MODELING THE OCCURRENCE OF FOUR CEREAL CROP APHID SPECIES IN IDAHO

John W. Merickel
Univerity of Idaho

Bahman Shafii
University of Idaho, bshafii@uidaho.edu

Sanford D. Eigenbrode
University of Idaho

Christopher J. Williams
University of Idaho

William J. Price
University of Idaho

See next page for additional authors

Recommended Citation

Merickel, John W.; Shafii, Bahman; Eigenbrode, Sanford D.; Williams, Christopher J.; and Price, William J. (2015). "MODELING THE OCCURRENCE OF FOUR CEREAL CROP APHID SPECIES IN IDAHO," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1008

Author Information

John W. Merickel, Bahman Shafii, Sanford D. Eigenbrode, Christopher J. Williams, and William J. Price

MODELING THE OCCURRENCE OF FOUR CEREAL CROP APHID SPECIES IN IDAHO

John W. Merickel[1], Bahman Shafii[1,2,3], Sanford D. Eigenbrode[2], Christopher J. Williams[1], William J. Price[3]

[1]Department of Statistics
[2]Department of Plant, Soil, and Entomological Sciences
[3]Statistical Programs
University of Idaho, Moscow, ID 83844

Abstract

Idaho is ranked 5th in the United States in overall wheat production and makes over $500 million in profit annually from wheat. Many pests have detrimental effects on wheat; some of the most predominant ones are aphids. Four species of aphids having economic effects on wheat crops in Idaho are: *Diuraphis noxia*, *Metopolophium dirhodum*, *Rhopalosiphum padi*, *Sitobion avenae*. Predictive regression models could be useful for better understanding of the occurrence of these aphid species. Count data for the four species were collected over 17 years via suction traps at 12 locations in wheat fields throughout Idaho. Species specific nonlinear logistic growth models were fitted to each suction trap location to model the aphid accumulation process during the wheat growing season. The nonlinear model used was parameterized to provide inference on three main aphid characteristics, the onset of trapped aphid accumulation, the rate of increase in aphid accumulation, and the maximum accumulated abundance of trapped aphids. Suction trap locations were further aggregated into 5 environments using hierarchical clustering based on climate data. Species specific models were then fitted to each of the 5 environments. Within each environment, the maximum yearly aphid abundance was determined to have a lag (1) autocorrelation structure across years, indicating a biotic feedback. A full nonlinear logistic growth model was then fitted to the entire data set using dummy variable regression to investigate potential climatic environmental patterns in the aphid accumulation process. Predicted models were validated both externally and internally. External validation used suction trap locations in Idaho that were excluded from the model building process to assess the predictive capabilities of the specified models. Internal validation was conducted using bootstrap simulation of the residuals for each model. Statistical models similar to those developed in this study can aid in understanding and evaluating the dynamics of the abundance of cereal crop aphid species in Idaho.

Keywords: *Nonlinear Regression, Logistic Growth Models, Autocorrelation, Suction Traps*

I. Introduction

Idaho is well-known for its agriculture, in particular for its potato production. Idaho is also one of the biggest producers of cereal crops, in particular wheat, in which it is ranked 5th in the U.S.A. for production (Idaho Wheat Commission, 2014). Wheat production for Idaho averages approximately 100 million bushels with a total value of over $500 million, providing over 8,500 jobs (Idaho Wheat Commission, 2014). Because wheat is such an important agricultural product in Idaho and the Pacific Northwest, the management of wheat crops is of great concern.

Among the challenges for wheat production are insect pests, weeds and diseases. Some of the common pest insects include aphids, cereal leaf beetles (*Oulema melanopus*), thrips (*Thysanoptera*), and wireworms (*Elateridae*) (Bechinski, 1998). Aphids, one of the most harmful pests of wheat, are the focus of this study. Aphids can damage crops both directly through feeding, as well as indirectly through the transmission of destructive viruses such as *Barley yellow dwarf virus* (Araya et al., 1986). Several of the most common aphid pest species of wheat, found nearly worldwide, include: the bird cherry-oat aphid (*Rhopalosiphum padi* (L.)), the corn leaf aphid (*Rhopalosiphum Maidis* (F.)), the English grain aphid (*Sitobion avenae* (Fabricius)), the greenbug (*Schizaphis graminum* (Rondani)), the rose grain aphid (*Metopolophium dirhodum* (Walker)), and the Russian wheat aphid (*Diuraphis noxia* (Kurdjumov)) (Araya et al., 1986).

Producers employ several preventative strategies for these problem species. Essentially, all Idaho wheat producers plant weed-free (certified) seed on more than half of their commercial wheat acreage (Bechinski, 1998) and most producers also tend to plant pest-resistant varieties, when these are available. In addition, some producers alter their fall planting times to avoid peek aphid populations, and most spray pesticides to prevent establishment and spread of aphids in their crops (Bechinski, 1998). According to a survey conducted in 1998, about 80 percent of wheat growers use field scouting and thresholds to determine the need for pesticide applications, 60 percent claim their fields are monitored weekly during growing seasons, and only about 14 percent use forecasts from the aphid suction trap network (Bechinski, 1998). A better understanding of the populations of wheat pest species and how they fluctuate both within and across years could greatly advantage wheat producers of Idaho.

In an effort to monitor aphid population fluctuations and movement of invasive aphid species, a study was conducted in the Pacific Northwest (Halbert et al. 1990). The wingless forms of pest aphids are of greatest concern as they can reach high abundances and produce the greatest damage to crops. Nonetheless, these wingless forms result from colonization by winged individuals that migrate from their overwintering hosts to grasses and cereals during the early spring. For this reason, the study used a network of suction traps consisting of 28 sites across the Pacific Northwest to document aphid occurrence and accumulation. By using the data from this aphid suction trap network, a better understanding of the patterns and dynamics of occurrence for several pest aphid species may be developed. The four focal species for this study were the bird cherry-oat aphid (*R. padi*), the rose grain aphid (*M. dirhodum*), the English grain aphid (*S. avenae*), and the Russian wheat aphid (*D. noxia*) (see the subsequent chapters for additional

details regarding these species). This research utilizes data from 12 of these suction trap locations, all within the state of Idaho: Aberdeen, Arbon Valley, Burley, Kimberly, Lewiston, Moscow, Parma, Picabo, Ririe, Rockland, Soda Springs, and Tetonia. Aphid count data from these sites were obtained for the years 1986 through 2003, with the exception of 2002.

The objectives of the study are as follows: 1. Modeling the occurrence and underlying autocorrelation structure associated with aphid species abundances across time using the data collected from suction traps throughout Idaho; 2. Developing site-specific cumulative population growth models for individual species of aphids, determination of potential inter-annual variability, incorporating possible environmental correlates; and 3. Comparing individual or multiple species abundances over time and space as well as potential regional differences across Idaho.

II. Material and Methods

Source and Description of Data

The data used for this study consisted of two types: abiotic and biotic. The biotic data pertain to cereal aphid pests. In Idaho some are anholocyclic, meaning they feed only on grassy hosts year around, while others are holocyclic or host-alternating species. Host-alternating aphids typically overwinter on woody plants (the so-called primary host) and feed on grains and grasses in the summer (secondary host). Four cereal pest aphids that have been of economic concern (Pike et al., 1990) were selected for this study: *Diuraphis noxia*, *Metopolophium dirhodum*, *Rhopalosiphum padi*, and *Sitobion avenae*. Of these four species, *R. padi* and *M. dirhodum* are host-alternating species, while *S. avenae* and *D. noxia* live primarily on grains and grasses. *Rhopalosiphum padi* host alternates between chokecherry (*Prunus virginiana*) bushes in the winter and a variety of grain crops and grasses in the summer. *Metopolophium. dirhodum* alternates between rose bushes (*Rosa sp.*) in the winter and small grain crops and grasses in the summer. Upon emergence in the spring, the populations produce winged (alate) generations which then migrate to summer host plants of grains and grasses (Halbert et al. 1988). Host-alternating aphid species typically overwinter as eggs on their woody hosts. Year around, *S. avenae* and *D. noxia* colonize various small grains, annual and perennial wild grasses (Halbert et al., 1988).

In 1986, a network of traps was established in the PNW consisting of 27 sites throughout the states of Idaho and Oregon. The data for this study were from a subset of this trap network. Each sampling location consisted of one suction trap of 8 meters in total height. The traps used a fan to draw air down the 8 meter, 30 cm in diameter tube through a screen funnel where the insects were collected in a jar of ethylene glycol (Allison & Pike, 1988). The suction traps extended into the air above local insect populations to target migrating insect populations. Each of the suction traps was placed in a cereal grain-dominated field, primarily wheat. The PNW trap network was operated each year from May to November from 1986 through 2003 with the exception of 2002. During the periods of operation, each trap was serviced weekly.

Table 1. Table of PNW suction trap network data for all Idaho sites: Each cell represents the number of times a site was sampled per year. Empty cells indicate that no samples were taken.

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2003
Aberdeen	9	16	16	14	16	19	14	21	23	16	14	12	17	14	16	17	14
Arbon Valley	7	16	15	16	15	18	11	20	20	15	13	7	17	14	16	16	12
Bonners Ferry			17	10	19	19	8	14	11	11							
Burley	11	16	5	16	18	22	13	20	21	16	13	13	17	14	16	17	
Caldwell		15	17	16	19												
Conda								16	22	16							
Corvallis							18	20	22		3						
Craigmont			17	14	21	22	14	21	21								
Hermiston							13	16	20	7	11						
Holbrook		16	17	12	17												
Kimberly	12	16	16	11	13	15	12	21	22	15	14	13	13	13	16	13	11
Klamath Falls							18	21	22	6	12						
Lewiston			13	11	15	20	9	19	20	15	13	13	17		13	15	
Madras							18	21	22	8	10						
Moro							18	21	22	8	12						
Moscow	14	18	17	16	22	22	14	21	22	13	11	11	13	22	15	12	
Mountain Home		12	15	12	20												
Neeley	10	17	15	14	16												
Parma	14	18	17	16	22	22	13	19	21	16	14	13	16	14	16	16	11
Pendleton							14	21	23	8	10						
Picabo			14	11	22	22	13	20	22	16	13	13	13	9	12	17	13
Preston	9	16	16	15	22	10	11	19	22								
Ririe	3	14	16	10		18	12	20	20		13	10	17	12	15	14	
Rockland	14	16	16	16	17	21	13	21	20	16	12	13	17	14	16	17	13
Shelley	8	17	11	8													
Soda Springs	11	10	9	14	15	14	7	14	14		13	13	17	14	16	15	
Tetonia	14	14	13	16	20	22	14	20	22	16	14	9	17	14	17	17	

Low High

The subset of traps used for this study was selected for completeness of the sampling record and to focus on the dynamics of aphids in the state of Idaho. Of the 21 sites in Idaho, data from 12 sites were retained, each of which had continuous and consecutive collections of aphid data for a minimum of 13 years and a maximum of 17, as indicated by table 2 below.

Table 2. Table of final selected data: Each cell represents the number of times a site was sampled per year. Empty cells indicate that no samples were taken.

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2003
Aberdeen	9	16	16	14	16	19	14	21	23	16	14	12	17	14	16	17	14
Arbon Valley	7	16	15	16	15	18	11	20	20	15	13	7	17	14	16	16	12
Burley	11	16	5	16	18	22	13	20	21	16	13	13	17	14	16	17	
Kimberly	12	16	16	11	13	15	12	21	22	15	14	13	13	13	16	13	11
Lewiston			13	11	15	20	9	19	20	15	13	13	17		13	15	
Moscow	14	18	17	16	22	22	14	21	22	13	11	11	13	22	15	12	
Parma	14	18	17	16	22	22	13	19	21	16	14	13	16	14	16	16	11
Picabo			14	11	22	22	13	20	22	16	13	13	13	9	12	17	13
Ririe	3	14	16	10		18	12	20	20		13	10	17	12	15	14	
Rockland	14	16	16	16	17	21	13	21	20	16	12	13	17	14	16	17	13
Soda Springs	11	10	9	14	15	14	7	14	14		13	13	17	14	16	15	
Tetonia	14	14	13	16	20	22	14	20	22	16	14	9	17	14	17	17	

Low  High

The minimum record of 13 years of data was to support a robust time series analysis. Among the 12 Idaho sites, Lewiston had the minimum of 13 years of data (from 1988 to 2001, excluding 1999). Aberdeen, Rockland, Parma, Arbon Valley, and Kimberly each had data for 17 years, and Moscow, Burley, and Tetonia each had 16 years of data from the time period 1986 to 2001. Picabo had 15 years of data recorded from 1988 to 2003 and data from Soda Springs included 15 years, from 1986 to 2001, excluding 1995. Ririe had 14 years of data collected from 1986 to 2001, excluding years 1990 and 1995. Figure 1 shows a map of the final suction trap locations selected for subsequent analyses.

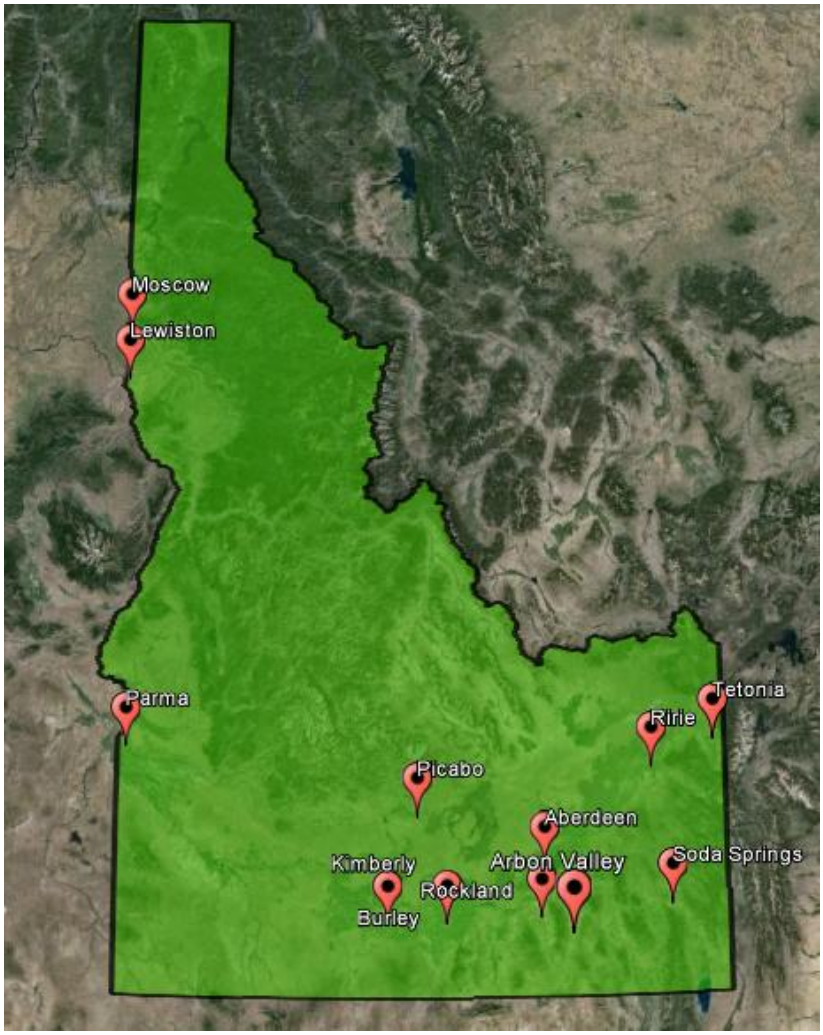

Figure 1. Map of the suction trap locations.

The suction traps were operated throughout the cropping season of spring wheat and winter wheat varieties (May to November). Approximately weekly during this period, samples were collected from the traps and mailed to the University of Idaho's Southwest Idaho Research Extension Center in Parma, to be sorted, and identified to species when possible.

The abiotic data consisted of climatic variables measured daily from 1986 to 2003 for each trap location. Directly measured climatic data were not available for some of the suction trap locations, so a gridded surface meteorological data model, developed by Abatzoglou (2011) was used to supplement the directly measured climatic data. These observed and modeled meteorological data contain daily measurements of: maximum temperature, minimum temperature, precipitation, wind speed, and wind direction. Calculations such as growing degree days were computed using these data. All the data used in this study were acquired through the Regional Approaches to Climate Change (REACCH) for Pacific Northwest Agriculture project, a USDA-funded Coordinated Agricultural Project. The aphid suction trap data is available at:

https://www.reacchpna.org/geoportal3/download?docUUID=erichs%2F%7B9DFCDF59-7F0D-4AE5-B16E-AABF1D8AE77A%7D.

Statistical Analysis

Data Management

In order to model the accumulation of aphid abundance through time, some data management was necessary. Within each aphid species, the raw aphid counts for each year at each sampling site were accumulated as follows: $x'_{ijk} = \Sigma_{l=1}^{k} x_{ijl}$ where x_{ijl} are the raw aphid counts at sample event time l, for year j and site i, and x'_{ijk} are the corresponding cumulative counts of aphids up to the k^{th} sampling event. The abiotic data were managed to develop a standardized growing degree-day scale for the intra-annual variability in the aphid accumulation process. Growing degree days measure the daily accumulation of average temperature relative to a base temperature of the subject of interest. In this case, both the host (wheat) and the organism (aphids) have a similar base temperature, considered here to be 4°C (Slafer & Rawson, 1995; Honek & Martinkova, 2004).

Cumulative growing degree days were calculated as follows:

$$g_{ijk} = \{\Sigma_{l=1}^{k} f(t_{ijl})\} \mid f(t_{ijl}) \geq 0 \;\; ; \;\; f(t_{ijl}) = \frac{t_{ijl_{high}} + t_{ijl_{low}}}{2} - 4 \qquad (1)$$

where g_{ijk} represents the cumulative growing degree day of the k^{th} sampling event at site i in year j, $t_{ijkl_{high}}$ and $t_{ijkl_{low}}$ represent the daily high and low temperatures, respectively for the l^{th} day of the year, and k being the day of the year corresponding to the k^{th} sample event. The value 4 represents the base temperature (C°). Growing degree days were calculated at each site beginning with January 1 ($l = 1$) for each day the average temperature was above the base temperature.

Nonlinear Model

A nonlinear logistic model was chosen to model the process of aphid abundance accumulation within a year. A logistic model form is parsimonious (has a reasonable number of parameters) while having relevant biological or ecological interpretations for the parameters. Similar models have been used for: the determination of cardinal temperatures in germination (Shafii and Price 2001), estimation of *Escherichia coli* growth at different temperatures (Fujikawa et al. 2004), and dose-response modeling (Price et al. 2012).

Initially, models were developed separately for each site-year-species combination as follows:

$$\tilde{x}'_{ijk} = \frac{m_{ij}}{1 + e^{-\beta_{ij}(g_{ijk} - L_{ij})}} \qquad (2)$$

Where $\tilde{x}'_{ijk}$ represents the estimated cumulative aphid abundance as defined above, for a given species, at a particular growing degree day (g_{jik}) as defined above, given parameters m_{ij}, β_{ij}, and L_{ij} within a year, j, and site, i. The parameter m_{ij} represents the theoretical maximum value of the aphid accumulation, and controls the upper asymptote of the "S" shaped curve. The parameter L_{ij} represents the value of cumulative growing degree days at which the rate of aphid accumulation is greatest, and is visually represented by the inflection point in the "S" shaped curve. Interpretation of L_{ij} is particularly important in dose-response studies as it represents the median lethal dose (Price et al., 2012). In our model, L_{ij} can also be interpreted as the value of cumulative growing degree days at which half of the accumulation of aphids has occurred. The parameter β_{ij} represents a rate related parameter for accumulation of aphid abundance.

Both least squares (Procedure NLIN) and maximum likelihood (Procedure NLMIXED) in SAS ver. 9.3 were used to estimate and assess the fit of Eq (2) to each data set. Procedure NLIN was used to fit the data using an iterative Gauss-Newton nonlinear least squares estimation routine, under the assumptions of uncorrelated, zero mean, homoskedastic, and normally distributed errors. While least squares estimation provided an initial assessment of the adequacy of Eq (2), a more appropriate approach was employed using Procedure NLMIXED to fit Eq (2) under the process of maximum likelihood estimation. Procedure NLMIXED estimated the parameters of Eq (2) by maximizing the likelihood function with a dual quasi-Newton algorithm (SAS Ver. 9.3 documentation). A general form of the likelihood is given as follows:

$$L\left(\theta_{ij}\middle|x'_{ijk}\right) \propto \Pi_{\mathrm{k=1}}^{\mathrm{n_{ijk}}} f\left(x'_{ijk}|\theta\right) \tag{3}$$

Eq (3) states that the likelihood of the data given a vector of parameters, is proportional to the product of the density function evaluated for each cumulative aphid count, where θ_{ij} is a vector of unknown parameters, x'_{ijk} is the observed k^{th} cumulative aphid count for a given site and year, and n_{ijk} is the total number of sampling events. Estimation proceeds by identifying the parameter values that maximize the likelihood assuming the mean of the distribution is the nonlinear function given by Eq (2). Several likelihood forms were evaluated, including: normal, negative binomial, and Poisson forms.

The model fit was first assessed by determining if the iterative estimation methods presented above converged successfully. The significance and adequacy of the parameter estimates were assessed using asymptotic 95% confidence intervals and inter-parameter correlations, respectively. The estimated model was also inspected visually by overlaying the predicted model on the observed data points. The ideal structure of this plot is to have the predicted model centered on the data and follow the pattern of the data well. In conjunction with the observed versus predicted plots, plots of the residuals were also assessed to determine if the assumptions of uncorrelated, normally distributed, zero mean, and constant variance of residuals were met.

Cluster Analysis

In order to develop environmental aphid accumulation models, a grouping of sampling sites into similar environments was carried out. Cluster analysis is a common multivariate technique used to find natural groupings in data where the observations within each cluster are similar, while the clusters are dissimilar to each other (Rencher and Christensen, 2012). We used clustering to group observations (aphid sampling sites) conditionally based on a set of covariates (abiotic climate data). The covariates used in the cluster analysis were minimum temperature, maximum temperature, average temperature, precipitation, elevation, latitude, wind speed, and wind direction. The SAS Procedure CLUSTER was used to conduct the cluster analysis by means of an agglomerative hierarchical clustering procedure. In this procedure, each site was initially considered as its own cluster, and then each pair of clusters closest to each other was merged repeatedly until one cluster was left. Both the mean and median clustering algorithms were used and each yielded the same results. For simplicity only the formula for the average method is displayed as follows:

$$D_{ab} = \frac{1}{N_a N_b} \Sigma_{i' \epsilon C_a} \Sigma_{j' \epsilon C_b} d(\boldsymbol{y}_{i'}, \boldsymbol{y}_{j'}) \quad ; \quad d(\boldsymbol{y}_{i'}, \boldsymbol{y}_{j'}) = \sqrt{\left(\boldsymbol{y}_{i'} - \boldsymbol{y}_{j'}\right)^2} \tag{4}$$

Where D_{ab} is the distance between clusters C_a and C_b, N_a and N_b are the number of observations in clusters C_a and C_b respectively, and $d(\boldsymbol{y}_{i'}, \boldsymbol{y}_{j'})$ is the Euclidean distance between the two observed vectors $\boldsymbol{y}_{i'}$ and $\boldsymbol{y}_{j'}$ of the two clusters.

The results from this hierarchical clustering procedure are typically displayed in a tree diagram or dendrogram, which shows all the steps in the hierarchical procedure and the corresponding distances (Rencher and Christensen, 2012). These techniques of clustering have been applied to data in many fields including medicine, criminology (Hartigan, 1975), geology, geography, economics, and market research.

Regional/Environmental Nonlinear Model

The nonlinear model mentioned previously was refitted to data clustered into environments. Prior to utilizing these clustered data, where each cluster covered multiple sites and years, it was necessary to standardize the data to a common scale. Scaling was achieved by dividing each cumulative aphid count of a given site (x'_{ijk}) by the maximum cumulative aphid count observed for that site, assessed across all available years. Hence, all data values over all sites and years were rescaled to proportional values between 0.0 and 1.0. Scaling the data this way was helpful in minimizing both the temporal variability present across the multiple years of each site, as well as site-to-site variability within a year. Model estimation was then carried out on data pooled across sites within each environmental region (previously identified through cluster analysis) using a maximum likelihood algorithm as described above (Procedure NLMIXED). Three potential likelihood forms were assessed for the scaled data: the beta, binary, and normal likelihoods.

Analysis of Autocorrelation

In regression analysis, it is particularly important to determine if the data contain an autocorrelation structure. When observations are measured for the same subject over time it is unreasonable to assume that the observations are independent. Time periods that are closer to each other are more likely to be similar than time periods that are farther apart (Fox, 2008). If an unaccounted for correlation structure exists in a regression analysis, the parameter estimates obtained will not be statistically efficient and their associated estimated standard errors will be biased. An autoregressive model can help mitigate these conditions. A general form of the autoregressive model with lag n (AR(n)) may be given as follows:

$$y_t{}' = \alpha_0 + \alpha_1\gamma_t + v_t \quad ; \qquad v_t = \phi v_{t-1} + \phi^2 v_{t-2} + \cdots + \phi^{n_t} v_{t-n_t} + e_t \qquad (5)$$

Eq (5) represents a simple linear regression model where $y_t{}'$ is the t^{th} response, α_0 is the intercept, α_1 is the regression coefficient, γ_t is the t^{th} ordered data observation, and v_t is the autoregressive error term. The second part of Eq (5) represents the autoregressive function of the error term of the simple linear model; where ϕ is the autoregressive coefficient, v_{t-1} is the error of the previous observation, and e_t is the independent and normally distributed random error term with zero mean, and constant variance, given n_t points in time. AR(n) models assume that observations closer in time are more correlated than observations farther apart in time. Eq (5) also satisfies the condition $|\phi| < 1$, and therefore, the autoregressive coefficient approaches zero for observations of increasing distance from one another in time.

Analysis of autocorrelation has been used in many areas of research such as marketing, economics, ecology and criminology. Autocorrelation correction in regression analysis was pioneered by Cochrane and Orcutt (1949). One example of analysis of autocorrelation is presented by Fox (2008) in which Canadian women's crime rate was analyzed over time.

Analyses to assess autocorrelation were carried out within the environmental groups resulting from the cluster analysis. In particular, the previously obtained estimates of parameter m_{ij}, which represents the maximum aphid count within a given year and site, was modeled using Procedure AUTOREG in SAS Ver. 9.3. Conditional heteroskedasticity at lag 1, lag 2, and lag 3 were assessed for each environment group across all years of available data.

Nonlinear Regression Analysis with Autocorrelation

To further adapt the model presented in Eq (2) and build a more comprehensive model, the temporal variation over years was incorporated into the nonlinear logistic growth model. The distribution of the response (relative cumulative aphid abundance) was considered to be normal for this model. Previously, Eq (2) was assessed using a Poisson density function to model the raw aphid counts for each year. The relative, cumulative abundance data, however, is scaled on a fine increment between 0.0 and 1.0, and it is reasonable to assume normality of the data within each sample time point. The model below was also assessed under the assumptions of the beta and binary distributions, but they yielded poor results and, hence, are not presented here.

Eq (2) was adapted such that the m_{ij} parameter accounted for an autocorrelation structure of lag 1. That is, the relative maximum aphid count for a given year and site is a function of the relative maximum aphid count of the previous year. The modified model in Eq (2) then becomes:

$$X'_{rjk} = \frac{M_{rj}}{1+e^{-\beta r\left(g_{rjk}-Lr\right)}} \quad (6)$$

Where M_{rj} is given as

$$M_{rj} = Int_{M_r} + AR1_r * M_{r,j-1} + e_r \quad (7)$$

In Eq (6), X'_{rjk} is the relative cumulative aphid abundance for the r^{th} region ($r = 1,2, \ldots, R$), the j^{th} year, and the k^{th} sampling event, the term g_{rjk} represents the growing degree day for the r^{th} region, j^{th} year, k^{th} sampling event, and M_{rj} is now the relative maximum aphid count for a given year. M_{rj} is an auto-correlated function, based on Eq (7), of the relative cumulative maximum aphid count of the previous year, $M_{r,j-1}$, an intercept term, Int_{M_r}, and an autoregressive coefficient, $AR1_r$. Int_{M_r} is the mean of the maximum relative cumulative aphid abundances for the region across all years. The $AR1_r$ term represents the degree that the maximum aphid count changes for a given year based on the previous year's maximum aphid count for the region (i.e. α_1 from Eq (5)). The e_r term is a random error associated with the M_{rj} parameter and is assumed to be a normally distributed random effect with zero mean and constant variance. This model was fitted to each environment across all sites and years of data within those environments.

Dummy Variable Regression Analysis

Dummy variable regression is a technique used to make inference on data that are both qualitative and quantitative (Fox, 2008). In this case, it was of interest to make inference on aphid prevalence while incorporating both spatial and temporal variation. The temporal variation was considered quantitative, while the spatial variation (environmental groups) was treated as a qualitative factor. Dummy variables were then assigned according to the environment for which the aphid data were recorded. By creating a dummy variable for environment, a full dummy variable regression model that incorporated all the data across all sites and years was specified. A simplified example of the dummy variable regression model is presented as follows:

$$w_r = D_1 X'_{1jk} + D_2 X'_{2jk} + \cdots + D_r X'_{rjk} + e \quad (7)$$

In the expression above, w_r represents the estimated response for the r^{th} region. The term D_r represents the dummy variable for the regression (0 or 1), and the terms X'_{rjk} represent the estimated reduced model for the r^{th} region. When the dummy variable D_r is equal to 1, the rest of the D_r terms in the regression are set to zero. This means that when modeling the effect of

X'_{1jk}, $D_1 = 1$, indicating the effect for region 1is present while the rest of the remaining regional effects are absent.

Likelihood ratio tests can then be carried out to determine if there is a substantial improvement in the likelihood from a reduced model form relative to the full model specification. Through this construct, a full-model dummy variable regression allowed for statistical inferences and comparisons of the parameters across the environments. For example, the full model allows comparison of the parameter estimate for the lag of one environment being equivalent to the corresponding estimate of a second environment.

Validation

Validation was conducted to assess the predictive capabilities of the constructed models. The model was validated externally using independent data as well as internally using bootstrapping of the residuals.

Internal Validation

The regression models were validated internally using a bootstrap simulation of the residuals of each initial fit. The bootstrap method is a simple computational method used to generate samples from an existing sample. The bootstrap method proceeds by sampling with replacement such that every observation in the initial sample has equal probability of being selected, thus it is possible to select a single observation multiple times (Efron & Tibshirani, 1986). This re-sampling procedure is designed to parallel the process by which the sample observations were drawn from the underlying population (Fox, 2008).

The bootstrap technique was used to generate new data sets for each region/environment and species combination to assess the fit of each model. The process proceeds by first fitting the region/environment model to the data which was scaled and accumulated as discussed previously. The residuals from these fits were then scaled back to their actual count values by multiplying them by the observed maximum for each site across all years of data at each respective site. The count values were then transformed back into un-accumulated counts such that the resulting values take the original data form.

Once the residuals were in original data form, they were then sampled with replacement using PROCEDURE SurveySelect (SAS Ver. 9.3), generating a new sample of the same size. These residual values were then randomly aligned and added to the predicted values generated from the initial fit of the respective region/environmental model. Negative simulated counts resulting from this step were set to zero. The new bootstrap simulated data were then re-accumulated and scaled as was previously done with the original data. The regional model (Eq(6), Eq(7)) was then fitted to the bootstrap data, and the resulting bootstrap residuals were stored. This process was repeated until $B = 1000$ bootstrap residual samples were achieved. These residuals were then examined to assess the model fitting process.

External Validation

External validation is used to determine how effective a model is at making predictions based on data that were not used to build the model. Before modeling, numerous aphid sampling sites were left out of the original analysis because of insufficient data for some years. Of the sites that were omitted from modeling, five having the most years of data were selected for external validation of the nonlinear environment/region level model.

Linear Discriminant Analysis

Linear discriminant analysis, a technique commonly used in multivariate statistics to differentiate between observations (Rencher and Christensen, 2012), was used to classify newly obtained aphid sampling sites into the environments that had been previously determined through cluster analysis. Given a classification into an environmental region, these new sites would allow for external validation of the estimated nonlinear aphid abundance model.

The purpose of a linear discriminant analysis is to develop a function of variables that most effectively separates the observations into the predefined groups. The linear discriminant analysis takes the form:

$$L_g = (2\pi)^{-\frac{p}{2}}\left|\eta_q\right|^{-\frac{1}{2}}e^{-0.5d_g^2(q)} \qquad ; \qquad d'^2_g(q) = (q-g)'\eta_q{}^{-1}(q-g) \tag{8}$$

Where L_g is the likelihood that a subject with a vector of observations, q, is classified into group g, p is the number of variables considered, and η_q is the pooled covariance matrix of the variables measured for the new group. The term $d'^2_g(q)$ is the squared Mahalanobis distance of the vector q to group g. Once the likelihood of membership in each group is calculated for a subject, that subject is then assigned to the group with which it has the greatest likelihood. The classification process is assessed by determining an error rate for the classifications; this is done by comparing the number of misclassified subjects to the total number of classifications.

Following discriminant analysis and classification of new validation sites into the previously defined environmental groups, predictions of relative aphid abundance were made. As was done in the model building process, plots of the predicted model overlaid on the observed data points were used for visual assessment of the predictive capability of the model. In addition, plotting the validation residuals allowed for assessment of the assumptions of uncorrelated, normally distributed, zero mean and constant variance residuals. This validation provided an assessment of the predictive accuracy of the regression model.

III. Results and Discussion

Nonlinear Model

The nonlinear model presented in Eq (2) was fitted to the cumulative aphid abundances for each site-species-year combination. As discussed previously, the model was first estimated using least squares in order to assess overall model adequacy, which was subsequently followed with a maximum likelihood estimation, where distributional forms such as the Normal, Poisson, and Negative Binomial likelihood forms could be assessed and evaluated. While the Negative Binomial likelihood provided an over-dispersed, discrete likelihood form that appropriately matched the count nature of the data, its implementation proved difficult in the estimation process due to the limited replications within sampling events. The Poisson likelihood, on the other hand, also matched the discrete count nature of the aphid data, but had no estimation problems. Because the data consisted of discrete values, the Normal likelihood was deemed less appropriate. An example of the nonlinear model fit using the Poisson likelihood (assuming a logarithmic link function) for the site Parma and aphid species *R. padi* in the year 1999 is given in Table 3 and Figure 2.

Table 3. Example of parameter estimates with approximate standard errors, significance, and asymptotic 95% confidence intervals for model fit (Parma data for *R. padi* in 1999).

Parameter Estimates for *R. padi* in Parma in 1999								
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
m	270.38	7.7073	14	35.08	<.0001	0.05	253.85	286.91
β	0.009247	0.000633	14	14.61	<.0001	0.05	0.007890	0.01060
L	1585.51	16.0805	14	98.60	<.0001	0.05	1551.02	1620.00

All parameter estimates in Table 3 are significantly different from zero. The estimated maximum cumulative aphid, *m*, count for 1999 is approximately 270 aphids. The growing degree day, *L* to reach 50% of this maximum is estimated to be 1585 growing degree units. Correlations of the parameter estimates are given in Table 4.

Table 4. Correlation matrix of the parameter estimates obtained from model fit to Parma data for *R. padi* in 1999.

Correlation Matrix of Parameter Estimates			
Parameter	***m***	***β***	***L***
m	1.0000	-0.3755	0.5761
β	-0.3755	1.0000	-0.6552
L	0.5761	-0.6552	1.0000

A commonly used acceptable range for inter-parameter correlation is <0.8 or >-0.8, but correlation is often of little concern when <0.99 or <-0.99 (Bates and Watts, 1988). All the inter-parameter correlations given in Table 3 are well within the smaller bounds of -0.8 and 0.8, indicating the three parameters are sufficiently un-correlated, and that there are no parameter redundancies in Eq (2).

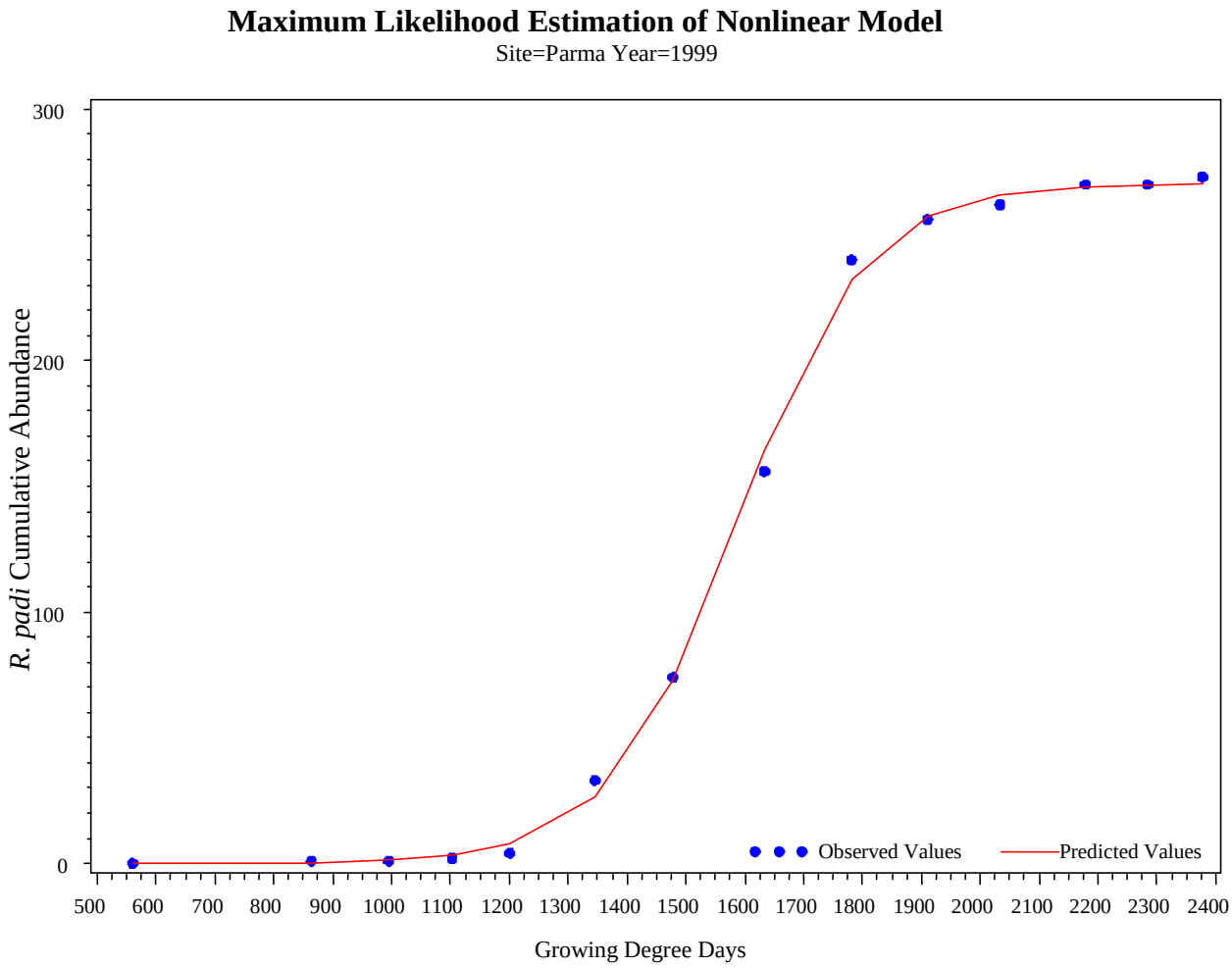

Figure 2. Example observed versus predicted plot (fit plot) of nonlinear model to data from Parma for *R. padi* in 1999.

The fitted curve in Figure 2 follows the observed data pattern well and the associated residual pattern in Figure 3 demonstrates a fairly random pattern with no extreme values.

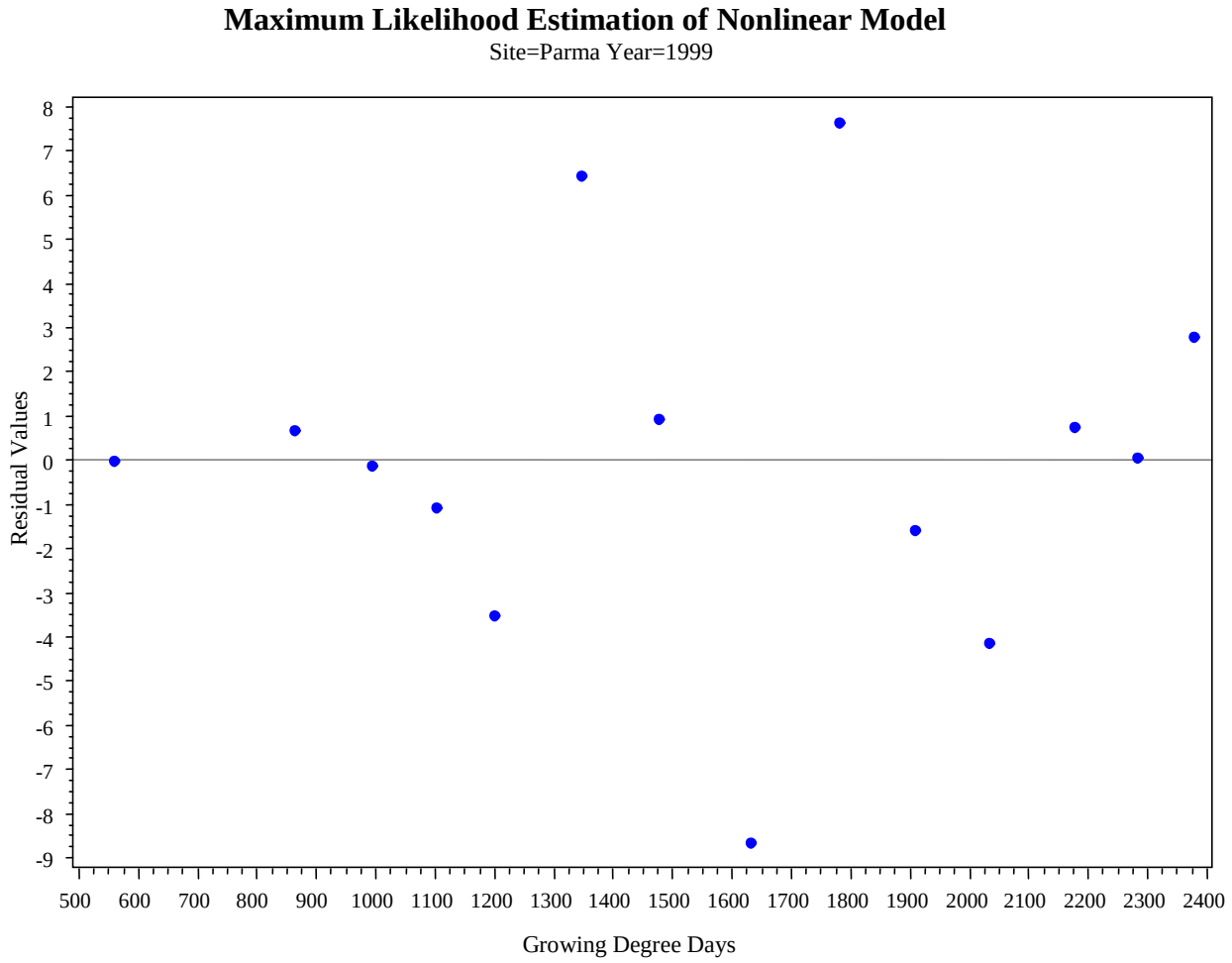

Figure 3. Example of residual plot from fit of nonlinear model to Parma data for *R. padi* in 1999. Each point on the plot represents a residual value.

The parameter estimates and diagnostics for the remaining 3 species for 1999 in Parma are available at: https://www.reacchpna.org/geoportal3/download?docUUID=erichs%2F%7B24BD99E9-5A12-4AF5-9539-EBD83E88DFDD%7D, accompanied by additional tables summarizing the site-year-species model fits. All of the remaining parameter estimates and diagnostics for all site-year-species combinations are also available at the above website. Additional matrix plots were also generated to summarize the model fits for each site-species combination for all years of data. An example of one of these summary matrix plots of model fits for species *R. padi* at Moscow is presented in Figure 4.

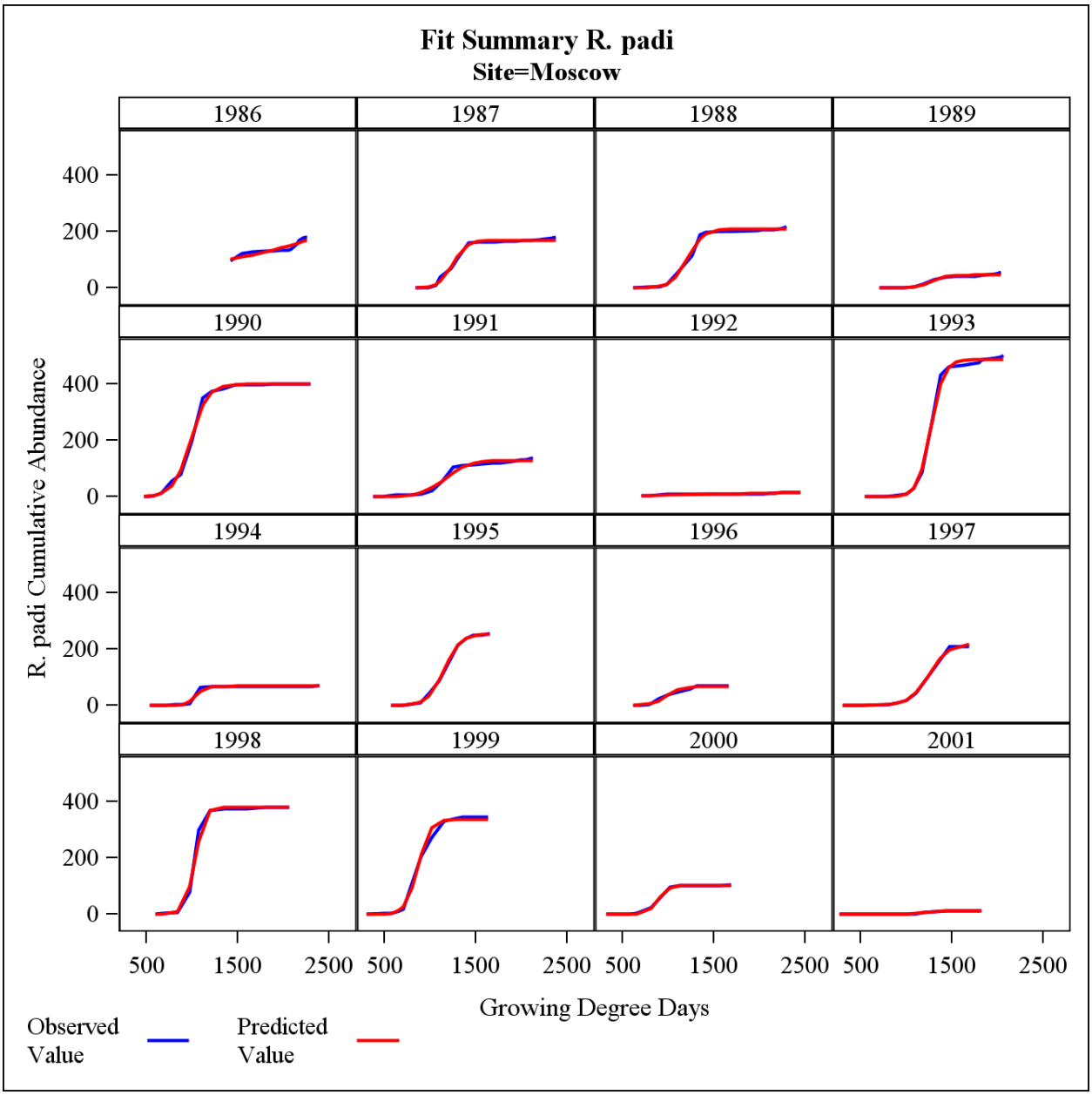

Figure 4. Example of fit plot summary of all years of data (Moscow, species *R. padi)*.

Figure 4 is useful for assessing model fits across years, showing the temporal variability in the maximum cumulative abundance across years, as well as the adequacy of the model fit for most cases. Estimation and diagnostics were assessed for all site-year-species combination for a total of 725 model fits. Additional matrix plots can be viewed at https://www.reacchpna.org/geoportal3/download?docUUID=erichs%2F%7B24BD99E9-5A12-4AF5-9539-EBD83E88DFDD%7D.

After all site-year-species combinations of data were fitted to Eq (2); each fit was classified as good, bad, or "non-estimable". The nonlinear model was considered a "good fit" for scenarios in which: the maximization algorithm converged, parameter estimates were uncorrelated; the fitted curve followed the data well; and there were no extreme residual values. The model was considered a "bad" fit when; parameter estimates were highly correlated, or the fitted curve did not follow the data, or when there were extreme residual values. Scenarios were considered non-estimable when maximum likelihood algorithm failed to converge, and no subsequent diagnostics could be carried out. Scenarios in which researchers sampled the suction trap through the entire growing season but counted less than ten aphids total were omitted from these evaluations. Of all the fitted models for *D. noxia, M. dirhodum, R. padi, and S. avenae,* 87%, 89%, 82%, and 74% of the fits respectively were considered "good." Overall, 83% were considered good. None of the sites, years, or species had particularly low percentages of good fits, and therefore all data (refer to Table 2) were considered suitable for further analysis.

Cluster Analysis

Cluster analysis was used to group sites based on environmental covariates and to investigate how environmental differences impact the aphid accumulation process (Figure 5). A clustering analysis was implemented using Eq (4) based on minimum temperature, maximum temperature, mean temperature, wind speed, wind direction, precipitation, elevation, and latitude.

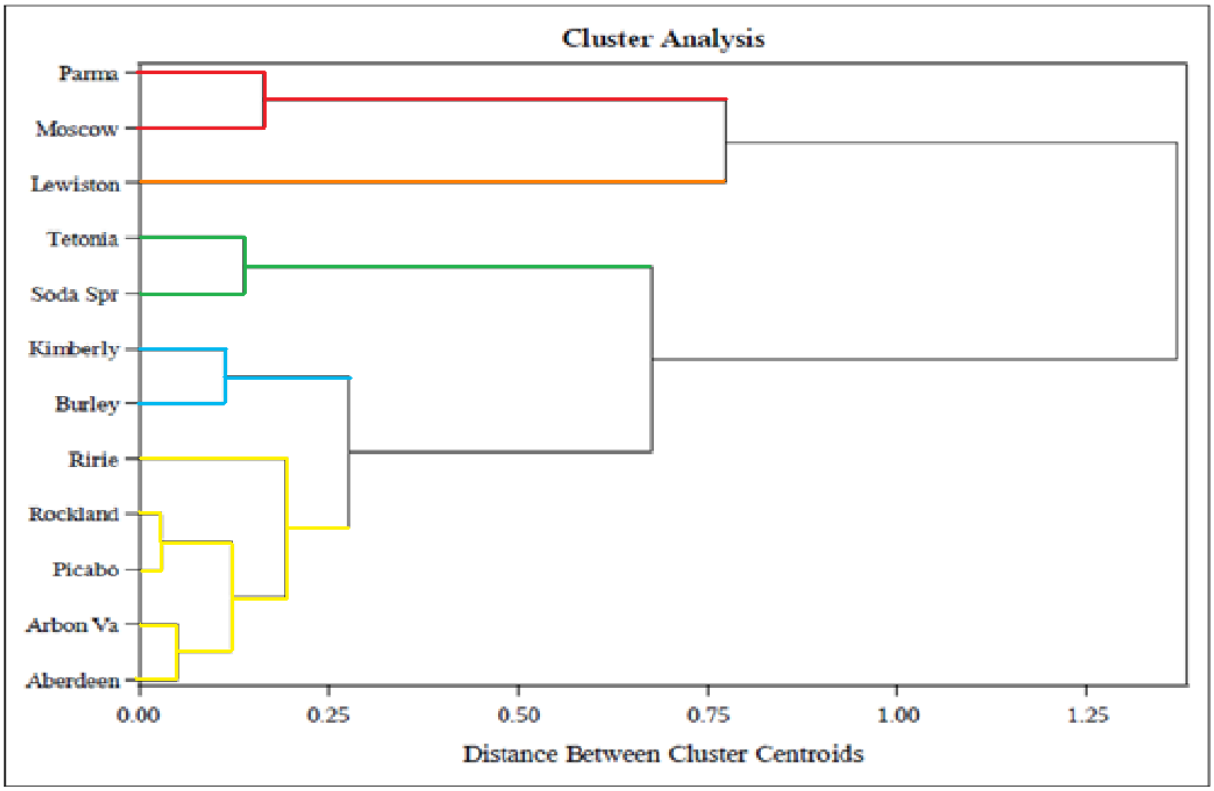

Figure 5. Dendrogram displaying results from hierarchical clustering algorithm used to group sites based on climatic data. Sites clustered together are indicated the same color.

The results displayed in Figure 5 show how the 12 sites were grouped into 5 clusters defined as follows: 1. Moscow and Parma, 2. Lewiston, 3. Tetonia and Soda Springs, 4. Kimberly and Burley, 5. Ririe, Rockland, Picabo, Arbon Valley, and Aberdeen.

From Figure 5, environment 2 is the only group with just one site (Lewiston); the rest of the groups contain at least 2 sites. Classifying Lewiston into a group by itself is not unexpected because it is the lowest point in Idaho (227m) with a distinctive, warmer climate than the rest of the state. Also, evident from the cluster analysis, no sites in the Southeastern region were grouped with sites in the Northwestern region. This is likely because Southeastern Idaho is significantly higher in elevation than the rest of the state. Also notable, the Moscow and Parma sites were classified together even though they are geographically distant. This classification is reasonable because Moscow (786m) and Parma (680m) are close in elevation in comparison to other Southeastern sites (all of which are over 1000m).

Regional/Environmental Nonlinear Model

To investigate the similarities in patterns of aphid accumulation from sites classified into a common cluster, the nonlinear model from Eq (2) was fitted to the aggregated data, utilizing clustered environments in place of sites. In order to mitigate variability across the sites and years within each region, the data were re-expressed in relative values between 0 and 1 as described previously. A Beta likelihood form, having variates restricted to the 0.0 to 1.0 range, was considered a natural choice for these data. The estimation process, however, was unstable when using the Beta, so a Normal likelihood form was selected as a reasonable approximation for the continuous scaled relative accumulation data within the 0.0 to 1.0 range. Initially, Eq (2) was fitted to each clustered environment, incorporating all years of data for the respective cluster sites. The parameter estimates of one such fit is presented in Table 5.

Table 5. Parameter estimates and corresponding, approximate t-values, p-values, and confidence intervals generated from the fitting of Eq (2) to data for environment 4, *R. padi*.

Parameter Estimates								
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
m	0.1260	0.02246	494	5.61	<.0001	0.05	0.08191	0.1702
β	0.01141	0.006172	494	1.85	0.0650	0.05	-0.00071	0.02354
L	1280.22	98.3585	494	13.02	<.0001	0.05	1086.97	1473.47

Table 5 shows that the standard errors of the parameter estimates are quite large. This is expected because the aggregated, clustered data encompasses more variability due to the presence of multiple sites and years within any given fit. Figure 6 gives an example fit, and visually shows how the model from Eq (2) fits the aggregated data.

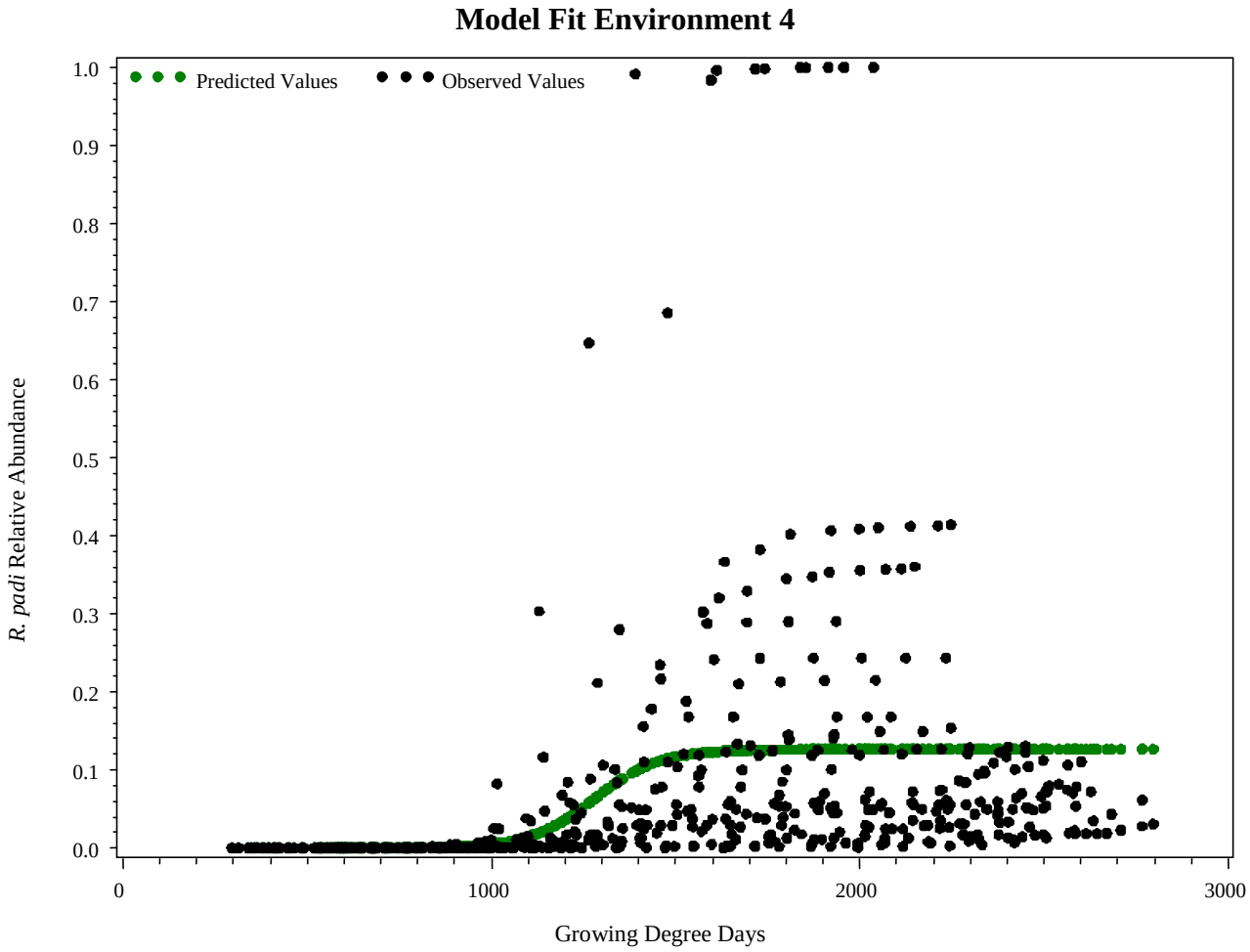

Figure 6. Example of nonlinear model fit to all years of data for species *R. padi*, environment 4 (Kimberly and Burley).

From Figure 6, it is clear that the model from Eq (2) does not account for all the variability of the aggregated data. The inadequacy of the fit is primarily due to site-to-site and year-to-year variability in aphid accumulation. More specifically, this unaccounted variation seems to occur primarily on the upper asymptote of the curve, which is parametrically controlled by the relative maximum parameter (M).

The temporal variability was investigated through autoregressive modeling of the correlation of estimates for the three parameters of the logistic growth model across years obtained previously. It was determined that for the most ecologically sensible interpretation, the dependencies at lag (1), lag (2) and lag (3) should be assessed, primarily in the maximum parameter, m. The remaining parameters β and L, showed less autocorrelation and had limited expectation to be biologically correlated across time. An example of the estimated

autocorrelations associated with the relative maximum parameter across years for environment 4, and species *R. padi* is given in Table 6.

Table 6. Autocorrelation estimates for the relative maximum parameter for region 4, *R. padi* at lag (1), lag (2) and lag (3) dependencies respectively.

Estimates of Autocorrelations			
Lag	**Covariance**	**Correlation**	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.0511	1.000000	|**********|
1	-0.0174	-0.339952	| *****| |
2	-0.00525	-0.102735	| **| |
3	-0.00355	-0.069477	| *| |

Table 6 shows that lag (1) had an estimate farthest from zero, indicating that the strongest correlation in time exists between $time_t$ and $time_{t-1}$, which is similar to the findings of Davis et al. (2014). The estimate for the autocorrelation at lag (1) is -0.339952 which means that the relative maximum at $time_t$ is inversely correlated with the relative maximum at $time_{t-1}$ by a factor of approximately 0.339952. Similar autoregressive analyses were conducted on all 20 of the environment-species combinations. Of the 20 combinations, 14 showed the strongest autocorrelation at a dependence of lag (1). Hence, a lag (1) autoregressive structure was imposed on the relative maximum parameter within the model (refer to Eq (6) and (7)). Corresponding plots and associated tables for the relative maximum parameter estimates of the remaining species-environment combinations can be viewed at: https://www.reacchpna.org/geoportal3/download?docUUID=erichs%2F%7B24BD99E9-5A12-4AF5-9539-EBD83E88DFDD%7D.

Nonlinear Regression Analysis with Autocorrelation

To account for the temporal variability in the relative maximum parameter across years, the model represented in Eq (6) was fit to each of the five environments. Table 7 gives the parameter estimates and corresponding approximate asymptotic 95% confidence intervals, standard errors, t-values and p-values of the model fit to environment 4, species *R. padi* (same data as Figure 6).

Table 7. Parameter estimates and corresponding, approximate t-values, p-values, and confidence intervals generated from the fitting of Eq (6) to data for environment 4, *R. padi*.

Parameter Estimates								
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
β	0.008437	0.000501	30	16.83	<.0001	0.05	0.007413	0.009460
L	1334.89	7.9965	30	166.93	<.0001	0.05	1318.56	1351.22
AR1	-0.1925	0.1870	30	-1.03	0.3114	0.05	-0.5744	0.1893
int_M	0.1958	0.05452	30	3.59	0.0012	0.05	0.08444	0.3071
ln_Var_M	-2.7969	0.2555	30	-10.95	<.0001	0.05	-3.3188	-2.2751

All the parameter estimates are significant in this case, except for the AR1 term. Although the AR1 term is not significant at the $\alpha = .05$ significance level, the AR1 was retained in the model because it is important for the model fit and essential to the model structure. Although the AR1 term is not significant in this estimation, the other two terms (int_M, and ln_var_M), which comprise the autoregressive aspect of the relative maximum, do show statistical significance. Due to the scaling of the data, all the values associated with the relative maximum parameter are very small; therefore the random error term was parameterized in a logarithmic form (ln_var_M) in order to stabilize the estimation process.

As with previous model estimations, it was important to assess the inter-parameter correlation. When there are more model parameters being estimated, the possibility of having higher inter-parameter correlation, i.e. redundancy within the model, increases. The resulting correlation matrix of the parameter estimates from Table 7 is presented in Table 8.

Table 8. Estimated parameter correlations from model fit to environment 4, species *R. padi*.

Correlation Matrix of Parameter Estimates					
Parameter	***β***	***L***	**AR1**	**int_M**	**ln_Var_M**
β	1.0000	-0.3462	0.007066	-0.02376	-0.05301
L	-0.3462	1.0000	-0.00765	0.02516	0.05395
AR1	0.007066	-0.00765	1.0000	-0.5790	-0.00069
int_M	-0.02376	0.02516	-0.5790	1.0000	0.002238
ln_Var_M	-0.05301	0.05395	-0.00069	0.002238	1.0000

The inter-parameter correlation presented in Table 8 shows that all the parameters estimated had low correlations, were well within the conservative bounds of -0.8 and 0.8, and the parameterization of the model was not redundant. In the case of fitting Eq (6), all inter-

parameter correlations satisfied the first criterion for all 20 environment-species combinations. Figure 7 shows the two dimensional fit plot of Eq (6) fit to region 4, and species *R. padi*.

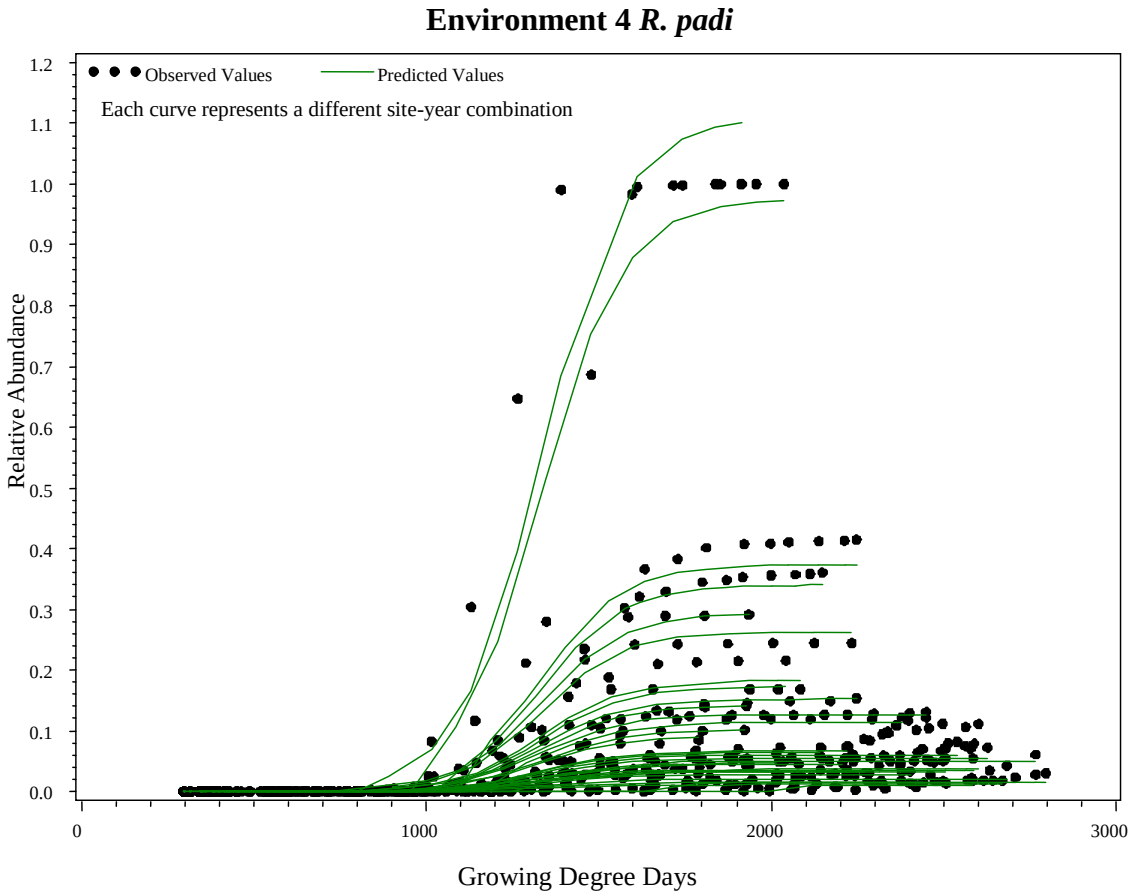

Figure 7. Nonlinear model with incorporation of autoregressive structure on the relative maximum parameter fit to environment 4, species *R. padi*

From Figure 7, it is evident that the autoregressive structure imposed on the relative maximum parameter successfully accounted for more variability in the upper end of the sigmoidal curve than did the other estimation. The 19 remaining 2D and 3D observed and predictive plots of the autoregressive-environmental model accompanied by the remaining 3 full-model parameter estimate tables can be viewed at the website indicated above.

The 3D surface of the model fit to the same data and the corresponding observed 3D surface are presented in Figures 8 and Figure 9, respectively. As another example of the fitted nonlinear model with autocorrelation, Figures 10 and Figure 11 show the 3D observed and predictive surfaces for species *M. dirhodum* environment 3, respectively.

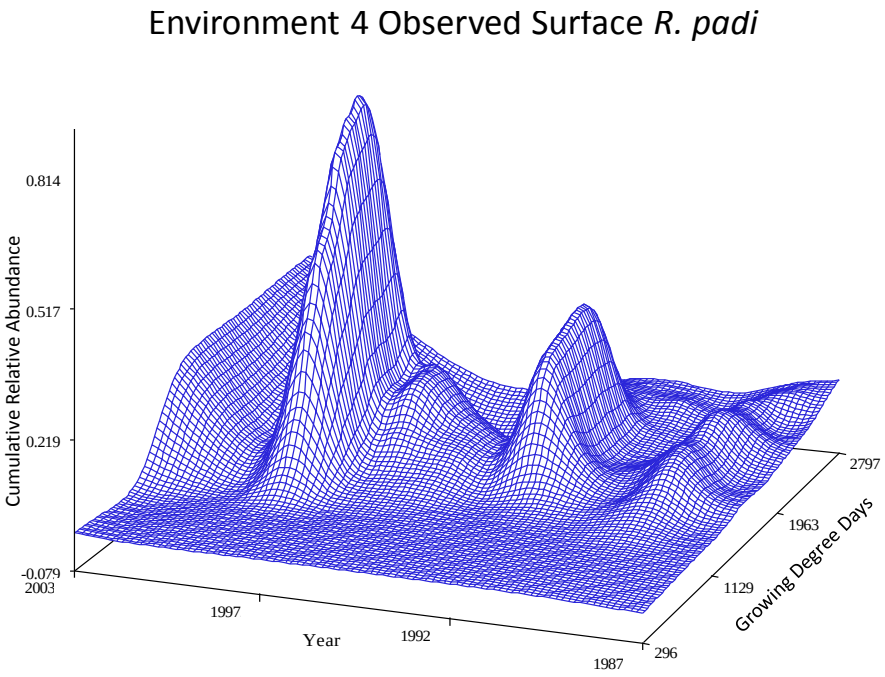

Figure 8. Observed surface of data from environment 4, species *R. padi*.

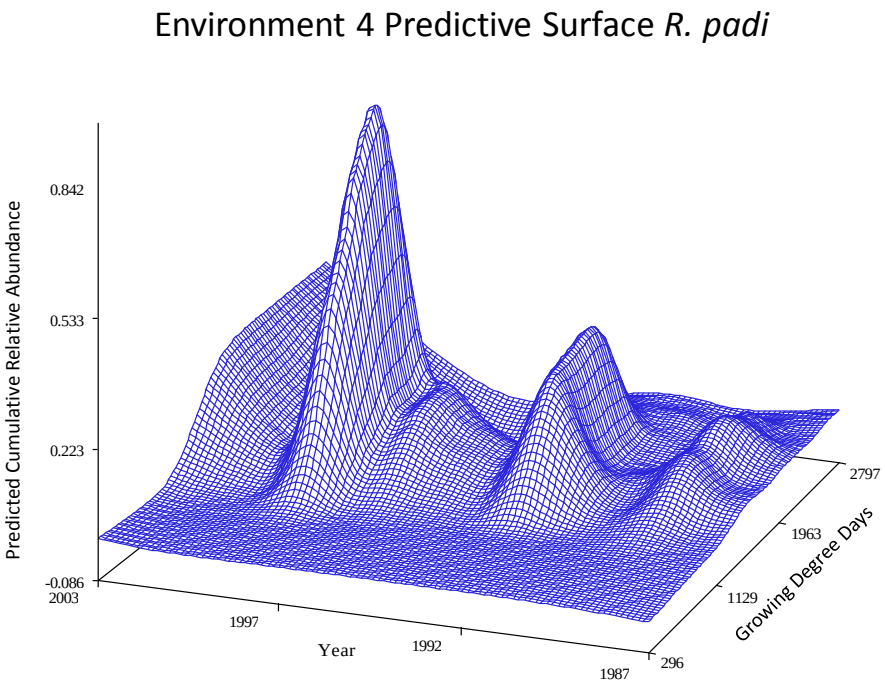

Figure 9. Example of predicted surface generated when fitting model from Eq (6) to data from environment 4, species *R. padi*.

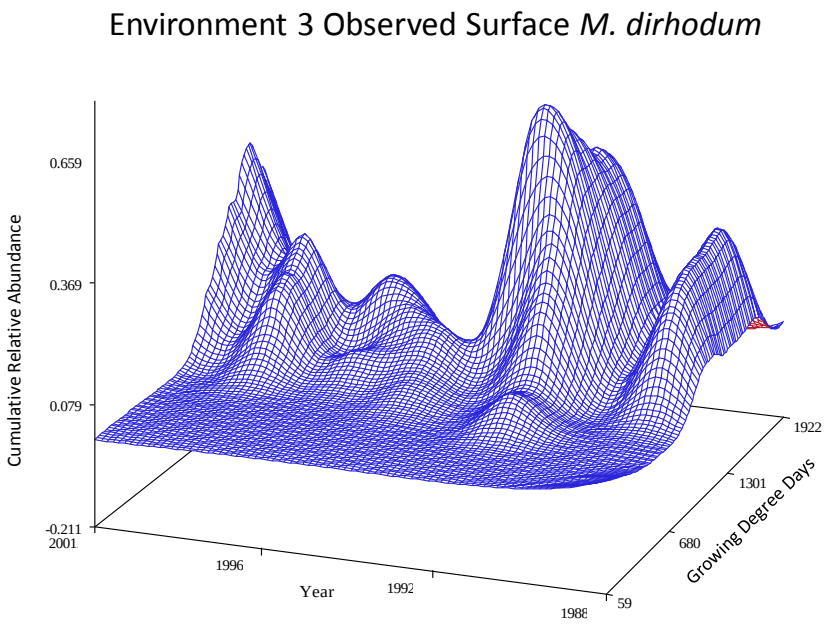

Figure 10. Observed surface of data data from environment 3, species *M. dirhodum*.

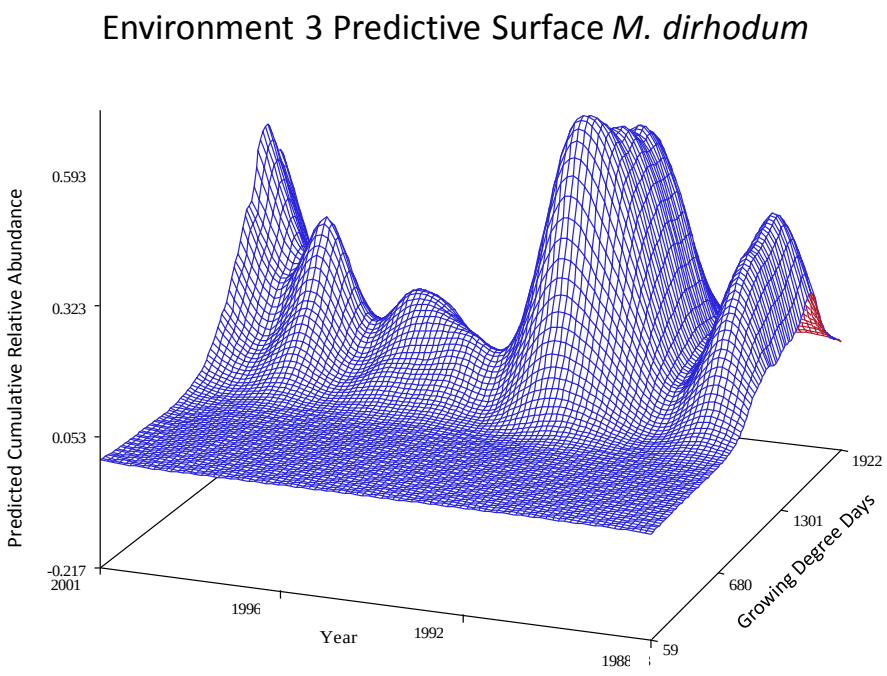

Figure 11. Example of predicted surface generated when fitting model from Eq (6) to data from environment 3, species *M. dirhodum.*

The surfaces presented within the respective panels of Figure 8 and Figure 9 are nearly identical, indicating that the model from Eq (6) is a good fit to the data. The 19 remaining observed and predictive surfaces of the autoregressive-environmental model can be viewed at the website mentioned previously. To further assess the adequacy of the fit, residual plots were also evaluated. Figures 12 and Figure 13 show the residuals obtained from the fitted model presented in Figure 8 and Figure 10, respectively.

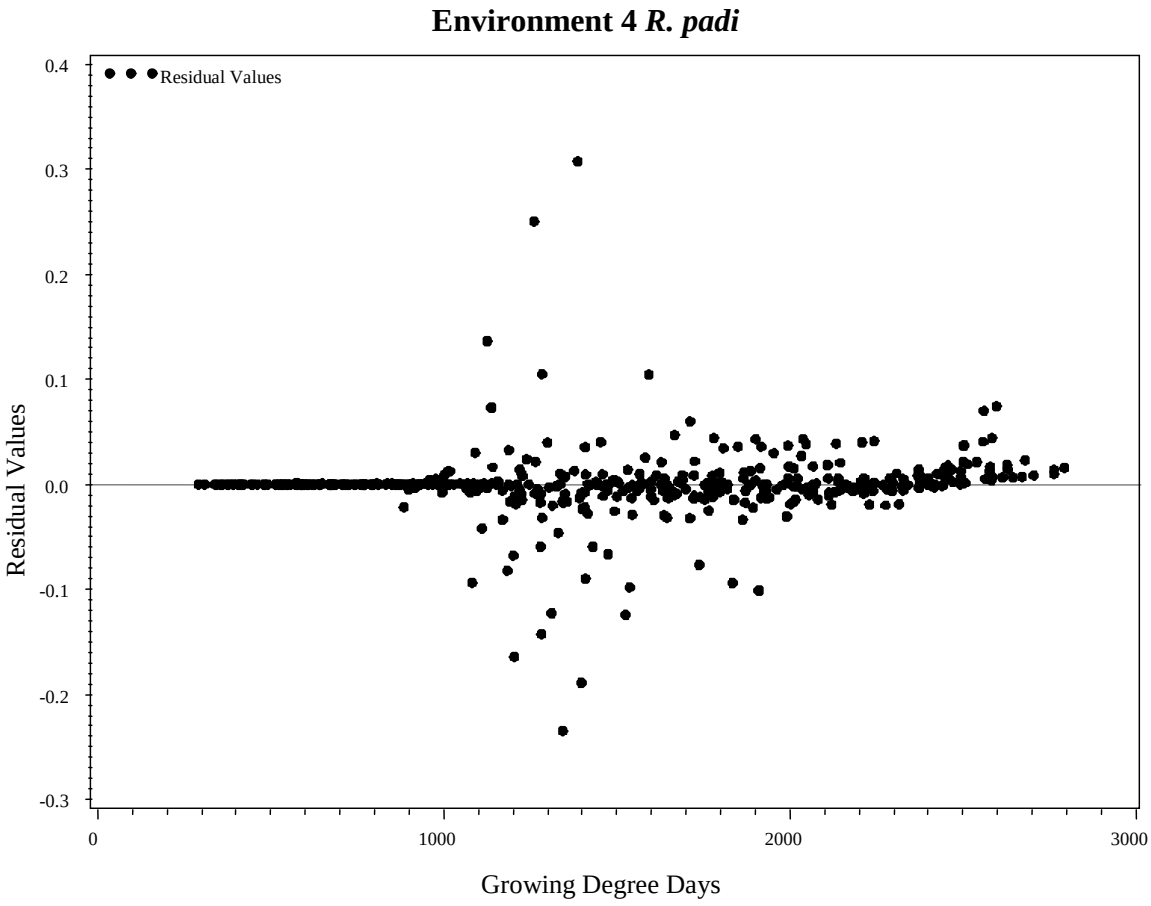

Figure 12. Residual plot generated from fitting Eq (6) to data for environment 4, *R. padi*.

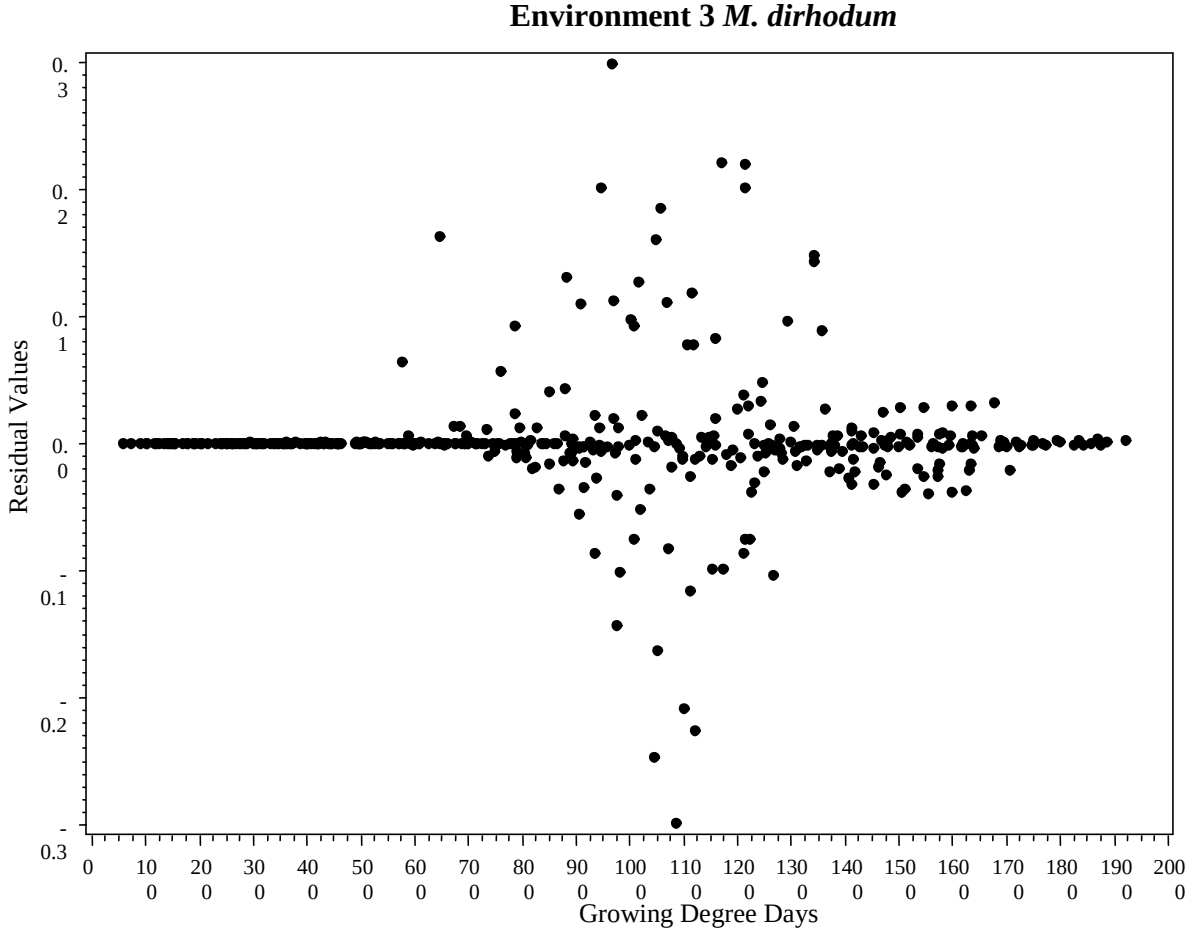

Figure 13. Residual plot generated from fitting Eq (6) to data for environment 3, species *M. dirhodum*

While the residuals show some patterning, the majority of values are close to zero and random in distribution. There is no sigmoidal shape to the residuals and relatively few extreme residual values.

Dummy Variable Regression Analysis

The purpose of conducting a dummy variable regression is to allow for the comparisons of the aphid accumulation process among environments. To enable these comparisons, a dummy variable was created for environment, resulting in 4 *full models* (one model for each species). Parameter estimates for the full, dummy variable model, fit to species *R. padi* are presented in Table 9 with their corresponding approximate standard errors, asymptotic 95% confidence intervals, t-values, and p-values. Parameter estimates for the remaining 3 full models fitted to the other species can be viewed at the REACCH website mentioned previously.

Table 9. Parameter estimates and corresponding approximate standard errors, t-values, p-values, and asymptotic 95% confidence intervals for full model fit to species *R. padi*. The suffix of each parameter refers to the environment number from which each estimate came.

Parameter Estimates								
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
$\beta 1$	0.004222	0.000238	176	17.75	<.0001	0.05	0.003753	0.004692
$L1$	1288.83	16.0431	176	80.34	<.0001	0.05	1257.17	1320.49
AR11	-0.2369	0.1922	176	-1.23	0.2194	0.05	-0.6162	0.1424
int_M1	0.4887	0.09868	176	4.95	<.0001	0.05	0.2940	0.6835
ln_Var_M	-2.3694	0.1124	176	-21.09	<.0001	0.05	-2.5912	-2.1477
$\beta 2$	0.008206	0.001035	176	7.93	<.0001	0.05	0.006164	0.01025
$L2$	1495.16	14.7948	176	101.06	<.0001	0.05	1465.96	1524.36
AR12	0.1036	0.3168	176	0.33	0.7442	0.05	-0.5217	0.7289
int_M2	0.2646	0.1343	176	1.97	0.0504	0.05	-0.00051	0.5296
$\beta 3$	0.006626	0.000531	176	12.48	<.0001	0.05	0.005578	0.007674
$L3$	1025.38	16.6614	176	61.54	<.0001	0.05	992.50	1058.27
AR13	-0.5727	0.1849	176	-3.10	0.0023	0.05	-0.9376	-0.2079
int_M3	0.5659	0.08340	176	6.79	<.0001	0.05	0.4014	0.7305
B4	0.008436	0.000878	176	9.61	<.0001	0.05	0.006703	0.01017
$L4$	1333.14	14.0142	176	95.13	<.0001	0.05	1305.49	1360.80
AR14	-0.1855	0.2319	176	-0.80	0.4248	0.05	-0.6431	0.2721
int_M4	0.1943	0.06765	176	2.87	0.0046	0.05	0.06078	0.3278
$\beta 5$	0.006430	0.000487	176	13.20	<.0001	0.05	0.005469	0.007391
$L5$	1243.38	15.0181	176	82.79	<.0001	0.05	1213.74	1273.02
AR15	-0.3982	0.1350	176	-2.95	0.0036	0.05	-0.6646	-0.1317
int_M5	0.3383	0.04839	176	6.99	<.0001	0.05	0.2428	0.4338

Within each full model, contrasts were conducted using likelihood ratio tests to compare various characteristics of the aphid accumulation process among environments. The motivation of the following contrasts arose from the natural geographic separation of environments provided by the cluster analysis. Environments 1 and 2 occupy the Northwestern part of Idaho, while environments 3, 4, and 5 occupy the Southeastern part of Idaho (refer to Figure 8). Therefore contrasts were conducted comparing characteristics of the aphid accumulation process for the

Northwestern environments to that of the Southeastern environments. Table 10 shows the contrasts of the onset parameter (L), the relative maximum parameters (AR1 and int_M), as well as the regression lines (common parameters β, L, AR1, and int_M), between the two Northwestern environments (1 and 2) and the three Southeastern environments (3,4, and 5) for species *R. padi*.

	Contrasts				
Species	**Label**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
D. noxia	Coincidence of Regression line: All Parameters NW vs SE	4	134	127.35	<.0001
D. noxia	Onset of Northwest Regions vs Southeast Regions	1	134	503.80	<.0001
D. noxia	Max Parms of Northwest Regions vs. Southeast Regions	2	134	0.87	0.4217
M. dirhodum	Coincidence of Regression line: All Parameters NW vs SE	4	161	123.56	<.0001
M. dirhodum	Onset of Northwest Regions vs Southeast Regions	1	161	474.26	<.0001
M. dirhodum	Max Parms of Northwest Regions vs. Southeast Regions	2	161	2.52	0.0836
R. padi	Coincidence of Regression line: All Parameters NW vs SE	4	176	50.58	<.0001
R. padi	Onset of Northwest Regions vs Southeast Regions	1	176	182.59	<.0001
R. padi	Max Parms of Northwest Regions vs. Southeast Regions	2	176	2.72	0.0690
S.avenae	Coincidence of Regression line: All Parameters NW vs SE	4	125	358.59	<.0001
S. avenae	Onset of Northwest Regions vs Southeast Regions	1	125	1380.84	<.0001
S.avenae	Max Parms of Northwest Regions vs. Southeast Regions	2	125	2.67	0.0734

Table 10. Contrasts of parameter estimates for environments 1 and 2 versus environments 3, 4, and 5. Row 1 for each species shows the contrast of all four key parameters (β, L, AR1, and int_M) between the two groups of environments. Row 2 for each species shows the contrast of the onset parameter (L) between the two groups of environments. Row 3 for each species shows the contrast of the relative maximum parameters (AR1 and int_M) between the two groups of environments.

In table 10, Row 1 for each species shows that the average of at least one of the 4 primary parameters of the Northwestern environments is significantly different from the average of at least one of the 4 primary parameters of the Southeastern environments for both species. Row 2 for each species shows that the average of the onset parameter for the Northwestern environments was significantly different from the average of the onset parameter for Southeastern environments at significance level $\alpha = .05$. Row 3 of the table for each species shows the average of both the relative maximum parameters for Northwestern environments was not significantly different from the average of both the relative maximum parameters for the Southeastern environments. The actual difference in the onset parameter between groups of

environments was also estimated to be 180, 228, 191, and 292 growing degree days for *D. noxia*, *M. dirhodum*, *R. padi*, *S. avenae* respectively.

The estimated difference in onset is positive, therefore it can be concluded the average onset is greater in terms of cumulative degree days for the Northwestern environments than the Southeastern environments. During the middle of the wheat growing season 200 growing degree days would be approximately 14 calendar days. This difference is relatively large and could indicate that aphid populations in Idaho have become relatively localized and are driven by local climatic factors. The implications of these contrasts are consistent with the inferences made by Halbert et al. in 1990 who indicated that suction trap collections reflect emigration of aphids from local colonies (20-50 miles from trap sites) rather than long distance migration.

Internal Validation

Internal validation was implemented through a bootstrap simulation of the residuals from each of the 20 species-environment models as described above. Once 1000 bootstrapped samples of residuals were obtained for each of the 20 models, summary statistics were calculated. Histograms and box-plots were also created to show the distribution of the residuals from each of the 20 models. Figure 14 gives an example of the residual distribution histogram for B=1000 bootstrapped residual samples obtained from the environment 4, species *R. padi*.

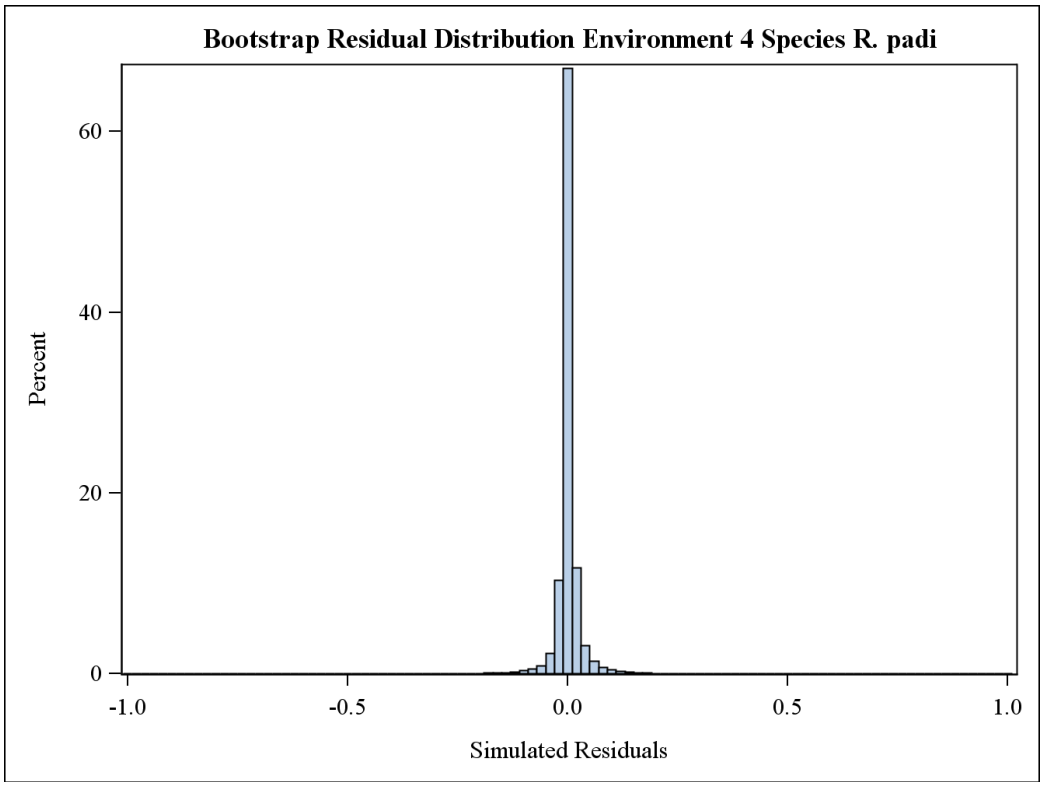

Figure 14. Histogram of the distribution of bootstrapped residuals for species *R. padi*, environment 4.

Figure 14 shows that the bootstrapped residuals from species *R. padi* in environment 4 were tightly distributed about zero. The mean and standard deviation for these residuals were 0.001092 and 0.02845 respectively. The first and third quartiles of the same residuals were -0.00451 and 0.00572 respectively. The remaining 19 environment-species combinations yielded similar results with residuals tightly distributed about zero. The results indicate that the model was stable and adequately fitted the simulated data sets. Residual distributions for the remaining 19 species-environment combinations can be viewed at the REACCH website.

External Validation

To validate each of the 4 full models, the Idaho suction trap sites that were not used in the analyses previously were classified into the environments created by the cluster analysis. The sites used included: Bonners Ferry, Caldwell, Conda, Craigmont, Holbrook, Mountain Home, Neely, and Preston. Only those sites with the most years of consecutive data were considered for inclusion in the external validation process. Even under this criterion, however, there were considerably fewer years available from the validation sites than in the primary sites.

Linear Discriminant Analysis

Before validation could proceed, the validation sites needed to be classified into the environments previously defined by the cluster analysis. A linear discriminant analysis (LDA) was therefore carried out to accomplish this task. LDA was performed under the assumption that the independent variables used to classify the sites (climate data) were approximately normally distributed with a common covariance structure. Because of the small number of observations, the assumption of normality was likely violated, and other non-parametric methods (k-nearest neighbor) were explored but yielded identical results, and therefore are not presented here. The LDA analysis was considered suitable to use with proportional prior probabilities (i.e. each site has an initial probability of .2 to be classified into each of the 5 environments).

The discriminant function was developed by using all climate data spanning 1986 to 2003 for each site (same variables as cluster analysis) and the corresponding environment memberships of each site as defined from the cluster analysis. The discriminant function was then used to classify the validation sites into the most appropriate environments. Table 11 shows the results of the LDA performed on the validation sites.

Table 11. Posterior probabilities of membership in corresponding environments resulting from linear discriminant analysis on validation data.

Posterior Probability of Membership in Environment							
Site	**Classified into Environment**		**1**	**2**	**3**	**4**	**5**
Bonners Ferry	**2**		0.0446	0.9554	0.0000	0.0000	0.0000
Caldwell	**1**		1.0000	0.0000	0.0000	0.0000	0.0000
Conda	**3**		0.0000	0.0000	1.0000	0.0000	0.0000
Craigmont	**1**		1.0000	0.0000	0.0000	0.0000	0.0000
Holbrook	**5**		0.0000	0.0000	0.0000	0.0000	1.0000
Mountain Home	**1**		1.0000	0.0000	0.0000	0.0000	0.0000
Neeley	**4**		0.0000	0.0000	0.0000	0.8483	0.1517
Preston	**5**		0.0000	0.0000	0.0000	0.0000	1.0000

The 5 rightmost columns of Table 11 show the posterior probabilities of membership in the corresponding 5 environments. For example, Neeley was determined to have an 85% probability of membership in environment 4 while also having a 15% probability of membership in environment 5. Table 11 shows Caldwell, Craigmont, and Mountain Home were classified into environment 1, Bonners Ferry into environment 2, Conda into environment 3, Neeley into environment 4, and Holbrook and Preston into environment 5. Because environments 1 and 5 both had multiple sites classified into them, the sites with the most data were selected for validation; Caldwell was selected for environment 1, and Preston was selected for environment 5. Typically the true membership of observations (sites) is known when performing a discriminant analysis, and therefore error rates can be calculated to assess the performance of the discriminant function. In this case, the true membership of the sites was unknown and consequently error rates were not calculable. Figure 15 shows the fit of the model for species *R. padi* in environment 4 to the data for Neeley.

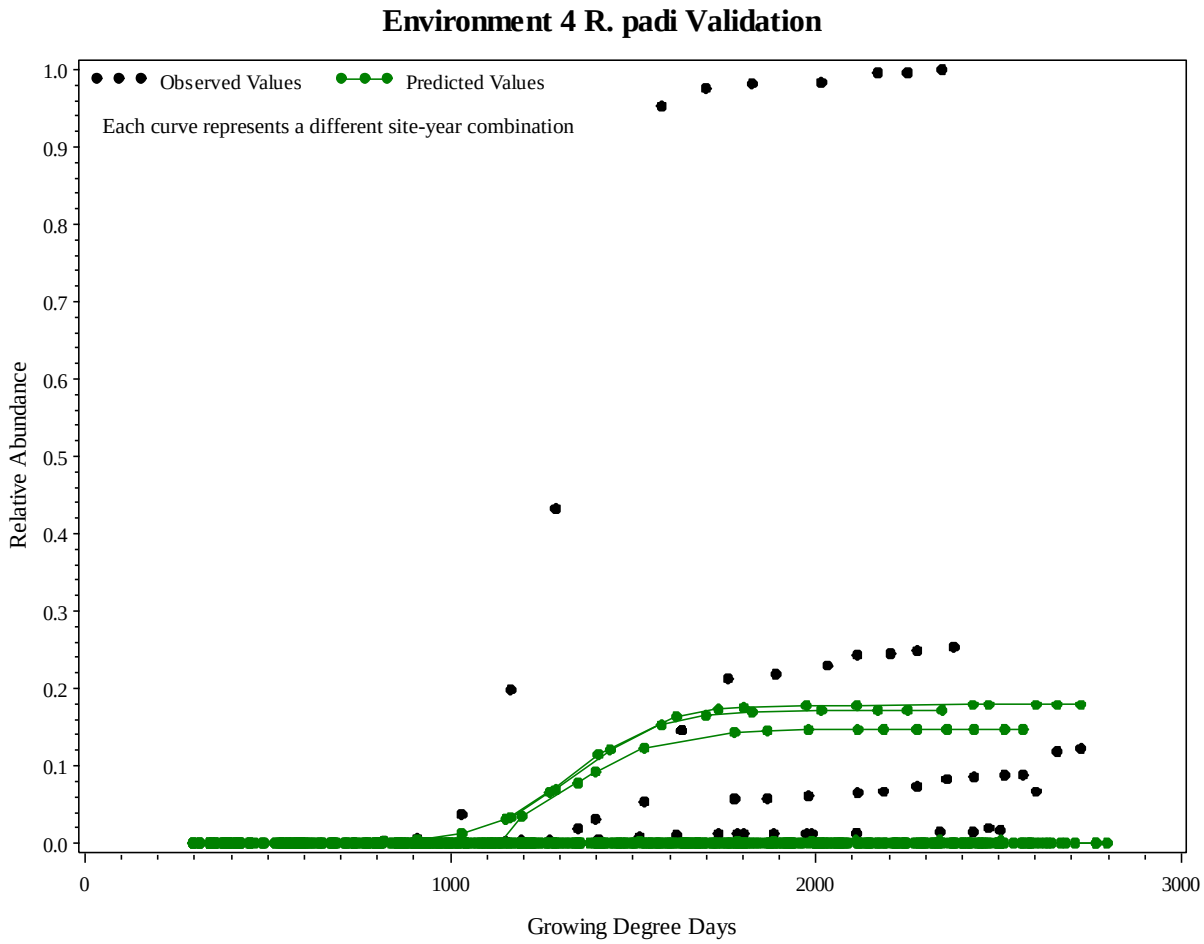

Figure 15. Fit plot of model fit to validation data for environment 4 (Neeley) species *R. padi*.

Figure 15 shows that the model for environment 4, species *R. padi* covers the data for Neeley containing smaller relative maximums, but does not cover the data from the year with the highest relative maximum. Because there are so few years of data from the validation sites, the maximum aphid count observed over the span of these data is likely not representative of the true maximum aphid count. Therefore when we scale the validation data, the data do not show the same patterns as the model building data which having between 13 and 17 years of data. It is also desirable to have a longer time series of data similar to that of the 12 sites selected for model building, because the relative maximum parameter is dependent on an autoregressive structure.

IV. Conclusions

Species-specific nonlinear regression models were developed to predict cumulative aphid abundances based on growing degree days and the relative maximum aphid abundance of the previous year. Models were developed for each combination of the 4 aphid species and 5 environments, resulting in a total of 20 predictive models. Internal validation was carried out using a bootstrap of the residuals for each of the 20 species-environment models and yielded tight residual distributions centered about zero. The results from the internal validation indicate that the modeling process implemented in this study provides unbiased estimates of predicted aphid abundances and can be applied to other datasets.

As of 1998, about 80 percent of wheat growers in Idaho implemented their pest management strategies based on field scouting thresholds of aphid abundances (Bechinski, 1998). The regression models developed in this study suggest the potential for modeling and generating forecasts that could decrease time and effort allocated to scouting prior to prediction of aphid movement. If these forecasts prove reliable, wheat producers could potentially time pesticide applications more accurately, alter planting effectively, and therefore, save money and time in the process.

In addition to the predictive capabilities of these nonlinear regression models, some ecological inferences could be made based on the parameter estimates. For example, it has been documented that host-alternating (holocyclic) cereal aphids can travel great distances between their winter and summer host plants (Bommarco et al., 2007), but it is not clear exactly how far for specific aphids and systems. This study detected significant regional differences in the onset parameter for each of the aphid species. The significant difference in onset suggests that aphid populations respond to temperature differently depending on the region, and therefore may be relatively local. Although it is difficult to statistically test these implications, it provides a motivation for future investigations. Furthermore, as a follow up to this study, one may consider modeling the cumulative aphid abundances based on Julian days, and conduct the same parameter estimate contrasts as reported in this study. Such an investigation would potentially provide insight into the magnitude of the migration of host-altering cereal aphids. It may also further our understanding of how the climates of the overwintering locations drive the aphid accumulation process compared to local climates of suction trap sites.

Although the modeling process described in this study was effective, there were limitations that should be considered when replicating this process on similar data. In this study, the suction traps were not operated consistently over the entire time period for all sites. For example, in 1986 the traps were not operated until August, because they were being assembled that year. There were also numerous sites that had years in which the traps were not operated at all. Because the data were scaled to proportions of the observed maximum aphid count for all years of data for each site, the gaps in the data could result in miscalculating these proportions. For subsequent analyses, it is recommended to use data consisting of consecutive years of data for as many years as possible. Also, suction traps do not capture the entire population of cereal aphids, but only those moving in the air column at the height of the traps. In general the numbers of aphids collected at suction traps are considered to be highly associated with the true total

number of aphids on a given crop (Bommarco et al., 2007). For future research, it is advised to incorporate other aphid collecting methods such as sweep netting and pan trapping in addition to the suction traps to obtain a sample that better represents the aphid population. Sampling via sweep netting targets established populations of aphids (rather than migrating aphids). Thus, integrating sweep netting with suction trapping could provide for a more complete understanding on the aphid accumulation process.

Proper assessment of the research limitations of this study and subsequent procedural adjustments could potentially enhance future research in this area. Because there were only individual suction traps set up at each sampling site, there were no replications within sampling events. In future studies, it is recommended to include multiple suction traps per sampling site, if programmatically feasible, to allow for a more appropriate likelihood form to be constructed for the data. Given proper replications (more than one suction trap per site), a negative binomial likelihood form may allow for a more sophisticated analysis of the site-year-species differences. Finally, when conducting the discriminant analysis to classify the validation sites into the environments determined by the cluster analysis, the accuracy of the classifications could not be assessed because the true memberships of the validation sites were unknown. For this reason, it is advised to interpret the results of classification methods, such as LDA, cautiously as the effectiveness of the method is relatively difficult to determine in this case.

V. References

Abatzoglou, John T. "Development of gridded surface meteorological data for ecological applications and modelling." International Journal of Climatology 33.1 (2013): 121-131.

Allison, David, and Keith. S. Pike. "An inexpensive suction trap and its use in an aphid monitoring network." J. Agric. Entomol 5.2 (1988): 103-107.

Araya, J.E., J.E. Foster, and S.G. Wellso. “Aphids as cereal pests.” Purdue University Agricultural Experiment Station Bulletin 509 (1986).

Bates, Douglas M., and Donald G. Watts. Nonlinear regression: iterative estimation and linear approximations. John Wiley & Sons, Inc., 1988.

Bechinski, E., and K. A. Loeffelman. 1998. IPM and the Idaho wheat industry – Results of grower surveys (1997). SA-8. University of Idaho Cooperative Extension System, Moscow, ID.

Bommarco, Riccardo, Simon Wetterlind, and Roland Sigvald. "Cereal aphid populations in non-crop habitats show strong density dependence." Journal of Applied Ecology 44.5 (2007): 1013-1022.

Cochrane, Donald, and Guy H. Orcutt. "Application of least squares regression to relationships containing auto-correlated error terms." Journal of the American Statistical Association 44.245 (1949): 32-61.

Davis, T. S., John Abatzoglou, Nilsa A. Bosque-Pérez, Susan E. Halbert, Keith Pike, and Sanford D. Eigenbrode. "Differing contributions of density dependence and climate to the population dynamics of three eruptive herbivores." Ecological Entomology 39.5 (2014): 566-577.

Efron, Bradley, and Robert Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." Statistical science (1986): 54-75

Fox, John. "Applied regression analysis and generalized linear models." Sage Publications, Second Edition. (2008).

Fujikawa, Hiroshi, Akemi Kai, and Satoshi Morozumi. "A new logistic model for Escherichia coli growth at constant and dynamic temperatures." Food Microbiology 21.5 (2004): 501-509.

Halbert, S., John. Connelly, and Larry. Sandvol. "Suction trapping of aphids in western North America (emphasis on Idaho)." Acta Phytopathol. Entomol. Hung 25 (1990): 411-422.

Halbert, Susan E, Larry E. Sandvol, and Guy W. Bishop. "Aphids Infesting Idaho Small Grain and Corn." Moscow, Idaho: University of Idaho, Cooperative Extension Service, Agricultural Experiment Station, College of Agriculture, (1988). Print.

Honek, A., and Z. Martinkova. "Host plant age and population development of a cereal aphid, Metopolophium dirhodum (Hemiptera: Aphididae)." Bulletin of entomological research 94.01 (2004): 19-26.

Pike, K. S., D. W. Allison, G. Low, G. W. Bishop, S. Halbert, and R. Johnston. "Cereal aphid vectors: A Western Regional (USA) monitoring system." World (1990).

Price, William J., Bahman Shafii, and Steven S. Seefeldt. "Estimation of dose-response models for discrete and continuous data in weed science." Weed Technology 26.3 (2012): 587-601.

Rencher, Alvin C., and William F. Christensen. "Methods of multivariate analysis." Wiley Series in Probability and Statistics.Third Edition (2002).

SAS Institute. 2011. SAS OnlineDoc, Version 9.2. Cary, NC: SAS.

Shafii, Bahman, and William J. Price. "Estimation of cardinal temperatures in germination data analysis." Journal of agricultural, biological, and environmental statistics 6.3 (2001): 356-366.

Slafer, G. A., and H. M. Rawson. "Base and optimum temperatures vary with genotype and stage of development in wheat." Plant, Cell & Environment 18.6 (1995): 671-679.

"Welcome to the Idaho Wheat Commission." Idaho Wheat Commission. N.p., n.d. Web. 16 Mar. 2015.