Crop–Climate Modeling Using Spatial Patterns of Yield and Climate. Part 1: Background and an Example from Australia

T. M. L. WIGLEY AND TU QIPU[1]

Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, U.K.

(Manuscript received 11 January 1983, in final form 20 July 1983)

ABSTRACT

A new technique in statistical crop-climate analysis, the direct linking of spatial patterns of crop yield and spatial patterns of climate, is explored. Yield and climate data from networks of crop reporting districts and meteorological stations are decomposed into orthogonal components using principal components analysis. Each yield component is then expressed as a function of the climate components using multiple regression. These regression equations are then combined to give an equation which relates interannual variations in the spatial patterns of yield to interannual variations in the spatial patterns of selected climate variables. The method is illustrated using wheat yield data from 59 crop reporting districts in southwestern Western Australia covering the period 1929–75. The regression models are calibrated using data for the period 1929–65 and the results are verified using data for the period 1966–75.. The climate contribution is shown to be highly significant, with winter precipitation being the most important variable. A single equation relating yield and climate patterns correctly reproduces the differing results obtained for separate parts of the study area by earlier workers. The influence of winter and autumn precipitation is nonlinear and, as a consequence, the study area divides into three zones: a high rainfall area where rainfall is generally more than optimum so that lower rainfall gives higher yields; a low rainfall area where rainfall is, on average, less than optimum so that positive rainfall anomalies are associated with higher yields; and an intermediate zone where average rainfall is close to optimum so that anomalies in either direction tend to suppress yields. Our analysis shows no evidence for any significant change in the sensitivity of wheat yields to climate in spite of a complete change in the variety of wheat cultivated.

1. Introduction

A voluminous and diverse literature exists on the relationships between crop yield and climate. While yield is certainly dependent on the weather conditions prevailing during, and sometimes prior to the growing season, expressing this dependence in terms of a reliable quantitative relationship, a crop–climate model, is a difficult task. The approaches available range from simulating the physiological growth of plants on daily or similar time scales using relationships derived from controlled experiments, up to the use of statistical "black box" relationships involving monthly or longer time scale weather variables. The former method is important in developing an understanding of the detailed causes of interannual yield variability, while the latter is more practical when seeking to understand or to predict the variations in larger-scale commercial yields over areas of 10^3 km^2 upwards. Monteith (1981) has given an excellent review of many of the physiological aspects of crop-climate relationships.

This paper deals with the development of statistical crop–climate models for regional-scale yields, with particular reference to wheat. Our main aim is to relate the spatial patterns of yield to spatial patterns of climate. The basic statistical technique applied is multivariate regression analysis. Although they have been stated on numerous occasions before (e.g., Katz, 1977) the problems which can be encountered are worth delineating.

a. Statistical problems

Most regression equations linking yield (the predictand) and climate (the predictor) involve a number of predictor variables. In such multiple regressions the statistical significance of a relationship is difficult to determine when, as is often the case in crop–climate modeling, the predictors are intercorrelated, or when the final equation involves only a subset of the original set of candidate predictors. The first problem, multicollinearity, may be reduced by careful *a priori* selection of predictor variables or by using techniques such as ridge regression (Haigh, 1977; Katz, 1979) or principal components regression (Briffa *et al.*, 1983). The second problem, elimination of candidate predictors, is important because, in crop–climate modeling, there is

[1] Permanent address: Department of Meteorology, Nanjing Institute of Meteorology, Nanjing, Jiangsu Province, People's Republic of China.

invariably a large set of candidate predictors available. The choice of predictors to form the model from a set of candidate predictors is usually made *a posteriori* on the basis of regression results (e.g., using some sort of elimination procedure such as step-wise regression). However, deciding just which predictors to retain is a difficult task (Hocking, 1976; Shibata, 1980) and estimating the true statistical significance of the result is equally daunting (Barnett and Hasselmann, 1979).

The only way to be confident of a regression equation in these circumstances is to test the equation on independent data. This is not often done satisfactorily in crop–climate modeling, generally because of a dearth of data. The fact that the relationship between crop yield and climate variables may be nonlinear only serves to compound these problems.

b. *Non-climatic factors*

Many yield time series show medium or long term changes which cannot be attributed to climate. The causes, generally lumped together under the heading "technology effects", may include changes in farming techniques, the effects of fertilizers or pesticides, the introduction of new varieties, expansion of the cultivated area into less favourable regions, and so on. Since there is no way to know *a priori* the overall influence of technology on yield, its effect is generally accounted for in some *ad hoc* way. For example, technology effects can be assumed to be linear, piecewise linear or quadratic in time and the crop–climate model developed using only the residuals from the appropriate fitted curve. Alternatively, the influence of technology could be filtered out using a high-pass filter. Any uncertainty in modeling nonclimatic factors must be transmitted to the residuals which are to be modeled in terms of climate and will limit the accuracy of crop–climate models.

In addition to these problems model results are limited by the complexity of the response of plants to weather. For example, short time scale weather events (such as a single frost) may significantly affect yield and it is rarely possible to account for such events in a crop–climate model[2]. Many crop–climate models use only monthly or longer time scale data so the influence of synoptic time scale events is only accounted for through their effect on the longer term averages. Finally, one of the most unsatisfactory aspects of a statistical crop–climate model is that it may not be interpretable in terms of the response of the crop to separate climate parameters. Thus, if candidate predictor variables are intercorrelated, a positive regression coefficient for a particular predictor does not necessarily mean that the response to that predictor will be positive if all other variables are held constant.

c. *Modeling approaches*

There is a hierarchy of regression models of increasing complexity which may be applied to crop–climate studies. As complexity increases, the statistical difficulties increase. The simplest models relate crop yield for a single location or region to a single climate variable (such as length of the growing season, total annual precipitation, etc.). These models generally explain little of the variance in crop yield, although they may give results which are of high statistical significance. At the next level single-location crop yield is related to a number of climate variables. This is the most common type of crop–climate model. Beyond this we could use more complex multivariate techniques to relate crop yields from a number of locations to a set of climate variables and develop an equation which has a number of variables on both the predictand and predictor sides. The aim of this paper is to explore this possibility by developing, through the use of principal components analysis, a crop–climate model in which the predictand is the spatial pattern of yield over a wide area.

Principal components analysis (PCA) has been used on numerous occasions in the analysis and modeling of crop yield data, but rarely for regression analysis involving the spatial patterns of both climate and interannual yield variations. Goodchild and Boyd (1975), Boyd *et al.* (1976) and Hill and Goodchild (1981) carried out a detailed investigation of wheat yields in Western Australia using PCA. In one of these papers (Boyd *et al.*) they examined the correlations between the first two yield components (based on the covariance matrix) and the long-term average spatial patterns of various growing season climate variables. Dennett *et al.* (1980) used PCA to identify regions in Europe where temporal yield and climate variations are coherent. Having done so, however, they then analysed crop–climate links separately in these coherent regions using traditional multiple regression techniques. Steyaert *et al.* (1978) used principal components of atmospheric pressure as predictors for regional average yield. Their regression equations have spatial patterns on the predictor side of the equation, but not on the predictand side.

In crop–climate modeling, the study most similar to our own is that of Mostek and Walsh (1981). These authors calculated principal components of corn yield, monthly temperature and monthly precipitation over the contiguous United States. Regression equations were then developed for the first yield component (which explained 31.1% of the total corn yield variance) as a function of selected climate principal components. Since the time series of PC1 amplitude variations must correlate highly with variations in the spatial mean

[2] Sakamoto (1978), in analyzing wheat yields in the U.S. Great Plains, accounted for such events by eliminating them from the model calibration period after a detailed year-by-year, region-by-region scrutiny of agricultural reports.

value of a variable, from a predictand viewpoint Mostek and Walsh's work is similar to that of Steyaert *et al.* In our own work we develop regression equations for the first three yield components and then recombine them to produce an equation explaining a large fraction of the overall spatial and temporal variability of yield. We know of only two other applications of the specific multivariate regression technique employed here, namely the studies of Kutzbach and Guetter (1980) and Jones *et al.* (1983) relating pressure patterns to patterns of temperature and precipitation.

We see two advantages in the use of this technique. First, it attempts to make optimum use of a large amount of data. When spatially distributed data on yield and climate are available simple averaging is inefficient since it obscures any spatial differences, and agglomeration irrespective of location cannot account for spatial differences in crop–climate response. The technique employed here retains spatial differences and leads to a model with spatially differentiated crop–climate relationships. Furthermore, principal components analysis, apart from extracting the dominant spatial patterns of variability, also serves to smooth the data spatially by relegating small-scale spatial details to the higher-order components. Multivariate principal components regression may, therefore, help to isolate the response "signal" from the small-scale "noise". Second, as Mostek and Walsh (1981) have pointed out, there may be practical advantages in developing crop–climate models which use large-scale climate patterns as predictors, since the reliability of and potential improvements in long-range weather forecasting are liable to be greatest for large-scale weather anomalies.

In the following sections of this paper we will describe the statistical technique of multivariate principal components regression (or spatial regression) and illustrate the method using yield data (per planted area) for wheat from southwestern Australia over the period 1929–75. We will begin by removing the technological influence, then calculate the principal components of crop yield and of climate, develop a multivariate regression equation using part of the data and testing it independently on the rest of the data, and finally examine the implications of our results.

2. Statistical description of yield and climate data

The study area is shown in Fig. 1. The region is divided into 59 shires for which yield data are available back to 1929. There is a considerable variation of climate over the region: mean annual precipitation ranges from around 280 mm in the east to 640 mm in the southwest, temperatures tend to become higher as one moves northward and away from the coast, and there are marked variations in annual evaporation and radiation (Waterhouse, 1969, quoted by Boyd *et al.*, 1976).

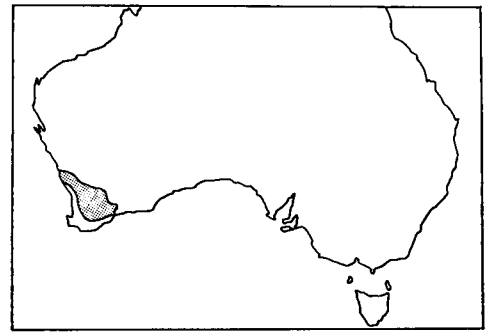

FIG. 1. General map of the study area.

Wheat yields show considerable variability both from year to year and decade to decade. Both short and longer term variations differ from shire to shire although most shires show a predominant upward trend which we attribute loosely here to the effects of technology. In order to analyze the influence of climate on year-to-year variations in yield the technology factor must be removed. We achieved this by using a low-pass filter to characterize the longer time scale yield changes for each shire and considering only the departures from the filtered values. Since there are no consistent similar long term variations in any of the climate variables used in this study, the long term yield changes can be identified with nonclimatic effects. There is some arbitrariness in the choice of filter (we used a 20-year moving quadratic), and since the yield departures depend to some degree on the filter used there must be a small but unavoidable uncertainty introduced into the analysis at this stage. In Fig. 2 we show three examples of the raw and filtered data. All subsequent analyses are based on departures from the long term trend for each shire. For simplicity we will refer to these yield departures simply as yields.

To characterize the spatial patterns of yield we performed a principal components analysis. First, however, we simplified the data using cluster analysis to group together neighbouring shires with similar yield variations [cf. Goodchild and Boyd (1975)]. This reduced the number of shires to 36 shire groups where the correlations within any shire group exceeded 0.90. Shire groupings are shown in Fig. 3a. We used normalized data for the PCA (equivalent to using the correlation matrix). Goodchild and Boyd (1975) and Hill and Goodchild (1981) have also analysed these yield data using PCA, but our work differs from theirs in two important aspects. First, we removed the technology effect from our data and, second, through normalization we effectively used the correlation matrix rather than the covariance matrix.

The first two yield principal component (PC) patterns, which account for 50.1% and 17.2% of the total yield variance respectively, are shown in Fig. 4 (see also Table 1). The first component (PC1) shows a rel-

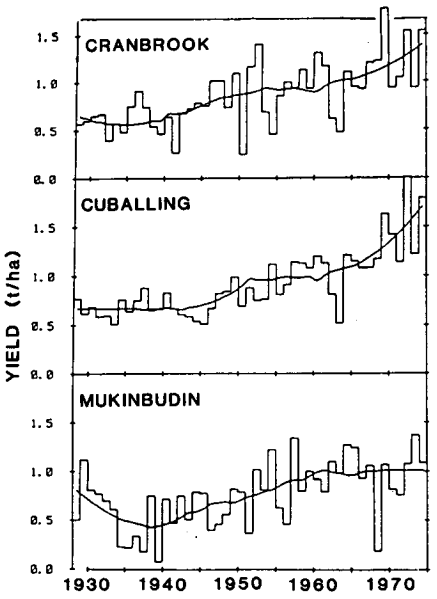

FIG. 2. Raw and filtered yield data (*t/ha*) for three shire groups. A, group 2, Cranbrook only; B, group 13, Cuballing, Narrogin, Pingelly, Wandering and Wickepin; C, group 20, Mukinbudin, Nungarin and Westonin.

atively flat distribution with all shires having loadings with the same sign and maximum loadings on shires in the center of the region. This component is representative of yield variations affecting the whole region similarly and the time series of PC1 amplitudes parallels the time series of regional-mean yields. PC2 characterizes out-of-phase variations in yield between the eastern and western parts of the region. Large amplitudes for PC2 occur in years where yield departures in the east and west have opposite signs. PC3 (not shown here) accounts for 9.3% of the variance and shows a north–south differentiation of yields. Higher order patterns were not considered.

Our basic climate data set was monthly temperatures from a network of 11 sites and rainfall from 22 sites (Fig. 3b). Other variables (radiation, evaporation, soil moisture levels, etc.) may be better as crop yield predictors, but these data are not readily available and do not span the whole period of our analysis. It is not practical to use the monthly data directly. Since the growing season in the southwest of the region spans April to November, retaining three principal components of rainfall and temperature for each month would give 60 climate variables, far too many for regression using only 47 years of data. We therefore divided the growing season into three periods, autumn (April and May), winter (June–August) and spring (September–November). This crude division has some obvious deficiencies, not least being the fact that the length of the growing season reduces considerably as one moves from southwest to northeast [Boyd *et al.*, 1976, Fig. 2(c)].

The seasonal climate data were then subjected to a principal components analysis using the correlation matrix. As with the yield data, most of the variance is explained by the first few components. Table 1 shows the distribution of variance for the first five components. In subsequent analyses we retained only the first three components. We have thus reduced the climate data set from 99 items [(11 temperature sites + 22 rainfall sites) × 3 periods] to 18 items [(3 temperature PCs + 3 rainfall PCs) × 3 periods].

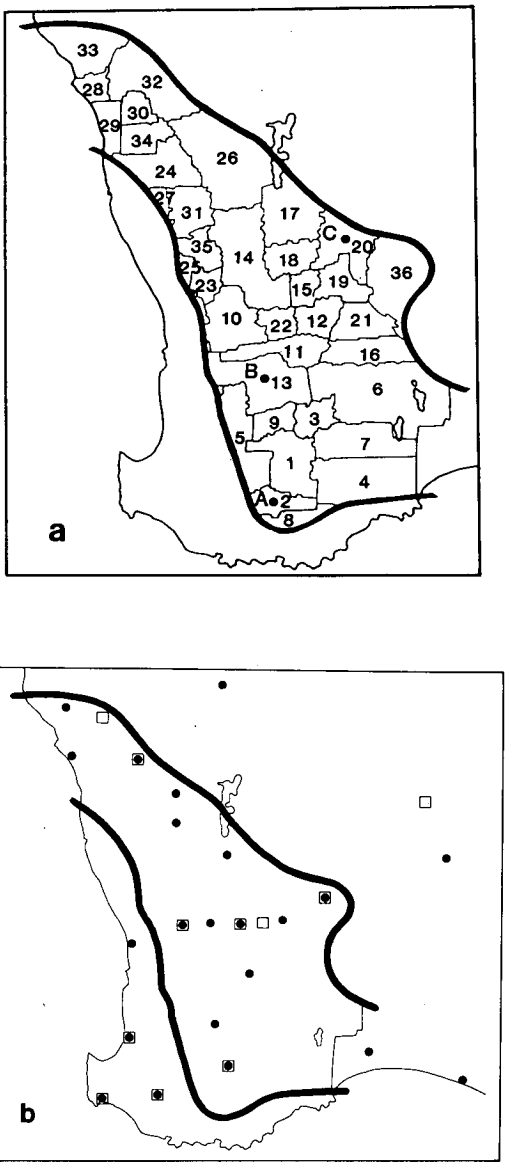

FIG. 3. (a) Shire group boundaries and numbers keyed to Table 3. A, B and C denote the shires whose yield time series are shown in Fig. 2. (b) Positions of rainfall (circle) and temperature (square) stations. More rainfall stations are used because of the greater spatial variability of rainfall.

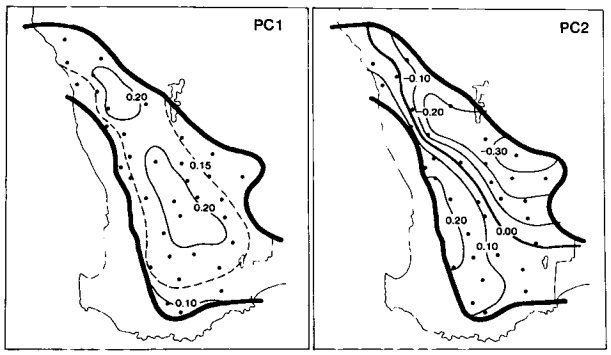

FIG. 4. The first two principal components of residual yield accounting for 50.1% and 17.2% of the total yield variance. Dots show approximate locations of the shire group centroids. Note that the loading at any point, if multiplied by the square root of the eigenvalue [$\lambda_1 = 36(0.501) = 18.04$; $\lambda_2 = 36(0.172) = 6.19$] gives the value of the correlation coefficient between the residual yield at that point and the principal component amplitude. Thus, for PC1, these correlations range between 0.30 (shire group 8) and 0.93 (shire group 14).

3. Regression analysis

To simplify discussion we will use the following terminology: yield departure (Y), rainfall (R), temperature (T), autumn (a), winter (w), spring (s), and 1, 2, 3 for first, second and third principal components. Following Mostek and Walsh (1981) we use vector notation to denote the component patterns, and the corresponding scalar symbols to denote the time-dependent amplitudes of these patterns. Lower case scalar symbols will be used to identify normalized values of the amplitudes. Some examples should clarify this terminology. If **Y1** denotes the first yield component pattern, the time series of amplitudes of **Y1** (i.e., the strength of this pattern in a particular year) is $Y1$[3] and $y1$ is the normalized value of $Y1$ (obtained by subtracting the mean and dividing by the standard deviation). Similarly, if **Tw2** denotes the second principal component of winter temperature, $Tw2$ denotes the time-dependent amplitude of **Tw2**, $tw2$ is the normalized value of $Tw2$ and so on.

Regression equations were derived for each of the first three yield component amplitudes using the normalized climate component amplitudes as predictor variables. A preliminary analysis of the data suggested that the link between yield and rainfall may be nonlinear (e.g., Boyd *et al.*, 1976, show in their Fig. 1 how maximum yield occurs roughly along the 400 mm annual isohyet with yield decreasing for both higher and lower rainfall). We therefore included the normalized squares of the amplitudes of PC1 for rainfall in each season as candidate predictors, making a total of 21 candidate predictors.

One of the potential problems with regression models for crop–climate response is intercorrelation of the climate variables which may lead to an unstable regression equation and/or a spuriously high correlation between observed and estimated predictand. Our chosen candidate predictors showed a number of statistically significant correlations. However, in only four cases did the coefficients of determination (r^2) exceed 20%. These were: $rs1 - ts1$, $r = -0.68$; $rw3 - tw1$, $r = -0.59$; $ta2 - ts2$, $r = -0.56$; and $rw1 - x$, $r = 0.47$ [where x here denotes the normalized form of $(Rw1)^2$].

An additional problem in assessing the significance of a regression equation arises when the final prediction equation involves only a subset of the candidate predictor variables chosen on the *a posteriori* basis of performance in the regression analysis (Barnett and Hasselmann, 1979). This point requires further elaboration since it affects, not only the results presented below, but many applications of regression analysis.

The strength of a multiple regression relationship can be judged using the sample multiple correlation coefficient R. However, R is not an unbiased estimator of the population multiple correlation coefficient. For an analysis based on N values of the predictand and k_0 predictor variables the expected value of R^2 when the population multiple correlation coefficient is zero is not zero but

$$E\{R^2\} = \frac{k_0}{N-1}. \tag{1}$$

Here $k_0/(N-1)$ represents a threshold value above which R^2 becomes meaningful. An alternative measure of the strength of a relationship is the adjusted-R^2, which is simply the amount of explained variance above $k_0/(N-1)$ expressed as a fraction of its maximum possible value: i.e.,

$$R_a^2 = \frac{R^2 - [k_0/(N-1)]}{1 - [k_0/(N-1)]}. \tag{2}$$

TABLE 1. Cumulative percentage of variance explained by the first five principal components of yield, rainfall and temperature.

Variable	Symbol	PC1	PC2	PC3	PC4	PC5
Yield departure	Y	50.1	67.3	76.6	80.0	83.1
Autumn rainfall (AM)	Ra	54.8	64.2	71.0	76.8	80.3
Winter rainfall (JJA)	Rw	57.3	67.3	73.5	78.7	82.7
Spring rainfall (SON)	Rs	49.8	63.3	70.8	76.1	80.3
Autumn temperature (AM)	Ta	81.7	90.8	94.2	95.8	97.1
Winter temperature (JJA)	Tw	79.1	87.8	91.6	94.7	96.7
Spring temperature (SON)	Ts	66.3	82.9	89.7	92.7	95.3

[3] A more precise notation would be to use $Y1_j$ to denote the amplitude of **Y1** in year j, but we have omitted the year subscripts on all PC amplitudes for simplicity.

R_a^2 has the property that $R_a^2 = 0$ when $R^2 = k_0/(N - 1)$ and $R_a^2 = 1$ when $R^2 = 1$. Suppose, now, that a regression analysis involving k_0 candidate predictor variables has been carried out and only the best k variables are retained (i.e., those that contribute most to the variance explained by all of the k_0 original variables). Since the choice of retained predictors has not been made randomly, but on the basis of their performance in the regression using *all* candidate predictors, we cannot assume that there has been a reduction in threshold value of R^2 from $k_0/(N - 1)$ to $k/(N - 1)$. This *a posteriori* selection makes it difficult to assess statistical significance, and the standard F-test using

$$F = \frac{(N - k - 1)R^2}{k(1 - R^2)} \tag{3}$$

with (k, N-k-1) degrees of freedom may give a spuriously high significance level for the k-predictor equation. The only way out of this dilemma is to test an equation using independent data.

In the present study we have 47 years of data and 21 candidate predictors. Because we consider verification of a model to be of considerable importance, we chose to use the first 37 years for calibration leaving 10 years for verification.

Regression equations were developed for $Y1$, $Y2$ and $Y3$ using a standard step-wise routine. At this stage a choice has to be made regarding the number of terms retained in the regression equations, a procedure sometimes referred to as "screening". A large number of terms will give a high calibration-R^2, but may not verify well, while a small number of terms may give an unacceptably low calibration-R^2. There are many objective screening criteria which may be employed; for example the equation could be chosen to maximize either the adjusted-R^2 or some related parameter (there is a wide range of choices, see e.g., Shibata, 1980; Klein, 1983), or one could decide to retain only those terms for which individual regression coefficients were significant at some chosen level (based on the usual t-test). We have used the latter method and produced two sets of equations using t-value cut-offs of 1 and 2. The results are summarized in Table 2.

Applying the conventional test to the F-values shows all results to be statistically significant at considerably better than the 1% level. Although the $t = 1$ cut-off equations explain more of the yield variance, the $t = 2$ equations are more significant. Since $Y1$, $Y2$ and $Y3$ explain, respectively, approximately 50.1%, 17.2% and 9.2% of the variance over the period 1929–75 we can estimate the overall explanatory power of the equations by assuming the same percentages apply over the period 1929–65. For the $t = 2$ equation the explained variance is approximately 0.501(50.1) + 0.769(17.2) + 0.328(9.2) = 41.4%, while the corresponding value for the $t = 1$ equation is 47.5%.

These estimates can be made more precise by using the regression equations to estimate, year-by-year, the yields for each of the 36 shire groups and comparing these estimates with the observed data. This can be done for both the calibration period, 1929–65, and the verification period, 1966–75. The procedure is as follows. For each year the observed amplitudes of the climate PCs are entered into the regression equations to give estimates of the yield amplitudes, viz. $\widehat{Y1}$, $\widehat{Y2}$ and $\widehat{Y3}$. Since the yield in shire i can be expressed as a sum of the products of the loadings (or weights) for shire i and the component amplitudes, these estimated amplitudes can be used to calculate the estimated yield for each of the 36 shire groups. The estimated yield for shire i is then

$$\hat{Y}_i \approx w1_i\widehat{Y1} + w2_i\widehat{Y2} + w3_i\widehat{Y3}. \tag{4}$$

where $w1_i$, $w2_i$ and $w3_i$ are the weights on shire i for **Y1**, **Y2** and **Y3**.

The results of these calculations are shown in Table 3 which lists the correlations between observed and estimated yields for each shire group for the calibration period and for the verification period for both the $t = 1$ and $t = 2$ models. The Table also shows the correlations between observed and estimated values of $Y1$, $Y2$ and $Y3$ and the overall correlation between observed and estimated yields. The latter values correspond to explained variances of 40.7% ($t = 2$) and 46.8% ($t = 1$) over the calibration period, essentially the same as the approximate values given above.

Although the spatial distribution of the performance of the models during calibration is of some interest, the most important results in Table 3 are those for the verification period. As invariably occurs in these types of study, verification performance is noticeably inferior to calibration performance. The results show that the $Y3$ equation is of little value—in spite of the fact the calibration results for $Y3$ were highly statistically significant! Performance of the $t = 1$ model is significantly worse than the $t = 2$ model for $Y1$, and not significantly different for $Y2$. Verification performance can be assessed relative to R_a^2, but some allowance must be made for the *a posteriori* choice of predictors (for example, by using k_0 and k in Eq. (2) to give rough bounds to R_a^2). Performance for $Y1$ is satisfactory, but $Y2$ is less satisfactory, although its verification result

TABLE 2. Calibration performance of the models for the first three yield principal components. R_a^2 has been calculated using the number of retained predictors.

	$t = 2$ cut-off				$t = 1$ cut-off			
Predictand	k	R^2	R_a^2	F	k	R^2	R_a^2	F
$Y1$	4	0.501	0.439	8.04	9	0.648	0.531	5.52
$Y2$	5	0.769	0.731	20.59	11	0.858	0.796	13.74
$Y3$	2	0.328	0.289	8.31	9	0.570	0.427	3.98

TABLE 3. Calibration and verification correlation coefficients for individual shire groups, yield principal components and overall.

Shire-group number	$t = 2$ model		$t = 1$ model	
	Calibration	Verification	Calibration	Verification
1	0.62	−0.03	0.73	−0.09
2	0.55	−0.01	0.62	0.13
3	0.58	0.02	0.60	−0.07
4	0.52	−0.20	0.61	−0.21
5	0.69	0.29	0.78	0.25
6	0.47	0.47	0.50	0.32
7	0.66	0.07	0.70	−0.05
8	0.35	0.16	0.34	0.40
9	0.59	0.32	0.60	0.22
10	0.71	0.19	0.71	0.15
11	0.69	0.22	0.66	0.18
12	0.65	0.16	0.69	0.11
13	0.72	0.21	0.71	0.23
14	0.67	0.46	0.64	0.31
15	0.72	0.70	0.73	0.44
16	0.61	0.51	0.62	0.34
17	0.73	0.72	0.88	0.60
18	0.79	0.84	0.85	0.73
19	0.80	0.77	0.83	0.69
20	0.74	0.80	0.87	0.79
21	0.64	0.49	0.53	0.36
22	0.60	0.31	0.60	0.15
23	0.66	0.04	0.68	0.06
24	0.60	0.65	0.66	0.52
25	0.45	−0.16	0.58	−0.24
26	0.73	0.67	0.82	0.47
27	0.57	0.22	0.51	0.16
28	0.59	0.47	0.67	0.48
29	0.58	0.69	0.66	0.69
30	0.67	0.41	0.72	0.36
31	0.72	0.45	0.78	0.30
32	0.70	0.49	0.78	0.35
33	0.59	0.50	0.69	0.41
34	0.62	0.49	0.71	0.38
35	0.70	0.34	0.71	0.34
36	0.68	0.86	0.80	0.83
$Y1$	0.708	0.560	0.805	0.339
$Y2$	0.878	0.524	0.926	0.563
$Y3$	0.573	−0.077	0.755	0.061
Overall	0.638	0.380	0.684	0.301

is still highly significant. At the individual shire-group level the $t = 2$ model performs better than the $t = 1$ model in 30 out of 36 cases[4]. The spatial pattern of verification performance shows that the models have little predictive value in the southernmost part of the study area (shire groups 1, 2, 3, 4, 7 and 8). The best results are obtained in the northeastern section (shire groups 15, 17, 18, 19, 20, 26 and 36). Considering the essential simplicity of the approach, a single model describing yield variations over a wide area using rather crude climate variables, we believe these results to be extremely promising.

[4] These results confirm our expectations based on the statistical significance of the calibration-R^2, but they also show (for $Y3$ particularly) that such significance tests can be misleading.

4. A composite model and its implications

The above analysis has shown that, after the effects of technological change have been accounted for, a significant percentage of the interannual variability of residual yield can be attributed to variations in climate. In order to examine the climatic influence more closely we need to make maximum use of the available data. We therefore developed a new model using all of the data. In the calibration/verification study the $t = 2$ model was demonstrably superior so we used a similarly strict criterion here for retaining predictor variables. Since the $Y3$ model did not verify, we only developed models for $Y1$ and $Y2$. The retained predictors were the first components of spring rainfall, autumn rainfall and winter rainfall for $Y1$ (the latter two appearing only as quadratic terms), and the first components of winter rainfall, autumn rainfall and spring temperature for $Y2$. The regression equations are:

$$\widehat{Y1} = 0.58Rs1 - 0.08[(Ra1)^2 - 0.82] - 0.06[(Rw1)^2 - 1.00], \quad (5)$$

$$\widehat{Y2} = -0.39Rw1 - 0.22Ra1 + 0.28Ts1. \quad (6)$$

(The quantities 0.82 and 1.00 associated with $(Ra1)^2$ and $(Rw1)^2$ in Eq. (5) are their respective mean values.) For $Y1$ the multiple correlation coefficient is 0.660 and for $Y2$ it is 0.776. These two equations when combined to estimate year by year residual yields of individual shire groups explain approximately 32% of the variance. Higher explained variance could be obtained by including more terms in the equations, by eliminating the southernmost stations from the analysis (the region where yield PCs 1 and 2 explain the lowest fraction of total yield variance; see Fig. 4), or, perhaps, by a more judicious choice of climate variables. Although we have used a rather strict criterion for acceptance of predictor variables to avoid over-calibration, it is apparent that, in our study area, a significant amount of the interannual variance cannot be explained by climate. However, the present results are of high statistical significance and clearly identify the most important features of the role of climate in determining yields over the whole study area, namely—

1) the major determining factor is rainfall, throughout the growing season. Temperature has a secondary effect in spring and, possibly, in autumn (*Tal* appears in the earlier analyses, but did not reach $t = 2$ significance in the analysis using all of the data).

2) the dominant influence is through large-scale variations in climate, since second and higher-order climate components play only minor roles in the regression equations.

3) the effect of rainfall on yield is nonlinear.

To discuss the overall influence of climate on yield the equations for $Y1$ and $Y2$ need to be combined.

The estimated yield for shire group i is

$$\hat{Y}_i = -0.08w1_i[(Ra1)^2 - 0.82] - 0.22w2_iRa1 - 0.06w1_i[(Rw1)^2 - 1.00] - 0.39w2_iRw1 + 0.58w1_iRs1 + 0.28w2_iTs1 \quad (7)$$

[obtained by substituting Eqs. (5) and (6) into Eq. (4) and ignoring $\widehat{Y3}$].

a. *Influence of winter rainfall*

Of the terms in Eq. (7) winter rainfall, through $Rw1$, is the most important. Since $Rw1$ explains over 57% of the total winter rainfall variance (see Table 1), and since the pattern of $Rw1$ is essentially flat with similar loadings over the whole study area, year-to-year variations in the amplitude of $Rw1$ correlate highly with total winter rainfall in all shire groups.[5]

The nonlinear aspect of the relationship between yield and $Rw1$ can be illustrated in two ways. First, we note that there must be a particular value of $Rw1$ in Eq. (7) which gives a maximum estimated yield. To find this value, we simply partially differentiate Eq. (7) with respect to $Rw1$ and equate the result to zero. This gives

$$Rw1 = -3.00\frac{w2_i}{w1_i}. \quad (8)$$

The optimum (i.e., yield maximizing) value of $Rw1$ (which corresponds closely to total winter rainfall) therefore depends on the values of $w1_i$ and $w2_i$; i.e., on the particular shire group. Eq. (8) defines the optimum value of $Rw1$ for any given shire group. For the western shires $w1_i$ and $w2_i$ are both positive, so that the optimum winter rainfall is negative and below the "normal" winter rainfall (which corresponds to $Rw1 = 0$). For eastern shires $w2_i$ is negative and $w1_i$ is positive so the optimum value of $Rw1$ here is positive and normal winter rainfall is less than optimum. An immediate consequence of these results is that, in western shires, any correlation between yield and winter rainfall should be negative, while in eastern shires the correlation should be positive. The line along which $w2_i$ is zero defines the region where the optimum winter rainfall is close to the normal value. This line (see Fig. 4b) is very close to the 400 mm annual rainfall isohyet. In this region, winter rainfall anomalies of either sign may cause a reduction in yield.

If the $Rw1$ part of Eq. (7) is isolated then its effect on yield can be calculated for each shire group for different values of $Rw1$. Fig. 5 shows these results for $Rw1 = \pm 4$ and ± 8, representing moderate and extreme

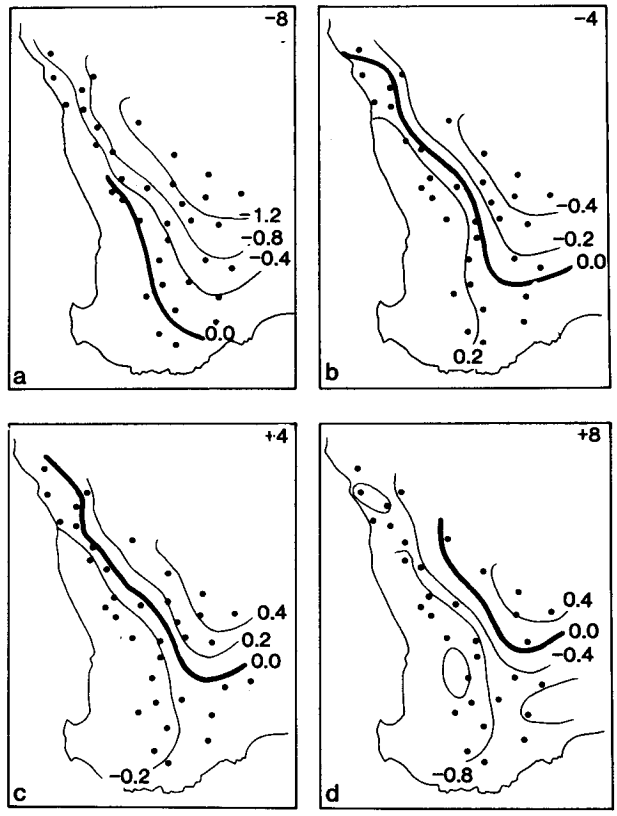

FIG. 5. The influence of winter rainfall on yield. The values ±8 and ±4 are amplitudes of the first principal component of winter rainfall and correspond to large and moderate rainfall anomalies (negative sign for negative anomaly). The zero line is the region of no yield influence, negative contours represent a reduction in yield and positive contours an increase in yield. Thus, for example, an extremely low winter rainfall (−8) would cause a reduced yield almost everywhere except in the far southwest where winter rainfall is always in excess of the optimum. Extremely high winter rainfall (+8) reduces yield in the southwest and increases yield in the northeast. In the northeast, however, even this extreme rainfall anomaly falls short of the optimum.

winter rainfall anomalies in both directions.[6] For very dry conditions (Fig. 5a) yields may be depressed over the whole region, except the far west and southwest where rainfall is sufficient to ensure that even a severe drought will not cause winter rainfall to fall below the optimum value. Moderately dry conditions (Fig. 5b) will tend to result in improved yields in the west and depressed yields in the east, while the opposite pattern (Fig. 5c) results from moderately wet conditions. A large positive winter rainfall anomaly (Fig. 5d) will expand the western region of depressed yields. In the far east, where winter rainfall is considerably below the optimum, even the most extreme positive anomaly (in terms of 20th century records) may not produce

[5] The correlation coefficient is $l1_i\sqrt{\lambda_1}$ where $l1_i$ is the loading on $Rw1$ at location i and λ_1 [=22(0.573)] is the corresponding eigenvalue. For stations within the wheat belt area, this correlation ranges from 0.67 to 0.92.

[6] e.g., if $Rw1$ amplitude is 8, the corresponding rainfall anomaly is $8/\sqrt{\lambda_1} = 2.25$ standard deviations.

maximum yields. Note that, in any individual year, these influences will tend to be obscured by other climatic and nonclimatic factors which affect yield, since the winter rainfall effect, while statistically significant, is not strong.

b. Autumn and spring rainfall

For autumn rainfall ($Ra1$) the influence on yield is also nonlinear and takes essentially the same form as for winter rainfall [see Eq. (7)]. The region where the optimum rainfall is close to normal rainfall again falls along the line where the loading on $Y2$ is zero. If diagrams similar to Fig. 5 are constructed for autumn the only difference from the winter case is that the gradients of yield anomaly from southwest to southeast are less. Yield variations are therefore less sensitive to autumn rainfall anomalies than to winter anomalies if anomalies are expressed in normalized terms. Equivalently, the spatial variability of the optimum autumn rainfall

$$Ra1 = -1.32 \frac{w2_i}{w1_i} \tag{9}$$

is considerably less in autumn than in winter.

Our results for spring rainfall point to a near-linear relationship with yield increasing as spring rainfall increases for all shire-groups. This follows directly from Eq. (7) when one observes that $w1_i$ is positive everywhere. The influence is strongest in the center of the study region. Such a relationship has been observed by Gentilli (1959) for the more arid part of the region. Our results show why this result has been detected in arid areas and not elsewhere. The reason is that, in most parts of the study area, the link with spring rainfall tends to be obscured by temperature variations. $Rs1$ is significantly negatively correlated with $Ts1$ ($r = -0.68$) and both $Rs1$ and $Ts1$ occur in the regression equation. In the wetter southwestern part of the region, where $w2_i$ and $w1_i$ are both positive, a positive spring rainfall anomaly will have a positive effect on yields directly, but a negative effect on yield indirectly through $Ts1$. In the arid northeastern part of the region, however, the rainfall and temperature effects reinforce each other because $w2_i$ and $w1_i$ have opposite signs. It is not surprising, therefore, that the link between spring rainfall and yield has only been observed in the arid sector.

c. Temperature influence

The only temperature term to appear in Eq. (7) is the first principal component of spring temperature. It is clear that temperature is a much less important variable than rainfall, and its influence must be partly obscured because of intercorrelations between rainfall and temperature. Note, however, that $Ts1$ and $Rs1$ arise independently in the overall yield equation through $Y2$ and $Y1$ respectively. Since spring rainfall and temperature effects tend to cancel out in the southwestern part of the region (wherever $w2_i$ is positive), it is only in the northeast where the influence may be significant. In this area higher temperatures should be associated with lower yields.

5. Conclusions

The main aim of this paper has been to develop and explore a new technique in statistical crop–climate analysis, the direct linking of spatial patterns of yield and spatial patterns of climate. The method, which is a further development of that employed by Mostek and Walsh (1981), decomposes both yield and climate data time series into principal components which are then related using standard multiple regression techniques. The equations for each yield component are then recombined to produce a single equation for the overall spatial patterns of yield. This method is similar to canonical correlation, although it is somewhat simpler conceptually. It is, however, more flexible in that it allows different climate–PC predictors to be used in the various yield–PC equations.

One of the main features of this method is that it effectively combines a number of otherwise separate results (such as a set of distinct crop–climate models for different subregions) into a single equation. In our example, a single equation reproduced all of the known (and apparently disparate) relationships between Western Australian wheat yield and climate and provided additional understanding of the way these relationships are linked and how they fit into a broader framework. Another advantage of using spatial patterns as predictand and predictor variables is that the noisiness of small-scale local variations, which may confuse otherwise broadly applicable results, is smoothed out.

From a statistical viewpoint all of the caveats which apply to simpler applications of regression analysis apply as well to spatial regression analysis. We have discussed in some detail the problem of determining the statistical significance of a regression equation, and emphasized the need to test equations on independent data. We encountered a situation where a highly significant relationship (for $\widehat{Y3}$) failed to verify on independent data. In crop–climate modeling adequate verification can be an insurmountable problem because yield data series are often too short to allow the analyst the luxury of reserving data for independent testing. The problem is compounded by the possibility that a model's failure to verify may, in fact, represent a real change in the way yields relate to climate, brought about by technological changes such as increases in irrigation, fertilizer or herbicide usage, changes in the amount of cultivated marginal land, changes in crop variety, etc. The fact that our Western Australian results

for the first two principal components do verify suggests that this particular crop–climate link has remained relatively stable, in spite of a complete change in wheat variety and significant changes in farming techniques over the study period. For our second yield component, however, the verification statistics (see Table 3) are not as good as expected for a completely stable relationship, and this could possibly reflect a second-order variation in crop–climate response resulting from technological change.

Hill and Goodchild (1981) suggest that climate was a more important factor prior to 1940 than subsequently. We have found no evidence for any significant decrease in the importance of climate after 1940. The difference here is partly a semantic one related to influences on different time scales. For many shires (see Fig. 2) the long-term technological trend (which we have factored out in our analysis) only becomes evident after about 1940. These trends would, therefore, tend to obscure the influence of climate after this date in Hill and Goodchild's analysis.

Our analysis of wheat yields in Western Australia sheds new light on the results of past studies. To summarize these earlier results, Gentilli (1959) found that winter precipitation and yields were positively correlated in high rainfall areas, Fitzpatrick (1970) found that these variables were negatively correlated in low rainfall areas, and Gentilli (1946, quoted by Boyd *et al.*; and 1959) found that spring precipitation was positively correlated with yields in low rainfall areas. We are able to explain these opposing correlation coefficients and also explain why the observed correlations are regionally specific: they arise from the nonlinear relationship between yield and winter rainfall and because of the intercorrelation between spring rainfall and temperature principal components. The nonlinear link between yield and rainfall is particularly important because it implies that yield is maximized with respect to rainfall variations roughly along the zero loading line for the second principal component pattern, a line which runs roughly along the 400 mm annual isohyet.

The strong and verifiable relationship between Western Australian wheat yields and climate that we have found appears to conflict with some of the conclusions reached by Goodchild and Boyd (1975), Boyd *et al.* (1976) and Hill and Goodchild (1981). For example, Boyd *et al.* (1976, p. 20) state that wheat production has ". . . a yield distribution not obviously related to rainfall". These differences arise largely because Goodchild and co-workers were not primarily interested in elucidating the crop–climate link. They chose, therefore, not to remove the technology trends; to the contrary, their analysis was particularly concerned with technological aspects and their first yield component goes some way towards isolating this effect. Their second yield component *did* correlate with the spatial pattern of mean precipitation, and the existence of such a link tends to support our own conclusions. This result is still consistent with Boyd *et al.*'s statement that ". . . the *major* component of variation in regional yields . . . is . . . not obviously related to any major climatic variable" (emphasis added). While our major component of yield variation (i.e., PC1) is climate-dependent, their's is not. The two PC1s differ because in our own analysis we used normalized data and chose to remove a technology factor *a priori*. Since a large part of the total yield variance is accounted for by the technology trends, the first component derived from the covariance matrix and using data which have not been detrended will reflect these technological factors. Nevertheless, we agree with Hill and Goodchild (1981) that climatic factors play a secondary role to socio-economic (i.e., technological) factors, but with the important distinction that these two factors operate on different time scales. Climatic factors are dominant on the interannual time scale, while socio-economic factors are of primary importance on medium to long time scales (i.e., $\geq$10 years).

As a final point, we must emphasize the inherent simplicity of our analysis. We have accounted for technological change in a relatively crude way (albeit certainly adequate given our desire to concentrate only on interannual variability), we have chosen not to consider explicitly a number of possibly important non-climatic factors (such as changing land-use, short time scale variations in fertilizer use, the effects of pests and disease, etc.), and we have used only the simplest and most readily available climate variables in our analysis. There is considerable scope for a more elaborate study, but even this simple application of spatial regression analysis is sufficient to demonstrate its elegance and value.

Acknowledgments. Our sincere thanks go to Dr. N. A. Goodchild, Biometrics Unit, Institute of Agriculture, University of Western Australia for providing the yield data used in this study and for his comments on our work, and to Dr. R. R. Brook and the Commonwealth Bureau of Meteorology, Melbourne, Australia for providing the climate data. Dr. R. A. Warrick and Dr. R. W. Katz also made useful and constructive comments on earlier versions of this paper. Tu Qipu is a visiting scientist in the U.K. under a British Council Fellowship Programme.

REFERENCES

Barnett, T. P., and K. Hasselmann, 1979: Techniques in linear prediction with application to oceanic and atmospheric fields in the Tropical Pacific. *Rev. Geophys. Space Phys.*, **17**, 949–968.

Boyd, W. J. R., N. A. Goodchild, W. K. Waterhouse and B. B. Singh, 1976: An analysis of climatic environments for plant-breeding purposes. *Aust. J. Agric. Res.*, **27**, 19–33.

Briffa, K. R., P. D. Jones, T. M. L. Wigley, J. R. Pilcher and M. G. L. Baillie, 1983: Climate reconstruction from tree rings: Part 1, Basic methodology and preliminary results for England. *J. Climatol.*, **3**, 233–242.

Dennett, M. D., J. Elston and R. Diego Q, 1980: Weather and yields of tobacco, sugar beet and wheat in Europe. *Agric. Meteor.*, **21**, 249–263.

Fitzpatrick, E. A., 1970: The expectancy of deficient winter rainfall and the potential for severe drought in the southwest of Western Australia. Misc. Publ. No. 70/1, Agronomy Dept., University of Western Australia.

Gentilli, J., 1946: Mimeographed tables and maps of rainfall and climate in Western Australia, and the rainfall-wheat relationship. Publ., Geography Dept., University of Western Australia. (Quoted by Boyd *et al.*, 1976.)

——, 1959: Weather and climate in Western Australia. Publ., W.A. Govt. Tourist and Publicity Bureau, W.A. Govt. Printer, Perth.

Goodchild, N. A., and W. J. R. Boyd, 1975: Regional and temporal variations in wheat yield in Western Australia and their implications in plant breeding. *Aust. J. Agric. Res.*, **26**, 209–217.

Haigh, P. A., 1977: Separating the effects of weather and management on crop production. Rep. to the Charles F. Kettering Foundation, 93 pp.

Hill, J., and N. A. Goodchild, 1981: Analysing environments for plant breeding purposes as exemplified by multivariate analyses of long term wheat yields. *Theor. Appl. Genet.*, **59**, 317–325.

Hocking, R. R., 1976: The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Jones, P. D., T. M. L. Wigley and K. R. Briffa, 1983: Reconstructing surface pressure patterns using principal components regression on temperature and precipitation data. *Proc. Second Int. Meeting on Statistical Climatology*, Lisbon, Instituto Nacional de Meteorologia e Geofísica, 4.2.1–4.2.8.

Katz, R. W., 1977: Assessing the impact of climatic change on food production. *Climatic Change*, **1**, 85–96.

——, 1979: Sensitivity analysis of statistical crop-weather models. *Agric. Meteor.*, **29**, 291–300.

Klein, W. H., 1983: Objective specification of monthly mean surface temperature from mean 700 mb heights in winter. *Mon. Wea. Rev.*, **111**, 674–691.

Kutzbach, J. E., and P. J. Guetter, 1980: On the design of paleoenvironmental data networks for estimating large-scale patterns of climate. *Quat. Res.*, **14**, 169–187.

Monteith, J. L., 1981: Climatic variation and the growth of crops. *Quart. J. Roy. Meteor. Soc.*,, **107**, 749–774.

Mostek, A., and J. E. Walsh, 1981: Corn yield variability and weather patterns in the U.S.A. *Agric. Meteor.*, **25**, 111–124.

Sakamoto, C., 1978: Reanalysis of CCEAI, U.S. Great Plains Wheat Yield Models. Center for Climatic and Environmental Assessment Tech. Note 78-3, NASA, Houston.

Shibata, R., 1980: Selection of the number of regression parameters in small sample cases. *Statistical Climatology*. S. Ikeda, E. Suzuki, E. Uchida and M. M. Yoshino, Eds., Elsevier, 137–148.

Steyaert, L. T., S. K. Le Duc and J. D. McQuigg, 1978: Atmospheric pressure and wheat yield modeling. *Agric. Meteor.*, **19**, 23–34.

Waterhouse, W. K., 1969: An agro-ecological survey of cereal production in the agricultural areas of Western Australia. M.Sc. thesis, University of Western Australia.