An Experiment in Nonlinear Statistical Weather Forecasting

EDWARD N. LORENZ[1]

National Center for Atmospheric Research,[2] Boulder, Colo. 80307

(Manuscript received 14 September 1976, in revised form 15 February 1977)

ABSTRACT

We inquire whether an empirical weather forecasting scheme can profitably incorporate a possible nonlinear relationship between observed predictands and predictors.

We analyze a set of twice-daily hemispheric 500 mb height fields into truncated series of spherical harmonics. From each set of spherical-harmonic coefficients, we predict the coefficients 24 h in advance by integrating the barotropic vorticity equation in spherical-harmonic form.

We then establish linear regression equations for predicting the same coefficients, using as predictors the coefficients which represent the observed height fields, and, in some instances, the numerically predicted height fields. We find that the empirical schemes which incorporate nonlinearity by using the numerically predicted fields perform considerably better than those which do not.

1. Introduction

Before the days of digital computers most operational weather forecasting was subjective. Ordinarily a forecaster would obtain a sequence of synoptic maps, each showing the values of the various weather elements at a network of stations at a single time, and would analyze the fields of weather elements into weather systems. He would then construct a prognostic map, or "prog," which would represent his estimate of the arrangement of the weather systems at some future time, say, 24 h ahead. From the prog he would estimate the next day's weather elements.

Sometimes the forecaster would base a prog partly on physical theory. For the most part, however, he would rely upon his familiarity with the manner in which weather systems typically behave. It was recognized at that time that objective methods of forecasting were possible in principle, but the amount of computing needed to implement them seemed to render them impractical.

With the advent of computers, objective operational forecasting became a reality. Among procedures based upon physical theory, the commonest, generally called "numerical weather prediction," consists of stepwise numerical integration of finite-difference approximations to the dynamic equations governing the atmosphere. Empirical procedures, often called "statistical weather prediction," include the use of linear regression, which requires the inversion of matrices of high order, or equivalent operations. Because numerical and statistical weather prediction use rather different mathematical techniques and because the underlying philosophies differ considerably, somewhat different groups of meteorologists have been attracted to them, and differing opinions as to their relative merits have been rife.

In an informal conversation in which the writer took part about 20 years ago, the question arose as to how the best system for producing an operational objective 24 h prog could be developed, if the system had to be ready within one year. We more or less agreed that the further improvements in numerical weather prediction to be expected in a single year would be small, and that the greatest immediate gains would come from an empirical scheme, in which the numerically produced prognostic charts or "numerical progs" would enter as predictors. Such a scheme was never established, perhaps because, fortunately, where was no demand even at the highest bureaucratic levels that a weather forecasting technique be made final within a year.

Nevertheless, the formulation of a scheme of this sort seemed to offer a good research topic. At that stage of computer development, the work involved in producing a large new set of numerical progs would have been prohibitive, and the logical source for the progs was the set already produced operationally in Washington. Any intent of the writer to pursue such research was abandoned when it was realized that, because the operational numerical forecasting model was being continually improved, no large collection of progs produced by one and same model was available.

Perhaps the study which most nearly approached the desired objective was one performed by Cooley

[1] Present affiliation: Department of Meteorology, Massachusetts Institute of Technology, Cambridge 02139.

[2] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

(1958). Here the predictors for future 1000 and 500 mb heights, in addition to present heights, were present values of Jacobians of height with vorticity or temperature, as they would appear in a two-level, geostrophic, numerical prediction model; using these as predictors in a linear regression scheme was equivalent to using a numerical prog produced with a single, uncentered, forward time step. The available computer power limited Cooley to a small number of forecasts, and the improvements yielded by the nonlinear terms (the Jacobians) were not obviously greater than those expected by chance.

Cooley's results seemed to support the claims of a number of devotees of statistical forecasting that nonlinearity need not be considered explicitly. Such a belief stemmed in part from a theoretical treatment of nonlinear prediction by Wiener (1956), which was apparently rather widely misinterpreted as implying that the performance of any nonlinear formula could be duplicated by a linear formula containing the same predictors. The appropriate interpretation of Wiener's verbally expressed result, which Lorenz (1973) has meanwhile converted into equations, is that by introducing a sufficient number of characteristic functions (functions which always assume the value 0 or 1) as *new* predictors, one can replace a nonlinear formula by a linear formula. No comparison of linear and nonlinear formulas using the same predictors is offered.

Recently the use of predictors chosen from numerically produced progs has formed an essential part of the Model Output Statistics (MOS) procedure currently used operationally by the National Weather Service (see Glahn and Lowry, 1972). Here also the predictions are made by linear regression. The predictands are weather elements, such as temperature and precipitation, at specific locations. An intermediate step of constructing a prog by linear regression would be superfluous. Whether the success enjoyed by the method is due to the *nonlinearity* in the numerical progs is not evident.

By now numerical progs are presumably better than progs based upon a combination of dynamic equations and linear regression would have been in the late 1950s. It may still be true, however, that if one now had to perfect a procedure for producing progs within one year, the best results could be obtained by basing an empirical scheme on today's numerical forecasts. During the same years computers have become powerful enough to produce economically a large number of numerical progs for a specific research problem. The purpose of this study is to reexamine the question as to whether the performance of a linear regression scheme, based on real weather data, can be improved by introducing nonlinear functions of the original predictors as additional predictors and, in particular, whether the appropriate additional predictors are the ones which would be suggested by commonly used numerical forecasting models.

2. Linear regression

In predicting by linear regression, the standard choice of a "best" formula is the one which minimizes the mean-square prediction error. This measure of goodness is chosen because other meaningful measures tend to make the mathematical treatment rather awkward. In practical weather forecasting there is no assurance that the best formula in this sense will be the best from the point of view of any specific user. However, the mean-square error is reasonably satisfactory for this study, whose principal purpose is not to develop new operational forecasting procedures but to examine the role of nonlinearity.

We desire formulas of the form

$$y=\sum_{i=0}^{M}\alpha_i x_i+\epsilon, \tag{1}$$

relating a predictand y to M predictors $x_1,\ldots,x_M$ and a prediction error ϵ. We have included a constant term α_0 in (1) in a concise form by introducing an additional "predictor" x_0 whose value is always unity. We wish to choose the coefficients $\alpha_0,\ldots,\alpha_M$ to minimize $\langle\epsilon^2\rangle$, where the angle braces denote an expected value, or an average over the joint population to which the predictand and the predictors belong. The coefficients would then satisfy the $M+1$ simultaneous equations

$$\sum_{j=0}^{M}\langle x_i x_j\rangle\alpha_j=\langle x_i y\rangle. \tag{2}$$

In practice we do not know the statistics of the population, and we generally estimate them by selecting a sample (the *dependent* sample) consisting of N' observed values of y and the corresponding values of $x_1,\ldots,x_M$. We then establish a formula

$$y=\sum_{i=0}^{M}a_i x_i+e \tag{3}$$

by choosing the coefficients $a_0,\ldots,a_M$ to minimize $\overline{e^2}$, where the overbar denotes an average over the dependent sample. The coefficients then satisfy the equations

$$\sum_{j=0}^{M}\overline{x_i x_j}a_j=\overline{x_i y}. \tag{4}$$

Instead of solving (4) as it stands we may introduce the predictors one at a time. To do this we let $y_0=y$ and $x_{i,0}=x_i$, and then let y_{k+1} and $x_{i,k+1}$ be the errors in predicting y and specifying x_i by means of x_0 and the first k actual predictors $x_1,\ldots,x_k$. (Algebraically a specification is identical with a prediction, but physically a specification does not involve a time lag.) For

additional conciseness we may denote the predictand y by x_{M+1}. The formula which replaces (3) and (4) is then

$$x_{i,k+1}=x_{ik}-(\overline{x_{kk}x_{ik}}/\overline{x_{kk}^2})x_{kk}. \tag{5}$$

(We shall omit the comma between two subscripts when each subscript is a single symbol.)

From (5) it follows that

$$\overline{x_{i,k+1}x_{j,k+1}}=\overline{x_{ik}x_{jk}}-\overline{x_{kk}x_{ik}}\;\overline{x_{kk}x_{jk}}/\overline{x_{kk}^2}. \tag{6}$$

Evaluation of successive values of $x_{i,k+1}$ from (5) requires evaluation of successive values $\overline{x_{i,k+1}x_{j,k+1}}$ from (6), but evaluation of the latter does not require evaluation of the former. Hence, if we have no immediate plans to use a prediction formula, and are interested only in how well it performs (i.e., if we care only about the values of $\overline{y_1^2}$, $\overline{y_2^2},\cdots$), repeated application of (6) is all that is needed.

Since x_0 is constant, y_1 is the departure of y from the sample mean, and $\overline{y_1^2}$ is the sample variance. The quantities $\overline{y_2^2},\cdots$ are successive residual variances, and the dimensionless quantity

$$\rho_k'=(\overline{y_1^2}-\overline{y_{k+1}^2})/\overline{y_1^2} \tag{7}$$

is often called the (sample) *reduction of variance* (using k predictors).

Because the dependent sample is necessarily of finite size, there may be within it a close resemblance between the predictand and some combination of the predictors which is not characteristic of the population. As a consequence, a strong physical relationship may be inferred when a weak one or none at all exists. A further consequence is that even if a physical relationship does exist, so that the desired formula (1) would perform acceptably when applied to new data, the formula (3) actually derived, being different, may perform poorly. The danger of inferring nonexistent relationships or putting worthless formulas to use may be reduced by choosing a second sample (the *independent* sample) of N'' values of y and the corresponding values of $x_1,\ldots,x_M$, and using this sample to test the formulas derived from the dependent sample.

For such a test we can compare the values of $\overline{\overline{y_1^2}}$, $\overline{\overline{y_2^2}}\ldots$, with $\overline{y_1^2}$, $\overline{y_2^2},\ldots$, where the double bar denotes an average over the independent sample. From (6) we find that

$$\overline{\overline{x_{i,k+1}x_{j,k+1}}}=\overline{\overline{x_{ik}x_{jk}}}-\overline{\overline{x_{kk}x_{ik}}}\;\overline{\overline{x_{kk}x_{jk}}}/\overline{\overline{x_{kk}^2}}$$
$$+(\overline{\overline{x_{kk}x_{ik}}}-\overline{x_{kk}x_{ik}}\;\overline{\overline{x_{kk}^2}}/\overline{x_{kk}^2})(\overline{\overline{x_{kk}x_{jk}}}-\overline{x_{kk}x_{jk}}\;\overline{\overline{x_{kk}^2}}/$$
$$\overline{x_{kk}^2})/\overline{\overline{x_{kk}^2}}. \tag{8}$$

The quantity $\overline{\overline{y_1^2}}$ is not strictly the variance of y within the independent sample, since y_1 is the departure of y from the mean of the *dependent* sample. We therefore prefer to call the quantity

$$\rho_k''=(\overline{\overline{y_1^2}}-\overline{\overline{y_{k+1}^2}})/\overline{\overline{y_1^2}}, \tag{9}$$

formulated analogously to ρ_k', the *reduction of error* (using k predictors).

Since the final term in (6) is always negative (or zero) when $i=j=M+1$, the dependent-sample error continually decreases (or does not increase) as k increases. However, Eq. (8), in addition to terms resembling those in (6), contains a final term which is always positive (or zero) when $i=j=M+1$. Hence the independent-sample error may actually increase.

In the Appendix we show that under suitable assumptions, when all M predictors are used

$$\langle\overline{e^2}\rangle=[(N'-M-1)/N']\langle\epsilon^2\rangle, \tag{10}$$

while, to a close approximation,

$$\langle\overline{\overline{e^2}}\rangle=[(N'-1)/(N'-M-2)]\langle\epsilon^2\rangle. \tag{11}$$

It is noteworthy that $\langle\epsilon^2\rangle$ is fairly well approximated by the geometric mean of $\langle\overline{e^2}\rangle$ and $\langle\overline{\overline{e^2}}\rangle$. The discrepancy between $\overline{e^2}$ and $\overline{\overline{e^2}}$ is likely to be inconsequential if M is small, but as M/N' approaches unity, the expected error tends toward zero within the dependent sample, while within the independent sample it tends toward infinity. It is therefore essential in practice to limit the number of predictors if the sample size cannot be indefinitely increased.

A special variant of the procedure of Eqs. (6) and (8) is popularly called the *screening procedure* (cf. Miller, 1962). Instead of introducing the successive predictors in their original order, we introduce at each step the predictor which yields the greatest additional reduction of variance. Effectively we renumber the predictors, letting x_k be the predictor which maximizes $\rho_k'-\rho_{k-1}'$. In the present study we shall insist that the first selected predictor still be x_0, so that $\overline{y_1^2}$ will still be the dependent-sample variance.

The screening procedure offers a method of reducing M/N' by reducing M; ideally we can terminate the selection when the additional reduction of variance is no greater than that expected by chance. The latter amount, however, is difficult to estimate, since a greater additional reduction of variance should result from introducing the best remaining predictor than from introducing a randomly chosen remaining predictor. We suggest circumventing this particular difficulty by terminating the procedure when there is *no* additional reduction of *error*.

3. The data

Our data were derived from synoptic analyses of the height of the 500 mb surface over the Northern Hemisphere, prepared twice daily by the National Meteorological Center (NMC) in Washington. Prior to the present study, Leith (1974) had analyzed the height fields for 10 years 1963–72 into series of spherical harmonics; these series took the form

$$z(\lambda,\phi,t)=\sum_{n=0}^{L}\sum_{m=0}^{n}[C_{m,n}(t)\cos m\lambda+S_{m,n}(t)\sin m\lambda]\times P_n^m(\sin\phi). \quad (12)$$

Here λ, ϕ and t are longitude, latitude and time, respectively, z is the 500 mb height, and P_n^m the associated Legendre function (or Legendre polynomial, if $m=0$) of degree n and order (or wavenumber) m, normalized so that its global mean square is unity. The series were truncated triangularly at $L=18$. Again we shall usually omit the comma between two subscripts when each is a single symbol or digit.

Since the original analyses actually terminated near 20°N, they were made hemispheric by linear interpolation to a constant value at the equator. The field of z over one hemisphere does not determine C_{mn} and S_{mn} uniquely, and the representations were made unique by assuming the Southern Hemisphere to be a mirror image of the Northern. This choice makes C_{mn} and S_{mn} vanish when $n-m$ is odd. The cofficients S_{0n} are undefined by (12), and may be equated to zero or disregarded altogether. Thus each height field is represented by 100 coefficients C_{mn} and 90 coefficients S_{mn}.

For the present study the data were augmented by 500 mb heights for the single month December 1962, similarly analyzed into spherical harmonics. From the full data set we then extracted 10 "winter seasons," each consisting of 100 successive days beginning 1 December. The 380 000 numbers consisting of the 190 values of C_{mn} and S_{mn} twice daily on each of the 1000 days constitute half of our data set.

The other half consists of numerically predicted values of C_{mn} and S_{mn}. From each set of simultaneous values of C_{mn} and S_{mn} we first estimated the values A_{mn} and B_{mn} in a spherical-harmonic analysis of the vorticity field

$$\nabla^2\psi(\lambda,\phi,t)=\sum_{n=0}^{L'}\sum_{m=0}^{n}[A_{mn}(t)\cos m\lambda+B_{mn}(t)\sin m\lambda]\times P_n^m(\sin\phi), \quad (13)$$

by means of a form of the geostrophic equation

$$g\nabla^2 z=2\Omega\nabla\cdot(\sin\phi\nabla\psi). \quad (14)$$

Here ψ is a streamfunction, and g and Ω are the acceleration of gravity and the earth's angular velocity.

We then made numerical predictions of A_{mn} and B_{mn}, 24 h in advance, basing the predictions on the barotropic vorticity equation

$$a^2\partial(\nabla^2\psi)/\partial t=-J(\psi,\nabla^2\psi)-2\Omega\partial\psi/\partial\lambda, \quad (15)$$

where a is the earth's radius and J denotes a Jacobian with respect to λ and $\sin\phi$. Finally we converted the prognostic values of A_{mn} and B_{mn} into prognostic values of C_{mn} and S_{mn} by inverting the procedure used to obtain A_{mn} and B_{mn} originally.

Care must be taken in truncating the series (13) to insure that the transformation (14) between height and vorticity is reversible. Since z is an even function of latitude and $\sin\phi$ is odd, $\nabla^2\psi$ must be odd, whence A_{mn} and B_{mn} will vanish if $n-m$ is even. To include the proper number of coefficients A_{mn} and B_{mn}, we must truncate at $L'=L+1$.

From standard formulas involving spherical harmonics (cf. Jahnke and Emde, 1945), we find that (14) transforms, with an exception to be noted, into

$$C_{mn}=-\gamma^{-1}(\alpha_{m,n+1}A_{m,n+1}+\alpha_{m,n}A_{m,n-1}), \quad (16)$$

$$S_{mn}=-\gamma^{-1}(\alpha_{m,n+1}B_{m,n+1}+\alpha_{m,n}B_{m,n-1}). \quad (17)$$

Here $\gamma=g/(2\Omega a^2)$, and

$$\alpha_{m,n}=n^{-2}(4n^2-1)^{-\frac{1}{2}}(n^2-m^2)^{\frac{1}{2}} \quad (18)$$

is defined for $n>0$ and $m\leqslant n$.

For $m>0$, (16) and (17) are easily reversed. Since $\alpha_{m,m=0}$, we find that

$$A_{m,m+1}=-\gamma C_{mm}/\alpha_{m,m+1}, \quad (19)$$

after which $A_{m,m+3},\ldots$ may be evaluated in succession, while $B_{m,m+1},\ldots$ are determined by analogous formulas.

For $m=0$, derivation of (16) with $n=0$, and hence of (19), would involve divisions by zero. Physically C_{00} represents the global average 500 mb height, while A_{01} represents the solid rotation component of motion, and there is no reason why these should be related geostrophically, nor for that matter, why C_{00} should be related to any feature of the vorticity field. We therefore obtain reversible formulas by setting $A_{0,L+1}=0$, and using (16) to evaluate $A_{0,L-1},\ldots$ in reverse order. The vorticity equation thus yields no prediction for C_{00}, and in our 24 h progs we have predicted C_{00} to retain its initial value.

There are a number of possible procedures for integrating the vorticity equation (15) in spherical-harmonic form. In our procedure we invert ∇^2 and perform the horizontal differentiation using standard spherical-harmonic formulas. For the multiplications we convert each spherical-harmonic series into Fourier series at each of L equally spaced latitudes in the Northern Hemisphere. We then multiply the Fourier series, truncate the products and convert back to spherical harmonics. Comparison of our procedure with other methods yielded results identical to six decimal places. With a higher horizontal resolution it would

have been more economical to convert the Fourier series to grid-point values, multiply the latter and convert back to Fourier series, than to multiply the Fourier series.

For advancing in time we use the four-cycle scheme introduced by Lorenz (1971), with a time increment $\delta t = 90$ min. Sample 24 h forecasts repeated with 45, 60 and 90 min increments yielded almost indistinguishable results. Completion of a 24 h forecast on the CDC 7600 computer at NCAR requires about 0.7 s.

There are three principal reasons why the numerically predicted values of C_{mn} and S_{mn} may differ considerably from reality. First, the initial height fields from which they are predicted have been subjected to much interpolation and extrapolation, and in regions where observations are sparse, considerable invention. Second, the geostrophic equation used to transform the heights to vorticities and the predicted vorticities to predicted heights does not hold perfectly. Finally, the barotropic vorticity equation is not the true governing equation. It is therefore hardly to be expected that the prognostic values of C_{mn} and S_{mn} will, without further modification, constitute good predictions. It is nevertheless hoped that they may contain some of the proper nonlinear combinations of the initial values of C_{mn} and S_{mn} to serve as useful predictors in a linear regression scheme.

4. Results

Let us call the 500 mb height maps at 0000 and 1200 GMT on any given day the *past map* and the *present map*, respectively, and let us call the 24 h numerical prognostic maps prepared from these maps the *past prog* and the *present prog*. Let us call the height map which follows the present map by 24 h the *future map*.

In the regression schemes which we shall examine, the predictands will always be the 190 variables C_{mn} and S_{mn}, divided by $2^{\frac{1}{2}}$ if $m>0$, taken from the future map. The division makes the sum of the squares of the predictands equal to the hemispheric mean-square height. The predictors will be values of C_{mn} and S_{mn} chosen from the present and past maps and the present and past progs. In addition, the "predictor" $x_0=1$ will be included.

In each scheme the values during the first seven winter seasons will be chosen as the dependent sample, while the last three seasons will comprise the independent sample. Since there are no data available to predict for the first day in each season, $N'=693$ for the dependent sample, while $N''=297$ for the independent sample. When comparing our results with Eqs. (10) and (11) we must recognize that 99 successive daily observations generally do not constitute 99 independent observations, whence, effectively, N' is less than 693.

TABLE 1. Hemispheric mean-square errors E' for dependent sample and E'' for independent sample in 24 h prediction of 500 mb height, using special prediction schemes. Units are m².

Scheme	Procedure	E'	E''
1	Climatology	10376	10080
2	Persistence	3908	3569
3	Numerical	4231	3879

TABLE 2. As in Table 1 except using schemes where predictors are chosen from present map. M is the maximum number of predictors (besides x_0) per predictand. Units are m².

Scheme	Procedure	M	E'	E''
4	All predictors	190	1435	3137
5	Screening	2	2530	2440
6	Screening	12	2277	2413

Each predictand is predicted individually by a linear regression formula. Expect when otherwise noted, the results presented for each prediction scheme consist of two numbers. The first of these is the sum over all predictands of the mean-square-prediction errors for the dependent sample; the second is the same sum for the independent sample. The computations are all based on Eqs. (6) and (8), and the units are m². The second number is the best available indicator of the relative usefulness of the corresponding prediction formula derived from our data set. According to (10) and (11), the mean of the two numbers is a possible indicator of the usefulness of a formula which could be derived by the same scheme if the data set were much larger.

Table 1 gives the results of a few basic schemes against which the remaining schemes may be compared. Scheme 1, often called "climatology," is a regression scheme where the only predictor is x_0. According to (3) and (4), the errors are those which would be made by predicting that each predictand will always assume its mean value over the dependent sample. The fact that in this scheme, and some of the following ones, the second number is actually smaller than the first seems to indicate only that the last three seasons were somewhat less variable than the first seven.

Schemes 2 and 3 use no regression at all. The future map is simply predicted to be the present map in Scheme 2, and the present prog in Scheme 3. We see that persistence gives much better predictions than climatology—a result which is typical for prediction of large-scale features at a 24 h range. Perhaps surprisingly, the numerical prog, used as a prediction, is inferior to persistence.

Table 2 deals with a few regression schemes where the predictors are restricted to the present map (and x_0). In Scheme 4, all 190 predictors are used for each predictand. There is considerable improvement over climatology and persistence, but the discrepancy between the samples is striking and, in view of the large value of M, should be expected.

In Fig. 1 the upper curves show the successive mean-square-prediction errors for the two samples as the predictors are introduced one by one. Here the predictors have been arranged in order, beginning with C_{00}, so that C_{mn} precedes S_{mn}, C_{mn} and S_{mn} precede $C_{m+2,n}$, and C_{mn} and S_{mn} precede $C_{0,n+1}$ or $C_{1,n+1}$. The necessary continual decrease in dependent-sample error is accompanied by an increase in independent-sample error after about 90 predictors are introduced. According to (10) and (11), the observed spreading of the curves is consistent with an approximate ratio $M/N'=0.35$; the 693 observations in the dependent sample thus appear to constitute somewhat more than 500 independent observations.

The appearance of as many as 90 useful predictors in the scheme may seem surprising, and it actually results from the summing of errors. For a single predictand only a few predictors appear to be useful, but different predictors are useful for different predictands. In the lower curves in Fig. 1 the errors are summed only over the predictands C_{26} and S_{26}. Again the discrepancy between the samples shows a general growth, but the errors in both samples drop suddenly when certain key predictors are introduced, notably C_{24}, S_{24}, C_{26} and S_{26}.

These observations strongly suggest that the screening procedure, applied separately to each predictand, should yield favorable results. In Scheme 5 the screening procedure has been used, and has been terminated after selecting two predictors (besides x_0). Scheme 6 is similar except that 12 predictors are chosen; this choice seems to maximize the reduction of error. It is evident that screening has yielded far more useful results than routine inclusion of all the predictors. Continuing the procedure beyond two predictors seems to yield genuine but rather small improvement.

The success of the screening procedure in this case is apparently due to the presence of a few very useful predictors, together with many virtually useless ones. Some of the latter may appear to be good within the dependent sample, but still not as good as the genuinely good ones, so that the procedure will first select predictors which are genuinely good even if not genuinely the best. In a data set where the useful predictors are more numerous but make smaller individual contributions, the screening procedure would be more likely to choose a poor predictor. Some selection method based on physical considerations would then be in order. Even in the present instance, physical considerations not be disregarded.

If the earth had no geographical irregularities such as continents and oceans, the probability of occurrence of a particular sequence of weather situations would presumably equal the probability of the same sequence displaced through any given longitudinal angle. It would then follow that the separate variables C_{mn} and S_{mn} would be uncorrelated with one another at any time lag, except for variables having the same wavenumber index m. It seems likely that in the real world the correlations between variables with different wavenumber indices may still be weak. Good results may therefore be anticipated when the predictors are restricted to those having the same value of m as the corresponding predictand. It is noteworthy that the useful predictors found for C_{26} and S_{26} all had wavenumber index 2.

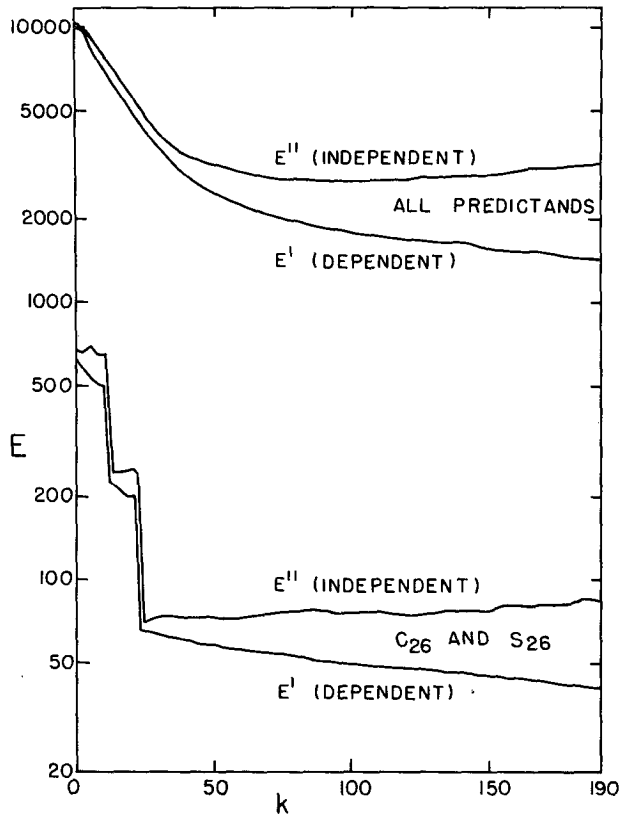

FIG. 1. Upper curves: hemispheric mean-square errors E' for dependent sample and E'' for independent sample in 24 h prediction of 500 mb height, as represented by 190 spherical-harmonic coefficients, using k spherical harmonic coefficients chosen from present map as predictors, as k increases from 0 to 190. Lower curves: similar mean-square errors E' and E'' in predicting coefficients C_{26} and S_{26} in spherical-harmonic representation of 500 mb height. Units are m^2.

In the schemes in Table 3 the suggested restriction on m is imposed. Scheme 7 uses all appropriate predictors for each predictand; the number varies from 18, when $m=1$ or 2, to 2, when $m=17$ or 18. The results are every bit as good as those obtained by screening among all predictors.

Schemes 8 and 9 show the results of screening among those predictors with the appropriate value of m, and terminating after two predictors, or after 12 (where 12 are available). These schemes are as good as Schemes 5 and 6, where all predictors are screened, but show no obvious advantage over Scheme 7. Apparently the restriction on m has already reduced M/N' sufficiently.

TABLE 3. Hemispheric mean-square errors E' for dependent sample and E'' for independent sample in 24 h prediction of 500 mb height, using schemes where predictors have same wavenumber index as predictands and are chosen from present map. M is the maximum number of predictors (besides x_0) per predictand, n the degree index of predictor and n' the degree index of predictand. Units are m².

Scheme	Procedure	M	E'	E''
7	All predictors	18	2410	2407
8	Screening	2	2544	2437
9	Screening	12	2418	2402
10	$n=n'$	2	2571	2446
11	$n=n'$ or $n'-2$	4	2516	2416

In Scheme 10 the predictors are restricted to have the same wavenumber index *m and* degree index n as the corresponding predictand, so that there are only two predictors (or one, if $m=0$) for each predictand. The results are comparable to Scheme 8 which screens for two predictors. Somewhat weaker restrictions on n might produce superior results; as one possibility in Scheme 11, we have allowed the degree index of the predictor to equal that of the predictand, or to be 2 less. The results compare well with Schemes 7 and 9.

Although, among Schemes 5–11, some may be genuinely superior to others, the principal conclusion is that all give comparable results for the independent sample. It seems unlikely that any other scheme using predictors taken from the present map will perform appreciably better.

Having established these initial results, we proceed to the principal topic of this investigation—the usefulness of nonlinearity in a statistical weather forecasting scheme. It is utterly impractical to introduce all quadratic functions of the original predictors as additional predictors, let alone higher degree polynomials or transcendental functions, even if screening is subsequently to be used. Instead, as already indicated, we introduce special nonlinear functions by using the coefficients C_{mn} and S_{mn} from the nonlinearly produced progs as predictors in a linear scheme. With both the present map and the present prog available together as predictor maps, it is computationally inconvenient to use all 380 available predictors or to screen from all of them. In view of our initial results, we have confined our attention to prediction schemes where each predictor has the same wavenumber index as the corresponding predictand.

TABLE 5. As in Table 3 except that predictors are chosen from present and past maps.

Scheme	Procedure	M	E'	E''
17	All predictors	36	1870	1879
18	Screening	4	2038	1902
19	Screening	12	1922	1871
20	$n=n'$	4	2051	1894
21	$n=n'$ or $n'-2$	8	2007	1878

Tables 4–6 are similar in format to Table 3, and differ only in the allowable predictor maps. In the first scheme in each table, all predictors having the same wavenumber as the predictand are used. The second scheme uses screening, selecting twice as many predictors as the number of predictor maps, while the third uses screening and chooses 12 predictors. The fourth scheme uses all allowable predictors with the same degree index as the predictand, while the fifth allows the degree index of the predictor to equal that of the predictand or fall short by 2.

In Table 4 the predictor maps are the present map and the present prog. There is some variation among the results of the different schemes, but as a group they are in a completely different class from those appearing in Table 3; it is inconceivable that some combination of predictors chosen from the present map alone, which we may have overlooked, could yield mean-square errors for the independent sample as small as those appearing in Table 4. We conclude not only that the nonlinear terms are important in short-range forecasting, but also that suitably devised empirical schemes can capture the nonlinear effects.

Unlike numerical forecasting, the most effective statistical weather forecasting schemes generally include both past and present observations as predictors. Since the past is presumably nonlinearly related to the present, use of the past and present should capture some of the physical nonlinearity even in a linear scheme. This observation seems to be partly responsible for the belief that explicit nonlinearity is not needed.

Table 5 is like Table 4 except that the predictor maps are the present and past maps; the present prog is

TABLE 4. As in Table 3 except that predictors are chosen from present map and present prog.

Scheme	Procedure	M	E'	E''
12	All predictors	36	1421	1402
13	Screening	4	1563	1421
14	Screening	12	1462	1398
15	$n=n'$	4	1603	1435
16	$n=n'$ or $n'-2$	8	1551	1398

TABLE 6. As in Table 3, except that predictors are chosen from present and past maps and present prog.

Scheme	Procedure	M	E'	E''
22	All predictors	54	1278	1291
23	Screening	6	1427	1313
24	Screening	12	1356	1284
25	$n=n'$	6	1461	1298
26	$n=n'$ or $n'-2$	12	1417	1280

not used. We see that using the present and past maps is indeed an improvement over using the present map alone, but that the past map is no substitute for the present prog. The mean-square errors in Table 5 are fairly uniform, and again it does not appear that any overlooked combination of predictors from the present and past maps could have yielded errors as small as those in Table 4.

Table 6, where the predictor maps are the present and past maps and the present prog, shows improvements over Tables 4 and 5, and thus reveals that the useful information contained in the past map is different from that in the present prog. The mean-square errors are uniformly low, and correspond to reductions of both variance and error of about 87%. Our conclusion that nonlinear effects can profitably be incorporated into an empirical scheme is reinforced.

5. Further investigations

In the results so far presented, all of the predictors have been chosen from the present map, the present (numerical) prog and the past (12 h old) map. It seems reasonable that appreciably greater reductions of error might be obtained by choosing some of the predictors from additional maps. If this is so, the appropriate maps have escaped our discovery, although minuscule improvements are easily found.

A natural possibility would be to use the past prog in addition to the present prog and the present and past maps. Using predictors with the same indices as the predictand, this plan succeeded in reducing the mean-square errors of 1461 and 1298 m², obtained in Scheme 25, only to 1444 and 1280 m². Another possibility would be to use the present and two past (12 and 24 h) maps together with the present prog (reducing the usable sample size from 99 to 98 days per season). This choice replaced the values in Scheme 25 only by 1450 and 1298 m².

The 100-day period beginning 1 December is not uniformly wintry, and it might appear profitable to remove the seasonal trend from the data before making the forecasts. The trend during the winter alone is probably fairly well approximated by a quadratic function of time, and is therefore easily removed by introducing into each regression scheme two additional predictors $x_0'=t$ and $x_0''=t^2$, where t is the elapsed time since the most recent 1 December. Table 7 compares Schemes 1, 10, 15 and 25, as presented in Tables 1, 3, 4 and 6, with similar schemes where the seasonal trend has been removed. Scheme 1a rather than Scheme 1 might properly have been called "climatology." For the dependent sample the improvement yielded by x_0' and x_0'' is small; for the independent sample it is nonexistent. Apparently much of the seasonal trend is captured without x_0' and x_0'' by virtue of being contained in other predictions; what remains is weak,

TABLE 7. Hemispheric mean-square errors E' for dependent sample and E'' for independent sample, in 24 h prediction of 500 mb height, using Schemes 1, 10, 15, 25, as presented in Tables 1, 3, 4, 6, and similar Schemes 1a, 10a, 15a, 25a where seasonal trend has been removed. Units are m².

Scheme	Predictors	E'	E''
1	None (climatology)	10376	10080
1a	None (climatology)	9949	10004
10	Present map	2571	2446
10a	Present map	2557	2459
15	Present map, present prog	1603	1435
15a	Present map, present prog	1590	1443
25	Present map, present prog, past map	1461	1298
25a	Present map, present prog, past map	1450	1303

and does not behave in the last three years as in the first seven.

The fact that the numerical prog, used as a prediction instead of a predictor, performs more poorly than persistence suggests that a better numerical forecasting equation could have been chosen in the first place. One such equation is the barotropic vorticity equation with the so-called lambda correction (cf. Bolin, 1955); this correction takes into partial account the influence of divergence, and tends to slow down the numerically predicted but unobserved rapid westward progression of the longest waves.

In spherical coordinates the new equation may be derived from the vorticity equation

$$a^2\partial(\nabla^2\psi)/\partial t = -J(\psi,\nabla^2\psi) - 2\Omega\partial\psi/\partial\lambda - 2\Omega\nabla\cdot(\sin\phi\nabla\chi), \quad (20)$$

the vertically integrated mass-continuity equation

$$\partial z_0/\partial t = -H\nabla^2\chi, \quad (21)$$

and Eq. (14), where χ is a divergent-velocity potential, H an atmospheric scale height and z_0 the height of an isobaric surface near the earth's surface, say 1000 mb. Assuming a linear relation between z_0 and z, the equation becomes

$$\partial[a^2\nabla^2\psi - \Lambda^2\nabla\cdot\sin\phi\nabla(\nabla^{-4}\nabla\cdot\sin\phi\nabla\psi)]/\partial t = -J(\psi,\nabla^2\psi) - 2\Omega\partial\psi/\partial\lambda, \quad (22)$$

where Λ is a dimensionless constant whose most suitable value is near 10, and ∇^{-4} is the inverse of ∇^4. Since the nonlinear terms in (22) are the same as those in (15), the numerical solution in terms of spherical harmonics requires only slightly more computing time. From the second map on each day of the 10 winter seasons, we have produced an alternative numerical prog, using the lambda correction with $\Lambda=10$.

With the new numerical prog as a prediction instead of a predictor, we obtain mean-square errors of 3129 and 2756 m², as opposed to 4230 and 3879 m² obtained with the old numerical prog. Despite this striking improvement, the new prog performs very little better

than the old one when used as a predictor. Thus, for example, the values 1461 and 1298 m² in Scheme 25, obtained when the predictors are the present and past maps and the past prog, are reduced only to 1444 and 1291 m² when the lambda correction is introduced. Evidently the errors which are partially removed by the lambda correction, such as errors in longitudinal phase, can be almost equally well removed by regression, after the prog has been produced.

Our principal results may seem to conflict with those of Cooley (1958) who found the nonlinear terms to be of but minor value. We should recall, then, that Cooley's results refer to the nonlinear terms in the instantaneous time derivative, rather than the nonlinear influence upon the change over a finite time interval. It is of interest to see, in the case of our data set, how much is contributed by the first time derivative, as given by (12), and also how much more is contributed by a few of the higher time derivatives.

A convenient way to determine the derivatives is to perform several successive, uncentered, forward time steps. If the initial map is called map 0, and the maps at the ends of the successive steps are called maps 1, 2, ..., the first derivative is a linear combination of maps 0 and 1, while if the time step is not too large, the second derivative is closely approximated by a linear combination of maps 0, 1 and 2, etc.

Accordingly, using in turn the second map on each day as map 0, we have computed maps 1–4 with a 90 min time step. Table 8 shows the result of using successive maps, and hence successive derivatives, as predictor maps, and letting the predictors have the same indices as the predictand. We see that although maps 0 and 1, and hence the present map and the first time derivative, do not perform as well as the present map and the present prog (Scheme 15), the time derivative is much more useful than in the case of Cooley's data. When map 2 is added, virtually all the useful information in the present prog seems to be captured, while the use of three or more time derivatives gives superior results. Use of maps 0–4 and the past map yields errors of 1411 and 1281 m², as compared with 1461 and 1298 m² for Scheme 25.

TABLE 8. Hemispheric mean-square errors E' for dependent sample and E'' for independent sample, in 24 h prediction of 500 mb height, using schemes where predictors have same wavenumber and degree indices as predictands and are chosen from successive maps produced by uncentered forward time steps. Units are m².

Predictor maps	E'	E''
0 (Scheme 10)	2571	2446
0, 1	1740	1623
0, 1, 2	1566	1428
0, 1, 2, 3	1519	1387
0, 1, 2, 3, 4	1505	1381
0, present prog (Scheme 15)	1603	1435

Although a scheme with three time derivatives uses more predictors, the numerical integration needed to obtain the time derivatives, i.e., three time steps, is small compared to the 16 steps used to obtain the 24 h prog. As an operational procedure, with a more highly refined model than the barotropic vorticity equation, use of time derivatives rather than a numerical prog would be far more economical, unless, as is the case in most of today's operations, the numerical prog is to be produced in any case.

There is one more matter which requires careful consideration. We have noted that the maps used to prepare the progs have been subjected to much interpolation and extrapolation. These same maps are used to verify the forecasting schemes.

In analyzing a map subjectively, a forecaster will naturally base his analysis in a region of sparse observations upon what he thinks is happening there, and when evidence to the contrary is absent, this is likely to be his earlier forecast for the region. This practice has been carried over into operational objective analysis procedures so that, where observations are not plentiful, the analysis will be based upon a combination of observations and earlier forecasts. We must therefore consider the possibility that the nonlinear terms have proven useful in our study not because of any true nonlinear behavior but because the verification map analyses are influenced by nonlinear forecasts. Certainly the NMC analyses are not based upon forecasts made with the barotropic vorticity equation, but some of the important nonlinear terms in the operational model are similar to those in the barotropic equation.

A sure test of this possibility would require a reanalysis of the 500 mb charts by an objective scheme not incorporating any forecasts. Such a task is beyond the scope of this study. Meanwhile, we can compare the performance of our linear and nonlinear schemes in local regions where the data are plentiful, and where the analyses are presumably fairly reliable. If the nonlinear schemes perform noticeably better than the linear schemes there, our conclusion that the nonlinear terms are useful should be on firmer ground.

We have therefore selected 10 grid points in middle latitudes for local verification of the schemes. These are listed in Table 9. Most of the points near cities are in regions of dense observations; the oceanic points, particularly those in the Pacific, are in regions of sparser observations. As an auxiliary data set, we have obtained the original (i.e., prior to the spherical-harmonic analysis) 500 mb height at each selected grid point, for the second observation on each day of the 10 winter seasons.

In our first test, we predict the values of C_{mn} and S_{mn}, using various schemes where the predictors have the same indices as the predictand. By suitably combining the predicted values we obtain predicted heights

TABLE 9. Grid points used for local verification of statistical forecasts. A negative longitude is west.

Point	Latitude (°N)	Longitude (°E)	Nearby city or general location
1	50	0	Le Havre, France
2	55	40	Moscow, U.S.S.R.
3	30	115	Hankow, China
4	35	140	Tokyo, Japan
5	40	170	W. Pacific
6	45	−150	E. Pacific
7	40	−105	Denver, U.S.A.
8	35	−80	Charlotte, U.S.A.
9	30	−55	W. Atlantic
10	50	−30	E. Atlantic

at the selected grid points. We use the auxiliary data set for verification only.

Table 10 shows the results of using Scheme 1 (climatology) and Schemes 20 and 25, where the present map and past map and then the present prog are introduced as prediction maps. Only at the points in China and Japan does the present prog fail to contribute strongly to the prediction. The greatest contribution is in the eastern Atlantic, where the data may be suspect; nevertheless, in the United States and particularly in Europe the improvement yielded by the nonlinear terms is comparable to that for the hemisphere as a whole. We conclude tentatively that the predictive capability of the nonlinear terms is real.

As a second test, we attempt to predict the grid-point heights directly from the values of C_{mn} and S_{mn}, without the intermediate step of predicting future values of C_{mn} and S_{mn}. Since there is no value of m or n associated with an individual grid point, we face the possibility of many useful predictors, and with a data set of this size, must limit our choice by some procedure such as screening.

One might expect this test to yield slightly greater reductions of error than the previous one, since it uses the auxiliary data set both in establishing the formulas and in verifying them, while the former test effectively uses grid-point heights which are reconstructed from the spherical-harmonic analysis in establishing the formulas. This expectation would indeed be realized if all available predictors were used in both tests. The tests would then, in fact, yield identical results if the grid-point heights could be perfectly reconstructed. However, the former test actually restricts the predictors to those having the same indices as the predicted coefficients, while the latter restricts the predictors by screening.

Table 11 shows the results of predicting the heights at grid-point 7, in the western United States, using two schemes, one of which screens from the present and past maps while the other screens from the present and past maps and the present prog. We find that in either scheme the reduction of error tends to level off

TABLE 10. Mean-square errors E' for dependent sample and E'' for independent sample at selected grid points, in 24 h prediction of 500 mb height, using schemes where predictors have same wavenumber and degree indices as predictands, and are chosen from indicated maps. Comparison values for hemisphere are from Tables 1, 20, 25. Units are m².

	Predictor maps					
	None (climatology)		Present map past map		Present map past map past prog	
Point	E'	E''	E'	E''	E'	E''
1	21595	24164	4212	4315	2752	2808
2	22758	19566	4131	3983	2606	2602
3	2895	2159	816	543	779	547
4	12714	11594	2011	2169	1814	1954
5	24061	18660	5833	3891	4043	2615
6	31002	35578	8204	8551	6401	6590
7	11823	12136	2381	2335	1761	1715
8	14548	14288	2915	2714	2247	1934
9	7122	4307	1715	1301	1398	1022
10	39039	37529	6855	5796	4406	3480
Hemisphere	10376	10080	2051	1894	1461	1298

after about 30 predictors are selected (a much larger number than when the predictands are C_{mn} and S_{mn}), while the proportional gain from including the present prog among the predictors is as great as in the first test. There is therefore no reason for abandoning our conclusion that the nonlinear terms are useful. It is of interest that in both schemes the first four selected predictors are C_{08}, S_{17}, $C_{2,12}$ and C_{39}, but, while all of these are selected from the present map in the first

TABLE 11. Mean-square errors E' for dependent sample and E'', for independent sample, at grid-point 7 (40°N, 105°W), in 24 h prediction of 500 mb height, using schemes where predictors are screened from indicated predictor maps. Comparison values from predictions of spherical-harmonic coefficients are from Table 10. M is the number of predictors selected. Units are m².

	Predictor maps			
	Present map past map		Present map past map present prog	
M	E'	E''	E'	E''
0	11823	12136	11823	12136
2	7992	8482	7636	8007
4	6299	7004	5666	5727
6	5571	6945	4893	4929
8	4906	6559	4286	4689
10	4403	6192	3883	4220
12	4024	5482	3511	4048
14	3737	5170	3104	3872
16	3528	4865	2840	3816
18	3328	4654	2666	3703
20	3177	4422	2533	3593
22	3045	4223	2422	3422
24	2934	4378	2306	3218
26	2830	4295	2200	3222
28	2742	4151	2105	3082
30	2659	4184	2030	3119
Coeff.	2381	2335	1761	1715

scheme, S_{17} and C_{39} are selected from the present prog in the second.

Further examination reveals that the errors level off at values far above those obtained in the first test. Evidently we have attempted to screen from a data set where each of a large number of useful predictors individually contributes a relatively small amount, while in the first test each of a small number of useful predictors contributes a large amount. In the present test there is a greater probability that a poor or mediocre predictor will, within the dependent sample, appear better than a good predictor and will be chosen in its stead. We note also, by comparison with Fig. 1, that the discrepancy between the samples when 30 predictors have been chosen by screening is as great as when 100 predictors are chosen essentially at random.

We are forced to conclude that within a data set of this size (and by usual standards our set is not small), it is not possible to deduce the best obtainable formulas for predicting future grid-point heights from present spherical-harmonic coefficients, without performing some intermediate step. In our test this step has consisted of predicting future spherical-harmonic coefficients. Perhaps equally good predictions could have been made if it had consisted of reconstructing present heights at judiciously chosen grid points to use as predictors.

6. Conclusions

Our experiment has yielded positive results. Without question, the linear regression formulas for predicting tomorrow's 500 mb heights, derived from a portion of our data, and using today's heights and suitably chosen nonlinear functions of today's heights as predictors, perform far better, when applied to the remainder of our data, than do similar formulas where the nonlinear functions are not included as predictors. The improvement yielded by the nonlinear functions may have been exaggerated by the map-analysis procedure but it nonetheless appears genuine.

Although the usefulness of the progs as predictors in our experiment is consistent with the positive skill (see, Klein and Glahn, 1974) of the MOS forecasts, it was not on that account assured in advance. The MOS method relies upon the screening procedure. As we have just seen, if a set of predictors is replaced by another set which includes some which are *individually* more closely related to the predictand, the screening procedure can yield better results, even if the new predictors are linear functions of the old ones. We cannot say without a further test whether the power of the MOS technique results mainly from capturing the nonlinearity or mainly from rendering the screening procedure more effective. Our own results are not so closely tied to the screening procedure; the superiority of Scheme 25 over Scheme 20, for example, where the screening procedure is not involved, is not noticeably different from the superiority of Scheme 24 over Scheme 19.

It is tempting at this point to maintain that our prediction errors are so small that our procedure, based upon a crude barotropic vorticity equation, performs comparably to present-day operational procedures. A word of caution is therefore needed. Our mean-square errors are those in predicting only that part of the height field which is represented by the spherical-harmonic analysis, truncated triangularly at wavenumber 18. The portion of the field which is not resolved by the spherical harmonics is not predicted at all; in a fair comparison, we should add the total variance of this portion to the mean-square errors which we have tabulated. We should also note that since the circulation is continually undergoing extended-period fluctuations, it is dangerous to compare a numerical mean-square error made by one procedure during one period with one made by another procedure during a different period.

Beyond our principal result, we have uncovered evidence suggesting that if a regression scheme using numerical progs as predictors is to be made operational, it might be profitable to use progs made for several times which closely follow the present, thereby eliminating much of the computing otherwise needed to produce the progs. Finally, our computations point toward some conclusions which might have been anticipated. First, if it is certain that the number of predictors is small compared to the number of *independent* observations of each, an independent sample becomes less essential. Second, the screening procedure is likely to prove most satisfactory when just a few of the predictors being screened can yield a nearly optimum prediction, rather than when the optimum formula contains many predictors, each making a small individual contribution. Selection of the predictors on physical grounds, when this can be done, is to be preferred to screening.

Acknowledgments. The writer has benefited greatly from frequent discussions of the various aspects of this study with Dr. C. E. Leith. Thanks are also due to Dr. Leith and Mr. R. L. Jenne for their aid in securing the necessary data.

APPENDIX A

Expected Mean-Square Errors

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be matrices of N' and N'' rows and $M+1$ columns whose elements are the values of the predictors (including x_0) in the dependent and independent samples, respectively, and let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be matrices of N' and N'' rows and one column whose elements are the values of the predictand in these samples. Let $\mathbf{A}_1$ be a matrix of $M+1$ rows and one column whose elements are the prediction coefficients, determined from the dependent sample, and let $\mathbf{E}_1$

and $\mathbf{E}_2$ be matrices similar to $\mathbf{Y}_1$ and $\mathbf{Y}_2$ whose elements are the prediction errors when the formula determined from the dependent sample is applied respectively to the dependent and independent samples. The mean-square prediction errors are then given by

$$N_1'\overline{e^2}=\mathrm{tr}(\mathbf{E}_1^{\mathrm{T}}\mathbf{E}_1), \tag{A1}$$

$$N_2'\overline{\overline{e^2}}=\mathrm{tr}(\mathbf{E}_2^{\mathrm{T}}\mathbf{E}_2), \tag{A2}$$

where the superscript T denotes a transpose and tr denotes a trace or diagonal sum. Eq. (3), applied to both samples, becomes

$$\mathbf{Y}_1=\mathbf{X}_1\mathbf{A}_1+\mathbf{E}_1, \tag{A3}$$

$$\mathbf{Y}_2=\mathbf{X}_2\mathbf{A}_1+\mathbf{E}_2, \tag{A4}$$

while (4) becomes

$$\mathbf{A}_1=(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{Y}_1. \tag{A5}$$

We consider first the special case where there is no true relation between the predictand and the predictors. This would occur, for example, if the data had been chosen from a set of random numbers. Without loss of generality we may let $\langle y\rangle=0$, using the angle braces as in Section 2 to denote an expected value. Noting first that the order of the operations $\langle\ \rangle$ and tr is interchangeable, second that the trace of the product of two or more matrices is unaltered by moving the first factor in the product to the last position, and third that the expected value of the product of unrelated quantities equals the product of the expected values, we find from (A1)–(A5) that

$$N'\langle\overline{e^2}\rangle=\mathrm{tr}\langle\mathbf{Y}_1^{\mathrm{T}}\mathbf{Y}_1\rangle - \mathrm{tr}[\langle\mathbf{X}_1(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\rangle\langle\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}}\rangle], \tag{A6}$$

$$N''\langle\overline{\overline{e^2}}\rangle=\mathrm{tr}\langle\mathbf{Y}_2^{\mathrm{T}}\mathbf{Y}_2\rangle - \mathrm{tr}[\langle\mathbf{X}_1(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\rangle\langle\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}}\rangle]. \tag{A7}$$

If the separate members of the dependent sample and also those of the independent sample are independently chosen,

$$\langle\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}}\rangle=\langle y^2\rangle\mathbf{I}_{N'}, \tag{A8}$$

$$\langle\mathbf{Y}_2\mathbf{Y}_2^{\mathrm{T}}\rangle=\langle y^2\rangle\mathbf{I}_{N''}, \tag{A9}$$

where the subscripts denote the order of the identity matrix, whence

$$\langle\overline{e^2}\rangle/\langle y^2\rangle=(N'-M-1)/N', \tag{A10}$$

$$\langle\overline{\overline{e^2}}\rangle/\langle y^2\rangle=[N'+q_{N',M+1}(M+1)]/N', \tag{A11}$$

where $q_{N',M+1}$ is defined by

$$\langle(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\rangle\langle\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\rangle=q_{N',M+1}\mathbf{I}_{M+1}. \tag{A12}$$

Here we have substituted $\langle\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\rangle$ for $(N'/N'')\langle\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\rangle$, after which the size N'' of the independent sample completely cancels out.

The expected value of an inverse may be crudely approximated by the inverse of the expected value, in which case $q_{N',M+1}$ in (A11) becomes unity. In certain instances this approximation leads to a serious underestimate of $\langle e^2\rangle$. The proper value of $q_{N',M+1}$ depends upon the joint probability distribution of the predictors. In Appendix B we show that if this distribution is Gaussian,

$$q_{N',M+1}=N'/(N'-M-2), \tag{A13}$$

whence

$$\langle\overline{\overline{e^2}}\rangle/\langle y^2\rangle=(N'-1)/(N'-M-2). \tag{A14}$$

We consider (A14) preferable to (A11) (with $q_{N',M+1}=1$) even when the predictors do not have a Gaussian distribution. In any event, the formulas differ appreciably only when M/N' is not small, when prediction should not be attempted anyway. Actually, the factor $N'-1$ in (A14) should be replaced by the inconsequentially smaller factor $N'-1-(2/N')$, because x_0, being constant, cannot also be Gaussian.

For the general case we replace the assumption of no true relation between y and the predictors by the less restrictive assumption of no true relation between ϵ and the predictors. We replace Eq. (3) by

$$\epsilon=\sum_{i=0}^{M}(a_i-\alpha_i)x_i+e, \tag{A15}$$

derivable from (1) and (3). The derivation of (A11) and (A14) may then be repeated step by step, replacing y and α_i, or the matrices which represent them, by ϵ and $a_i-\alpha_i$. The result is Eqs. (10) and (11).

APPENDIX B

Expected Inverses of Covariance Matrices

Let $n>m+1$, and let $\mathbf{X}$ be an n-row, m-column matrix whose elements are chosen randomly and independently from a Gaussian distribution with mean 0 and variance 1. Let $\mathbf{C}=\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Obviously $\langle\mathbf{C}\rangle=n\mathbf{I}_m$, while $\langle\mathbf{C}^{-1}\rangle=r_{n,m}\mathbf{I}_m$ for some scalar $r_{n,m}$ which we wish to determine.

For $m=1$ the single element of $\langle\mathbf{C}^{-1}\rangle$ is the expected reciprocal of a quantity chosen randomly from a chi-square distribution with n degrees of freedom, i.e.,

$$r_{n,1}=1/(n-2). \tag{B1}$$

For $m>1$ it is sufficient to determine the expected value of a single diagonal element, say,

$$(\mathbf{C}^{-1})_{11}=\begin{vmatrix}C_{22} & \dots & C_{2m}\\ C_{m2} & \dots & C_{mm}\end{vmatrix}\Big/\begin{vmatrix}C_{11} & \dots & C_{1m}\\ C_{m1} & \dots & C_{mm}\end{vmatrix}. \tag{B2}$$

Let $\mathbf{H}$ be an $n\times n$ orthonormal matrix, i.e.,

$$\mathbf{H}\mathbf{H}^{\mathrm{T}}=\mathbf{H}^{\mathrm{T}}\mathbf{H}=\mathbf{I}_n, \tag{B3}$$

with the last (nth) column of $\mathbf{H}$ proportional to the last (mth) column of $\mathbf{X}$. The last column of $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is

then proportional to the last column of $\mathbf{H}^T\mathbf{X}$, so that

$$\mathbf{H}^T\mathbf{X}=\mathbf{Y}+\mathbf{Z}, \tag{B4}$$

where $\mathbf{Y}$ contains only zeros except in its last row and $\mathbf{Z}$ contains only zeros in its last row and its last column. Thus

$$\mathbf{C}=(\mathbf{X}^T\mathbf{H})(\mathbf{H}^T\mathbf{X})=\mathbf{A}+\mathbf{B}, \tag{B5}$$

where $\mathbf{A}=\mathbf{Y}^T\mathbf{Y}$ is of rank 1 and $\mathbf{B}=\mathbf{Z}^T\mathbf{Z}$ contains only zeros in its last row and its last column. Hence for $i=1$ or 2, if $l=m-1$,

$$\begin{vmatrix} C_{ii} & \dots & C_{im} \\ C_{mi} & \dots & C_{mm} \end{vmatrix} = A_{mm} \begin{vmatrix} B_{ii} & \dots & B_{il} \\ B_{li} & \dots & B_{ll} \end{vmatrix}. \tag{B6}$$

Because $\mathbf{H}$ is orthonormal the non-zero elements of $\mathbf{Z}$ have the same statistical properties as all of the elements of $\mathbf{X}$. It follows from (B2) and (B6) that the expected value of $(\mathbf{C}^{-1})_{11}$ is the same as if $\mathbf{X}$ had had $n-1$ rows and $m-1$ columns, i.e.,

$$r_{n,m}=r_{n-1,m-1}. \tag{B7}$$

Repeated application of (B7) shows that

$$r_{n,m}=1/(n-m-1). \tag{B8}$$

In the application to Appendix A, $n=N'$, $m=M+1$, $n/r_{n,m}=q_{N',M+1}$, and $\mathbf{X}=\mathbf{X}_1$. If the predictors are originally correlated, they may be replaced at the start by uncorrelated linear combinations without altering the mean-square errors.

REFERENCES

Bolin, B., 1955: Numerical forecasting with the barotropic model. *Tellus*, **7**, 27–49.

Cooley, D. S., 1958: Statistical forecasting operators based on dynamical equations. *Tellus*, **10**, 331–341

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

Jahnke, E., and F. Emde, 1945: *Tables of Functions*. Dover, 382 pp.

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227.

Leith, C. E., 1974: Spectral statistical-dynamical forecast experiments. *Proc. Intern. Symp. Spectral Methods in Numerical Weather Prediction*, GARP Programme on Numerical Experimentation, Rep. No. 7.

Lorenz, E. N., 1971: An N-cycle time-differencing scheme for stepwise numerical integration. *Mon. Wea. Rev.*, **99**, 644–648.

——, 1973: Predictability and periodicity: A review and an extension. *Preprints 3rd Conf. Probability and Statistics in Meteorology*, Boulder, Colo., Amer. Meteor. Soc., 1–4.

Miller, R. G., 1962: *Statistical Prediction by Discriminant Analysis. Meteor. Monog.*, No. 25, Amer. Meteor. Soc., 54 pp. (see Appendix).

Wiener, N., 1956: Nonlinear prediction and dynamics. *Proc. Third Berkeley Symp. Mathematical Statistics and Probability*, Vol. 3, University of California Press, 247–252.