Journal of Hydrology (NZ) 55 (1): 65-68 2016

NOTE

Preliminary evaluation of New Zealand seasonal river flow forecasts

Earl Bardsley

Faculty of Science and Engineering, University of Waikato, Private Bag 3105, Hamilton. Corresponding author: e.bardsley@waikato.ac.nz

Abstract

In 2001 The National Institute of Water and Atmospheric Research (NIWA) initiated monthly forecasts for mean river flows in the coming three months. The forecasts are presented as specified probabilities that the flows will be in the lower, middle or upper discharge terciles for a given region. An evaluation of forecast skill for a small data set comprising the first 26 months of monthly forecasts indicates that flow forecasts with highest probability have a 50% chance of being correct. This is an improvement over the 33% success rate expected from random chance, but still not very helpful for practical decision making. On the other hand, there was an 80% chance of a correct prediction that the least probable outcome would *not* happen, which should be of practical value if confirmed from analysis of the complete prediction record to date.

Introduction

The New Zealand National Institute of Water and Atmospheric Research (NIWA) initiated monthly river flow forecasting in 2001, comprising forecasts of regional mean flows for the coming three months. For this purpose New Zealand was divided into six flow forecasting regions corresponding to the zones used for climatic forecasting: northern North Island (NNI), south western North Island (SWNI), eastern North Island (ENI), northern South Island (NSI), south western South Island (SWSI) and Eastern South Island (ESI). The flow forecasts are presented in the form of probabilities assigned to coming regional mean river flow being in the lower, middle, or upper tercile of the long-term regional discharge range.

The forecasts referenced in this study were created from consensus among various experts. The three-month regional flow forecasts of lower, middle or upper terciles for the time of year were made each month taking into account climate forecasts (i.e., rainfall) for each region and the initial wetness conditions at the start of the forecast period. A NIWA hydrologist participated in a consensus forecast meeting of the climatologists to establish the climate forecasts, and then provided information to other hydrologists to discuss and form a consensus forecast of regional soil moisture and river flows. Further information is given by Pearson (2008).

As with any sequence of forecasts, it is of interest to make an evaluation of forecast skill. This preliminary study considers the first 26 months of 3-month flow forecasts for the six prediction regions, from November 2001.

Methodology

Forecast verification is a subject area in its own right – see, for example, Hamill and

Juras (2006), Joliffe and Stephenson (2012), and Ebert *et al.* (2013). The focus in this paper is primarily on the simpler topic of forecast invalidation, or checking that forecast successes are better than might be expected from random chance.

In this regard, categorical forecasts like the NIWA tercile probabilities present something of a problem for forecast evaluation because it can be difficult to verify that a forecast has failed or succeeded. For example, if there was 1/3 assigned probability for each tercile then any flow outcome would be consistent with the forecast.

Various approaches have been suggested for evaluation of categorical probabilistic forecasts – see, for example, Zhang and Casey (2000), Maia *et al.* (2007), and Sohn and Park (2008). One approach is to convert the tercile probability forecasts to binary forecasts. That is, a sufficiently high probability assigned to a particular flow outcome constitutes a forecast that can either succeed or fail depending on the actual discharge outcome. Random reordering of the time sequence of the forecasts can then be applied to check whether the number of successful forecasts (hits) is greater than chance.

Applying this method failed to reveal discharge forecasting skill in any of the six regions (Bardsley, 2015). However, restricting the analysis to only a subset of binary sequences with high probability of outcome means that many forecasts are excluded from analysis because sometimes no combination of two tercile probabilities yields high probability of outcome. That is, the effective data set is reduced and this limits the power of the test to detect skill.

An alternative test method is presented here that yields more useful results from the same data set. For any given tercile forecast, there is a 1/3 probability that randomly relocating the maximum assigned probability will result in a match (hit) to whatever flow was forecast (low, medium, or high flow). Likewise, there is a 2/3 probability that a random relocation of the lowest assigned probability will *not* match whatever was forecast. A mismatch in the latter situation is defined here also as a hit because the non-occurrence of an event assigned a low probability of occurrence is also a forecast success in the tercile framework. In both the match and mismatch cases, evidence of forecasting skill is the occurrence of a greater number of hits than might be expected by random chance.

The above reference to randomly relocating assigned probabilities is with respect to the order of the three assigned probabilities in a given forecast. For example, a single forecast might be in the form of the assigned probabilities 0.3, 0.6, and 0.1 for mean river flow in the coming three months to be in the lower, middle, or upper flow tercile, respectively. The three possible random relocations for the maximum assigned probability are therefore 0.6, x , x; x, 0.6, x; or x, x, 0.6, where x denotes 0.3 or 0.1. For checking outcomes of maximum probability forecasts, data were excluded for those cases where there were two equal assigned maximum probabilities, such as 0.2, 0.4, 0.4.

Having now obtained N forecasts with no cases of equal maximum assigned probability, a random relocation of the specified maximum (or minimum) probability of outcome implies that the number of chance hits corresponds to a random variable from a binomial (N,p) distribution, where p = 1/3 or 2/3 for the maximum or minimum cases respectively. Therefore, if a given sequence of N forecasts produces K hits then a confirmation of no of forecasting skill is finding a sufficiently large value of z, being the probability that random chance will yield a number of hits greater than or equal to K . That is, $z = 1 - \mathrm{F}(N,K\text{-}1, p)$, where F(.) is the binomial cumulative distribution function with p being either 1/3

or 2/3, depending respectively on whether maximum or minimum assigned probabilities are being investigated. If $K = 0$ then $z = 1.0$. If z is sufficiently large, say greater than 0.05, then it can be said that no forecasting skill has been detected.

This type of test is simply a means to check whether forecasts are better than random chance, as defined by random relocation of the maximum or minimum assigned probabilities. As is well known, a forecast method may have a high level of statistical significance but still have such a low hit rate as to be of no practical value; see, for example, Bardsley (2015) and cited references.

Results

The results of the two tests are shown in Tables 1 and 2, for maximum and minimum probability forecasts, respectively. It is evident from Table 1 that the maximum probability forecast outcome has a hit rate of about 50%. While a 50% success rate may be of limited value for decision making there is evidence of some degree of skill in that the 50% value is rather more than the random chance hit rate of 33%. There is also some suggestion of possible regional variation in forecasting accuracy but the small data base does not permit more detailed analysis.

Table 1 – Regional and combined-region results of success rates for the forecast mean flow to match the highest assigned probability. Values of z calculated for $p = 1/3$. See text for definitions.

	Combined	N NI	SW NI	E NI	N SI	SW SI	E SI
N	129	22	24	21	20	19	23
K	63	9	12	8	10	10	14
z	0.0002	0.30	0.07	0.40	0.09	0.07	0.006
Hit %	49	41	50	38	50	53	61

Table 2 – Regional and combined-region results of success rates for the forecast mean flow not to match the lowest assigned probability. Values of z calculated for $p = 2/3$. See text for definitions.

	Combined	N NI	SW NI	E NI	N SI	SW SI	E SI
N	114	22	23	24	25	23	24
K	141	17	17	18	21	19	22
z	<0.001	0.21	0.31	0.26	0.05	0.08	0.005
Hit %	81	77	74	75	84	83	92

The results of Table 2 are of some interest, with the suggestion that the overall hit percentage is around 80%. This is up somewhat from the expected 67% hit rate by random chance. The improvement to 80% is useful because this is possibly a sufficiently high probability of a non-outcome on which to make economic decision related to regional river flows over the coming three months. It could be useful, for example, to have confidence in a forecast that lower tercile flows will not occur in a coming summer period in a given region.

Conclusion

The results from this preliminary study suggest that for all regions the tercile probabilities for coming 3-month flows will be helpful for decision making if the least probable outcome needs to be relied on to

not occur with some degree of certainty. On the other hand, the maximum probability forecasts would appear to be less helpful to aid decisions relating to coming river discharges.

It is emphasised, however, that the analysis here was based on a short time series near the initiation of the forecasts. As a working hypothesis it might be assumed that those involved in the forecasts develop enhanced skill over time, learning from the accumulated experience of both successful and unsuccessful forecasts. This would be an interesting topic for further work.

Finally, there is an obvious need for a more detailed statistical investigation using the entire discharge forecast data set to the present. This would enable conclusions to be made on such factors as regional variations in forecast accuracy, demonstrating forecasts to be an improvement over temporal correlation, accuracy of predicting the time when a run of high or low flows will end, accuracy of forecasting coming extreme conditions, and correlating outcomes to the strength of the assigned probabilities. A similar detailed study would be helpful for the climate forecasts on which the discharge forecasts depend.

Acknowledgement

I am grateful to Charles Pearson (NIWA) for provision of the forecasting data considered here. Thanks also go to colleagues at the 2015 New Zealand Meteorological Society Conference for correcting an error in an earlier version of the manuscript. The conclusions and views expressed in this paper are the author's alone.

References

Bardsley, W.E. 2015: Note on the hypergeometric distribution as an invalidation test for binary forecasts. *Stochastic Environmental Research and Risk Assessmen*t DOI 10.1007/s00477-015-1071-z

Ebert, E.; Wilson, L.; Weigel, A.; Mittermaier, M.; Nurmi, P.; Gill, P.; Göber, M.; Joslyn, S.; Brown, B.; Fowler, T.; Watkins, A. 2013: Progress and challenges in forecast verification. *Meteorological Applications* 20: 130-139.

Hamill, T.M.; Juras, J. 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* 132: 2905-2923.

Jolliffe, I.T.; Stephenson, D.B. (*eds.*) 2012: *Forecast verification: a practitioner's guide in atmospheric science*. 2nd edition. Wiley-Blackwell.

Maia, A.H.N.; Meinke, H.; Lennox, S.; Stone, R. 2007: Inferential, nonparametric statistics to assess the quality of probabilistic forecast systems. *Monthly Weather Review* 135: 351-362.

Pearson, C.P. 2008: Short- and medium-term climate information for water management. *WMO Bulletin* 57: 173-177.

Sohn, K.T.; Park, S.M. 2008: Guidance on the choice of threshold for binary forecast modelling. *Advances in Atmospheric Sciences* 25: 83-88.

Zang, H.; Casey, T. 1999: Verification of categorical probability forecasts. *Weather and forecasting* 15: 80-89.

Manuscript received 23 May 2015; accepted for publication 5 November 2015