Forecasting Models For Invasive Species (Draft)

Gotelli et al.

6/23/2020

## Introduction

A major research focus in ecology is forecasting how species and communities will change in response to climate warming and other environmental drivers (). Such shifts can potentially bring together new combinations of species that previously have not co-ccurred before (); the geograpnic ranges of many non-native species do seem to be changing recent decades. Some of these species have strong direct effects on native assemblages as predators (), competitors (), and parasites (), as well as indirect cascading effects through habitat modifications ().

Although a variety of forecasting models are currently used to predict changes in geographic ranges (), most of them rely on three basic data layers. The first is a set of geographic coordinates that specify the locations of current populations of the species. These records come from data portals such as GBIF (), iNaturalist (), and other sources of museum records and individual surveys in which the presence of species has been recorded.

The second data layer is a set of environmental variable and climate variables, especially measures of temperature and precipitation, that have been measured, interpolated, and then downscaled for the locations of the individual populations. These first two data layers can be analyzed with a variety of species distribution models (SDMs), including MaxEnt (), and more recent neural network () and random forest models (). These algorithms seek to estimate and optimize the suitability of the habitat in each pixel of the landscape as a function of environmental variables. The fitted model can then be used to forecast the probability of occurrence (or habitat suitability) at any location on a (terrestrial map) as a function of contemporary climate variables in that location.

The third data layer is a forecast of future environmental conditions, often derived from climate change models (), at different locations. This forecasting layer for future conditions can be combined with the species distribution model that was fit with contemporary climate variables to estimate suitability of future conditions for species occurrences ().

In spite of the diversity of methods for analyzing data of this form, they all suffer from two important limitations. The first limitation is that these models assume that populations are in equilibrium with their environmental conditions (). In this equilibrium state, contemporary population occurrences would represent locations that are suitable for populations to persist. However, the pace of climate change has accelerated in recent decades (), and even short-lived species may no longer be in equilibrium with their current abiotic environment. Indeed, many recent studies have emphasized that rates of spread of species may not be able to keep pace with the projected “climate velocity”, and many models of climatic debt () emphasize a disequilibrium between the current composition of a community, and the composition that would be predicted assuming an equilibrium with climate ().

The second limitation is that most species distribution modeling is based on presence-only data. In the absence of widespread standardized surveys, we have only the location of the occurrences of a species, but not locations in which absences have been explicitly recorded. Without true absences, most algorithms impute “pseudo-absences” by randomly placing absence points (10,000 in the default settings for MaxEnt) uniformly within an arbitrarily bounded domain that encompasses the recorded species occurrences. But for many groups of organisms (especially plants and invertebrates), many of the locations where pseudoabsences are imputed probably contain the species, but have not been properly sampled. The location of these pseudoabsences has a very large affect on the outcome of the model (), but often the domain boundaries provided by users are arbitrary and can simply reflect the original raster that contained the occurrence records.

These limitations have made the forecasting of both native and non-native species highly uncertain (). In this paper, we propose an alternative framework for forecasting the distribution of non-native species, that does not depend on the spatial pattern of population presences and absences, does not assume that populations are in equilibrium with their environment, and does not assume that the presence of a population is a good indicator of the climatic niche of a species.

Instead, we argue for species distribution modeling that is based on estimates of short-term population growth rates derived from demograpic data (). Basic life-table data describe rates of birth, individual growth, and death as a function of the age or size or stage of the organism (). These data can be collected in a variety of ways, and can be converted to an estimate of r (). In a model of exponential population growth, r is the intrinsic rate of increase, measured in units of time/individual\*individual.

Alternatively, when demographic rate data are not available, r can be estimated from sequential population censuses as loge(N(t+1)/N(t)). With either life table or population growth data, r is a natural metric for the potential for population persistence in different environmental conditions. If r is greater than 0.0, instantaneous birth rates exceed instantaneous death rates, and the population is able, at least in the short-term, to persist and grow. If r is less than 0.0, instantaneous death rates exceed instantaneous birth rates, and the population will, in the long-run, decline to extinction.

Here, we use examples from laboratory studies, field experiments, and field observational studies to illustrate three case studies of invasive species for which we forecast population growth rates at sites across a map of the continental U.S. These analyses are based only on demographic data, and do not use species occurrence records to construct a forecast.

In each case, we find that the models generate realistic maps of habitat suitability when fitted with contemporary climate layers. These predictions based on demographic data are either comparable to, or better than, published predictions based on traditional species distribution models.

## Methods

The general framework for these analyses is simplified and adapted from Gotelli amd Ellison (2002, 2006). Three general equations describe the relationship between demography, environmental variables, and geographic location:

The function establishes the relationship between population growth rate and the environmental conditions . In an experimental study, a life table response experiment (Caswell 200x) would be used to rear organisms under constant conditions. Each level of temperature or some other continuous environmental variable would represent a separate, replicated treatment.

Alternatively, for a field based population study, a researcher would measure population size at two consecutive time steps, and the associated environmental conditions over that same interval. For both the experimental and non-experimental data, the fitted function describes the relationship between and . For an environmental variable , we propose the simplest model for predicting should be

In a least squares regression model, this is a quadratic function with fitted constants , , and . We favor this model for theoretical reasons: for any environmental variable, there should be an optimal, intermediate level of that maximizes population growth rate. Population growth rate decreases as the environmental variable moves away from the optimal value. If there is a range of values of within which is positive, this range will include the peak. Although we have presented this framework in the context of a single environmental driver, the model can be easily extended with multiple regression to incorporate 2 or more drivers, such as precipitation and temperature.

The function describes the relationship between the environmental variable and the location of a point in the landscape, specified by its geographic coordinates . This data layer is usually created from smoothed and interpolated climate data (), and is rarely modeled by the investigator. The data layer may be available for past, current, or future conditions based on climate forecasting and hind-casting models ().

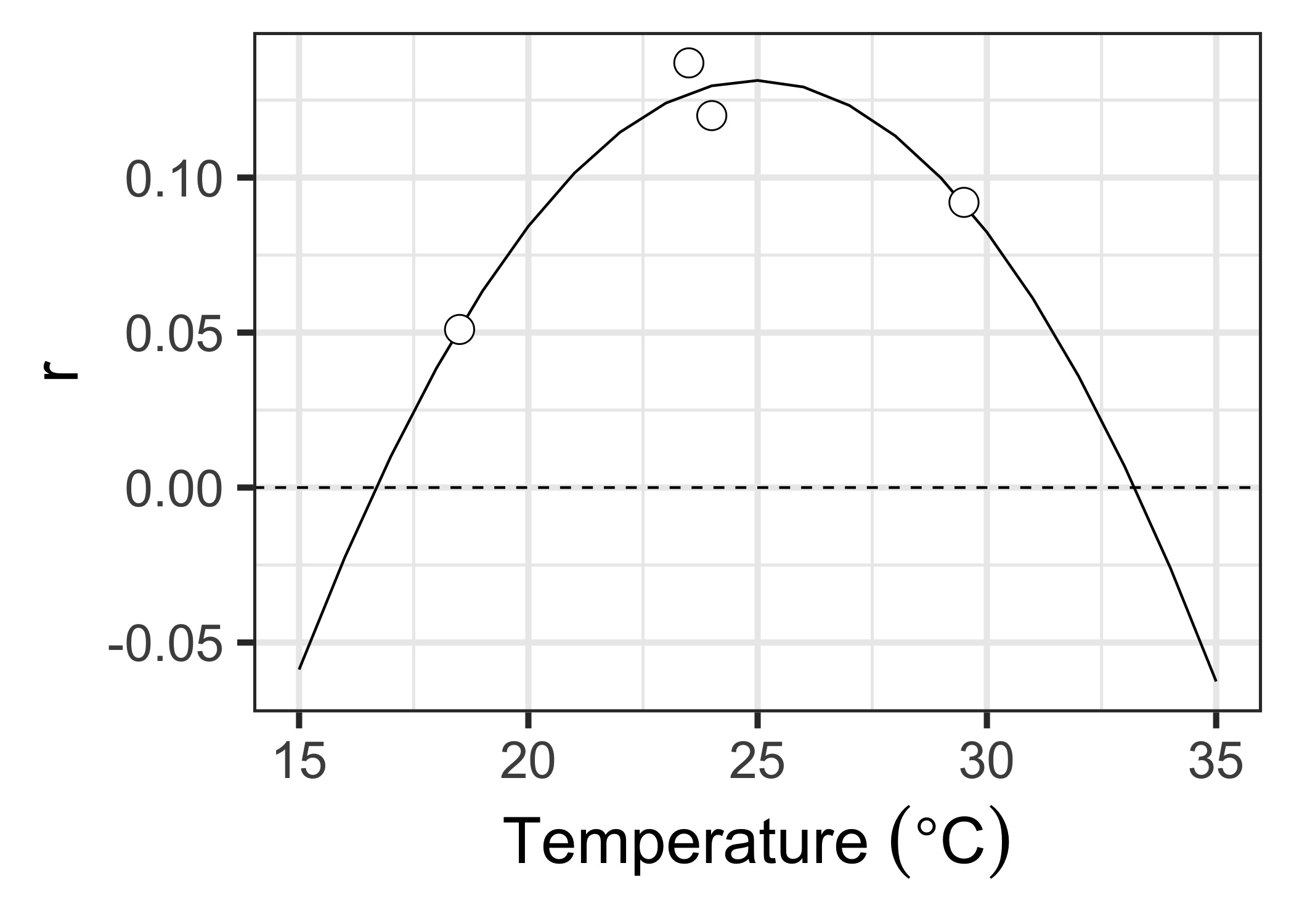
The final step is to estimate r for each pixel in the landscape by combining the calculations from (1) and (2). r can be interpreted as a continuous variable representing the ability of the population to achieve short-term population growth, given the set of environmental conditions at location . is an unbounded continuous variable, and the more positive the value of , the higher the growth rate, and presumably the better the environmental conditions are for population growth. A critical value in this continuum is , which is the dividing line between increasing and decreasing population growth rate. Modeling r as a function of environmental variables is an empirical realization of Hutchinson’s () definition of the niche: an -dimensional hypervolume describing a combination of variables for which .

## Case Study #1: Medfly

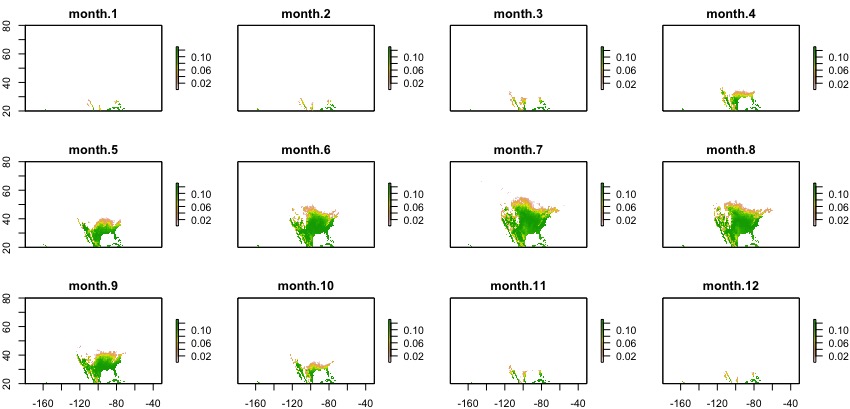
We begin with the simplest kind of data for demographic modeling: life table analysis of organisms reared im constant laboratory conditions. Vargas et al. (2000) describe a lab study in which cohorts of the Mediterranean fruit fly Ceritata capitata were raised in one of four different temperature treatments. These tephritid flies are native to sub-Saharan Africa, and are serious agricultural pests throughout the world.

In each of the four treatments in the Vargas et al. (2000) experiments, cohorts of eggs were reared at at different minimum and maximum temperatures, which were maintained on a daily 9h:9h cycle, with 3 hours of ramping between the maximum and minimum. For our analyses, we used the average of the maximum and minimum to characterize the treatment. Vargas et al. (2000) present many different demographic rates from these experiments. We had hoped to model age-specific survivorship and fecundity in each treatment as a function of temperature. However, the raw data for these life table analyses were not presented. Instead, we have used the average population growth rate r, reported in Table 6 of Vargas et al. as the response variable.

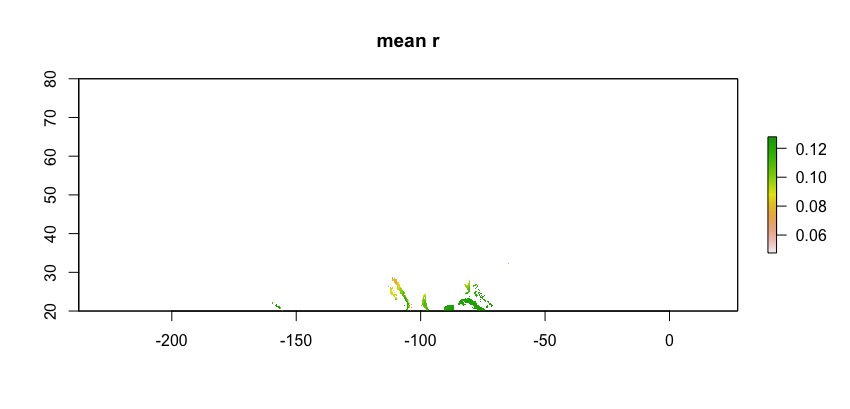
Figure 1 shows the relationship between population growth rate and average temperature for each of the 4 treatments. The smooth black curve is the best-fit quadratic regression line. This model predicts a range of viable temperatures for population growth between 16.7 and 33.2C. The model predicts maximum population growth rate at a temperature of 24.9C, which is close to estimates of an earlier study of maximum net reproductive rate at a constant temperature of 24C (Vargas et al. 1997).



Next, we used this prediction function to estimate medfly population growth rate across the continental U.S. For these predictions, we used current WorldClim values for average monthly temperature. Figure 2 shows the pixels in which r > 0 for each month of the year, based on the prediction function in Figure 1. There is wide spatial variation, with potential medfly distributions covering much of North America in the warmer months, but being absent from the continental U.S. during most of the colder months.



Finally, we assumed for populations to be self-sustaining, they would need to have r > 0 for all 12 months of the year. Figure 3 illustrates this pattern, and predicts that the medfly would be restricted to the southern most tip of the Florida peninsula and the Florida Keys.



A less restrictive model (not shown) would consider viable pixels to be those in which the average population growth rate across all 12 months is greater than 0. This model would extend the predicted distribution of the medfly up the Florida peninsula, and down a narrow coastal band of the Gulf Coast. However, this model seems unrealistic because the medfly generation time is < 40d and because the cold months occur consecutively in the winter season, so that population growth that occurred during the summer could not sustain local populations through the cold winter months.

How does this forecast from a demographically-based species distribution compare to the actual distribution of the medfly? Surprisingly well. Medfly populations in the continental U.S. are currently found only in southern Florida, which nicely matches our model predictions. Medfly has also been eradicated from agricultural areas in southern California, but high levels of watering may have made the microclimate there more comparable to the warm moist habitat of southern Florida.

In contrast, MaxEnt, GARP, and other species distribution models based on occurrence records of the medfly in its native range of sub-Saharan Africa and other introduced sites substantially over-estimate its range in North America. These models do not incorporate the acute temperature sensitivity of medfly population growth rate to temperature, and may also be based on occurrences in places that experience less seasonal variability in temperatures than continental North America.

## Case Study #2: *Bromus*

As an important alternative to traditional life-table experiments, lambda can be estimated from short-term field measurements of abundance. Specifically, if abundance is measured in the same plot in two consecutive time periods, then . If abiotic variables such as temperature or precipitation can be measured at the plot scale, these can be used to generate predictive functions as in Equation 2.

We illustrate this method with 7 years of annual census data for cheatgrass….(*need full methods section here from Michael and Alden*). The data consist of measurements from experimental snow-fence plots. Although some plots were measured for several consecutive years, there are not enough data for a formal time series analysis. Therefore, we treated each pair of observations on a single plot in two consecutive years as an independent data point.

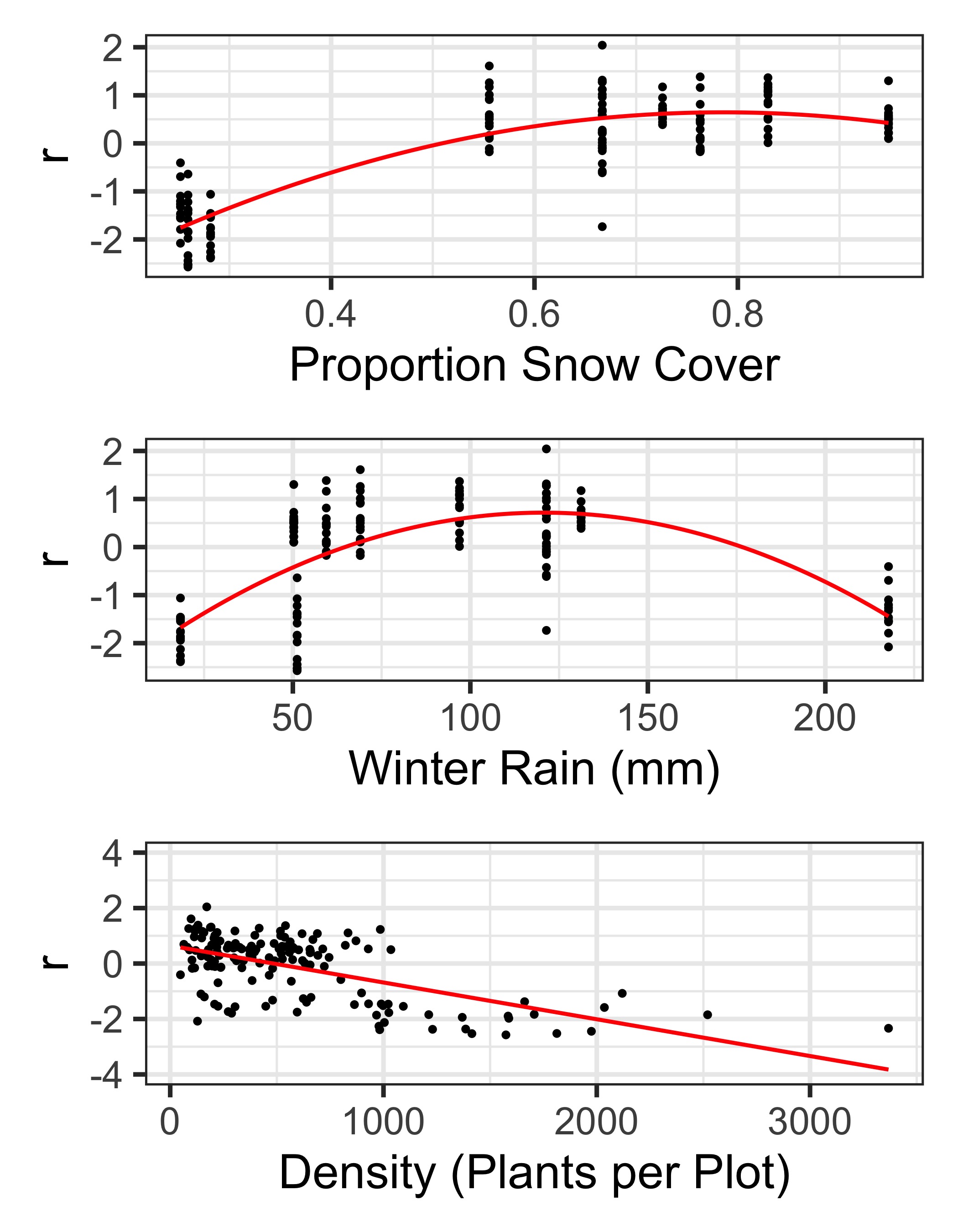
For covariates, we analyzed the following abiotic variables: growing season snow, growing season rain, growing season total precipitation, growing season degree-days, winter snow, winter rain, winter total precipitation, winter degree days with no snowpack, winter degree days, snow cover.

We analyzed the relationship between these variables and r using linear and quadratic regression models. For most variables, adding a quadratic term substantially improved the r2 and model fit, The quadratic coefficients were usually negative, implying a unimodal peak of population growth at an intermediate level of the variable.

We also included initial population density as a potential covariate. If population growth is density-dependent rather than exponential, then growth rate should decline with increasing density. In the simplest logistic growth model, the relationship between r and population size is linear. We confirmed this linear relationship for the cheatgrass data, with no significant improvement from adding a quadratic term. For this logistic growth model, maximum population growth rate is achieved as population size approaches 0. The intercept of this model estimates the maximum population growth rate as = 0.642016 individuals/individual/year, which corresponds to a doubling time of 1.08 years. However, many individual plots had recorded estimates much higher than this. From this model, we can also estimate the carrying capacity of the population as the population size for which = 0. With a fitted linear slope of -.001326, this estimated is individuals per plot.

We selected variables for the final model by evaluating the statistical significance of the individual coefficients, and changes in the adjusted r2, which imposes an AIC penalty on complex models with many parameters. We also tried to include interaction terms between the predictor variables, but these never improved the model fit, so they were not retained.

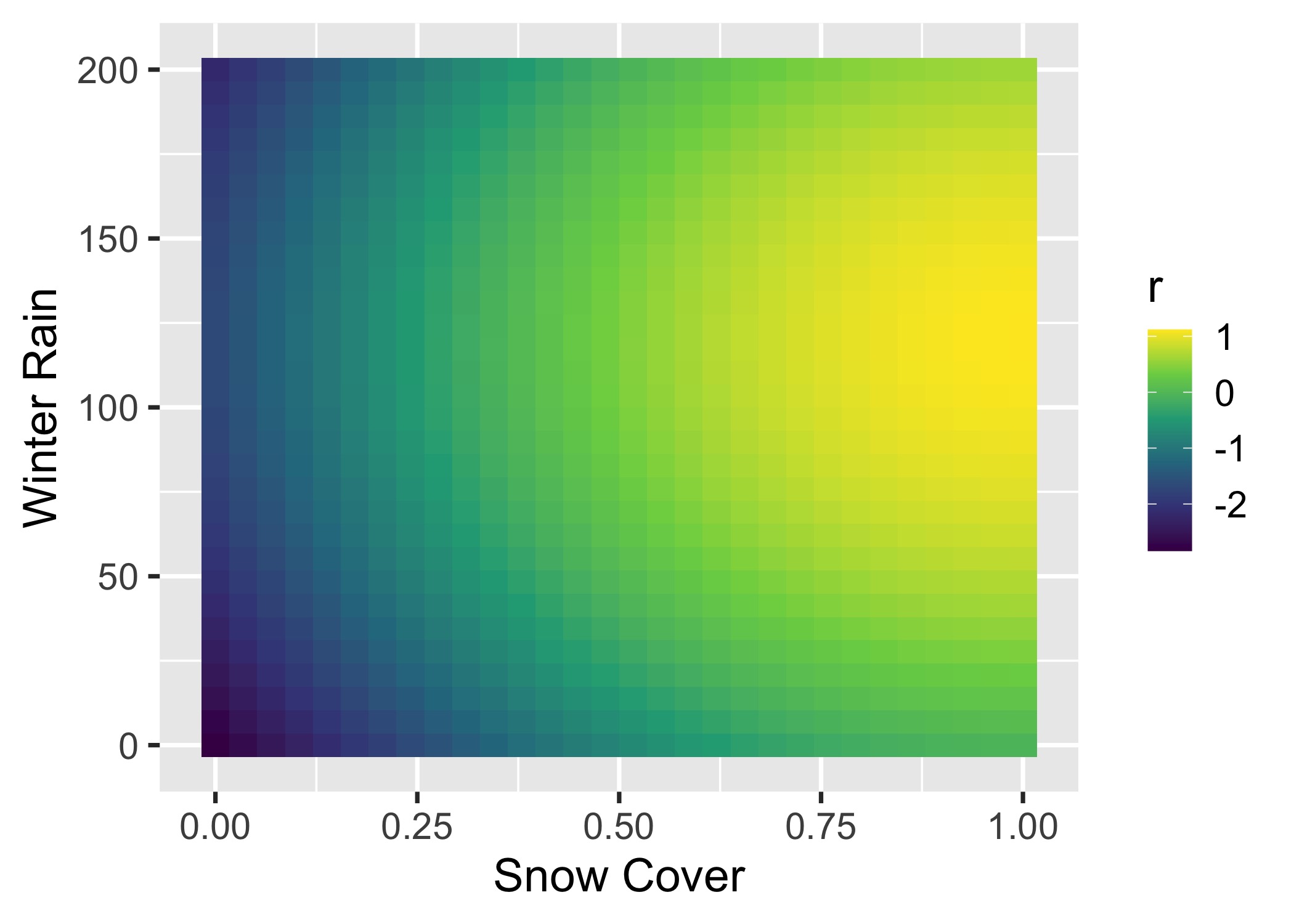
The final best-fitting model included quadratic functions for snow cover and winter rain, and a linear function for initial density (Figure 4):



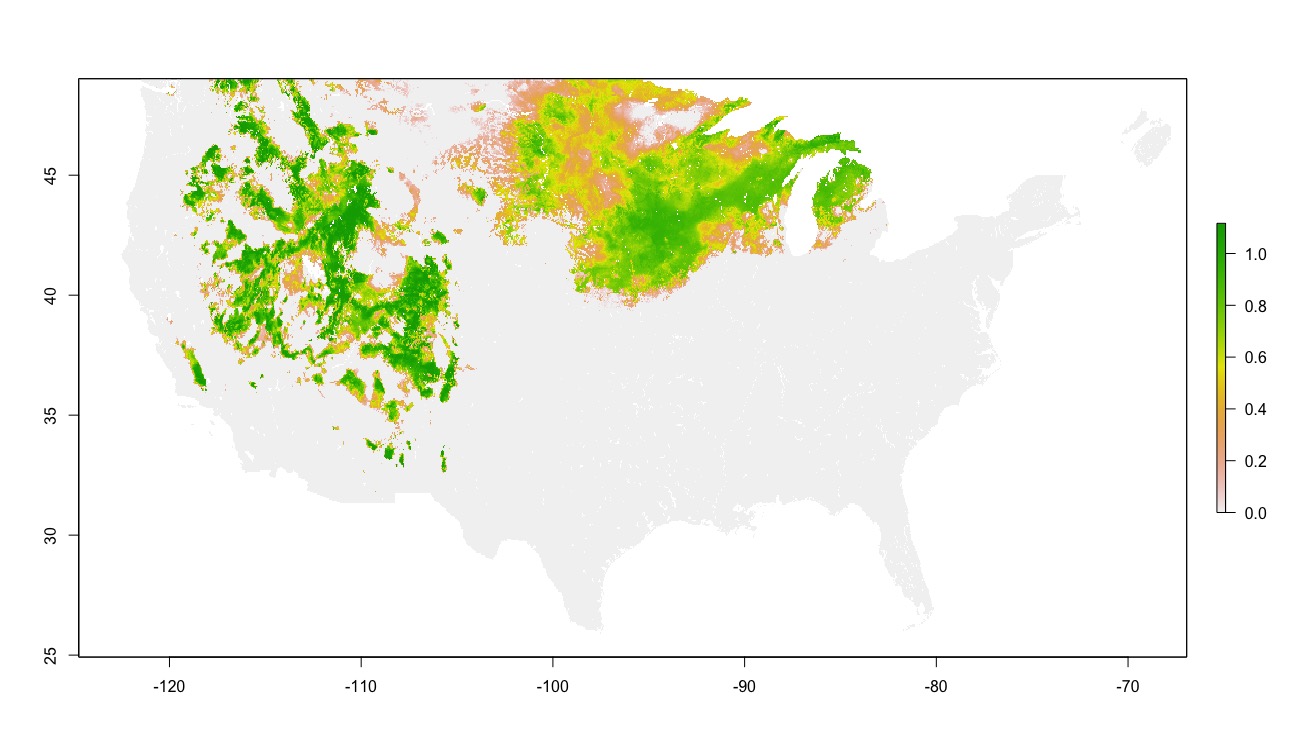
The final model with these three variables explained 79.1% of the variance in *r*. All of the model coefficients were statistically significant, except for the quadratic term for snow cover (p = 0.12).

For purposes of forecasting, we used mapped data layers for snow cover and winter rain. We set population size to a constant of 0, so that population growth rate was estimated for an empty landscape in which new populations might become established. Although density was set as a constant for the forecasts, it is still important to have it in the model so that the coefficients for snow cover and winter rain are properly estimated.

Figure 5 illustrates a heat map in which r is plotted as a function of snow cover and winter rain. This plot is a visualization of the Hutchinsonian niche, because it indicates combinations of abiotic factors for which for colonizing populations.



Finally, we apply the model using data layers for contemporary snow cover and winter rain. The model predictions are in line with the current distribution of Bromus… In contrast, more traditional SDMs of Bromus based on species occurrence records suggest that… (\*need input from Michael and Alden here)



These