QUANTIFYING HOW CLIMATIC FACTORS AFFECT VARIATION IN PLANT DISEASE SEVERITY: A GENERAL METHOD USING A NEW WAY TO ANALYZE METEOROLOGICAL DATA

STELLA MELUGIN COAKLEY and LARRY R. McDANIEL

Department of Biological Sciences, University of Denver; Research Location: National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307, U.S.A.*

and

ROLAND F. LINE

ARS, U.S. Department of Agriculture, Pullman, WA 99164, U.S.A.

Abstract. A general method is presented for analyzing how climatic conditions affect plant disease severity. An example of its application is given for the analysis of stripe rust (caused by *Puccinia striiformis*) data on winter wheat cultivar Gaines and climatic data collected at Pullman, WA. for 1968–1986. A computer program WINDOW was written to identify the climatic factors most highly correlated with disease. This program is designed to utilize meteorological data for an entire growing season of a crop as well as to include climatic conditions preceding planting. This program uses an iterative process to examine variable-length segments of meteorological data in a more exhaustive analysis than previously possible. Climatic factors considered include: mean maximum, minimum, and average temperature; total and frequency of precipitation; consecutive days with and without precipitation; accumulation of negative and positive degree days; and number of days with extreme temperature events. Variables that were highly correlated with disease were the basis for regression models that were developed to predict disease severity index for each of the three cultivars. Two- and three-variable models explained, respectively, 75 and 76% of the variation in disease from year to year. Predictions (which could be made early enough in the growing season to allow application of chemical control) were evaluated on the basis of whether years with severe disease were accurately predicted. Models were validated using Allen's PRESS statistic and by application to new data. The method is potentially applicable to studies of how climatic conditions affect the populations or productivity of other types of organisms.

1. Introduction

For plant disease to occur, a susceptible host must be available to a pathogen capable of infecting it (or to an insect vector carrying the pathogen), and the environment must be simultaneously favorable for all of these organisms and their required interactions. Plant disease epidemics can significantly reduce potential crop yield. For such to occur, there must be a large population of

* This research was supported by a National Science Foundation Grant (ATM 85-03115), Climate Dynamics Program, Division of Atmospheric Sciences.

Climatic Change **12** (1988) 57–75.

susceptible hosts and virulent pathogens, and the environment must be favorable for a sufficient time at frequent enough intervals for development of widespread disease. Disease development stops if the environment is limiting for either the host or pathogen. The climatic factors that most frequently limit host-pathogen interactions and potential crop yields are temperature and moisture.

Much is known about how ambient meteorological factors affect disease development and how climatic factors limit disease (Rotem, 1978). Past emphasis has largely been on using controlled environments to determine the optimum conditions for specific stages of disease development. Automation of data collecting and processing has enabled determination of micrometeorological conditions in the field that are associated with disease development throughout a growing season (Coakley, 1985; Jones *et al.*, 1984; Sutton *et al.*, 1984; Teng and Rouse, 1984). Nevertheless, such in-field monitoring is apt to be used only for research purposes or where high-value crops warrant the costs of specific predictive systems.

For the most part, daily meteorological data routinely collected by the National Oceanic and Atmospheric Administration (NOAA) have been ignored for use in epidemiological studies because the data were not considered to accurately represent the microclimatic conditions associated with disease development. However, if one wants to study either year-to-year variation in disease severity over the short term (5–20 yr) or how longterm climatic variation may affect disease occurrence, such data is the only available. Since the historical weather data collected under the NOAA-National Weather Service Cooperative Observer Program give long time series for numerous locations (Coakley, 1987), these were chosen for our investigation of the interaction between disease and climatic variation.

In the first phase of our research, we determined that the frequency and severity of stripe rust epidemics (caused by *Puccinia striiformis* West.) on winter wheat (*Triticum aestivum* L. em Thell) in the Pacific Northwest (PNW) varied in direct relationship to climatic variation (Coakley, 1978, 1979).

In the second phase, statistical models were developed to predict stripe rust severity on three cultivars at Pullman, WA (Coakley and Line, 1981a, b). These models were modified and subsequently used to predict disease severity at four other locations in the PNW (Coakley *et al.*, 1982). The Pullman models were not as accurate at the other locations, therefore, regional models were developed for each cultivar using data from four locations (Coakley *et al.*, 1983).

The regional models were successfully verified in 1983 when the model version based on both winter and spring temperatures correctly predicted disease intensity for the three cultivars at five locations (including one location that was not included in the model development) (Coakley and Line, 1984). These models are used in the Pacific Northwest by researchers and extension personnel to predict stripe rust and as a basis of guidelines for use of fungicides.

However, when the method of analysis for stripe rust was used to analyze

how climatic factors affect a different disease, Septoria tritici blotch (caused by *Mycosphaerella graminicola*) on wheat in Indiana, it was necessary to develop new procedures. The analytical approach for Septoria was to develop a method for exhaustively analyzing climate-disease interactions that would allow quantification of how climatic variation affects plant disease epidemics. A statistical model that predicted Septoria tritici blotch severity on the basis of temperature and precipitation variables resulted from the initial development of the computer program WINDOW (Coakley *et al.*, 1985).

This new approach for analyzing how climatic conditions affect disease occurrence is presented with an example of how the method was used to analyze stripe rust severity data and climatic data at Pullman, WA. Only data on one wheat cultivar is presented with the intent that this example will facilitate researchers in other disciplines in transferring the analysis method to their own data bases. Results were similar for the analysis of two other cultivars grown in the Pacific Northwest (Coakley *et al.*, 1988). The method includes a computer program WINDOW that identifies the climatic factors that are most highly correlated with the disease, the development of statistical models that show the relationship of climatic factors to plant disease severity, and validation of the models. This methodology should be applicable to investigations of the effect of climatic variation on populations of other organisms. Although our entire method is described, the unique part of it is the iterative program WINDOW we use to identify the meteorological factors most highly correlated with disease severity; the emphasis of our description of methods is on how WINDOW is used.

2. Data Base

Daily meteorological data for August 1967–July 1986 for Pullman, WA (latitude 46°46′ N, longitude 117°12′ W, elevation 775 m) were obtained from the National Climatic Data Center, Asheville, NC. The data were for minimum and maximum temperature, and total precipitation.

Stripe rust severity (percentage of the total leaf and glume surface covered by rust) and stage of growth were recorded for the winter wheat cultivar Gaines at Pullman, WA. A single value for disease severity along with infection type was recorded for each 1.5–3.0 m long row of wheat ($\geqslant$100 plants) at various stages of plant growth. Disease severity was converted to a 0–9 disease index (DI) (Table I). When data were not available for plant growth stage 8 (dough stage) (Zadoks *et al.*, 1974), the DI for growth stage 8 was estimated by extrapolation from the DI at growth stage 7 (milk stage), as described by Coakley *et al.*, (1983, Table III). For each year, data were collected at one to four locations in Pullman; the data in Table I represents an average of all locations available for a given year.

TABLE I. Stripe rust severity index on winter wheat cultivar Gaines at Pullman, Washington.

Year	Actual disease index[a]	Predicted disease index[b]	
		Model I	Model II
1968	6.50 E	4.98	4.77
1969	2.00	2.68	2.50
1970	3.00	2.79	2.99
1971	5.75	4.83	5.22
1972	4.00 E	4.82	4.84
1973	3.50	2.73	2.34
1974	3.00	2.01	2.56
1975	6.25	5.89	5.82
1976	6.50	5.94	5.94
1977	0.00	2.10	1.69
1978	6.25 E	6.41	5.90
1979	3.00	1.84	2.18
1980	5.50	5.42	5.04
1981	7.50 E	7.44	7.79
1982	2.00	3.71	3.57
1983	5.67	6.33	6.68
1984	6.38	6.88	6.95
1985	2.00	1.52	2.27
1986	2.50	6.44	5.48
Mean	4.52	4.49	4.50

[a] Disease index (DI) was recorded at growth stage 8 (dough stage) except where indicated by the letter "E"; those were recorded at stage 7 and extrapolated for stage 8 as described in Coakley *et al.*, 1983. The 0–9 scale disease index (DI) is based on converting percent disease intensity to DI where 0 = 0% disease, 1 = <1%, 2 = 1–5%; 3 = 6–20%, 4 = 21–40%, 5 = 41–60%, 6 = 61–80%, 7 = 81–95%, 8 = 96–99%, and 9 = >99%.

[b] Predictions were made with Gaines models I and II described in Table IV.

3. Identifying Limiting Climatic Factors [WINDOW]

WINDOW is a Fortran program developed during this research project to identify the climatic variables that are most highly correlated with disease data. This section describes the way in which WINDOW is used to analyze for correlation between disease and climatic factors.

3.1. *Selection of Climatic Factors*

The number of meteorological variables that can be considered at one time is flexible. Variables considered for all diseases include: precipitation frequency, total precipitation, total consecutive days with precipitation (CDWP), total consecutive days without precipitation (CDWOP), mean maximum, mean mini-

mum, and mean average temperature. Other variables are added depending on the disease being studied. For analysis of stripe rust data, the following were also used: accumulation of positive and negative degree days from a 7 °C base (that temperature is the optimum for both *in vivo* germination of urediospores and infection of the wheat plant (Sharp, 1965); degree days were calculated as described (Coakley and Line, 1981a)); total consecutive days with minimum temperature less than 7 °C; total days (TD) with average temperature less than 0 °C; and TD with maximum temperature greater than 25 °C.

3.2. *Selection of Time Periods to Examine*

Day of Year.
In WINDOW, all references to calendar dates are made to day of the year (DY), in which 1 January = DY 1 and 31 December = DY 365, except in leap years when it equals DY 366 (Stone, 1983).

Beginning Date and Duration
For this analysis, the starting date was 29 July (DY 210); ending date was on 28 July of the following year, prior to harvest in August at Pullman. Each year, the plots at Pullman were planted in the first two weeks of October.

Length of Window
The next step is to set the number of days (Window length) for which the climatic data are averaged or summed; the WINDOW program can look simultaneously at nine subsets of each Window set, the first subset being the full-length Window and the other eight being progressively smaller subsets (Figure 1). During development of WINDOW, Window lengths of 105 to 25 days were examined, but Windows longer than 65 days did not substantially improve results. Window subsets are initially set from 65 to 25 days in length, with each subset five days shorter than the previous one (e.g., Figure 1, Window P). In subsequent runs, the Window subsets may be set only one day shorter than the previous one in order to identify the time period most highly correlated with disease severity. Figure 1 shows an example of how the data is sequentially examined: the first Window begins on DY 5, and ends on DY 69; data arrays are built for each of the nine subsets of Window A for each of the 12 climatic variables selected and then the Window is advanced. Data arrays are then built for the subsets of Window B which begin on DY 10 and end on DY 74. This is repeated until data arrays are built for Windows A–P. The amount that each Window is advanced is a variable. We use increments of one to five days, with five days used initially. In Figure 1, the last Window (P) begins on DY 80, and ends on DY 144. In an actual analysis, WINDOW is used to examine the data for the growing season in three segments. In Segment I, the first Window begins on DY 210 and the last begins on DY 365. In Segment II (part of which is

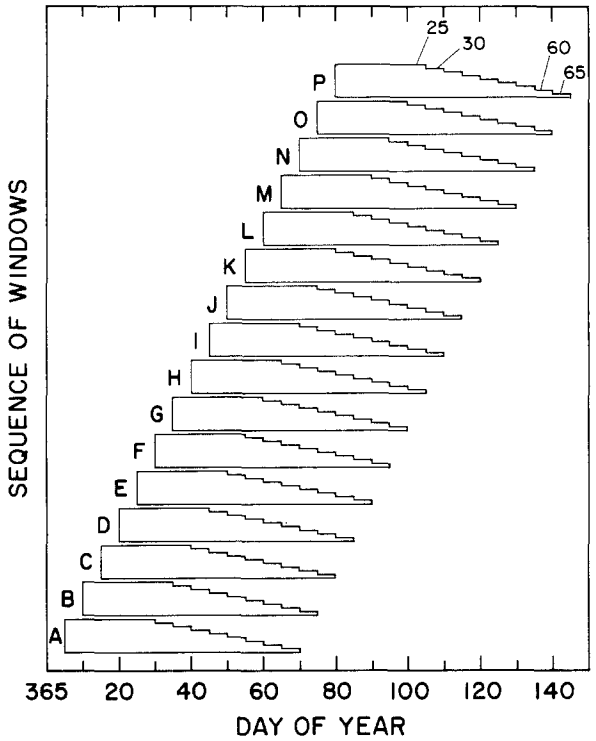

Fig. 1. Example of how meteorological data are considered by the WINDOW program. Window A has nine subsets; each start on DY 05 and are 65, 60, 55, ..., 25 days in length. After the meteorological data are assembled for Window A, the Window is advanced 5 days to Window B, where all subsets begin on DY 10. This is repeated for Windows C–P; Window P begins on DY 80 and ends on DY 144.

shown in Figure 1), the first set begins on DY 5 and the last set begins on DY 140. In Segment III, the first set begins on DY 145 and the last set begins on DY 205.

Defining Parameters

A list of parameters is specified for each run of WINDOW and allows the researcher to readily change the conditions of analysis. Table II lists those used for the analysis of the data on Gaines. The parameter list is where the base temperature for calculating degree days (e.g. negative degree days are calculated from a base of 7°), and the condition for counting the number of days less than or greater than some variable are set.

Building Data Arrays

Data arrays are built for each year for each Window's nine subsets according to the parameters and selected variables. For each variable, WINDOW either calculates a mean (e.g., mean maximum temperature), counts (e.g., frequency of

TABLE II: Example of a list of parameters to be set in WINDOW program.

Parameter	Example of current set value
Number of years in file	NYIF = 17
Number of variables	NFAC = 12
Beginning Window Date	IBW = 005
Last Window Date	LW = 140
Increment of Window	INC = 05
First year in file	IFIRST = 68
Last year in file	LAST = 84
Degree day base	POS = 7
Breakpoint for temperature	
Number of days < 0 °C	TDAVT = 0
Number of days < 7 °C	TDMNT = 7
Number of days > 25 °C	TDMXT = 25
Number of years with missing disease data	MISSIN = 0
Number of Subsets in each Window	NOFW = 09
Lists Value of each factor for each Window and Subset	LIST = NYIF + 6
Label for Listing	PLACE = 'PULLMAN'
Cultivar	CULTI = 'GAINES'

precipitation), or sums a cumulative total (e.g., positive degree days). The method of counting consecutive days is based on Shaner and Finney (1976), and only sequences of two or more days that meet the specified criteria are considered; that is, two such consecutive days count as one period, three consecutive days count as two periods, etc. These periods are then summed for the Window subset being examined; for example, in a subset of 25 days with sequences of 5, 3, and 4 days without precipitation, consecutive days without precipitation (CDWOP) would be counted as 4, 2, and 3 days, respectively, for a sum of 9 CDWOP.

Correlation of Disease Data with Meteorological Variables

Once the data arrays are built for the meteorological variables and the disease data, correlation analysis is done by WINDOW to determine any relationships between the two sets of data. Although a listing could be made for the value of each factor for each Window and its subsets, we conserve paper and our time by printing out only data arrays for Windows with a correlation coefficient that is significant at $P \leqslant 0.05$ for at least one variable. The results from an analysis are examined for a pattern of increasing and then decreasing correlation coefficients as the Window advances, and correlation coefficients with $P \leqslant 0.01$ are selected first. Figure 2 gives an example of a print-out for the Total Precipitation (TPREC) factor which begins on DY 75 (This is Window set O in Figure 1). In Figure 2, a peak area occurs in the Window subset that begins on DY 75 and ends 25 days later (on DY 99). TPREC is correlated ($r = 0.71$) with disease severity at $P \leqslant 0.01$. A similar correlation exists between disease and TPREC for the Window subset that begins on DY 70 and ends on DY 94.

******************** TOTAL PRECIPITATION ********************

THIS WINDOW START IS 75 AND WILL BE INCREMENTED BY 5 THE LAST WINDOW START IS 140 FIRST WINDOW WAS 5

DURATION OF WINDOW	65	60	55	50	45	40	35	30	25
	11.69	11.69	8.89	8.51	8.21	7.41	6.85	4.84	3.78
	3.80	3.65	3.60	3.60	3.59	3.43	3.42	3.34	3.26
	10.12	9.05	9.04	9.04	8.99	7.04	5.26	4.37	2.92
	5.10	5.10	5.10	4.21	4.21	3.73	3.20	2.41	2.31
	12.39	9.43	8.33	8.28	8.28	8.28	7.24	6.86	6.68
	9.39	8.30	8.16	4.84	4.84	4.82	4.43	4.22	3.33
	3.83	3.83	3.62	2.10	1.46	1.46	1.46	0.96	0.96
	9.68	8.81	7.06	7.06	7.06	6.43	5.99	5.69	4.34
	9.86	9.84	8.88	8.60	7.94	7.76	6.49	5.52	5.19
	10.82	10.82	10.08	10.08	10.03	9.98	8.76	6.91	4.88
	5.00	3.78	2.74	2.44	1.40	1.40	1.40	1.17	0.66
	10.89	8.81	7.90	7.80	7.72	6.48	5.84	4.49	4.39
	14.11	14.10	14.10	11.69	10.06	9.81	7.14	6.44	5.28
	9.44	8.02	6.52	5.94	4.41	4.41	4.13	3.85	2.98
	14.55	12.94	12.94	12.73	11.31	11.08	9.22	9.22	8.76
	9.34	8.27	8.26	8.24	7.75	7.22	7.22	7.15	2.65
	9.91	9.30	9.27	5.96	5.95	5.78	4.74	4.74	4.46
	11.39	10.86	10.63	10.24	8.82	8.82	8.12	7.92	6.83
SUB SET MEAN	9.39	8.53	8.01	7.23	6.70	6.35	5.53	5.02	4.11
SUB SET STD DEVIATION	3.14	2.92	3.01	3.03	2.87	2.77	2.27	2.22	2.03
CORRELATION	0.38	0.38	0.36	0.39	0.41	0.47	0.51	0.49	0.71
SIGNIFICANCE LEVEL	1.00	1.00	1.00	1.00	1.00	1.00	0.05	0.05	0.01

Fig. 2. Example of a print-out that shows the correlation coefficients between total precipitation (TPREC) and disease severity for a Window that begins on DY 75 and has subsets 65, 60, ..., 25 days long (DURATION OF WINDOW). The first 18 lines of data are TPREC (cm) for each Window subset for the years 1967 to 1984, respectively. The mean and standard deviation for the 18 years are given as SUB SET MEAN and SUB SET STD DEVIATION. SIGNIFICANCE LEVEL (P) is printed out for $P \leqslant 0.05$. The highest correlation between TPREC and disease severity is 0.71 ($P \leqslant 0.01$) for the Window subset 25 days in length.

********** TOTAL PRECIPITATION **********

THIS WINDOW START IS 73 AND WILL BE INCREMENTED BY 1 THE LAST WINDOW START IS 79 FIRST WINDOW WAS 71

DURATION OF WINDOW	29	28	27	26	25	24	23	22	21
	4.07	4.07	4.04	4.04	4.04	4.04	4.03	4.03	4.02
	3.59	3.59	3.59	3.21	2.96	2.96	2.88	2.78	2.78
	4.24	4.24	2.92	2.92	2.92	2.92	2.03	2.03	2.03
	3.63	3.63	3.53	3.00	3.00	3.00	2.67	2.67	2.67
	6.92	6.74	6.74	5.60	5.60	5.60	5.12	5.12	5.12
	5.36	5.36	5.36	5.14	5.13	4.87	3.75	3.75	3.75
	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.95
	5.69	4.39	4.34	3.94	3.94	3.94	3.91	3.40	3.28
	5.19	5.19	5.19	4.96	4.95	4.94	4.58	4.43	4.15
	5.04	5.04	4.89	4.81	4.81	4.66	4.66	4.66	4.66
	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
	4.49	4.49	4.49	4.49	4.49	3.81	3.33	3.25	2.59
	5.61	5.28	5.28	5.07	5.07	3.93	3.93	3.93	3.92
	5.24	4.78	4.38	4.37	4.22	3.99	3.99	3.99	3.99
	8.77	8.77	8.76	7.49	7.48	7.47	7.03	7.03	7.01
	4.35	3.69	3.69	3.69	3.64	3.16	3.03	2.98	2.93
	5.59	5.48	5.31	5.31	5.31	5.31	5.31	5.31	5.31
	8.78	8.27	8.22	7.54	7.53	6.61	6.51	6.50	6.50
SUB SET MEAN	4.95	4.74	4.61	4.30	4.28	4.05	3.79	3.73	3.67
SUB SET STD DEVIATION	2.08	2.02	2.05	1.80	1.81	1.70	1.67	1.67	1.68
CORRELATION	0.63	0.68	0.70	0.71	0.70	0.73	0.75	0.75	0.72
SIGNIFICANCE LEVEL	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01

Fig. 3. Example of the print-out for TPREC in which the precise time period is identified that gives the highest correlation with disease severity. See the legend for Figure 2 for the explanation of the print-out, except that this Window begins on DY 73 and subsets are $29, 28, \ldots, 21$ days in length. The subsets that end 22 and 23 days later ($r = 0.75$, $P \leq 0.001$) are the most highly correlated with disease severity. A significance level of 0.00 represents $P \leq 0.001$.

The next step is to identify the most precise time period for a variable that will give the highest correlation with disease severity. To do this for our example in Figure 2, the first Window in the next analysis was set to begin on DY 71, the

TABLE III: Correlation between meteorological factors and stripe rust index on Gaines wheat at Pullman, Washington. The correlation coefficients (r) are significant at $P \leqslant 0.01$ except when 'a' follows the r, $P \leqslant 0.001$.

Factor[a]	Window increment					
	5-day			1-day		
	Begin Date[b]	Length	r	Begin Date	Length	r
MMAX	365	(25)	0.72	004	(21)	0.71
	115	(65)	−0.73a			
MMIN	360	(25)	0.67			
	120	(65)	−0.62			
MAVE	360	(25)	0.70			
	120	(65)	−0.62			
PDD	115	(65)	−0.76a			
NDD	360	(25)	−0.70			
DLOC	360	(25)	−0.79a	363	(22)	−0.81a
				001	(24)	−0.74a
DG25C	115	(65)	−0.85a			
TPREC	75	(25)	0.71	073	(23)	0.75a
PFREQ	65	(50)	0.66	069	(49)	0.67
	95	(60)	0.69	073	(28)	0.66
	140	(40)	0.66	079	(69)	0.68
				080	(38)	0.64
				095	(59)	0.70
CDWP	65	(50)	0.63			
	80	(65)	0.62			
	95	(60)	0.68			
	140	(40)	0.66			
CDWOP	65	(60)	−0.64			
	80	(35)	−0.63			
	140	(45)	−0.62			

[a] Factors are: MMAX = mean maximum temperature in °C; MMIN = mean minimum temperature; MAVE = mean average temperature; PDD = positive degree days; NDD = negative degree days; DLOC = total days that the average temperature was < 0 °C; DG25C = total days that the maximum temperature was > 25 °C; TPREC = total precipitation in cm; PFREQ = total number of days with precipitation; CDWP = total consecutive days with precipitation; and CDWOP = total consecutive days without precipitation.

[b] The Window began on this day of the year and was (x) days in length; e.g. MMAX 365 (25) was a Window that took the mean maximum temperature for 25 days beginning on day of year 365 (31 December) and ending on 24 January.

last Window to begin on DY 79, and the Window increment to advance 1 day at a time. The length of the Window subsets were set at 21, 22, 23, ..., and 29 days to span the 25-day subset which had the peak correlation in the previous analysis. Figure 3 gives the array with the highest correlation for TPREC; the Window begins on DY 73 and is 23 days long ($r = 0.75$, and $P \leqslant 0.001$). The actual values of TPREC are printed out for each year along with the mean and standard deviation for all years with disease data (Figure 3).

When the Windows were advanced in 5-day increments, twenty-one variables were found to be correlated ($P \leqslant 0.01$) with disease severity (Table III). From Table III, variables were selected for multiple regression analysis to determine the mathematical form of the relationship between the meteorological variables and disease severity. Selection was restricted to those factors that ended by DY 155 (4 June). This restriction was made to develop models for use in predicting disease in time for application of chemicals for control if needed. After selection, the exact Window was identified for each variable by using Windows that were advanced in one-day increments (Table III); the variables listed on the right-hand side of Table III were used in regression analysis.

4. Model Development

Following selection of the meteorological factors to be included in model development, both meteorological and disease data are evaluated for normality to ensure that parametric methods are appropriate. Since the data for the stripe rust analysis were normally distributed, parametric methods were used. Variables are plotted against time to determine whether disease or meteorological factors are time dependent; no relationship between time and variables was found in the analysis of the stripe rust data.

4.1. *Regression Analysis*

The discussion of the regression analysis is presented in a brief form because this part of our method is not unique to our research. However, because this method is often misused, we have included what we believe are the more important points to be considered in a study such as ours. Consulting with a statistician is advisable if a researcher is not familiar with regression analysis.

Regression analysis is used to determine if a linear relationship exists between the variables identified by WINDOW (independent x-variables) and disease at a specified time (dependent y-variable). The Statistical Analysis System (SAS) procedures are used for the analysis and include REG, RSQUARE, and STEPWISE (SAS, 1985). The meteorological x-variables used are listed in Table III. The disease index for the previous year was included as an x-variable; however, the correlation between the previous year's (PYS) and the current year's disease severity was low and since the inclusion of PYS in models did not significantly

improve disease predictions, PYS was not further considered. The dependent variable (y) was disease index (Table I) for 1968 through 1984. Data for 1985 and 1986 were used for model validation. The SAS procedure RSQUARE is used to evaluate all possible models up to a maximum of three independent variables. A greater number of variables could have been used but we choose to limit the number to ensure simple equations that make biological sense. Because of the relatively few years of data ($n = 17$), it is important to limit the number of x-variables. The number is also limited because meteorological variables are often not independent of each other and it is necessary to limit a given model to ones that are not highly correlated with each other. RSQUARE provides the R^2 and adjusted-R^2 for each of the models evaluated. For Gaines, 175 models were considered (120 three-variable, 45 two-variable, and 10 one-variable models). The SAS procedure STEPWISE uses four regression methods for generating models: forward, backward, stepwise and maximum R^2 improvement. The best models from STEPWISE are evaluated along with those listed from RSQUARE.

4.2. *Model Evaluation*

A number of two- and three-variable models with the high adjusted-R^2 are selected for evaluation according to the rules we set. We choose to examine each of the models rather than to set default limits in the program because we believe that this way we are able to select the most useful model which may not have the highest R^2. The variable identities are considered, and models that have two overlapping or highly correlated variables are excluded from further evaluation. For example, one three-variable model included both DLOC 001 and MMAX 004; these two variables are both measures of temperature for time periods that overlap and are highly correlated with each other ($r = 0.91$). A correlation matrix is printed out to allow evaluation of correlation between all combination of x- and y-variables.

The SAS procedure REG is used to develop the models that are selected. The models are evaluated for their performance in prediction for the years included in the model development. The mean disease index for Gaines was 4.52, whereas the mean predicted disease index was 4.49. The standard errors of the predictions are examined in order to select models that minimize these errors. Regression coefficients are examined for stability of sign; if the sign contradicts what is known about the biology of the pathogen, the variable may be excluded. Studentized residuals are plotted against predictions and time. The plots are examined for patterns, trends, or clustering. Ideally, the residuals will appear as a random scatter plot. Non-random residuals can be used as a diagnostic of model deficiencies, e.g., whether a transformation or a quadratic factor is needed, or whether non-linear regression techniques are appropriate (Armitage, 1971; Daniel and Wood, 1980; Montgomery and Peck, 1982). For the Gaines

TABLE IV: Models for predicting stripe rust disease index ($\hat{y}$) on Gaines winter wheat at Pullman, WA. Regression coefficients (β), meteorological variables[a] (X), adjusted-R^2 (Adj-R^2), standard error of β [$s(\beta)$], Variation Inflation Factor (VIF), and Range of (X) for each model are listed.

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$					
Gaines ($n = 17$)					
I $\hat{y} = -2.188 + 0.270$[PFREQ 095] $+ 0.454$[MMAX 004]					Adj-$R^2 = 0.75$
$s(\beta)$	1.413	0.064	0.106		
VIF	0	1.078	1.078		
Range	(X)	16–31 days	−4.23–6.25 °C		
II $\hat{y} = -1.334 + 0.182$[PFREQ 095] $+ 0.407$[MMAX 004] $+ 0.311$[TPREC 073]					Adj-$R^2 = 0.76$
$s(\beta)$	1.546	0.095	0.110	0.250	
VIF	0	2.440	1.224	2.748	
Range	(X)	16–31 days	−4.23–6.25 °C	0.69–7.03 cm	

[a] Meteorological variables are defined in Table III.

models selected (Table IV), plots of the residuals appeared as random scatter plots.

The variance inflation factor (VIF) measures the effect of multicollinearity between variables on the variances of estimated coefficients and is another measure of model stability. If a VIF > 5, the associated regression coefficients are poorly estimated (Draper and Smith, 1981; Montgomery and Peck, 1982).

The models are also evaluated on the basis of the accuracy of the predictions made. For example, whenever the stripe rust disease index is >5.5, disease is severe and the economic feasibility of the application of chemical control would

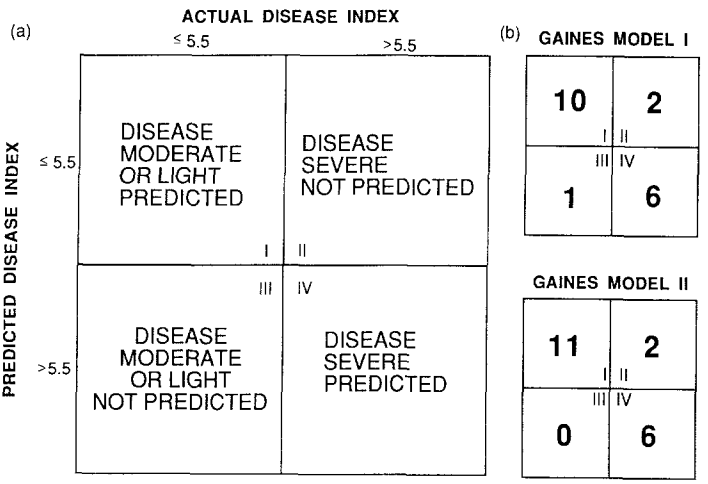

Fig. 4(a). A contingency table used to evaluate the accuracy of disease predictions relative to actual disease index. In quadrant I and IV, actual disease and predicted disease occurrence are in agreement. In quadrant II, an underprediction of disease occurs; in quadrant III, an overprediction of disease is made. (b) Actual number of years which fall in each of four quadrants (I, II, III, IV) defined in Figure 4a. Model I and II correspond to the two- and three-variable model equations given in Table IV.

be considered on a case-by-case basis. Whenever the disease index is $\leqslant 5.5$, disease severity is moderate or light, and application of chemical control would be less cost-effective. The models are evaluated using a contingency table (Figure 4a) for how accurately they predict severe disease.

Based on all criteria, one two- and one three-variable model were selected as giving the best predictions of disease index (Table IV).

5. Model Validation

Model validation is essential for determining how it will function in its intended use. Techniques used include: (1) analysis of model coefficients and predicted values; (2) data splitting (Draper and Smith, 1981; Montgomery and Peck, 1982; Snee, 1977), and (3) collection of new data to check model predictions.

Examination of the regression coefficients and standard errors of the estimates (Table IV) shows only how well the model predicts the data set from which it was developed (Teng, 1981, 1985). All coefficients had VIF's less than 2.75 (Table IV), which indicates that the coefficients were properly estimated and stable.

Allen's PRESS (Predicted Error Sum of Squares) statistic is used for model validation. This statistic is a form of data splitting (Draper and Smith, 1981) and can be calculated using the SAS procedure REG. PRESS is calculated in the following way: An observation, for example i, of n data points is deleted and the regression model is fitted to the remaining $n-1$ data points. This model is used to predict the withheld observation y_i which is then called the predicted $\hat{y}_{(i)}$. The prediction error for this point i is $e_{(i)} = y_i - \hat{y}_{(i)}$. The first observation is returned to the data set and the procedure is repeated for each observation $i = 1, 2, \ldots, n$ resulting in a set of n deleted residuals $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$. The PRESS statistic is thus defined by Montgomery and Peck (1982) as the sum of squares of the n deleted residuals as in

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2. \tag{1}$$

The PRESS statistic was calculated for each regression model evaluated. An example is given for Gaines model I (Table V). The magnitudes and signs of the β-coefficients are stable for all models. The prediction error for each year shows how well the model based on 16 yr of data predicts for that year.

The models were also validated by making predictions for 1985 and 1986, years that were not included in model development, and comparing the predicted disease ($\hat{y}$) with observed disease (y) (Table I). The models accurately predicted disease index on all cultivars in 1985. In 1986, the two-variable Gaines model predicted a disease index that was significantly higher than that which occurred. A significant overprediction is when severe disease is predicted,

TABLE V: Calculation of Allens's PRESS statistic for the model $\hat{y} = -2.187 + 0.270X_1 + 0.454X_2$, where $\hat{y}$ = predicted disease index X_1 = precipitation frequency for 5 April to 2 June, and X_2 = mean maximum temperature for 4 to 24 January; y = observed disease index.

Year	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$	$\hat{y}=\beta_0+\beta_1(X_2)+\beta_2(X_2)$[a]
1968	6.50	4.73	1.77	3.13	−2.670 + 0.291(20) + 0.408(3.87)
1969	2.00	2.77	−0.77	0.59	−1.986 + 0.264(19) + 0.441(−0.61)
1970	3.00	2.74	0.26	0.07	−2.327 + 0.276(16) + 0.452(1.43)
1971	5.75	4.77	0.98	0.96	−2.269 + 0.272(22) + 0.447(2.35)
1972	4.00	4.89	−0.89	0.79	−2.031 + 0.265(21) + 0.468(2.91)
1973	3.50	2.60	0.90	0.81	−2.588 + 0.286(17) + 0.455(0.71)
1974	3.00	1.80	1.20	1.44	−2.573 + 0.283(18) + 0.483(−1.48)
1975	6.25	5.85	0.40	0.16	−2.130 + 0.267(25) + 0.451(2.89)
1976	6.50	5.87	0.63	0.40	−2.156 + 0.268(24) + 0.444(3.59)
1977	0.00	2.53	−2.53	6.40	−1.132 + 0.232(17) + 0.418(−0.69)
1978	6.25	6.43	−0.18	0.03	−2.209 + 0.272(25) + 0.458(4.04)
1979	3.00	1.09	1.91	3.65	−2.092 + 0.253(22) + 0.564(−4.23)
1980	5.50	5.40	0.10	0.01	−2.143 + 0.268(28) + 0.457(0.08)
1981	7.50	7.42	0.08	0.01	−2.160 + 0.269(28) + 0.453(4.52)
1982	2.00	3.88	1.88	3.53	−2.124 + 0.275(22) + 0.423(−0.11)
1983	5.67	6.61	−0.94	0.88	−1.959 + 0.259(21) + 0.500(6.25)
1984	6.38	7.15	−0.77	0.59	−2.682 + 0.296(31) + 0.443(1.49)
			PRESS = 23.45		

[a] Coefficients are estimated for each year based on $n-1$ observations (16 yr). For example, in 1968, observations for 1969 to 1984 were used to estimate the coefficients and the resulting equation was used to predict disease index ($\hat{y}$) for 1968.

but only light disease occurred; in contrast, a significant underprediction would be one in which light disease was predicted but severe disease occurred. The 1985 and 1986 x-variables were examined for each model and compared with the range given in Table IV; all fell within the range of values for the x-variables used in model development.

Figure 4b gives the summary of the predictions made for all years using Models I and II. Based on actual (y) and predicted ($\hat{y}$) disease, the upper left quadrant (I) represents when both y and $\hat{y}$ were $\leqslant 5.5$ and disease was light or moderate. Quadrant II represents when severe disease was not predicted, but it did occur. Quadrant III is when severe disease was predicted by actual disease was light or moderate. Quadrant IV indicates the years in which severe disease was predicted and occurred ($\hat{y}$ and $y > 5.5$); in these cases, chemical control would have been recommended and justified on the basis of actual disease index. The models for stripe rust were evaluated for accuracy without considering the relative values of y and $\hat{y}$. For example, with Gaines model II, in one year the actual y was 5.75 whereas $\hat{y}$ was 5.22; although this underprediction is reflected in Quadrant III of Gaines I (Figure 4b), this small difference would probably be negligible with respect to the decision to control or not control disease.

On the basis of adjusted-R^2, the three-variable models showed only a slight improvement over the two-variable models. However, the only advantage of

using a two-variable rather than a three-variable model is that there is one less variable to calculate, a minimal advantage unless the variables have to be calculated by hand rather than by computer.

On a statistical basis, misses in predictions are expected whenever regression models are used; however, the reason for the misses and the nature of the misses are important. The models presented were built on meteorological variables that occurred early enough in the growing season to be useful in making predictions at a time when disease control is still possible. The incorrect predictions of most concern were those in which the actual disease was higher than that predicted – in theory, a grower could suffer greater economic losses by not applying control when it is needed than by applying control that is not needed. To improve predictions, all variables under 5-day increments in Table III – regardless of when they ended – were subsequently considered in development of a model for Gaines. A large improvement was obtained by including a factor for the total number of days that the maximum temperature was greater than 25 °C (DG25C). The three-variable model for Gaines that includes DG25C has the formula:

$$\hat{y} = 5.940 - 0.256\ [\text{DG25C } 113] + 0.309\ [\text{MMAX } 004] + 0.039\ [\text{PFREQ } 80] \quad (2)$$

where DG25C is summed for 66 days beginning DY 113. The Adjusted-R^2 is 0.88, as compared with 0.76 for the early-season, three-variable Gaines model (Table IV). The predictions for 1985 and 1986 were 3.69 and 2.96, respectively, and were accurate predictions for the 2.0 and 2.5 disease indices that actually occurred. When this new model was evaluated for percent accuracy, 11 yr fell in quadrant I and the remaining 8 yr fell in quadrant IV. There were no cases in which disease was significantly over- or underpredicted. Unfortunately, because DG25C depends on data from 23 April to 27 June, this model has limited value in making predictions for disease control unless a reliable long-term temperature forecast becomes available for the month of June.

6. Discussion and Conclusion

The method presented has been developed and tested by analyzing data for two foliar diseases of winter wheat caused by two different fungal pathogens. Most of the variation in disease severity that occurs from year to year can be explained by variation in climate. The models we developed can be used to predict disease severity early enough in the growing season for chemical control when economically beneficial. Economical factors that a grower must consider include the size of field, the potential yield reduction caused by the disease, and cost of application relative to the projected value of the wheat crop.

This research differs most significantly from previous studies on the effects of climatic conditions on disease in the way in which the meteorological data are

analyzed. In our earlier studies (1978 to 1983), meteorological data was examined on a monthly, seasonal, or in some cases, weekly basis. However, there is no biological reason for an organism to respond to climatic data on a calendar basis, and it is not surprising that attempts to relate biological events to such climatic data have been frequently futile. Although stripe rust occurrence was quantified on this basis, attempts to relate Septoria leaf blotch to monthly or seasonal data failed. The method that we developed (Program WINDOW) to analyze the meteorological data in variable-length segments relative to disease severity adds a new dimension to the possibility of quantifying the populations or productivity of many different organisms relative to climatic conditions. Such quantifications are especially valuable for considering the possible effects that climatic variation may have on the occurrence of an organism.

A long-term objective of our research has been to develop the methods necessary to evaluate the effects of past and future climatic variation on the occurrence of important diseases on agricultural crops. Earlier attempts to develop a regional model for stripe rust were successful (Coakley *et al.*, 1984) and research is currently underway to develop a general method for regionalizing statistical models such as we have described in this paper.

Where data are available for non-meteorological variables, e.g., planting date, emergence date, heading date, fertilizer, irrigation etc., it would be appropriate to include these factors in model development. It is important that meteorological data for an entire growing season, and perhaps longer, be analyzed to ensure that any significant meteorological factors are identified. It is probable that organisms with life cycles of a year or less will reflect changes in climatic conditions more readily than those that mature over several years. This makes studies of pathogens and insects particularly attractive because there are multiple life cycles in a single growing season of the host, and these cycles are frequently limited by climatic conditions. The methodology described should also be applicable to quantification of the most important meteorological conditions necessary for maximum crop yields.

In the development of an equation to show the relationship between dependent and independent variables, using an infinite number of data sets would give the most robust model. However, only a finite number of years are available and the size of the sample must be considered when evaluating the model developed. An equation may fit a few data points very closely as indicated by a high adjusted-R^2, but when tested on new data, it may not accurately predict disease. Our results suggest that a minimum of 8 years of data be used for model development. For locations in the Pacific Northwest with eight to ten years disease data, adjusted-$R^2 > 0.90$ for models are common, but the models frequently do not accurately predict disease for years not included in model development (Coakley and McDaniel, unpublished). Statistical models can be used validly only when then new variables are within the range of the variables included in model development. With only a few years of data, there is a greater chance that

future years will have conditions outside those used in model development. Predictions may be accurate outside the range of variables included in model development, but this must be constantly monitored. If extreme events occur and model predictions are inaccurate, it may be necessary to reformulate the models using the new data.

Two- and three-variable models were compared using the adjusted-R^2 statistic rather than R^2 because the adjusted-R^2 takes into account the number of variables in the model. R^2 always increases when a variable is added even when the variable does not contribute to improving the model. Comparison of models with different numbers of variables is facilitated by use of adjusted-R^2.

The models described herein are similar to earlier models for stripe rust (Coakley *et al.*, 1982, 1984) in that both have a winter temperature factor comparable to that used in the earlier models, as well as a spring precipitation factor.

Gaines model I and II explain, respectively, 75 and 76% of the variation in disease from year to year, which are essentially the same as that explained by the single factor model described in Coakley *et al.* (1982). To evaluate whether these models were an improvement over the one previously used in the PNW (Coakley *et al.*, 1983), predictions for years not included in model development (1973, 1975, 1978, and 1981) were made using the Gaines model described in Coakley *et al.* (1983) and the Gaines model I (Table IV). When Figure 4a was used to evaluate the predictions, the Gaines model I correctly predicted whether or not disease was severe in all four years. The earlier model correctly predicted whether or not disease was severe in only two of the years. Hence, the new models developed using WINDOW are considered to be superior to those previously used.

In conclusion, the method described here should be of general use for analyzing how climatic conditions affect disease occurrence. They should also be useful in evaluation of how climatic conditions affect other organisms, such as needed if accurate assessment is to be made of how projected climatic variation will affect future agricultural production.

Acknowledgments

Acknowledgment is made to the National Center for Atmospheric Research Boulder, CO 80307, U.S.A. which is supported by the National Science Foundation, for supplying computer time and research space for this research. The authors thank Nancy Mielinis for editorial review of the manuscript.

References

Armitage, P.: 1971, *Statistical Methods in Medical Research*, Blackwell Scientific Publ., Oxford, 504 pp.

Coakley, S. M.: 1978, 'The Effect of Climate Variability on Stripe Rust of Wheat in the Pacific Northwest', *Phytopathology* **68**, 207–212.

Coakley, S. M.: 1979, 'Climate Variability in the Pacific Northwest and its Effect on Stripe Rust Disease of Winter Wheat', *Climatic Change* **2**, 33–51.

Coakley, S. M.: 1985, 'Describing and Quantifying the Environment', *Plant Dis.* **69**, 461–466.

Coakley, S. M.: 1987, 'Historical Weather Data: Its Use in Epidemiology', in K. Leonard and W. Fry, (eds.), *Plant Disease Epidemiology*, Vol. II, Macmillan Publishing Co. (in press).

Coakley, S. M. and Line, R. F.: 1981a, 'Climatic Variables that Control Development of Stripe Rust Disease on Winter Wheat', *Climatic Change* **3**, 303–315.

Coakley, S. M. and Line, R. F.: 1981b, 'Quantitative Relationships Between Climatic Variables and Stripe Rust Epidemics of Winter Wheat', *Phytopathology* **71**, 461–467.

Coakley, S. M. and Line, R. F.: 1984, 'Validation of Regional Models for Predicting Stripe Rust on Winter Wheat', *Phytopathology* **74**, 871–872 (abstr.).

Coakley, S. M., Boyd, W. S., and Line, R. F.: 1982, 'Statistical Models for Prediction of Stripe Rust on Winter Wheat in the Pacific Northwest', *Phytopathology* **72**, 1539–1542.

Coakley, S. M., Boyd, W. S., and Line, R. F.: 1984, 'Development of Regional Models that Use Meteorological Variables for Predicting Stripe Rust Disease on Winter Wheat', *J. Climate and Appl. Meteor.* **23**, 1234–1240.

Coakley, S. M., Line, R. F., and Boyd, W. S.: 1983, 'Regional Models for Predicting Stripe Rust on Winter Wheat in the Pacific Northwest', *Phytopathology* **73**, 1382–1385.

Coakley, S. M., Line, R. F., and McDaniel, L. R.: 1988, 'Predicting Stripe Rust Severity on Winter Wheat Using an Improved Method for Analyzing Meteorological and Rust Data', *Phytopathology* (in press).

Coakley, S. M., McDaniel, L. R., and Shaner, G.: 1985, 'Model for Prediction of *Septoria tritici* Blotch Severity on Winter Wheat', *Phytopathology* **75**, 1245–1251.

Daniel, C. and Wood, F. S.: 1980, *Fitting Equations To Data: Computer Analysis of Multifactor Data*, John Wiley & Sons, N.Y., 458 pp.

Draper, N. R. and Smith, H.: 1981, *Applied Regression Analysis*, John Wiley & Sons, New York.

Jones, A. L., Fisher, P. D., Seem, R. C., Kroon, J. C., and Van DeMotter, P. J.: 1984, 'Development and Commercialization of an In-field Microcomputer Delivery System for Weather-driven Predictive Models', *Plant Dis.* **68**, 458–463.

Montgomery, D. C. and Peck, E. A.: 1982, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York.

Rotem, J.: 1978, 'Climatic and Weather Influences on Epidemics', in J. G. Horsfall and E. B. Cowling (eds.), *Plant Disease, How Disease Develops in Populations*, Academic Press, New York, Vol. 2, pp. 317–337.

SAS Institutes Inc.: 1985, *SAS User's Guide: Statistics Version 5, 1985* Edition, Cary, NC.

Shaner, G. and Finney, R. E.: 1976, 'Weather and Epidemics of Septoria Leaf Blotch of Wheat', *Phytopathology* **66**, 781–785.

Sharp, E. L.: 1965, 'Prepenetration and Postpenetration Environment and Development of *Puccinia striiformis* on wheat', *Phytopathology* **55**, 198–203.

Snee, R. D.: 1977, 'Validation of Regression Models: Methods and Examples', *Technometrics* **19**, 415–428.

Stone, J. F.: 1983, 'On Julian Day Notation for Meteorological Conditions', *Agric. Meteor.* **29**, 137–140.

Sutton, J. C., Gillespie, T. J., and Hildebrand, P. D.: 1984, 'Monitoring Weather Factors in Relation to Plant Disease', *Plant Dis.* **68**, 78–84.

Teng, P. S.: 1981, 'Validation of Computer Models of Plant Disease Epidemics: A Review of Philosophy and Methodology', *Journal of Plant Diseases and Protection* **88**, 49–63.

Teng, P. S.: 1985, 'Construction of Predictive Models: II. Forecasting Crop Losses', in C. A. Gilligan (ed.), *Advances in Plant Pathology, Mathematical Modelling of Crop Disease*, Academic Press Inc. New York, Vol. 3, pp. 179–206.

Teng, P. S. and Rouse, D. I.: 1984, 'Understanding Computers: Applications in Plant Pathology', *Plant Dis.* **68**, 539–543.

Zadoks, J. C. and Konzak, C. F.: 1974, 'A Decimal Code for the Growth Stages of Cereals', *Eucarpia Bull.* 7, 12 pp.

(Received 17 April, 1987; in revised form 3 September, 1987)