., Henderson *et al.* (2002).

The length of the SALPEX96 period is shorter than is ideal for a reliable model calibration. However, it has the merit that during this period the rainfall information was produced on the same basis as that used for the forecasts, i.e., it was hourly data produced using the same quantitative precipitation model and on the same 5-km spatial grid. Furthermore, the SALPEX96 rainfalls come from "analysis runs", i.e., the simulations use global weather forecasts that are regularly adjusted back to measured information, so that chaotic divergence of the simulation with time is minimal. Thus the model-generated rainfall is a realistic simulation of what was likely to have occurred for the entire SALPEX96 period.

The principle disadvantage of the data is that the storm activity is focused on the Southern Alps. For other parts of the country the events in the record are not particularly significant, so that there may be insufficient information in the data for good model calibrations.

Simulated rainfall often underestimates that which would be measured (Henderson *et al.*, 2002). To compensate for underestimation, all the rainfall, whether used for calibration or forecasting, was adjusted by multiplying it by a constant factor. The factor was based on a comparison of the total simulated rainfall over the basin with the total measured runoff and estimated evaporation for the whole SALPEX96 period. The calculation is only approximate but it serves to counteract a known problem with QPFs in New Zealand.

One risk with increasing the rainfall over a comparatively short time span such as the SALPEX96 period is that a multiplier may compensate for a poor estimate of the initial storage conditions in the catchment at the start of the calibration period. Since we have no way to accurately estimate the actual initial conditions in a catchment, this is a risk we have had to take. It may explain why multipliers greater than unity were found beneficial to the calibrations of the Clutha and Haast, catchments where we would have expected a unit value to be appropriate. These

catchments drain from the Southern Alps, where Henderson *et al.* (2002) and Copeland *et al.* (2000) have shown that quantitative precipitation estimates on a 5-km grid closely match the actual rainfall.

Forecast results

The "end-of-forecast period" runoff and the corresponding "volume-of-flow" categories are discussed first. For the forecasting of the flood peak and its timing there are fewer results available for comparison, as often the model might predict a peak in a forecast interval when none actually occurred during that period, or the converse situation could occur. In each of the figures used to present the results, the Transformed Skill Score values from equation 2 have been plotted

Table 3 – Details of the river basins used to assess the skill of flood forecasts.

Basin name	Site number	Basin area (km^2)	Rank of basin area	Quality of model calibration*	Maximum time shift used in the Variable Shift tests (h)	Flow exceeded 10% of the time (m^3/s)	Rainfall factor
Awatere at Awapiri	60203	987	3	Good	3	27	1.0
Clutha at Clyde	75213	12018	14	Medium	12	850	1.5
Grey at Dobson	91401	3830	11	Good	7	760	1.0
Haast at Roaring Billy	86802	1020	4	Medium	3	390	1.6
Mangakahia at Titoki Bridge	46626	798	2	Good	3	50	2.4
Mataura at Seaward Downs	77519	5109	12	Very Poor	8	130	1.0
Mokau at Totoro Bridge	40708	1046	5	Poor	4	68	1.4
Motu at Houpoto	16501	1393	6	Good	4	170	1.0
Motueka at Woodstock	57009	1750	8	Good	5	140	1.0
Ruamahanga at Waihenga	29202	2340	9	Good	5	160	1.7
Tukituki at Red Bridge	23201	2380	10	Extremely Poor	5	55	1.0
Waipaoa at Kanakanaia Cableway	19716	1580	7	Very Poor	4	65	1.0
Waitara at Tarata	39501	725	1	Medium	3	120	2.7
Whanganui at Paetawa	33301	6643	13	Good	8	440	2.1

* Calibration quality is assessed by the value of the Coefficient of Determination.
Good 90-80%, Medium 80-70%, Poor 70-60%, Very Poor 60-50%, Extremely Poor <50%.

Table 4 – Percentage of positive skill scores in each forecast category. The rows in bold give the statistic being assessed and the forecast lead times with respect to the latest flow update, i.e., **18-h** indicates that 6 hours into a 24-hour rainfall forecast the flow information was updated. Results are reported for flows where (i) no adjustment was made to the latest flow data, (ii) where the modelled and measured data were linearly scaled to match at the last time flow data became available, and (iii) where the forecast flows were shifted in time and linearly scaled using the best match to the flows over the preceding period before each forecast.

End of forecast flow	**24-h**	**18-h**	**12-h**	**6-h**
No adjustment	50	50	50	50
Scaling only	57	21	43	43
Shifting and Scaling	50	43	50	64
Volume				
No adjustment	71	64	64	43
Scaling only	64	29	43	43
Shifting and Scaling	57	43	50	50
Peak flow				
No adjustment	44	25	57	67
Scaling only	56	22	71	67
Shifting and Scaling	44	44	88	86
Time to peak				
No adjustment	44	13	29	0
Scaling only	44	22	14	0
Shifting and Scaling	44	44	25	38

Table 5 – Ranks of the positive skill scores in each forecast category in Table 4, where a rank of "1" indicates the best result in a category/

End of forecast flow	**24-h**	**18-h**	**12-h**	**6-h**
No adjustment	2=	1	1=	2
Scaling only	1	3	3	3
Shifting and Scaling	2=	2	1=	1
Volume				
No adjustment	1	1	1	2=
Scaling only	2	3	3	2=
Shifting and Scaling	3	2	2	1
Peak flow				
No adjustment	2=	2	3	2=
Scaling only	1	3	2	2=
Shifting and Scaling	2=	1	1	1
Time to peak				
No adjustment	1=	3	1	2=
Scaling only	1=	2	3	2=
Shifting and Scaling	1=	1	2	1

against catchment area, and there are separate plots for each set of forecast lead-times.

Tests on the Transformed Skill Score values showed that the assumptions needed to calculate confidence limits would not be justified. Consequently, the results have been summarised in terms of visual trends. Table 4 gives a summary of the results in terms of the relative number of results showing a positive skill score in each category of results. Table 5 summarises the results in terms of the rank of each result with respect to the other results in the same category. The number of results in Tables 4 and 5 differ from those shown in Figures 2-5 because some data were outside the plot range of the figures in both the positive and negative directions.

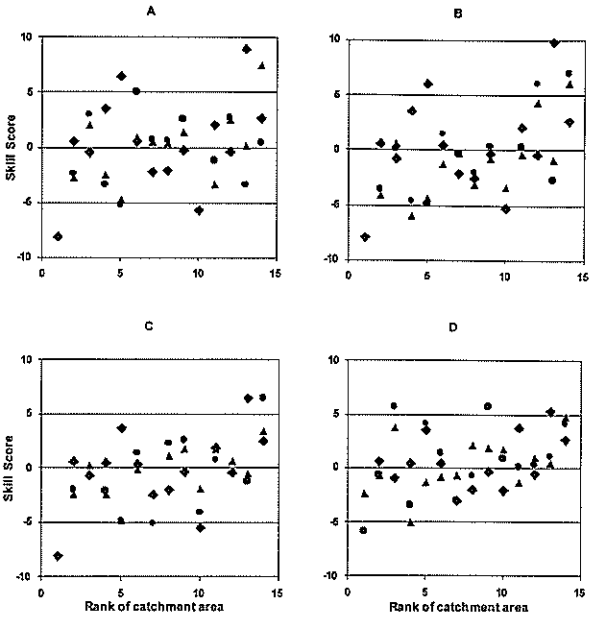

Figure 2 – Transformed Skill Score versus Rank of catchment area for the end-of-period forecast flow. See Table 3 for catchment names that correspond to each rank. Diamond-shaped symbols are for forecasts with no adjustment to the latest flow data, triangles are for forecasts with matching of the flow values at the forecast time, and circles are for forecasts with matching of the flow values and temporal shifting of the forecast. For Plot A, the forecast is based on a 24-hour ahead rainfall and flow information at the start of the forecast, B uses the same rainfall but flow information has been updated 6 hours into the forecast, C is the same as B but 12 hours into the forecast and D is the same as B but 18 hours into the forecast.

Flow at the end of the forecast period

Figure 2 shows how well the system forecast the flow in each river at the end of the each forecast period.

In general, plots of Transformed Skill Score versus rank of catchment area for end-of-period forecast flow (Figs. 2A-2D) show positive skill scores for the larger catchments. However, no clear picture emerges as to whether the two adjustment schemes generally improve the forecasts or not. Occasionally there is a spectacular improvement for an individual catchment for a particular forecast horizon, e.g., for the Mataura River, scaled and shifted forecasts 12 hours into a 24-hour forecast have a Transformed Skill Score value of 20, which is well in excess of the axis maximum for Figure 2C. It may be that flow data at the start of the flood forecast process added little of value, since at earlier times the hydrograph at the Mataura station may have shown no indication of the coming flood, i.e., the "12-hour" forecasts are the first to show the hydrograph rising. However, there is no similar improvement at the "6-hour" forecast, and so we must conclude that the "12-hour" forecast result is fortuitous.

Throughout the results, the Waitara River forecasts are consistently worse than those for other rivers in the sample. Although the results for the Tukituki catchment were also

initially poor, they improve considerably, but still show negative skill scores as updated flow information was introduced into the forecasts.

No discernible trend with respect to the efficacy of the adjustment strategies is apparent in any of the plots of Transformed Skill Score versus rank of catchment area for end-of-period forecast flow (Figs. 2A-2D).

Volume of flow over the forecast period

Figure 3 shows how well the volume of water over each forecast period was forecast. There is one value in this set of results for each value discussed in the preceding section. However the results show somewhat different trends. In general, plots of Transformed Skill Score versus rank of catchment area for forecast volume of flow (Figs. 3A-3D) show more positive skill scores than negative ones, with a trend towards decreasing skill with catchment size. Also, for the forecasts made with the least information, those labelled "24 hour ahead forecasts", the adjustment schemes seem to have been generally beneficial. Surprisingly, as more information is introduced into the forecasting process, the skill at forecasting the volume of runoff over the forecast period and the efficacy of the adjustment schemes seem to deteriorate.

As with the runoff results there is a wider range of skill scores at smaller catchment areas, but this may be simply a reflection of the fact that there are more smaller catchments in the sample.

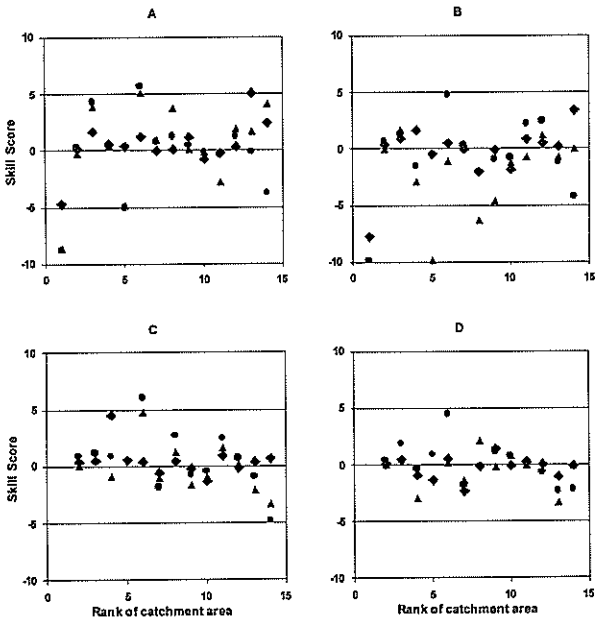

Figure 3 – Transformed Skill Score versus Rank of catchment area for forecast volume of flow. Symbols and meaning of plots A, B, C, and D are the same as for Figure 2.

Peak flow

Transformed Skill Score versus rank of catchment area for forecast peak flows are shown in Figures 4A-4D. As already noted, there are fewer results for the forecast peak flows than for the end-of-period flow values; this makes it more difficult to summarise any trends in the results. For the larger catchments, peak flow is forecast with a generally positive skill score. However, adjustment of the forecast to the latest measured flow data can make the forecast reliability more variable. Some of the results for adjusted Clutha River forecasts show an almost perfect skill score (TSS=50), whereas others lose whatever skill there was in the unadjusted forecast. In general, unadjusted forecasts for the larger basins seem more robust, in the sense of giving a more consistent indication of the quality of the forecast. For the smaller catchments, adjustment of the initial forecast seems to lead to better forecasts.

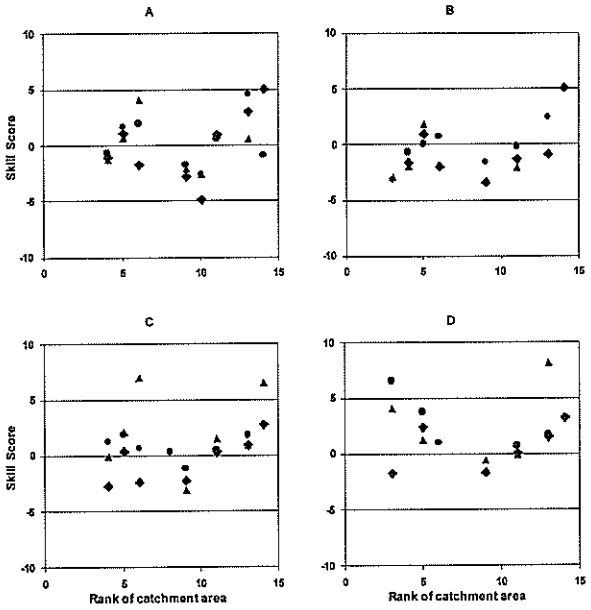

Figure 4 – Transformed Skill Score versus Rank of catchment area for forecast peak flows. Symbols and meaning of plots A, B, C, and D are the same as for Figure 2.

Time to peak flow

For each peak flow forecast there is a corresponding forecast for time to peak flow. Figures 5A-5D show that the forecast skill of when a flood will peak seems to decrease with increasing catchment size. However, the majority of the results in this category have negative skill scores.

Timing problems in a flood forecast generally begin with the global weather model, which sets the boundary conditions for the meso-scale weather models. Global models, typically using a spatial scale of the order of 100-200 km, can introduce large timing errors. For example, slow-moving fronts can take many hours to traverse a global model grid cell. A small error at the global scale can lead to a larger error at the basin scale. We would expect timing errors to reduce greatly with the use of data assimilated into the meso-scale weather model, as improved timing at smaller spatial scales is an intended outcome of data assimilation research.

Discussion

Unadjusted results

The rainfall-to-runoff model component of a flood forecasting system that is based on rainfall forecasts preserves mass, i.e., what rainfall is fed into the model is accounted for in terms of runoff and storage changes. Any error in the rainfall forecast will thus be reflected in the forecast runoff hydrograph. However, the severity of the forecast errors will depend

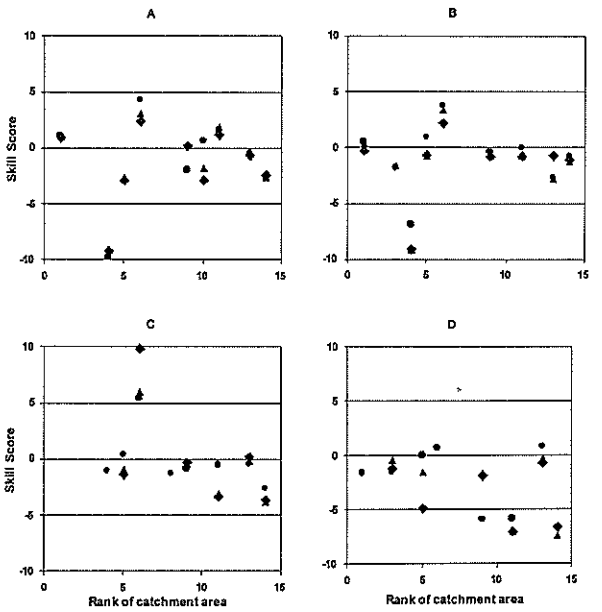

Figure 5 – Transformed Skill Score versus Rank of catchment area for forecast time of peak flow. Symbols and meaning of plots A, B, C, and D are the same as for Figure 2.

on the quantity being forecast. The volume of runoff over a time period of 24 hours is probably the easiest quantity to forecast with some skill since, over the time that the volume is accumulated, overestimated flows in one interval can be compensated for by underestimates in a later interval. However, if the length of the forecast period is reduced, the opportunity for compensation between intervals becomes less. This can be seen in the plots of Transformed Skill Score versus rank of catchment area for forecast volume of flow (Figs. 3A-3D) and Table 4. The results given in this study confirm this expectation.

The next easiest quantity to forecast ought to be the peak flow, since this also has a storage element to its calculation. As the runoff from individual sub-basins of the model are combined and routed down the channel network, overestimates from some sub-basins may be offset by underestimates from other sub-basins. The larger the catchment, the greater will be the chance of this happening, so it is not surprising that generally positive skill scores occurred for the larger catchments, with an area rank of 10 or greater, in this aspect of the study.

Forecasting the flow at the end of the forecast period, and forecasting the time when a river will peak are the more difficult aspects of flood forecasting to get right, because they are more dependent on the correct timing of the rainfall. While some compensating errors may cancel out in the forecast of the end-of-forecast period runoff, these are not as great as in the case of forecasting the flood peak, where time is a separate factor in assessing the forecast performance. Consequently, these two components of forecast assessment are likely to be more sensitive to errors in the rainfall forecast. The results of this study confirm this—while some skill is shown in the forecast of end-of-forecast period flows, the results for the time to peak flow were generally disappointing. However, assessing the time of a flood peak to within an hour or so may not be of great operational concern for rivers whose peak flows last over several hours. Particularly for a large catchment, use of an hourly time step to assess the quality of the timing of a forecast is a fairly stringent test. Better timing skill scores would have been achieved using a time resolution of, say, three hours.

As the plots of Transformed Skill Score versus rank of catchment area for end-of-period forecast shown in Figures 2A to 2D progressively include flow data that is closer to the end of the forecast period, the skill scores for the end-of-period runoff should show a gradual increase, and this is the case.

In some ways, the end-of-period runoff results conflict with those for the volume of runoff. This may arise because, as the end of the forecasting period is approached, the measured and forecast flows converge, whereas the opportunity for volumetric errors to compensate for one another lessen, as fewer data are being used. Collischonn *et al.* (2005) noted, in connection with the use of a coupled rainfall/rainfall-to-runoff modelling system in an area subject to convective storms, QPF-based forecasts were often worse than assuming that the rainfall ceased after the forecast time. One explanation for this could be poor spatial location and/or timing of rainfall from convective cells. Although our storms were frontal rather than convective, temporal and spatial errors could still lead to some forecasts giving worse results, i.e., receiving a lower skill score, than a "no rain" forecast.

Scaled results

Tables 4 and 5 show that simple scaling of a forecast to the last measured flow generally leads to a poorer performance than making no adjustments at all. In part this may be connected with correct estimation of the volume of runoff, since with simple scaling no attempt is made to preserve the volume of flow. Use of simple scaling also depends

strongly on the accuracy of the single measured flow value to which the forecast is adjusted. Any error in the latest flow could lead to errors in both the forecast peak and the end-of-period flow forecast. Since simple scaling can be sensitive to errors in the measured flow—a fact supported by the generally poorer results in the percentages and ranks of positive skill scores in each forecast category (Table 4 and Table 5)—it is not recommended as a reliable way to adjust forecasts.

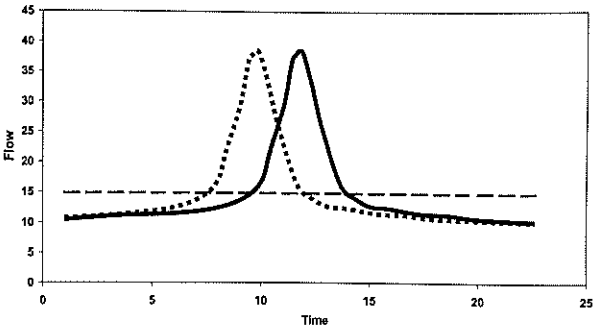

Figure 6 – The value of a forecast versus the skill of a forecast. The graph displays the results of a forecast made at time = 1. The dotted line is the forecast, the broken line is the mean of the data, and the solid line is the subsequent measured hydrograph. A perfect forecast would match the solid black line and have a Transformed Skill Score (TSS) of $+\infty$. If the mean value of the data is taken to be the forecast, the TSS of the forecast is zero. However, the dotted forecast, which subjectively would be considered to be useful even if the timing is wrong, has a TSS of –0.59!

Shifting and scaling results

While open to the same criticism as simple scaling because of the failure to preserve the forecast volume of flow, the simultaneous adjustment of both the timing and magnitude of flows leads to generally better forecasts. Here the volumetric error is possibly less severe, as there is some opportunity for underestimates to be compensated for by overestimates in the adjustment calculation. Results from shifting and scaling forecasts are comparable to those from the unadjusted simulations for the end-of-period flow. They are worse for the volume of flow, as would be expected; but better for both the peak flow and the time-of-the-peak flow (see Table 5).

As more flow information is introduced into the adjustment process, the forecast skill of the peak flow increases considerably. On balance the application of shifting and scaling seems to have a beneficial effect on forecast skill.

In an effort to provide an objective analysis, we have focused on a numerical skill score as the basis for assessing forecast performance. In general our forecasts do not show great skill. However, as the graph of the value of a forecast versus the skill of a forecast (Fig. 6) shows, a forecast with a low skill score can have value. This is a subjective assessment. Thus while our results lean towards pessimism with respect to skill, this does not mean that the forecasts are without value.

Conclusions

A series of comparative flood forecast have been made under controlled conditions for rivers across New Zealand with a range of basin sizes in an attempt to objectively quantify the degree of skill of the forecasts. To ensure that the forecasts were comparable, additional information that might be included in a particular forecast scheme, e.g., access to upstream flow data, had to be ignored since such data were not available for all catchments. Individual forecast systems can capitalise on such extra information and would likely be more skilful than the results presented here. The generally low skill scores presented here probably reflect the short period of uneventful data available

for model calibration and the likely spatial and temporal errors in the forecast rainfall fields. So, although the results show that flood forecasting using the combination of meso-scale rainfall forecasts as input to a rainfall-to-runoff model has some skill for larger catchments, there is still room for improvement. The improvements need to be centred on a better calibration of the rainfall-to-runoff models, improved rainfall forecasts, and more sophisticated adjustment schemes able to assimilate a wide variety of data available in some basins and to adjust the states of the model to better track catchment wetness.

By not using the individual advantages of particular basins, the results may be unduly pessimistic. However, the results qualitatively show that the forecast skill is best for volumes of runoff, but worse for the time to peak flow. Forecast skill generally increases with catchment size. Flood peak estimation by an unadjusted model can be improved by shifting and scaling the forecast hydrograph using measured flows just prior to the forecast time.

Acknowledgements

The authors would like to thank the following organisations for the supply of meteorological or river flow data for this project: Contact Energy Limited, Environment Canterbury, Environment Southland, Greater Wellington Regional Council, Marlborough District Council, the Meteorological Service of New Zealand, Northland Regional Council, Taranaki Regional Council, the United Kingdom Meteorological Office, and the New Zealand Foundation for Research, Science and Technology (FRST) (contract C01X0010). The research and development of the modelling systems used was funded by the FRST under contract C01X0218.

References

Collischonn, W.; Haas, R.; Andreolli, I.; Tucci, C.E.M. 2005: Forecasting River Uruguay flow using rainfall forecasts from a regional weather-prediction model, *Journal of Hydrology 305(1-4):* 87-98.

Copeland, J.C.; Wratt, D.S.; Ibbitt, R.P.; Henderson, R.D. 2000: Forecasting riverflow with linked meteorological-hydrological models: The rain engine. Presented at Fresh Perspectives: a Joint Conference of New Zealand Hydrological Society, Meteorological Society of New Zealand and New Zealand Limnological Society, Christchurch, November 2000.

Henderson, R.; Turner, R.; Thompson, S.; Ibbitt, R.; Gray, W. 2002: Validation of mesoscale forecasts using flow as a measure of catchment-averaged rainfall. Poster presented at International Conference on Quantitative Precipitation Forecasting, Reading University, UK, September 2002.

Ibbitt, R.P.; Henderson, R.D.; Copeland J.; Wratt D.S. 2000: Simulating Mountain Runoff with Meso-scale Weather Model Rainfall Estimates: A New Zealand Experience, *Journal of Hydrology 239 (1-4)*: 19-32.

Pearson, C.P.; Jordan, R.S. 1991: New Zealand flood forecasting procedures and reliability, Publication No. 25, Hydrology Centre, Department of Scientific and Industrial Research, Christchurch, New Zealand.

Pielke, R.A.; Cotton W.R.; Walko R.L.; Tremback C.J.; Lyons W.A.; Grasso L.D.; Nicholls M.E.; Moran M.D.; Wesley D.A.; Lee T.J.; Copeland J.H. 1992: A comprehensive meteorological modeling system - RAMS. *Meteorology and Atmospheric Physics 49*: 69-91.

Wilks, D.S. 1995: *Statistical Methods in the Atmospheric Sciences: an Introduction.* Academic Press, U.S.A.

Manuscript received 12 July 04; revised; accepted for publication 5 July 05