P1.7

COMBINING AUTOMATED AND HUMAN PREDICTIONS: THE RESULTS OF A 1000-DAY REAL-TIME TRIAL

Harvey Stern*
Bureau of Meteorology, Melbourne, Vic., Australia.

"Consider mechanically integrating judgmental and statistical forecasts instead of making judgmental adjustments to statistical forecasts …Judgmental adjustment (by humans) of (automatically generated statistical forecasts) is actually the least effective way to combine statistical and judgmental forecasts … (because) judgmental adjustment can introduce bias (Mathews and Diamantopoulos, 1990) (see also, Stern (1996), who documents forecaster over-compensation for previous temperature errors) …The most effective way to use (human) judgment is as an input to the statistical process … Cleman (1989) reviewed over 200 empirical studies on combining and found that mechanical combining helps eliminate biases and enables full disclosure of the forecasting process. The resulting record keeping, feedback, and enhanced learning can improve forecast quality" (Sanders and Ritzman, 2001).

1. INTRODUCTION

Woodcock et al. (2008) present the results of combining a set of *automatically generated* Day-1 to Day-6 minimum and maximum temperature forecasts with a corresponding set of *official* forecasts prepared for Australian capital cities in 2006. They suggest that most of the *combined* forecasts are better than the corresponding *official* forecasts.

There is an increasing interest in the question of what might be the appropriate future role for the human in the forecast process. Computer-generated forecasts are unable (by themselves) to fully replicate the decision-making processes of human forecasters. Similarly, human forecasters are unable (by themselves) to optimally integrate into the forecasting process, guidance from computer-generated predictions.

However, there is the accepted mathematical concept that two or more inaccurate but independent predictions of the same future events may be combined to yield predictions that are, on the average, more accurate than either of them taken individually (Thompson, 1977). Automated and human forecasts might be expected to "bring to the table" different knowledge sets, and this suggests the development of a weather forecasting system that mechanically combines human and computer-generated predictions.

2. PURPOSE

Sanders and Ritzman (2001) highlight the difficulty associated with utilising (human) judgment as an input to the statistical process 'when the (human) forecaster gets information at the last minute'. The purpose of the present paper is twofold:

(1) To describe the development of a system that mechanically combines judgmental (human) forecasts (derived with the benefit of knowledge of all available computer generated forecast guidance) and computer generated forecasts guidance and to evaluate the accuracy of the new set of forecasts and to compare it with the accuracy achieved by the judgmental (human) forecasts; and,

(2) To draw the attention of readers to the results of a 1000-day real-time trial (conducted from 20-8-2005 to 15-5-2008) of a knowledge based system that mechanically integrates (combines) automatically generated and official predictions. The system yields a graphical product that depicts all of the elements included in a public weather forecast (Figure 1).

Although space is too limited here to present details of the combining process for the prediction of all weather elements, the process of integrating human and automated forecasts is briefly illustrated for Probability of Precipitation estimates in Figure 2.

---

*Corresponding author address: Dr Harvey Stern, Bureau of Meteorology, Box 1636, Melbourne, Victoria, 3001, Australia.
Email: h.stern@bom.gov.au

3. RESULTS

The approach was first evaluated in a *hindcast* mode by Stern (2005a), who showed that the process of combining human (official) and automated forecasts had the potential to yield a set of predictions that is far more accurate than either set taken separately. The human (official) forecasts explained 42.3% of the variance of the observed weather (rainfall amount, significant weather, minimum temperature, and maximum temperature), whilst, by itself, the automated forecast set explained 43.2% of the variance. Stern (2005a & b) showed that adopting a combining strategy had the potential to lift the overall percentage variance explained to 50.2% (Figure 3).

It is considered that because a 'real-time' trial of a methodology involves evaluating forecasts that are generated *prior* to the event, the results of such a trial possesses greater validity than if the new methodology had been evaluated in an *hindcasting* mode (even with the application of sophisticated cross validation techniques).

Subsequently, detailed analyses of the accuracy of forecasts generated during a *real-time* trial commencing 20 August 2005 were presented in a series of papers by Stern (2006, 2007a, b, c, d & e; 2008a, b & c). After one year, the results demonstrated that the combined forecasts did indeed have the potential to substantially improve upon the existing (official) product (Table 1 and Figure 4).

Since the first year of the trial, the mechanically combined forecasts generated during the *real-time* trial have continued to perform strongly as testified by verification statistics derived from the 1,000 Melbourne Day-1 to Day-7 forecast sets generated by combining human and computer predictions between 20 August 2005 and 15 May 2008.

For example, the accuracy of the 14,000 Melbourne Day-1 to Day-7 minimum and maximum **temperature** predictions so generated has been increased through agency of the mechanical integration process, with the Mean Square Error (MSE) of the mechanically integrated forecasts being 0.81 deg C lower than the MSE of the corresponding human (official) product.

Similarly, the accuracy of the 7,000 Melbourne Day-1 to Day-7 **rainfall** forecasts so generated has also been increased by means of the mechanical integration process, mechanically integrated forecasts of whether or not it was going to rain being correct 6.6% more often than the corresponding human (official) product.

Furthermore, the accuracy of the 7,000 Melbourne Day-1 to Day-7 **thunderstorm** forecasts so generated has also been increased by means of the mechanical integration process, the Critical Success Index (CSI) of the mechanically integrated forecasts of thunderstorms being 3.6% higher than that of the corresponding human (official) product.

The accuracy of the 7,000 Melbourne Day-1 to Day-7 **fog** forecasts so generated has also been increased by means of the mechanical integration process, albeit only slightly, the CSI of the mechanically integrated forecasts of fog being 0.9% higher than that of the corresponding human (official) product.

The verification of the 1,000 Melbourne Day-1 to Day-7 forecast sets refers to an *overall* evaluation undertaken on the forecast performance with all lead times taken together. Nevertheless, even when the evaluation was undertaken with lead times taken separately, a lift in accuracy occurred in most instances.

4. VERY LONG RANGE FORECASTS

Since, 20 August 2006, **very long range** forecasts have also been generated by combining computer predictions with climatology (climatology was used, given the absence of very long lead time human forecasts).

Verification over a one-year period to 19 August 2007 (Stern, 2008b), revealed that Day-8 forecasts so generated explained 11.2% of the variance, Day-9 forecasts explained 7.2% of the variance, and Day-10 forecasts explained 3.4% of the observed variance. However, for these very long range day-to-day forecasts, the variance explained was mainly for the temperature components.

Specifically for **Day-8**, Quantitative Precipitation Forecasts (QPFs) explained 4.2% of the observed variance, whilst Minimum Temperature Forecasts (MINFs) explained 17.9% of the observed variance and Maximum Temperature Forecasts (MAXFs) explained 17.5% of the observed variance.

For **Day-9**, QPFs explained 3.1% of the observed variance, whilst MINFs explained 10.4% of the observed variance and MAXFs explained 10.0% of the observed variance.

For **Day-10**, QPFs explained 0.9% of the observed variance, whilst MINFs explained 7.7% of the observed variance and MAXFs explained 4.6% of the observed variance.

5. FUTURE WORK

That the system also generates forecasts for 55 other localities in Victoria's Central District creates the potential for automated digital representation of the distribution of various weather elements across the District (Fig 5).

This potential future work fits in nicely with current cooperation between the Bureau of Meteorology and NOAA, the National Oceanic and Atmospheric Administration that is resulting in Australia implementing the US software system, the Graphical Forecast Editor (GFE). The GFE enables forecasters to provide a digital representation of weather (Commonwealth of Australia, 2006).

6. CONCLUSION

The results of a 1000-day *real-time* trial of a system that mechanically integrates (combines) *automatically generated* and *official* predictions have been presented.

The results demonstrate the potential benefit to be gained were one to adopt Sanders and Ritzman's (2001) proposal to "consider mechanically integrating judgmental and statistical forecasts (the new methodology proposed here) instead of making judgmental adjustments to statistical forecasts (the existing methodology)", and to operationally implement a system based upon the new methodology, such as the knowledge based system described in the present paper.

7. REFERENCES

Armstrong, J. S., 2001: Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers.

Cleman, R. T., 1989: Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5, 559-583 (refer to Armstrong, 2001, 411).

Commonwealth of Australia, 2006: Official Committee Hansard Joint Standing Committee on Treaties. Reference: Treaties tabled on 28 March, 20 June and 8 August 2006 Monday August 14 2006 Canberra by Authority of the Parliament.

Mathews, B. P. and Diamantopoulos, A., 1990: Judgmental revision of sales forecasts: effectiveness of forecast selection, *Journal of Forecasting*, 9, 407-415 (refer to Armstrong, 2001, 411).

Sanders, N. R. and Ritzman, L. P., 2001: Judgmental adjustment of statistical forecasts (refer to Armstrong, 2001, 405-416)

Stern, H., 1996: Statistically based weather forecast guidance. Ph. D. Thesis, School of Earth Sciences, *University of Melbourne*, subsequently published in 1999 as Meteorological Study 43, *Bureau of Meteorology, Australia.*

Stern, H., 2005a: Defining cognitive decision making processes in forecasting: a knowledge based system to generate weather graphics, *21st Conference on Weather Analysis and Forecasting; 17th Conference on Numerical Weather Prediction*. Amer. Meteor. Soc., Washington, DC, 1-5 Aug., 2005.

Stern, H., 2005b: Generating quantitative precipitation forecasts using knowledge based system, *17th BMRC Modelling Workshop*, Melbourne, Australia, 3-6 Oct., 2005.

Stern, H., 2006: Combining human and computer generated weather forecasts using a knowledge based system. 22nd Conference on Interactive Information and Processing Systems, Atlanta, Georgia, USA 27 Jan. - 3 Feb., 2006.

Stern, H., 2007a: Increasing forecast accuracy by mechanically combining human and automated predictions using knowledge based system 23rd Conference on Interactive Information and Processing Systems, San Antonio, Texas, USA 14-18 Jan., 2007.

Stern, H., 2007b: Employing weather derivatives to assess the economic value of high-impact weather forecasts out to ten days - indicating a commercial application. 22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical

Weather Prediction, Park City, Utah, USA 25-29 Jun., 2007.

Stern, H., 2007c: Improving forecasts with mechanically combined predictions. Bulletin of the American Meteorological Society (BAMS), June 2007, 88:850-851.

Stern, H., 2007d: The future role of humans in weather forecasting. Australian Meteorological and Oceanographic Society (AMOS) 2007 Annual Conference, Adelaide, 5-8 Feb., 2007.

Stern, H., 2007e: The future role of humans in the weather forecasting process - to provide input to a system that mechanically integrates judgmental (human) and automated predictions? 5th Conference on Artificial Intelligence Applications to Environmental Science, San Antonio, Texas, USA 14-18 Jan., 2007.

Stern, H., 2008a: Fog and thunderstorm forecasting in Melbourne, Australia. 24th Conference on Interactive Information and Processing Systems, New Orleans, Louisiana, USA 20-24 Jan., 2008.

Stern, H., 2008b: Does society benefit from very long range day-to-day weather forecasts? Symposium on Linkages among Societal Benefits, Prediction Systems and Process Studies for 1-14-day Weather Forecasts, New Orleans, Louisiana, USA 23 Jan., 2008.

Stern, H., 2008c: Improving wind forecasts by mechanically combining predictions. Australian Meteorological and Oceanographic Society (AMOS) 2008 Annual Conference, Geelong, 29 Jan. - 1 Feb., 2008.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. Mon. Weath. Rev., 105, 228-229.

Woodcock, F., Elliot, G., and Setek, M., 2008: Comparison of official and operational consensus forecasts of daily extreme temperatures in 2006. Aust. Met. Mag., 57, 95-108.

Table 1 At the conclusion of the first year of the real-time trial, enhanced forecast accuracy is demonstrated for various weather elements (from Stern, 2007c).

Element	Verification parameter	Human (official)	Combined
All elements	% variance explained	33.40	41.30
Rain or no rain	% correct	70.10	76.80
Rain amount	RMS error ($mm^{0.5}$)	1.05	0.97
Min temp	RMS error (°C)	2.39	2.27
Max temp	RMS error (°C)	2.82	2.49
Thunder	Critical Success Index (%)	17.90	21.60
Fog	Critical Success Index (%)	15.50	17.80

Figure 1 Mechanically integrated forecast for Melbourne 5-9-2008 to 14-9-2008.

Day & Date	Morning	Afternoon	Min Temp (deg C)	Max Temp (deg C)	Precip Amount (mm)	Precip Prob (%)	9am Wind/ 3pm Wind Melb Apt (km/hr)
Fri-5-9-2008	Partly Cloudy.	Partly Cloudy.	7	18	0	27	N 15 N 25
Sat-6-9-2008	Possible Shower.	Partly Cloudy.	8	18	0	43	SW 8 S 15
Sun-7-9-2008	Shower.	Shower.	9	16	0.8	56	SW 15 S 15
Mon-8-9-2008	Mist.	Cloudy.	7	16	0	39	N 8 S 15
Tue-9-9-2008	Cloudy.	Shower.	9	18	1.4	56	N 35 N 25
Wed-10-9-2008	Shower.	Shower.	8	16	2.9	61	WSW 35 S 35
Thu-11-9-2008	Shower.	Shower.	9	17	1.8	63	WSW 25 S 25
Fri-12-9-2008	Possible Shower.	Possible Shower.	9	18	0	46	N 25 N 25
Sat-13-9-2008	Windy.	Windy.	9	18	0	41	N 45 N 35
Sun-14-9-2008	Shower.	Shower.	10	18	2.4	64	N 55 N 45

Figure 2 The process of integrating human and automated forecasts for Probability of Precipitation (PoP) estimates:

Firstly, the estimate from a statistical model (62%) is averaged with the implied estimate from the NOAA Global Forecasting System (GFS) of 100% to yield 81%;

Secondly, this 81% outcome is then averaged with a previous estimate (generated 'yesterday') by the combined system (of 65%) to yield 73%; and,

Thirdly, this 73% outcome is then averaged with the implied estimate from the human (official) forecast (of 47%) to yield 60% (from Stern, 2006).

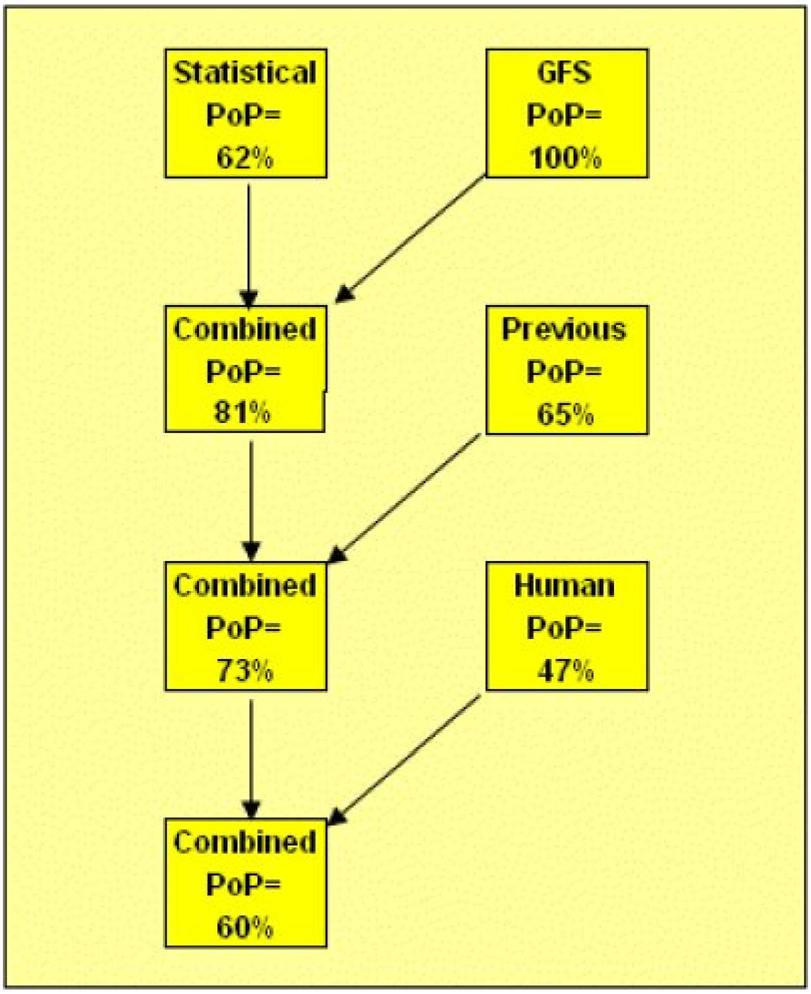

Figure 3 Lifting the accuracy of forecasts (% variance explained) by adopting a combining strategy (from Stern, 2007a)

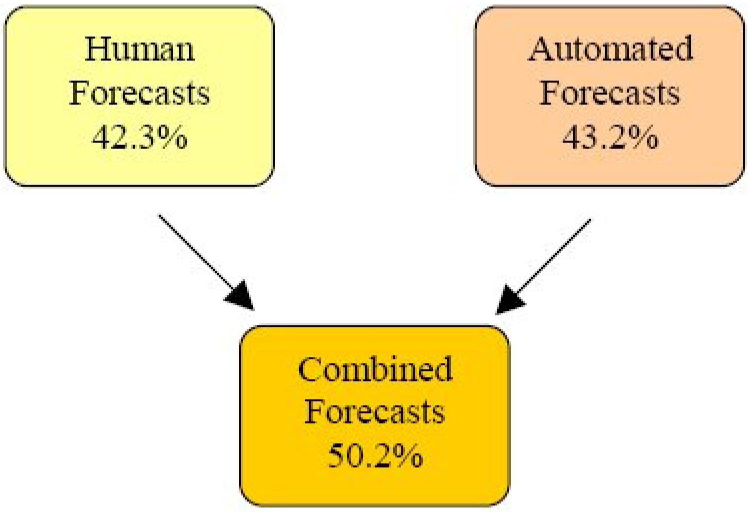

Figure 4 At the conclusion of the first year of the real-time trial, enhanced forecast accuracy is demonstrated for various lead times (from Stern, 2007c).

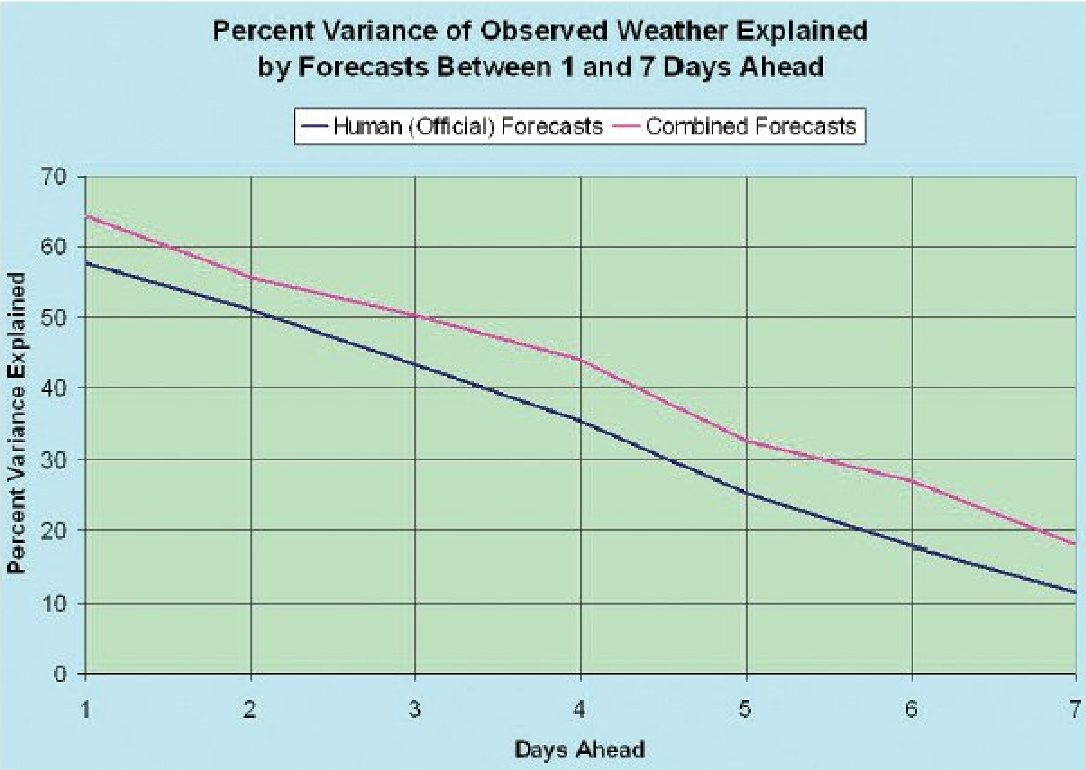

Figure 5 Analysis of minimum temperatures (ºC) that were forecast for 16-6-2008 across the portion of the Central District to the east and southeast of Melbourne. Note the relatively mild temperatures (~ 4ºC) predicted for the area around the shores of Port Phillip Bay, and also over the Dandenong Ranges in the upper left section of the map. Relatively colder temperatures (~ 1ºC) are suggested in the valleys surrounding the Dandenong Ranges.