Mausam, (1989), **40**, 2, 175-180

551.509.331 : 551.524.36(547)"324"

Prediction of winter minimum temperature of Pune by analogue and regression methods

Y. E. A. RAJ

Meteorological Office, Pune

(*Received 7 October 1987*)

सार — शीतकाल में पुणे के न्यूनतम तापमान के पूर्वानुमान के लिए वस्तुनिष्ठ पूर्वानुमान योजनाएं (स्कीम) विकसित की गई हैं। ये योजनाएं बहुसमाश्रयण, अनुरूप और उपरोक्त दोनों तकनीकों के सम्मिश्रण पर आधारित की गई हैं। 24, 15 और 12 घंटे की पूर्वानुमान के समाश्रयण समीकरण को विकसित करने के लिए प्राग्वक्ताओं का चयन करने के लिए 12 आदेशकों (प्रीसक्रिप्टर्स) की छान-बीन की गई है। अनुरूप और अनुरूप एवं बहुसमाश्रयण की तकनीकों के आधार पर पूर्वानुमान योजना को विकसित करने के लिए समाश्रयण समीकरणों के लिए चुने गए प्राग्वक्ताओं का उपयुक्त रूप से प्रयोग किया है। 1116 दिनों के वृहत आंकड़ा नमूने से योजनाएं विकसित की गई हैं और 403 दिनों के एक स्वतंत्र नमूनों में इनका परीक्षण किया गया है। योजनाओं ने काफी प्रोत्साहक परिणाम दिए हैं। विभिन्न प्राग्वक्ताओं और विभिन्न योजनाओं के सापेक्ष महत्व का विवेचन किया गया है ।

ABSTRACT. Objective forecasting schemes to forecast the winter minimum temperature of Pune have been developed. The schemes have been based on multiple regression, analogue and a combination of the above two techniques. As many as 12 prescriptors have been screened to select predictors for developing regression equations for 24, 15 and 12- hr forecast schemes. The predictors chosen for the regression equations have been suitably used to develop forecast schemes based on the technique of analogue and analogue *cum* multi-regression. The schemes have been developed from a large data sample comprising of 1116 days and have been tested in an independent sample of 403 days. The schemes have given very encouraging results. The relative importance of the various predictors and of the different schemes have been discussed.

1. Introduction

Pune, located in the meteorological subdivision of Madhya Maharashtra of India at 18° 32′ N, 73° 51′ E in the eastern side of the Western Ghats at an elevation of 559 m a.s.l. is known for its salubrious climate. It represents the climate of north Madhya Maharashtra fairly well. This region is prone to low temperatures in winter. Advance warning of cold wave conditions and frost are, therefore, useful to the public especially for the farmers growing grapes, which are cultivated in abundance in this region. The months of December and January are the coldest in winter which stretches from November to February.

Day to day variation of maximum and minimum temperatures (T_x and T_n) of a station is generally controlled by processes such as net flux of terrestrial radiation, transport of heat by turbulence in the air, latent heat transfer, heat conducted within the soil to or from the earth's surface, total insolation etc (Haltiner and Martin 1957) . Reference to literature shows that forecasting schemes for T_x and T_n are generally based on statistical techniques. Klein and Hammons (1975) suggest as many as 43 meteorological parameters as prescriptors, *i.e.*, potential predictors for forecasting T_n or T_x. As to how to choose prescriptors for a forecasting scheme has been lucidly explained in Lund (1955).

An attempt has been made in this paper to evolve objective forecasting schemes for the T_n of Pune for the months of December and January. We intend usage of multiple regression and analogue techinques and a combination of both to achieve this objective. WMO (1966) and Kendall and Stuart (1968) give detailed description of regression analysis. Raghavendra (1956), Sinha (1957), Srinivasan and Hashim (1967), Banerji and Chowdhury (1972) have all used various techniques to evolve forecast schemes of T_x/T_n of certain Indian stations.

In India the observations of T_n are taken daily morning at 0830 IST. We propose to evolve three forecast schemes for (*i*) 24-hr forecast based on the 0830-hr charts, (*ii*) 15-hr forecast based on the observations available up to 1730 hr and (*iii*) 12-hr forecast based on the data available up to 2030 hr of the previous day.

2. Formulation of forecasting schemes based on regression method

2.1. *Selection of prescriptors*

Let Y, the T_n of the n^{th} day denote the predictand. We define two prescriptors based on advection of temperature as under. The 0530 hr, 0.9 km a.s.l. wind field of the $(n-1)^{\text{th}}$ day was studied and the place, say P from where the air parcel would reach Pune approximately after 24-hr was roughly located. The 0830 hr charts were then referred and the value of T_n and its anomaly at P were interpolated. These were denoted by A and A'. Similarly reference to the wind field at 0.9 km of 1730 hr and the distribution of T_x yielded B and B'. Along

TABLE 1

Mean and SD of prescriptors (P) and Cor (P, Y)

Prescriptor P	Cor (P, Y)	Mean of P (°C/Octas)	SD of P (°C/Octas)
Y_1	0.83	10.8	3.2
Y_2	0.65	10.8	3.2
C_1	0.47	0.7	1.6
C_2	0.36	0.7	1.6
A	0.50	14.9	3.4
A'	0.75	—0.2	2.4
X	0.19	29.2	1.7
D	0.72	9.2	4.1
C	0.64	1.1	1.7
B	0.20	29.3	1.7
B'	0.20	—0.1	1.7
X_1	0.51	20.6	3.0

with these four derived parameters another eight prescriptors which suggested themselves were selected The 12 prescriptors thus selected are :

Y_1, Y_2 : T_n of the $(n—1)^{th}$ and $(n—2)^{th}$ days

C_1, C_2 : No. of octas of low/medium clouds at 0830 hr of $(n—1)^{th}$ and $(n—2)^{th}$ days.

A, A' : Advection parameters at $(n—1)^{th}$ day morning

X : T_x of the $(n—1)^{th}$ day

D, C : Dew point temperature and No. of octas of clouding at 1730 hr of the $(n—1)^{th}$ day.

B, B' : Advection parameters at $(n—1)^{th}$ day evening

X_1 : Temperature at 2030 hr of the $(n—1)^{th}$ day.

2.2. *Data*

The Dec-Jan period of 1961-84 and Jan 1985 has been chosen as the period of the study. In this the 18-year period 1961—78 has been earmarked as the developmental period (DP) and the remaining as the test period (TP). Thus, the DP and TP consist of 1116 and 403 days respectively. The values of A, A', B and B' for the DP and TP were extracted from the daily weather charts. The values of all the other prescriptors were directly extracted from the publications/records of India Met. Dep. Table 1 gives the means and standard deviations (SD) of the prescriptors and the predictand and also the correlation coefficients (CCs) between them.

2.3. *Some properties of the predictand*

The T_n of Pune of Dec-Jan during the DP had a mean of 10.8° C and SD of 3.2° C. The histogram of the frequency distribution is exhibited in Fig. 1. The distribution is highly skewed with a modal value of 9.5° C and skewness coefficient (*i.e.*, 3rd moment/SD^3) 0.57. The lowest value of T_n of 2.7° C was recorded on 18 Jan 1968 whereas the highest value of 22.8° C was observed on 10 Dec 1965. Harmonic analysis of the series $Y : \{ y_i, i = 1, N \}$, where $N = 1116$ and y_i the value of T_n on the i^{th} day of Dec-Jan counted from 1 Jan 1961 showed no significant periodicity. This conforms with the observation made in India Met. Dep. (1974) that heat/cold waves do not occur with any regular pattern. The auto CC remained positive up to a lag of 14 days. The formula suggested in Brooks and Corruthers (1953) for computation of effective length of a time series, yielded a reduction factor of 0.14, thus reducing the original series length of 1116 to an effective length of 145. The testing of the series for trend by the method of least squares did not reveal any trend. The daily means of all the 62 days of Dec-Jan, each computed from 18 values varied between 9.5° and 11.5°C and were not significantly different from the population mean of 10.8° C. It is, thus, concluded that the series $Y : \{y_i, i = 1, N\}$ is a stationary series without cycles and is random apart from persistence.

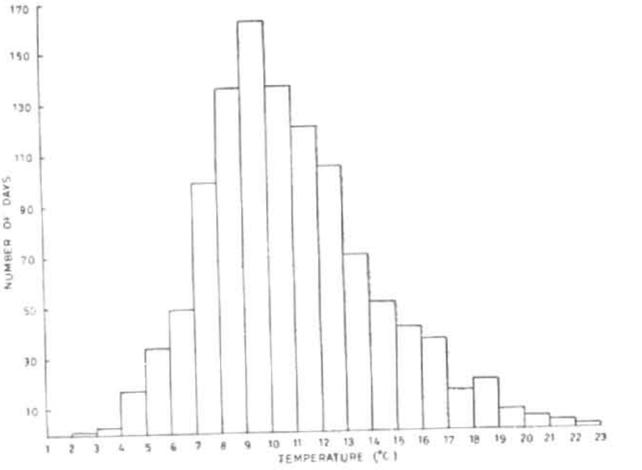

Fig. 1. Histogram of minimum temperature distribution of Pune for Dec-Jan 1961-78

2.4. *Screening of prescriptors*

We first discuss briefly the methodology to test the significance of a CC. Here the enigma lies in the non-randomness of the sample. The sample size of 1116 is obviously an over estimation in significance tests whereas the effective sample size of 145 (Sec. 2.3) for the predictand, if applied to CCs involving other prescriptors could be an under estimation. We, therefore, set the critical value of a CC at 0.1 which is significant at 5% level if based on a sample size of 400.

For screening the prescriptors, we adopted the forward elimination method suggested in WMO (1966). Bansal and Datta (1974) have used similar standard packages. Significant predictors were selected in a stepwise fashion. The screening was stopped as and when none of the absolute values of the partial CCs (PCCs) exceeded 0.1. Here the PCC of a prescriptor is the CC between the prescriptor and the predictand when the influence of all other prescriptors considered has been eliminated.

Table 2 gives the results obtained from the screening procedure for 24, 15 and 12 hr forecast schemes. The predictors appear in descending order of importance. The PCCs and the partial regression coefficients (PRCs), the variance accounted by each predictor, the multiple CC (MCC) and the expected standard error are also given. We discuss the results of screening for the three schemes, one by one.

TABLE 2

Results of forecast schemes based on regression

	24-hr				15-hr				12-hr			
	P	*VA*	PCC	PRC	*P*	*VA*	PCC	PRC	*P*	*VA*	PCC	PRC
	Y_1	69.2	0.56	0.74	Y_1	69.2	0.40	0.49	Y_1	69.2	0.39	0.48
	A_1	3.8	0.36	0.39	A'	3.8	0.27	0.27	A'	3.8	0.26	0.26
	Y_2	0.7	—0.16	—0.15	C	2.7	0.31	0.36	C	2.7	0.29	0.34
					D	1.3	0.25	0.14	D	1.3	0.25	0.14
					X	1.3	0.14	0.15	X	1.3	0.09	0.10
					B	0.3	0.11	0.12	X_1	0.4	0.14	0.08
					Y_2	0.2	−0.10	−0.08	B	0.3	0.11	0.12
									Y_2	0.2	—0.10	—0.08
Constant term of the equation				4.54				—3.23				—3.08
Total percentage variance explained		73.7				78.8				79.2		
MCC		0.86				0.89				0.89		
SE (°C)		1.65				1.48				1.46		

P—Predictor, *VA*—Percentage variance accounted by *P*, PCC/MCC—Partial/Multiple CC, PRC—Partial Regression Coefficient SE—Standard Error

TABLE 3

Performance of the forecasting schemes based on regression during the test period

	24-hr	15-hr	12-hr
NC	75.4	80.7	81.9
NPC	19.6	16.1	15.4
NW	5.0	3.2	2.7
SC_1	0.30	0.45	0.48
SC_2	0.38	0.55	0.58
SE(°C)	1.69	1.47	1.43

NC, NPC, NW—Percentage of number of correct, Partially correct, Wrong Forecasts

SC_1, SC_2—Skill Scores, SE—Standard Error

2.4.1. *24-hr forecast scheme* — Six prescriptors Y_1, Y_2, C_1, C_2, A and A' were available for this scheme. Of these C_1, C_2, and A' were eliminated, Y_1, A' and Y_2 were the predictors for Y in that order. The MCC was 0.86 explaining 73.7% of the variation of Y.

2.4.2. *15-hr scheme* — For this 11 prescriptors, all but X_1 of the 12 considered were available. The selected seven predictors were Y_1, A', C, D, X, B and Y_2. The MCC was 0.89 explaining 78.8% of the variation, an improvement of 5.1% from the 24-hr scheme.

2.4.3. *15-hr scheme* — All the 12 prescriptors were available for this scheme. Y_1, A', C, D, X, X_1, B and Y_2 were the eight selected predictors. The MCC obtained was 0.89 explaining 79.2% of the variation.

2.5. *Performance of the schemes in the test period*

The schemes evolved in Sec. 2.4 were tested in the 403 days TP of Jan '79-Jan '85. For the purpose of forecast verification the following methodology was followed. Let d be the change of T_n of n^{th} day compared to that of $(n—1)^{th}$ day. The change d is classified as little change if d=0, slight rise if $d = 1$, rise if $d = 2$ or 3, appreciable rise if $d = 4$ or 5, marked rise if $d = 6$ or 7 and large rise if $d \geqslant 8$. The classifications slight fall, fall, appreciable/marked/large falls are similarly defined for the corresponding negative values of d. The forecast and the observed values of T_n are classified as per above. The forecast is taken as correct (C) if the observed class coincides or falls within either of the adjacent classes of the forecast class. It is taken as partially correct (PC) if the forecast is out by two stages. Otherwise it is termed as wrong (W).

With the above methodology, the 24, 15 and 12-hr schemes were tested in the TP. The percentage of C, PC and W forecasts are given in Table 3. It is seen that the percentage of C forecasts are 75.4, 80.7 and 81.9 for the above schemes, thus showing a steady increase with decreasing forecast period. The wrong forecasts form a negligible 2.5% only for 12-hr forecasts.

The skill scores of the schemes were computed by two methods. Those days of the TP for which $d = -1, 0, 1$ were taken as the expected number of C forecasts based on temperature persistence and the skill score as defined in Panofsky and Brier (1968) was computed. This was denoted by SC_1. Alternatively the skill score was computed by constructing the forecast contingency table (*loc cit.*). This was denoted by SC_2. The values of SC_1 and SC_2 for the schemes are given in Table 3. It is seen that $SC_2 > SC_1$ for a given scheme. For the 12-hr scheme we have $SC_1 = 0.48$ and $SC_2 = 0.58$. These figures clearly demonstrate the efficiency of the schemes.

The ability of the schemes to predict large changes of temperature was also examined. It was found that correct forecasts in respect of 15 and 12-hr schemes were approximately 60% when $d = 4$ or more, *i.e.*, for classes of apreciable/marked/large rise/fall. This shows the ability of the systems to predict sudden fall of minimum temperature leading to cold wave conditions.

2.6. *Scope for further improvement*

In order to find out whether inclusion of non-linear terms of the prescriptors (such as Y_1^2, $Y_1 Y_2$ etc) would lead to any improvement. several such terms from the set of predictors were formed and added as additional prescriptors and the whole process of screening was gone through. It was found that addition of non-linear terms neither increased the MCC nor improved the percentage of correct forecasts. Panofsky and Brier (1968) also support this inference. Thus, it is clear that the system cannot be improved unless fairly independent and preferably physically significant prescriptors are added. Radiation parameters and moisture parameters such as moisture depth, precipitable water content etc, if incorporated, might offer some improvement.

3. The analogue method of forecasting

Forecasting by the philosophy of analogues was also tried for the three schemes. Datta and Gupta (1975) have used techinques based on analogue in tracking tropical storms. Let for the $(n-1)^{\text{th}}$ day, the value of T_n be y_1. Let $k_1, k_2, \ldots\ldots\ldots, k_m$ be m days prior to and are analogues to the $(n-1)^{\text{th}}$ day. Now if $z_1, z_2, \ldots\ldots, z_m$ are the T_n values of $(k_1+1), (k_2+1), \ldots\ldots, (k_m+1)^{\text{th}}$ days then the forecast for T_n for the n^{th} day, is :

$$Y = \frac{1}{m} \sum_{1}^{m} z$$

However, what we mean by an analogous day has to be defined. For this, we have relied upon the results of screening regression of Sec. 2. The predictors for each scheme have already been listed according to the order of their importance (Table 2).

Let $\{P_i, i=1, L\}$ be the predictors for a given scheme arranged in descending order of importance and let p_i be the value of P_i, for the $(n-1)^{\text{th}}$ day say D of the test period. Further let p_{ij} be the value of P_i for the j^{th} day of the DP. If $p_{ij} = p_i$ for every i we can take the j^{th} day as analogue to D. However, equality is obviously too stringent a condition and so we set that the j^{th} day of the DP is analogous to D if :

$$p_i - \epsilon_i \leqslant p_{ij} \leqslant p_i + \eta_i \qquad (3.1)$$
$$(1 \leqslant i \leqslant L,\ 1 \leqslant j \leqslant N)$$

The values of ϵ_i and η_i have to be determined for each predictor P_i after taking into consideration its dispersion and other distribution characteristics. They need not be constant for a given predictor but could vary in its range. In the neighbourhood of the averages, ϵ_i and η_i should preferably be smaller and far away from the mean they should be larger. Smaller values of ϵ_i and η_i may give better analogues but the number of such days will be small, whereas larger values will yield several days of poor analogues. Taking all these aspects into consideration, the values of ϵ_i and η_i were fixed after critically studying the frequency distribution, dispersion, skewness etc of the predictors.

Evidently, a day to day search of the DP is required to pick out the analogous days. This was accomplished by suitable computerisation. If there were at least 25 analogous days $Y = (1/m)\ \Sigma\ z$ was computed and was taken as the forecast value. If there were not 25 such days the least important predictor of the scheme, *i.e.*, P_L was disregarded and DP was searched again by setting $L = L - 1$ in (3.1). This process was continued untill atleast 25 days of analogues were available. However, when $L = 1$ no condition was imposed in the minimum number of analogous days required.

Table 3 supplies the results of the forecasting schemes based on analogue. It is seen that the performances of the schemes are only slightly below that of the multi-regression schemes. Over all, the results appear satisfactory.

4. Analogue-*cum*-multi-regression

This techinque aims at a slight improvement of the analogue technique discussed in the previous section. Herein we use the m days of analogue to form a multiple regression equation and base our forecast thereon. This appears to be a slightly sophisticated way of forecasting compared to the analogue method, inasmuch as the multi-regression equation gives due allowances for the deviations of p_{ij} s from p_i s. Wood Cock (1980) has utilised this method to forecast temperatures over Australia. The Multiple-regression Correlation Coefficient (MCC) obtained gives a very good indication of the homogeneity of the analogous days. When the MCC is not significant this homogeneity could be considered as inadequate. Testing the significance of MCC posed a problem as a set of analogous days can hardly be considered as random. We, therefore, tested the significance of the MCC at a relatively sharper level of 1% . If the MCC was not significant, the least important predictor was omitted and the whole operation was repeated as in Sec. 3. As less number of predictors means more analogous days the MCC could be expected to become significant after a few such steps. This type of methodology to quantify the extent of analogue has been used by Lund (1963 a, b) in deriving standard map patterns,

TABLE 4

Performance of schemes based on analogue and analogue *cum* regression during the test period

	Analogue			Analogue *cum* regression		
	24-hr	15-hr	12-hr	24-hr	15-hr	12-hr
NC	69.8	79.2	81.6	73.7	79.6	80.2
NPC	22.1	16.1	13.6	21.6	16.2	15.4
NW	8.1	4.7	4.7	4.7	4.2	4.4
SC_1	0.12	0.40	0.47	0.24	0.40	0.45
SC_2	0.25	0.51	0.57	0.35	0.50	0.55
SE (°C)	1.72	1.56	1.53	1.66	1.56	1.56
ANP	2.9	5.8	6.5	2.9	5.8	6.5
AND	183	70	66	183	70	66

(NC, NPC, NW, SC_1, SC_2, SE—as in Table 3, ANP/AND—Average No. of Predictors/Days)

The results obtained are given in Table 4. The technique yielded slightly better results for the 24-hr forecast and almost identical results for the 15 and 12-hr forecasts compared to the analogue technique. The results are not appreciably different from those obtained from the multiple-regression techniques based on a large fixed sample (Table 3). Despite this, this method is intrinsically attractive and requires pursuance.

5. Discussions and conclusions

A close look at Tables 1 and 2 reveals some interesting features. The T_n of $(n-1)^{th}$ day, *i.e*,Y_1 remains the most important predictor of that of n^{th} day for all the schemes. The PCC corresponding to Y_2 is negative implying that given Y_1, Y and Y_2 are in fact negatively correlated even though simple CC between them is as high as 0.65. The predictor A' which is the advection temperature anomaly measured in the morning, holds its importance even in the 12-hr scheme. This shows the significance of morning temperature advection. A simple interpretation of this feature which could perhaps be of help in 24-hr forecasts of minimum temperature of any station is that 'the minimum temperature anomalies, rather than the minimum temperatures are advected by the 0.9 km a.s.l. wind'. That both PCC(D, Y) and PCC (C, Y) are significant shows that D supplies only a fraction of the information about the moisture depth over the atmosphere. It is seen that PCC (X,Y)== 0.14 (for 15-hr scheme) is not substantially lower than CC (X,Y)=0.19. This shows that even though the correlation between maximum and minimum temperatures is not highly significant, whatever information supplied by T_x is independent and so, important. The significance of B in 12 and 15-hr scheme points out to the importance of temperature advection in the evening also.

The study has shown that analogue and analogue *cum* multi-regression technique can be successfully employed in temperature forecasting. These are, perhaps, more appealing than the widely used multi-regression technique but need not always give better results or be cost effective. The analogue *cum* multi-regression method has several inbuilt byproducts such as the MCC and SE which help us to quantify the reliability of analogue before issuing the forecast. The SE of the forecast which varies day to day is an excellent indicator of the forecast accuracy and so forecast confidence intervals of varying widths can be formed day to day. This is in contrast with forecasts based on conventional regression equations which have a fixed SE. As analogue can be defined in any number of ways, techniques based on analogue offer enormous scope and it might be possible to detect and arrive at a package superior to conventional methods.

Acknowledgements

The author thanks the Deputy Director General of Meteorology (Weather Forecasting) for having provided the facilities to undertake this study. He also thanks S/Shri B.C. Rathod and Valsan Mathew for their help and Shri A.V. Pawar for typing the manuscript.

References

Banerji, A.K. and Chowdhury, A.B., 1972, Forecasting summer maximum temperature at Nagpur, *Indian J. Met. Geophys.*, **23**, p. 251.

Bansal, R.K. and Datta, R.K., 1974, A statistical method of forecasting the movement of cyclonic storms in the Bay of Bengal, *Indian J. Met. Geophys.*, **25**, pp. 391-396.

Brooks, C.E.P. and Carruthers, N., 1953, *Handbook of Statistical Methods in Meteorology*, London, pp. 322-328.

Datta, R.K. and Gupta, R.N., 1975, Tracking tropical storms in the Bay of Bengal and calculation of probabilities of its striking east coast by storm analogue technique, *Indian J. Met. Hydrol. Geophys.*, **26**, pp. 345-348.

Haltiner, G.J. and Martin, F.L., 1957, *Dynamical and Physical Meteorology*, McGraw Hill, London, pp. 125-141.

India Met. Dep., 1974, Heat and cold waves in India, Forecasting Manual, No. IV-6.

Kendall, M.G. and Stuart, A., 1968, *The Advanced Theory of Statistics*, Griffin, London, (Vol. II), pp. 278-374.

Klein, W.H. and Hammons, G.A., 1975, Maximum-minimum temperature forecasts based on model output statistics, *Mon. Weath. Rev.*, **103**, pp. 796-806.

Lund, I.A. and Wahl, E.W., 1955, An objective system for preparing operational weather forecasts, Air Force Surveys in Geophysics, No. 75—AFCRC.

Lund, I.A., 1963(a), Map pattern classification by statistical methods, *J. appl. Met.*, **2**, pp. 56-65.

Lund, I.A., 1963(b), Experiments in predicting sea level pressure changes for periods of less than twelve hours, *J. appl. Met.*, **2**, pp. 517-525.

Panofsky, H.A. and Brier, G.W., 1968, *Some applications of Statistics to meteorology,* University Park, Pennsylvania, pp. 117-118 and 191-208.

Raghavendra, V.K., 1956, Forecasting maximum temperature at Poona, *Indian J. Met. Geophys.*, **7**, p. 17.

Sinha, K.L., 1957, Prediction of minimum temperature at Delhi, *Indian J. Met. Geophys.*, **8**, p. 116.

Srinivasan, T.R. and Hashim, S.S., 1967, Statistical study of daily maximum temperature at Madras, *Indian J. Met. Geophys.*, **18**, p. 75.

WMO, 1966, Statistical Analysis and Prognosis in Meteorology, Tech. Note No. 71.

Wood Cock, F., 1980, On the use of analogues to improve regression forecasts, *Mon. Weath. Rev.*, **108**, pp. 292-297.