Long-Range Forecasting of Temperature and Precipitation with Upper Air Parameters and Sea Surface Temperature in a Multiple Regression Approach

By Ernest C. Kung and Hiroshi Tanaka

Department of Atmospheric Science, University of Missouri-Columbia, Columbia, Missouri 65211*

(Manuscript received 1 October 1984, in revised form 13 March 1985)

**Contribution from Missouri Agricultural Experiment Station Journal Series No. 9763*

Abstract

With a premise that the patterns of the general circulation undergo continuous variations at various time scales, experiments with a multiple regression scheme are conducted in the long-range forecasting of temperature and precipitation using upper air parameters and sea surface temperature for the period 1963–1983. The scheme uses five predictors; it also effectively eliminates the problem of multi-collinearity. The scheme gives a lead time of predictors two to eleven months preceding predictands. The forecast scheme performs the complete regression analysis as an annual updating procedure for the teleconnection of the large-scale flow patterns. In forecast experiments, the data from the forecast year are excluded from the data base of the entire study period in the regression analysis. This also ensures that the forecast experiments are completely independent of the data of the forecast year. The experiments indicate that the predictands properly selected over various sizes of large areas of the United States, Canada and the USSR show considerable forecast skill.

1. Introduction

Multiple regression is a useful empirical approach to long-range forecasting. It incorporates the relevant synoptic-statistical information into a linear prediction equation using observation data of the atmosphere, although the atmospheric processes are basically nonlinear. The knowledge obtained in the study of multiple regression forecasting also contributes to a better understanding of the behavioral pattern of the general circulation.

In our preceding reports (Kung and Sharif, 1980; Kung and Sharif, 1982; Kung, 1983) it was pointed out that for the practical purpose of long-range forecasting, the regression cannot be stabilized. Indeed, this is the basic nature of the general circulation, the patterns of which undergo continuous variations at various time scales. With this premise, Kung and Sharif (1982) developed, with upper air parameters and sea surface temperature, a multiple regression scheme for the long-range forecast of the Indian summer monsoon onset and rainfall. The forecast experiment for the period 1957–1977 showed the predicted onset date and seasonal rainfall to be very close to the recorded dates and rainfall.

In the 1982 scheme of Kung and Sharif, five or six regressors were selected from extensive correlation analyses involving upper air parameters over India and Australia and from sea surface temperature in the Indian region. Coefficients of regression equations were refitted for each forecast year during the experiment period 1957–1977 without involving the data of the forecast year. This was equivalent to a yearly updating of the regression coefficients to counter the problem of interannual transiency in the circulation patterns. In the forecast experiment it was indicated that a data period of less than 20 years is adequate for formulating the multiple regression forecast at a seasonal range. Whereas apparently encouraging results were obtained in the 1982 forecast experiments, some shortcomings of the scheme were also noted. The examination of the experimental results indicated that the regression scheme was not disturbed by the multi-collinearity among regressors. However, it would be more preferable to develop a scheme which minimizes the multi-collinearity as a consequential result of the analysis procedure rather than as a result of a partially subjective selection of the predictors. In the forecast experiments the entire data base was involved in the selection of predictors at the beginning, although the data of the forecast year were excluded in evaluating regression coefficients. For an objective forecast experiment, it is desirable to completely eliminate the data of the forecast year even if we are reasonably sure that their involvement in the selection of predictors is not a serious matter.

This paper extends our study of multiple regression forecasting to the seasonal range prediction of monthly temperature, monthly and seasonal precipitation and some related parameters of the general circulation in defined areas of the middle latitudes in the Northern Hemisphere. The series of studies by Namias (e.g., 1978, 1980, 1982) and available investigations by Barnett (1981), Harnack and Landsberg (1978), Nicholls and Woodcock (1981) and others suggest such an attempt is meaningful. In so doing, the 1982 forecast scheme has been improved to remedy the abovementioned shortcomings. The scheme effectively eliminates the problem of the multicollinearity by regressing the residual components of a predictand on predictors after deciding the first predictor in the screening process. The forecast experiments are performed for the period 1963-1983, and the selection of predictors and fitting of the regression coefficient are repeated for each year without involving the data of the forecast year. Thus, the forecast scheme performs the complete regression analysis as an annual updating procedure, and the forecast experiments are completely independent of the data of the forecast year. An attempt also is made in forecast experiments to explore various problems and possibilities associated with regression forecasting.

2. Data

The data used in this study cover the period from 1962 to 1980. They include monthly mean upper air observations at 152 stations, monthly mean sea surface temperature (SST) data, and monthly precipitation (F) data. To prepare these data, twice daily upper air observations and Navy monthly sea surface temperature records were obtained from the archives of the National Center for Atmospheric Research. The precipitation data were compiled from *Monthly Climatic Data for the World* of the National Oceanic and Atmospheric Administration.

The distribution of the 152 upper air stations is shown in Fig. 1. At each upper air station seven meteorological variables are averaged for monthly means at 700, 500, 300 and 200 mb. Monthly temperature is also obtained at the surface level and noted as Ts. The seven variables are the eastward wind component u, northward wind component v, geopotential height Z, temperature T, specific humidity q, kinetic energy k, and zonal/meridional flow indicator I, where $k=1/2(u^2+v^2)$ and $I=|u|/|v|$ (see Heddinghaus and Kung, 1980). Additionally, 16 zonal sections (S1 to S16) are defined as illustrated in Fig. 2. Each strip in the figure is in the middle of a 5-degree latitudinal section. The seven monthly variables along these strips are averaged to obtain the zonal mean values for the 16 sections.

The Navy SST data, which were given originally on the NMC octagonal grid over the Northern Hemisphere, are interpolated to a $10^\circ \times 10^\circ$ latitude-longitude grid. A total of 177 data points are available for SST. The compiled monthly precipitation data at the stations are totaled for seasonal precipitation data: December, January and February for winter; March, April and May for spring; June, July and August for summer; September,

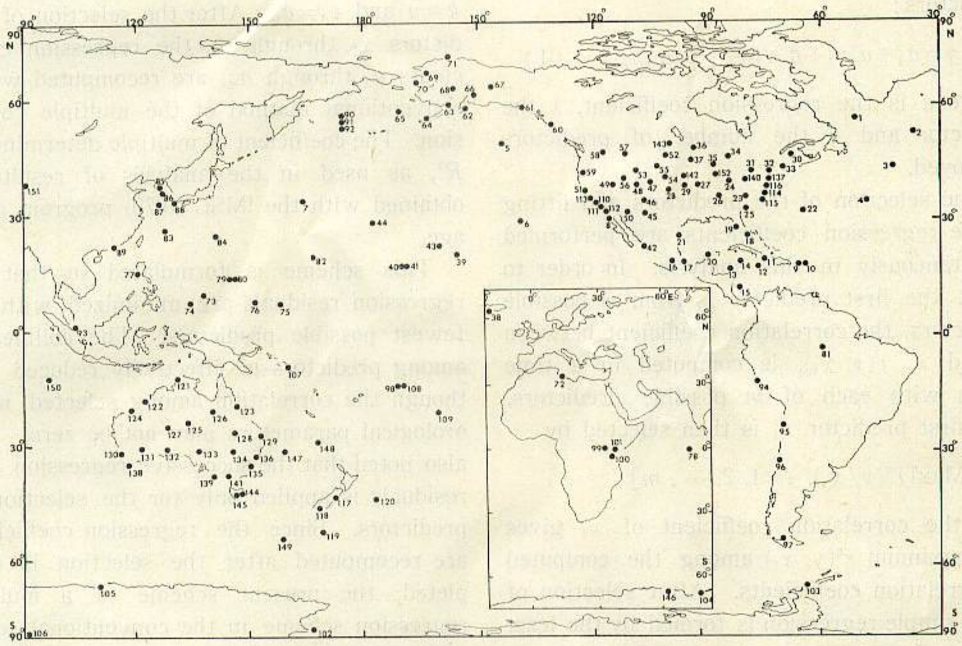

Fig. 1 Distribution of upper air stations.

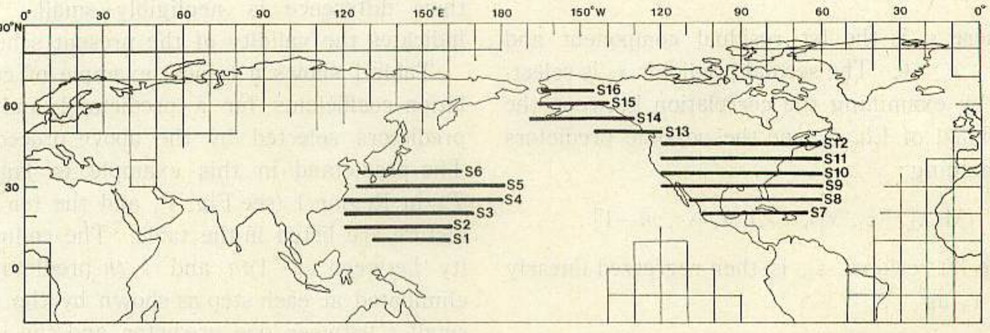

Fig. 2 Zonal sections representing 5-degree latitudinal zones for indicated sections.

October and November for fall. Physical units of meteorological parameters as used in this study are ms^{-1} for u and v, m for Z, °C for T and SST, g kg^{-1} for q, m^{2}s^{-2} for k, and mm for precipitation. The indicator I is non-dimensional.

The maintenance and variation of the general circulation are controlled by the processes whose nature is basically non-linear. The examination of covariances of meteorological fields is, thus, desirable to study the physical processes explicitly. However, for the purpose of empirical long-range forecast, the use of monthly mean values of standard observation and simple derivation is preferable. The covariance fields are related with the fields of linear parameters in the system of the general circulation. The purpose of the empirical regression scheme is to implicitly relate the observed fields to the eventual future states of the atmosphere without involving the explicit argument of the non-linear processes. For the practical use of predictors it is also more straightforward and convenient to avoid the use of covariances as predictors.

3. Scheme of regression analysis and forecast

In the general linear regression a predictand y is expressed as a function of n

predictors:

$$y=a_0+a_1x_1+a_2x_2+\cdots+a_nx_n \qquad (1)$$

where a is the regression coefficient, x the predictor and n the number of predictors employed.

The selection of the predictors and fitting of the regression coefficients are performed simultaneously in this analysis. In order to select the first predictor x_1 from m possible predictors, the correlation coefficient between y and x, $r(y, x_i)$, is computed in a time series with each of m possible predictors. The first predictor x_1 is then selected by

$$\mathrm{Max}[r^2(y, x_i),\ i=1, 2, \cdots, m],$$

i.e., the correlation coefficient of x_1 gives the maximum $r^2(y, x_i)$ among the computed m correlation coefficients. After selection of x_1, a simple regression is formed by the least squares fitting,

$$y=a_1x_1+\varepsilon_1 \qquad (2)$$

where ε_1 is the 1st residual component and $r(x_1, \varepsilon_1)=0$. The second predictor x_2 is selected by examining the correlation between the residual of Eq. (2) and the possible predictors remaining:

$$\mathrm{Max}[r^2(\varepsilon_1, x_i),\ i=1, 2, \cdots, m-1].$$

The 1st residual, ε_1, is then regressed linearly on x_2 as

$$\varepsilon_1=a_2x_2+\varepsilon_2 \qquad (3)$$

where ε_2 is the 2nd residual and $r(x_2, \varepsilon_2)=0$. a_2 is determined by the least squares method again.

Likewise, the selection of the k *th* predictor x_k is done through the same procedure:

$$\mathrm{Max}[r^2(\varepsilon_{k-1}, x_i),\ i=1, 2, \cdots, m-k+1].$$

The regression of ε_{k-1} on x_k is

$$\varepsilon_{k-1}=a_kx_k+\varepsilon_k \qquad (4)$$

where $r(x_k, \varepsilon_k)=0$. Summation of the single regression equations obtained will yield the multiple regression

$$y=a_1x_1+a_2x_2+\cdots+a_kx_k+\varepsilon_k \qquad (5)$$

where $k=n$, and Eq. (5) becomes Eq. (1) with $k=n$ and $\varepsilon_n=a_0$. After the selection of predictors x_1 through x_n, the regression coefficients a_1 through a_n, are recomputed with a conventional method of the multiple regression. The coefficient of multiple determination R^2, as used in the analysis of results, is obtained with the IMSL (1979) program package.

This scheme is formulated so that the regression residuals are minimized with the fewest possible predictors. The collinearity among predictors is effectively reduced even though the correlation among selected meteorological parameters may not be zero. It is also noted that the successive regression with residuals is applied only for the selection of predictors. Since the regression coefficients are recomputed after the selection is completed, the present scheme is a multiple regression scheme in the conventional sense. The comparison of the regression coefficients, as obtained during the selection, shows that their difference is negligibly small. This indicates the validity of the present scheme.

Table 1 shows a typical example of correlation coefficients for a predictand and ten predictors selected by the above procedure. The predictand in this example is January Ts in Region 1 (see Fig. 4), and the ten predictors are listed in the table. The collinearity between $(k-1)th$ and k *th* predictors is eliminated at each step as shown by the very small r between one predictor and the next. Typical of the regression analysis under the present scheme, x_1 and x_2 show a high correlation with predictand y; but others, except for x_{10}, show relatively low correlation with the predictand. Fig. 3 shows the coefficient of multiple determination R^2, and the mean square error MSE as functions of the number of predictors n for the same example of regression analysis as in Table 1. With four predictors, 94% of total variance is accounted for, and with five predictors, 99% is accounted for. The MSE in units of °C^2 is 0.19 and 0.09, respectively. More predictors will not gain in prediction. Thus, the number of predictors is set at $n=5$ in the forecast experiment of this study.

Forecast experiments for each predictand

Table 1 Simple correlation coefficients for a predictand and ten predictors. The predictand y is January Ts in Region 1. Predictors are identified at the top of the table with physical parameters followed by two subscripts, which give the location and level in mb.

	y	x_1 $U_{21,500}$	x_2 $I_{75,300}$	x_3 $k_{43,500}$	x_4 $I_{39,500}$	x_4 $u_{74,300}$	x_6 $q_{11,500}$	x_7 $k_{71,500}$	x_8 $k_{15,500}$	x_9 $k_{51,500}$	x_{10} $SST_{(40,130)}$*
y	1.00										
x_1	−0.78	1.00									
x_2	−0.51	−0.00	1.00								
x_3	0.16	0.03	0.17	1.00							
x_4	−0.13	−0.18	0.18	−0.10	1.00						
x_5	−0.31	0.13	−0.04	−0.30	0.13	1.00					
x_6	0.17	−0.26	−0.13	−0.42	0.46	0.11	1.00				
x_7	−0.02	−0.12	−0.04	−0.18	0.32	−0.01	0.25	1.00			
x_8	−0.22	0.18	0.20	0.36	0.09	−0.01	−0.15	0.04	1.00		
x_9	0.06	−0.25	0.04	−0.11	−0.07	0.41	−0.17	0.02	−0.21	1.00	
x_{10}	−0.50	0.48	0.19	0.16	0.00	0.39	−0.25	−0.43	0.14	0.13	1.00

* For SST( ) indicate latitude N and longitude E.

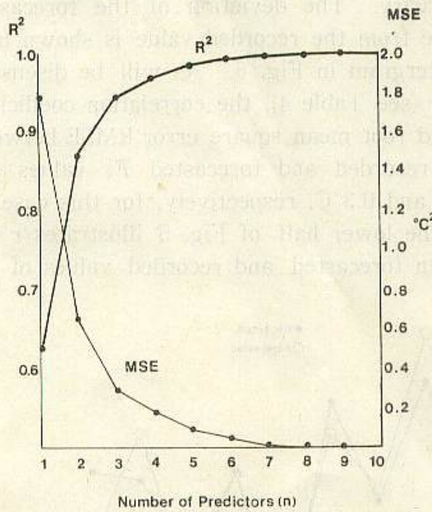

Fig. 3 Coefficient of multiple determination and mean square error as functions of the number of predictors in predicting January Ts in northeastern U.S. (Region 1).

are performed for each year from 1963 to 1980 using the dataset from 1962 to 1980. For each forecasting year, five predictors are selected by correlation analysis as described above. Data from the forecasting year are excluded. The regression coefficients in Eq. (1) are then determined by the least squares method, again excluding the forecasting year. The forecasted value of the predictand is evaluated using the regression equation obtained and the assumed real-time data which are the data of the forecasting year. Because data of the forecast year are not used, both in the selection of the predictors and in the determination of the regression coefficients, the forecast is completely independent of the data of the forecast time. Therefore, the conditions of real-time experiments are satisfied.

4. Forecast experiments

Predictands of large areas in forecast experiments include monthly surface temperatures Ts, monthly and seasonal precipitation P, kinetic energy k and flow indicator I for specified months and seasons. The areas over which these predictands are defined are shown in Fig. 4 as Regions 1 to 9. Regions 1 to 3 are defined in relation to the available upper air station in the area and represent the northeastern, mid-Atlantic and midwestern United States for Ts, k and I. Regions 4 to 9 are defined for P through graphical determination with the clustering program by Barr *et al.* (1976) utilizing precipitation data. Regions 4 and 5 represent the northcentral United States. Region 6 is central Canada. Regions 7, 8 and 9 are in the USSR, and Region 7 includes the Ukraine.

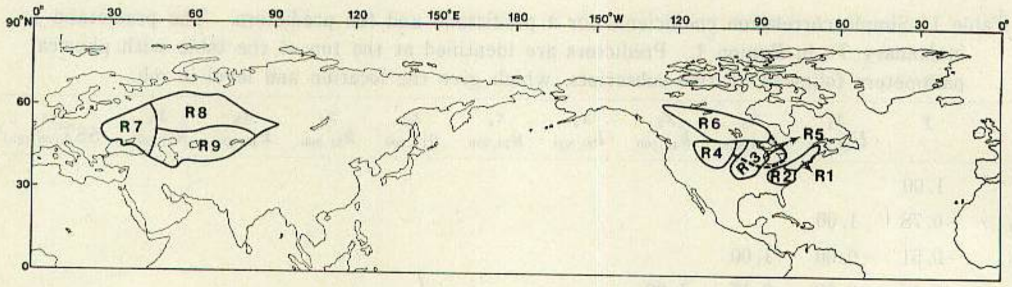

Fig. 4 Areas over which the predictands are defined.

Five predictors, x_1 to x_5, are selected for each predictand for individual forecast years from monthly mean upper-air meteorological variables and *SST* at all available data locations. With seven meteorological variables at four levels for stations and 16 zonal sections, there are 4704 possible predictors in the upper-air data group. There are an additional 177 possible predictors in the *SST* grid data, giving 4881 possible predictors for each month.

The lead time of a predictor month over the predictand month is the forecast time in the forecast experiment. For each predictand the lead is extended from one month (*i.e.*, the month immediately preceding the predictand month) to 12 months. The 4881 predictors for each forecasting month (*i.e.*, predictor month) are considered separately in choosing predictors x_1 to x_5. Subsequent comparison of regression forecasts may determine the best predictor month for the predictands.

An example of a forecasting experiment with high forecast skill is shown in the upper half of Fig. 5 for July Ts in the northeastern United States (Region 1) with predictors in February. The deviation of the forecasted value from the recorded value is shown in a scattergram in Fig. 6. As will be discussed later (see Table 4), the correlation coefficient r and root mean square error RMSE between the recorded and forecasted Ts values are 0.92 and 0.3°C, respectively, for this case.

The lower half of Fig. 5 illustrates r between forecasted and recorded values of Ts

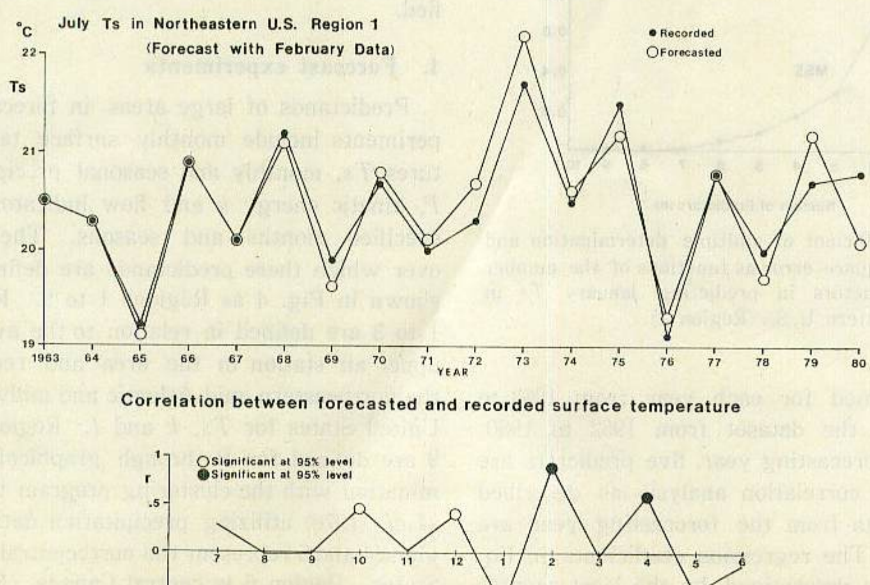

Fig. 5 Forecast experiment of July Ts in northeastern U.S. (Region 1).

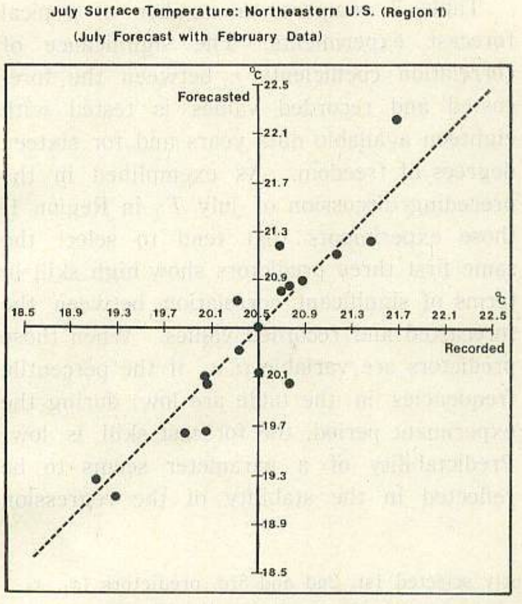

Fig. 6 Scattergram forecasted and recorded July Ts as in Fig. 5.

in the same forecast experiment with lead time of predictors varying from one month (June of the same year) to 12 months (July of the preceding year). As shown in the figure, July Ts in the northeastern United States is well-predicted using predictors in April and February. The high forecast skill is in excess of a 99% level in terms of r between the forecasted and recorded values. Since the correlation is highest for February, predictors in February are used in forecasting July Ts of Region 1.

Since the independent regression analysis is repeated for each forecast year, it will be pertinent to examine the consistency of the results of these analyses—particularly in view of the large number of possible predictors. Table 2 shows the five predictors selected for each forecasting year with the same example of July Ts in Region 1. As shown in the table, the 300 mb u at station 71 (Barrow,

Table 2 Five predictors selected for forecasting July surface temperature in the northeastern United States (Region 1) with February data set from 1963 to 1980. The bottom line is for the fitting with the entire data set. x_1 to x_5 are predictors. Location and level of the predictor are indicated in parentheses. Location is either a station as numbered or a zonal strip as numbered with S.

Year	x_1	x_2	x_3	x_4	x_5
1963	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	k (73, 300 mb)
1964	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)
1965	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	k (73, 300 mb)
1966	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)
1967	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)
1968	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	k (73, 300 mb)
1969	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	Z (42, 200 mb)	v (132, 500 mb)
1970	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)
1971	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)
1972	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	I (69, 500 mb)	I (28, 200 mb)
1973	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	T (59, 700 mb)	v (43, 700 mb)
1974	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	T (42, 700 mb)	k (50, 700 mb)
1975	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	I (17, 700 mb)	I (14, 700 mb)
1976	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	I (17, 700 mb)	I (110, 700 mb)
1977	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	I (17, 700 mb)	k (12, 700 mb)
1978	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	I (17, 700 mb)	v (145, 300 mb)
1979	u (71, 300 mb)	k (56, 700 mb)	v (136, 700 mb)	u (S1, 200 mb)	T (42, 700 mb)
1980	u (71, 300 mb)	I (149, 500 mb)	k (57, 700 mb)	T (82, 300 mb)	k (127, 700 mb)
Fitting	u (71, 300 mb)	k (56, 700 mb)	I (68, 700 mb)	q (S5, 700 mb)	I (139, 300 mb)

Alaska: 71.3°N, 156.8°W) is selected as the first predictor for all years. The 700 mb k at station 56 (Salt Lake City: 40.8°N, 112°W) and 700 mb I at station 68 (McGrath, Alaska: 63.0°N, 155.6°W) are selected as the second and third predictors, respectively, for all but a few years. The fourth and fifth predictors vary more than the first three, although the 700 mb q of zonal section S5 (30°N in the western Pacific Ocean) and 300 mb I at station 139 (Mt. Gamier: 37.8°S, 140.8°E) are more often selected. It is also noted that the most selected of these five predictors are the same as those selected by fitting of multiple regression with the entire dataset, and they are listed at the bottom of the table.

Table 3 compares a number of typical forecast experiments. The significance of correlation coefficients r between the forecasted and recorded values is tested with eighteen available data years and for sixteen degrees of freedom. As exemplified in the preceding discussion of July Ts in Region 1, those experiments that tend to select the same first three predictors show high skill in terms of significant correlation between the forecasted and recorded values. When those predictors are variable (*i.e.*, if the percentile frequencies in the table are low) during the experiment period, the forecast skill is low. Predictability of a parameter seems to be reflected in the stability of the regression

Table 3 Percentile frequencies of the most frequently selected 1st, 2nd and 3rd predictors (x_1, x_2 and x_3) in typical temperature and precipitation forecast experiments during the 18-year period. The correlation coefficient between the forecasted and recorded values is shown as r.

Predictand	r	x_1	x_2	x_3
Jan Ts in R1	0.48*	94%	44%	11%
Jan Ts in R1, 2	0.45*	88%	88%	17%
Jan Ts in R3	0.72**	100%	72%	72%
Jul Ts in R1	0.92**	100%	94%	88%
Jul Ts in R1, 2	0.33	17%	6%	6%
Jul Ts in R3	0.76**	100%	72%	61%
Jan k 300 mb in R1, 2, 3	0.78**	94%	17%	6%
Jul I 700 mb in R1, 2, 3	0.63**	72%	67%	28%
Jun P in R4	0.55**	88%	72%	22%
Jun P in R5	0.61**	94%	44%	28%
Jun P in R6	0.68**	100%	61%	33%
Jun P in R7	0.34	88%	28%	17%
Jun P in R8	0.49*	83%	67%	28%
Jun P in R7, 8, 9	0.57**	88%	56%	33%
Winter P in R4	0.63**	100%	56%	33%
Winter P in R5	0.50*	50%	50%	33%
Winter P in R6	0.81**	78%	78%	44%
Winter P in R7	0.72**	100%	33%	22%
Winter P in R8	0.66**	94%	61%	28%
Winter P in R7, 8, 9	0.68**	94%	44%	11%
Spring P in R7	0.54**	94%	78%	44%
Summer P in R7, 8, 9	0.67**	94%	78%	56%

Note: (1) ** significant at 99% level
* significant at 95% level
(2) R in predictands indicates single or combined regions as shown in Fig. 4.

analysis in this analysis-forecasting scheme. It may be stated that the forecast skill is high if the temporal variation of the predictand is affected by only a few dominant principal components, whereas it is low if many alternative components dominate. Of only five predictors selected from a pool of 4881 possible predictors, the stability of x_1—in terms of repeated selection of a parameter for x_1 during the forecast experiment—appears most related to the level of forecast skill. This is also true for x_2 and x_3 but to a lesser degree. For instance, January Ts in Region 1 shows a high stability of x_1, but that of x_2 and x_3 are relatively low; here r is only 0.48 although it is significant at a 95% level. In contrast to this, July Ts in Region 1 shows a stable selection of x_1, x_2 and x_3, and its forecast skill is the highest of the listed experiments in the table with $r=0.92$.

Fig. 7 illustrates the forecast experiments of Januarry Ts in the midwestern United States (Region 3) and 300 mb k in the eastern and midwestern United States (Regions 1, 2 and 3). The correlation between the forecasted and recorded values is $r=0.72$ for Ts with August predictors and $r=0.78$ for 300 mb k with September predictors. Both are significant at a 99% level (Table 3). Likewise, Fig. 8 shows July Ts for Region 3 and 700 mb I in Regions 1, 2 and 3 combined. The r between the forecasted and recorded values is 0.76 for Ts with May predictors and 0.63 for 700 mb I, both significant at a 99% level. In Figs. 5, 6, 7 and 8, it appears that Ts of large areas and large-scale flow characteristics may be predicted with high skill several months in advance if proper predictors are selected through regression analysis.

Forecast experiments for precipitation are conducted over large regions. Fig. 9 shows June precipitation in the northcentral United States (Region 4) and central Canada (Region 6) as examples of monthly precipitation forecast in North America. A large yearly

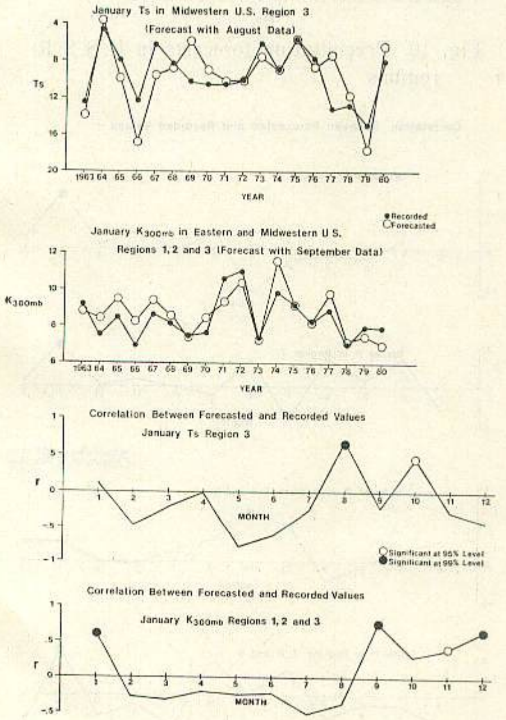

Fig. 7 Forecast experiments of January Ts in midwestern U.S. (Region 3) and January $k_{300\,\mathrm{mb}}$ in eastern and midwestern U.S. (Region 1, 2 and 3).

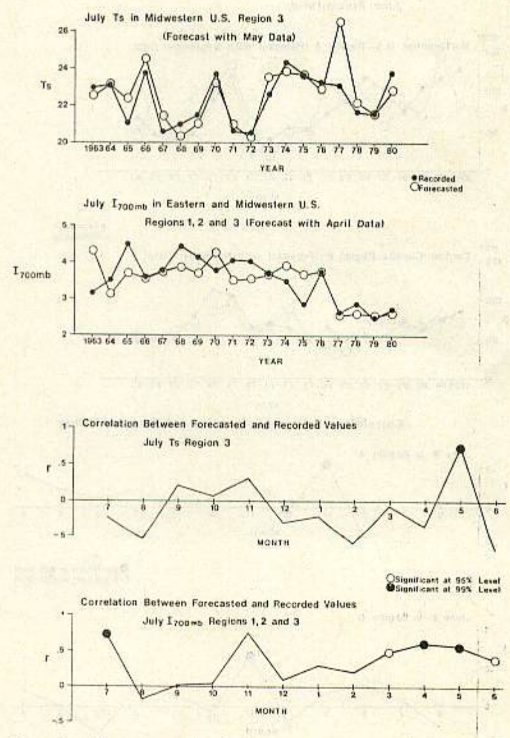

Fig. 8 Forecast experiments of July Ts in midwestern U.S. (Region 3) and July $I_{700\,\mathrm{mb}}$ in eastern and midwestern U.S. (Regions 1, 2 and 3).

fluctuation of June precipitation in these important agricultural regions is predicted with skill. The correlation between the forecasted and recorded values is $r=0.55$ for Region 4 with September predictors, and $r=0.68$ for Region 6 with November predictors. Both are significant at a 99% level. Figs. 10 and 11 illustrate examples of precipitation in regions of Russia. Seasonal precipitation in winter and spring in Region 7, as respectively forecasted with October and January predictors, shows high skill. Summer precipitation for Regions 7, 8 and 9 combined is forecasted using November predictors with high skill. Monthly precipitation for June for the same combined region using April predictors also shows high forecast skill The selection of the areas for forecast experiments is dominated by the availability of proper data and economic importance of the region. The optimal size of the region is yet to be decided in future experiments. However, as shown in these examples and in listings for Tables 3 and 4, the monthly and seasonal precipitation appear to be predictable for large regions with properly selected predictors through regression analysis.

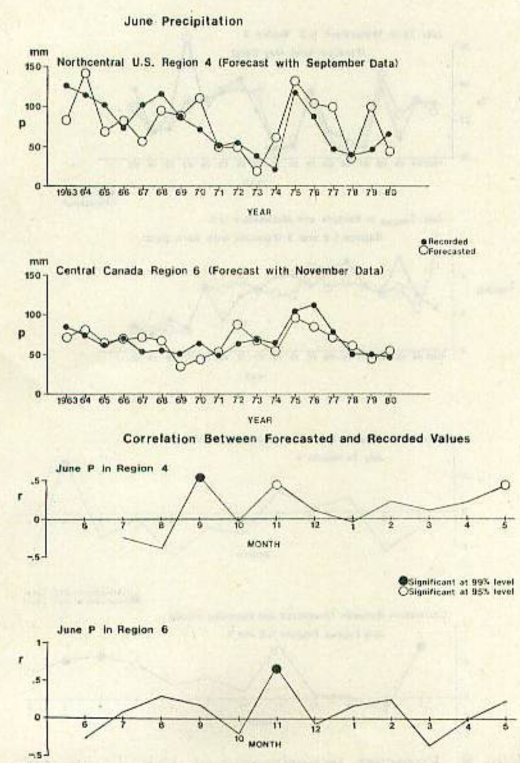

Fig. 9 Forecast experiments of June precipitation in northcentral U.S. (Region 4) and central Canada (Region 6).

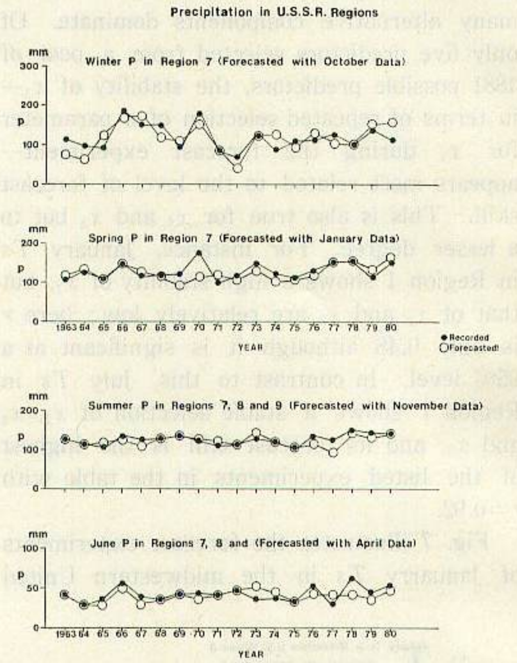

Fig. 10 Precipitation forecasts in U. S. S. R. regions.

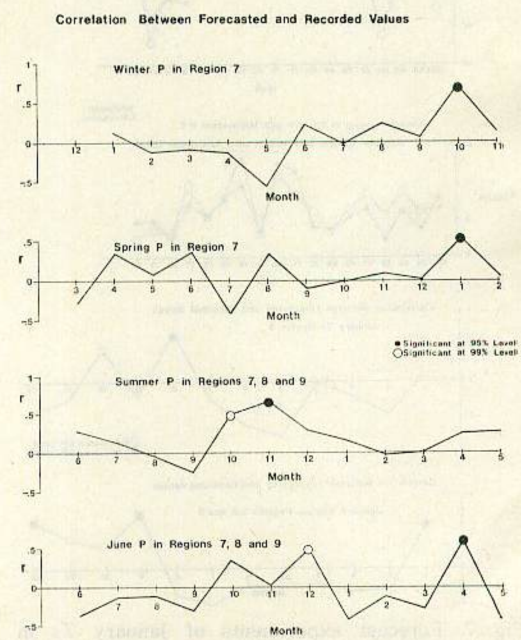

Fig. 11 Correlation between forecasted and recorded values for forecast experiments as in Fig. 10.

Table 4 Most frequently selected predictors, x_1 to x_5, in forecast experiments during the 18-year period and skill parameters. A predictor is identified by a physical parameter followed by two numbers, the location and the level in mb. For SST, latitude N and longitude E are indicated in parentheses.

Predictand	Predictor month	Lead Time (months)	x_1	x_2	x_3	x_4	x_5	r*	$RMSE$**	S	R^2
Jan Ts in R1	Oct	3	u: 21,500	I: 75,300	k: 43,500	I: 39,500	u: 74,300	0.48	1.9	12	0.971
Jan Ts in R1,2	Mar	10	I: 60,700	I: 123,700	u: 150,700	v: 150,300	SST (30, 210)	0.45	2.4	NS***	0.980
Jan Ts in R2	Aug	5	k: 51,700	u: 20,700	q: 134,700	v: S 1,700	u: 94,500	0.72	2.5	23	0.992
Jul Ts in R1	Feb	5	u: 71,300	k: 56,700	I: 68,700	q: S 5,700	I: 139,300	0.92	0.3	78	0.992
Jul Ts in R1,2	Mar	4	z: 43,700	q: 105,700	v: 95,700	I: 106,300	I: 117,200	0.33	1.1	NS***	0.951
Jul Ts in R3	May	2	T: 60,200	T: 115,300	k: 131,700	k: 129,700	SST (50, 150)	0.76	1.0	36	0.988
Jan k 300 mb in R1,2,3	Sep	4	T: 105,500	T: 123,300	I: S 4,700	T: 22,500	I: 48,500	0.78	81	54	0.982
Jul I 700 mb in R1,2,3	Apr	3	I: 12,200	v: 74,200	I: 121,200	v: 66,500	u: S 9,300	0.63	0.5	32	0.993
Jun P in R4	Sep	9	q: 8,700	u: 10,500	z: 13,300	q: 102,700	v: 11,500	0.55	31	4	0.972
Jun P in R5	Jan	5	q: 107,500	k: 115,300	I: 20,200	T: S 4,200	SST (50, 160)	0.61	22	21	0.970
Jun P in R6	Nov	7	q: 102,500	n: 147,500	z: 61,200	q: S 2,500	v: 148,200	0.68	14	41	0.982
Jun P in R7	Apr	2	v: 34,500	I: 49,300	u: 43,700	v: 73,700	I: 53,300	0.34	20	NS***	0.972
Jun P in R8	Aug	10	q: 111,500	v: 110,200	q: 146,500	I: 132,500	q: 10,700	0.49	15	15	0.978
Jun P in R7,8,9	Apr	2	u: 105,200	k: 132,500	v: 123,700	z: 46,500	u: 134,500	0.57	8	28	0.989
Winter P in R4	Jul	5	u: 124,200	u: S 10,500	I: 106,200	k: 117,500	v: 34,700	0.63	10	28	0.986
Winter P in R5	Jul	5	q: 45,700	k: 134,500	q: S 12,700	I: 66,700	q: 96,500	0.50	25	17	0.987
Winter P in R6	Jul	5	k: S 11,700	v: 70,500	T: 75,500	v: 21,300	q: 63,700	0.81	6	60	0.984
Winter P in R7	Oct	2	k: 31,700	T: 147,300	I: 146,700	T: 55,200	I: 13,300	0.72	26	36	0.986
Winter P in R8	Jan	11	v: 117,300	k: 48,700	I: 102,300	q: 148,500	T: 125,200	0.66	15	17	0.975
Winter P in R7,8,9	Oct	2	k: 31,700	I: 57,200	k: 129,200	SST (30, 150)	I: 21,300	0.68	12	38	0.980
Spring P in R7	Jan	2	k: 20,300	T: 147,200	SST (30, 190)	k: 131,500	k: 129,700	0.54	20	14	0.978
Summer P in R7,8,9	Nov	7	I: 125,200	k: 73,300	T: 115,200	T: 12,300	D: 129,700	0.67	13	33	0.988

* r is the correlation coefficient between the forecasted and recorded value.
** $RMSE$ is in the same physical units as the respective predictands.
*** NS: No skill (zero or negative).

5. Utilities of regression forecasting

The series of forecast experiments as listed in Table 3 are further examined in Table 4. The listing includes all examples used in this report. For each series of forecast experiments the table shows the predictand; predictor month and lead time; most frequently selected predictors for x_1 to x_5; correlation coefficient r between the forecasted and recorded values of the predictand; root mean square error of forecast RMSE; skill score S; and coefficient of multiple determination R^2. The skill score is defined as

$$S=\left(1-\frac{\mathrm{MSE}}{var}\right)\times 100 \tag{6}$$

where var is the variance of the recorded values of the predictand during the 18-year experiment period. The parameter S thus measures the size of the deviation of the forecasted from the recorded value during the experiment—making it useful, along with r, as additional information regarding forecast. The positive S value indicates existence of forecast skill, with 100 being the perfect forecast through the time series. If S is zero or negative there is no forecast skill.

As presented, the predictands properly selected over various sizes of large areas could be predicted. The regression scheme in this study also indicates considerable stability by repeated selection of the same parameters for the first few predictors from a very large group of meteorological parameters in a series of independent regression analyses. The R^2 values in Table 4 show that nearly all of the variance is accounted for with five predictors. Most of the series show significantly large r between forecasted and recorded values (also see Table 3) and positive forecast skill in terms of S which is often large. The results of k and I seem to indicate that properly defined time mean upper air parameters also may be adopted as predictands. The use of these predictands in conjunction with the traditional predictands Ts and P shall give useful information in forecasting at the seasonal range. Some series with small r and no skill score are shown in Table 4. Reorganization of the regions defining the predictands may improve these scores. Increase of upper air stations or redistribution of the data station network in defining the predictands and selecting predictors also may be considered to improve the forecast skill of those predictands.

This regression scheme is based on teleconnections of the large-scale flow characteristics with which the predictands and predictors are defined. The scheme gives a lead time of predictors two to eleven months preceding predictands. From what is listed in Table 4 and shown throughout this paper (Figs. 5, 7, 8, 9 and 11), this forecast time will vary from one predictand to another, and it may be a characteristic of the teleconnection on which this forecast scheme is based. Thus, predictors at different months might have to be taken for different predictands of the same month. However, it is difficult to state a specific pattern for selection of predictors only from the listing studied in this experiment. Despite the well established importance of SST as a predictor and its teleconnection with various phenomena, SST has not appeared as a major predictor in this series of forecast experiments. In the present scheme the variables which directly affect the predictand are selected as predictors. However, it does not deny the importance of SST as a controlling factor in the physical processes. It is most likely that the effect of SST is reflected in upper air parameters and thus implicitly represented by them in the regression procedures employed.

6. Remarks

The results of the present study indicate that the use of five predictors is sufficient. Repeated selection of the same parameters for the first few predictors of a predictand also should mean that teleconnections represented by these predictors are adequately utilized in the present multiple scheme of regression forecasting. Examination of the results listed in Tables 3 and 4 suggests regression forecasting on the basis of a single predictor may not be sufficient. The inadequacy of a simple regression scheme is further evidenced in that the first predictor accounts for only a limited portion of the total variance (see Fig. 3). In adopting the present procedure spread of the lead time of predictors among predictands will make the forecast operation complicated. The scheme deviates substantially from most synoptic long-range forecasting that considers simultaneous predictors. However, this specific feature of the scheme is in accord with the revealed regression feature of the atmospheric parameters.

Acknowledgements

The authors acknowledge the contribution of their former colleague P. H. Chan in the preparatory stage of this study. The technical assistance of S. J. Brooks, L. K. Gibson, R. L. Rees and K. R. Kreiner is sincerely appreciated.

This research was supported jointly by the National Science Foundation, and the National Oceanic and Atmospheric Administration under GARP Grant NSF ATM-8410487.

References

Barr, A. J., J. H. Goodnight, G. P. Sall and J. T. Helwig, 1976: A user's guide to SAS-76. SAS Institute, Inc., Raleigh, NC, 329 pp.

Barnett, T. P., 1980: Statistical prediction of North American air temperature from Pacific predictors. *Mon. Wea. Rev.*, **109**, 1021–1041.

Harnack, R. P. and H. E. Landsberg, 1978: Winter season temperature outlooks by objective methods. *J. Geophys. Res.*, **83**, 3601–3616.

Heddinghaus, T. K. and E. C. Kung, 1980: An analysis of climatological patterns of the Northern Hemispheric circulation. *Mon. Wea. Rev.*, **108**, 1–17.

IMSL, 1979: Reference manual, ed. 7. International Mathematical and Statistical Libraries, Inc., Houston.

Kung, E. C., 1983: Reply. *J. Clim. Appl. Meteor.*, **22**, 1134–1135.

——— and T. A. Sharif, 1980: Regression forecasting of the onset of the Indian summer monsoon with antecedent upper air conditions. *J. Appl. Meteor.*, **19**, 370–380.

——— and ———, 1982: Long-range forecasting of the Indian summer monsoon onset and rainfall with upper air parameters and sea surface temperature. *J. Meteor. Soc. Japan*, **60**, 672–681.

Namias, J., 1978: Multiple causes of North American abnormal winter 1976–77. *Mon. Wea. Rev.*, **106**, 279–295.

———, 1980: Some concomitant regional anomalies associated with hemispherically averaged temperature variations. *J. Geophys. Res.*, **85**, 1585–1590.

———, 1982: Anatomy of great plains protracted heat waves (especially the 1980 U. S. summer drought). *Mon. Wea. Rev.*, **110**, 824–838.

Nicholas, N. and F. Woodcock, 1981: Verification of an empirical long-range weather forecasting technique. *Quart. J. Roy. Meteor. Soc.*, **107**, 973–976.

高層観測資料と海面温度を用いた重回帰的手法による気温および降雨量の長期的予測

Ernest C. Kung and Hiroshi Tanaka

Department of Atmospheric Science, University of Missouri-Columbia

大気大循環の変動は種々の時間スケールから構成される連続的なものであるとの前提のもとで，1963年から1983年の期間を対象として，高層観測資料と海面温度を用いて気温と降雨量の長期的予測を行なう重回帰スキームの実験を行なった。このスキームは五つの説明変数を使用し，変数相互間の多重線形性によって生じる問題点を除去するものである。このスキームは被説明変数に２ヶ月から11ヶ月先行する説明変数を使用する。また予報スキームにおいては年ごとの係数の更新手続として，大規模な気流系におけるテレコネクションについて完全な回帰分析を行なう。予報実験においては予報年の資料は回帰分析における対象期間から除外した。これは予報実験が予報年自身の資料からは独立であることを確保するものである。予報実験の結果は，米国，カナダ，ソ連上の種々の面積について適切に選定された被説明変数に対し，かなりの予報能力を示した。