The Data

The data is about measured populations of fish from Tallgrass Prairie national preserve.

	Season	Normal	Flood_6Months	MultipleFloods_6Months	Drought_6Months	Flood_Drought_6Months	ToleranceCode	ScientificName	Freq
0	2001	1	0	0	0	0	I	Luxilus cardinalis	787
1	2001	1	0	0	0	0	I	Notropis percobromus	1
2	2001	1	0	0	0	0	I	Notropis topeka	7
3	2001	1	0	0	0	0	I	Notropis volucellus	1
4	2001	1	0	0	0	0	I	Pimephales tenellus	23
...	...	...	...	...	...	...	...	...	...
355	2019	0	0	1	0	0	T	Pimephales notatus	6
356	2019	0	0	1	0	0	T	Pimephales promelas	1
357	2019	0	0	1	0	0	T	Semotilus atromaculatus	75
0	2011	0	0	1	0	0	I	Luxilus cardinalis	0
1	2012	0	0	1	0	0	I	Luxilus cardinalis	0

Figure 1: A Portion of the Data

The data contains columns for the year, whether flooding or drought occurred six months before the measurement of population, the species, the species' tolerance for poor water quality and human disturbance, and the population of that species in the park that year.

I	Intolerant
M	Moderately Tolerant
T	Tolerant
U	Unknown

Figure 2: Tolerance Code Meanings

I took this dataset, and fit the same Bayesian model to the full dataset and four subsets based on the four different tolerance code, with the goal of comparing how different environmental factors might possibly affect the fish populations, and whether those factors would have more effect on species of different tolerance codes.

The Model

To explore this data, I built a Bayesian model that accounts for the flooding and drought data.

It assumes the population of a species of fish for any given year can be modeled as a Normal distribution, with a mean consisting of a linear combination of several other normally distributed variables, and a variance modeled by a half-cauchy distribution.

Once this model is fitted to a dataset using PyMC3, it can take the drought/flood conditions for a year and predict the average number of fish of a single species in the population.

```
# Define priors
sigma = HalfCauchy("sigma", beta=10, testval=1.0)
intercept = Normal("Intercept", 0, sigma=20)
flood_coeff = Normal("Flood_6Months", 0, sigma=20)
multiflood_coeff = Normal("MultipleFloods_6Months", 0, sigma=20)
drought_coeff = Normal("Drought_6Months", 0, sigma=20)

# Define likelihood
likelihood = Normal("y", mu=intercept - flood_coeff * fish_data.Flood_6Months
                                      - multiflood_coeff * fish_data.MultipleFloods_6Months
                                      - drought_coeff * fish_data.Drought_6Months,
                    sigma=sigma, observed=fish_data.Freq)
```

Figure 3: Code for the Model

The Results

After that model was built, I then used the NUTS sampler to fit it to different subsets of the population data.

Intercept

Intercept is the variable in the model that is not based on any of the input values. It can be considered a sort of "baseline" value that represents the expected number of fish in a sample, before any of the climate conditions are considered.

Since each species has one tolerance category, the difference in this value only represents the difference in mean population count for all the species in a dataset across all years. These results tell us that the species categorized as "Tolerant" had, on average, twice as many fish in all the years' samples for a tolerant species as there were for a species in the "Intolerant" category.

This data point will be useful for helping compare the impact of climate events on different categories.

Classification	Mean Value
All Data	59.748923116347434
Tolerant	57.2367494043198
Moderately Tolerant	46.19413763532205
Intolerant	29.272990398236775
Unknown	0.5872680087324352

Flood_6Months Normalized

This coefficient represents the impact of the value of the flood_6months column, which is 1 if there was one flood within six months before the sample was taken, or zero otherwise.

What I mean by "Normalized" for these other coefficients: In order to compare the mean value of the coefficient between data sets, I divided the coefficient by the intercept variable. That means that the values in this table are a way of comparing the magnitude of the coefficient, rather than just the raw number.

Here we see all the values are negative. Since the model assumed this coefficient would have a negative impact on the expected number of fish in a sample, it seems like this variable actually has a positive impact on the expected number of fish in all but the "Unknown Tolerance" categories.

Classification	Mean Value
All Data	-0.29895795460160207
Tolerant	-0.3565970811676861
Moderately Tolerant	-0.4508351508215578
Intolerant	-0.3167720036562346
Unknown	0.5159925869893939

MultipleFloods_6Months Normalized

This value was also normalized to compare between categories. It represents the impact of the value of the MultipleFloods_6Months column, which is 1 if there was more than one flood within six months before the sample was taken, or zero otherwise.

Since all the values are positive, this variable had a negative impact on the expected number of fish in a sample. There doesn't seem to be a correlation between the tolerance of the species and the impact of multiple floods, although we again see that the species with unknown tolerance are most severely impacted.

Classification	Mean Value
All Data	0.4026912471802705
Tolerant	0.23574510933068404
Moderately Tolerant	0.09375774066204416
Intolerant	0.31714123248177567
Unknown	0.6050900233014277

Drought_6Months Normalized

This value was also normalized to compare between categories. It represents the impact of the value of the Drought_6Months column, which is 1 if there was a drought within six months before the sample was taken, or zero otherwise.

Again, since the values are positive, this variable had a negative impact on the expected number of fish in a sample. This time there is a correlation between tolerance category and impact, but it seems like more tolerant species are more affected by the drought. We also see that unknown tolerance species are most impacted. Perhaps the unknown ones are even less tolerant?

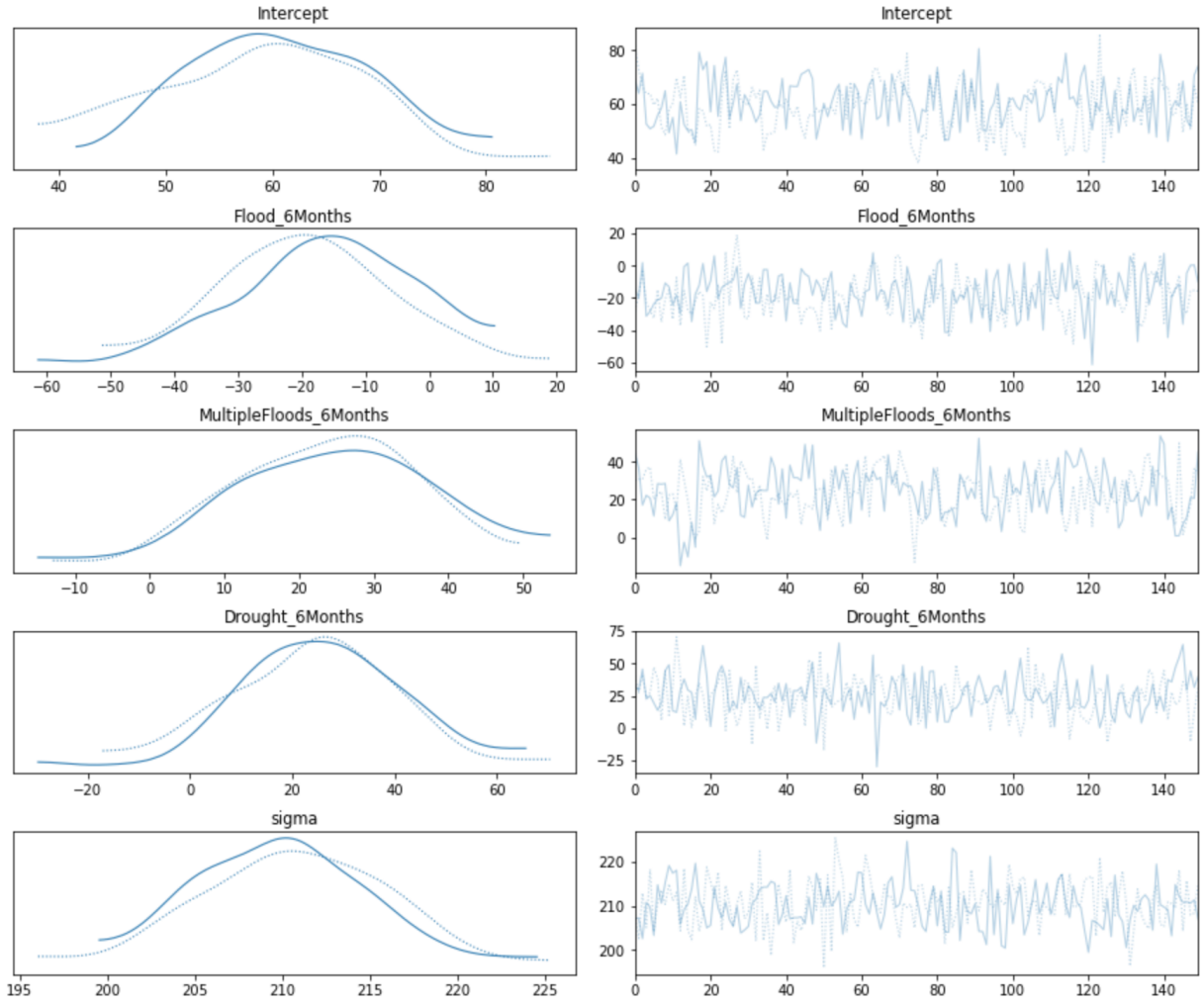
Classification	Mean Value
All Data	0.41801142986850914
Tolerant	0.30268707720812876
Moderately Tolerant	0.21181955553293932
Intolerant	0.1140208128016585
Unknown	0.8181050847388662

A1 Full Results

The following pages have the trace charts for the variables in the model, for the entire dataset and the subsets I compared. The charts on the left side show the probability distribution for the variable at the end of the sampling, the charts on the right side show the mean of the variable during the trace. The left side charts are particularly useful because on top of visually depicting the most likely value for the variable in the model, they also show confidence in the value.

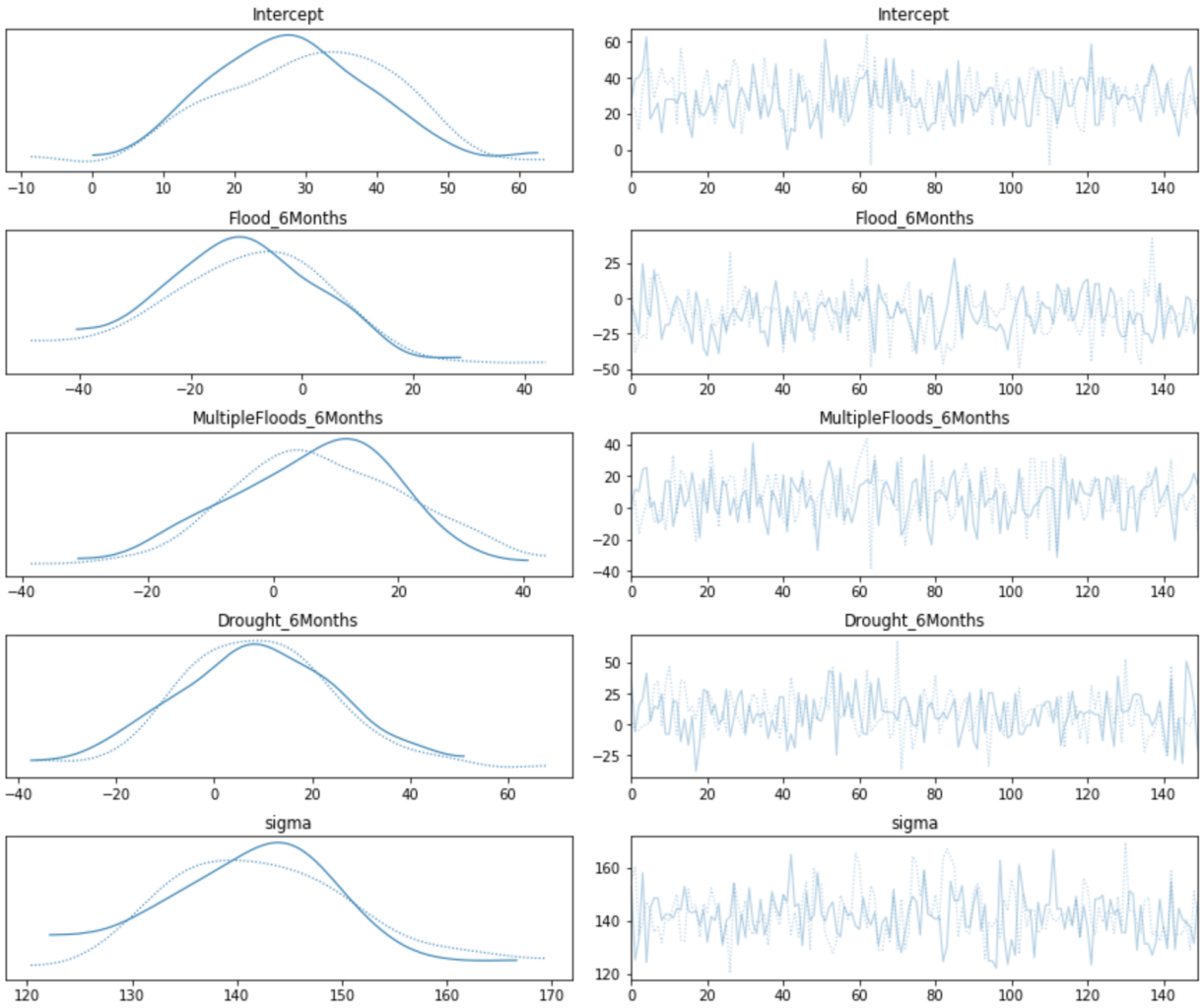
Data for the entire population:

```
Intercept 59.748923116347434
Flood_6Months -17.86241584451161
MultipleFloods_6Months 24.06036836740004
Drought_6Months 24.97573278496801
Flood_6Months Normalized -0.29895795460160207
MultipleFloods_6Months Normalized 0.4026912471802705
Drought_6Months Normalized 0.41801142986850914
```



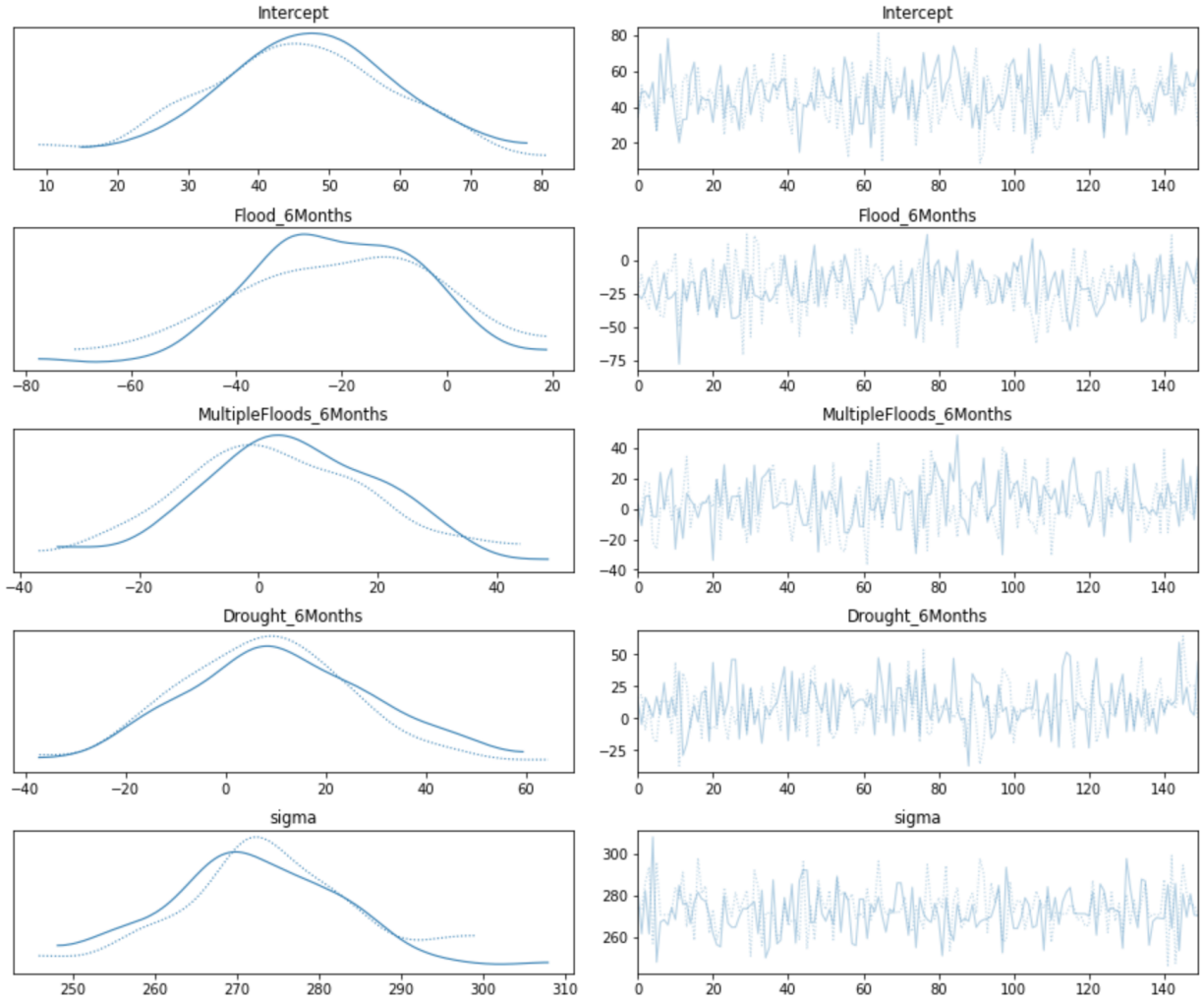
Data for Intolerant Species

```
Intercept 29.272990398236775
Flood_6Months -10.438662933060936
MultipleFloods_6Months 6.9009643218683925
Drought_6Months 8.860555904783906
Flood_6Months Normalized -0.3565970811676861
MultipleFloods_6Months Normalized 0.23574510933068404
Drought_6Months Normalized 0.30268707720812876
```



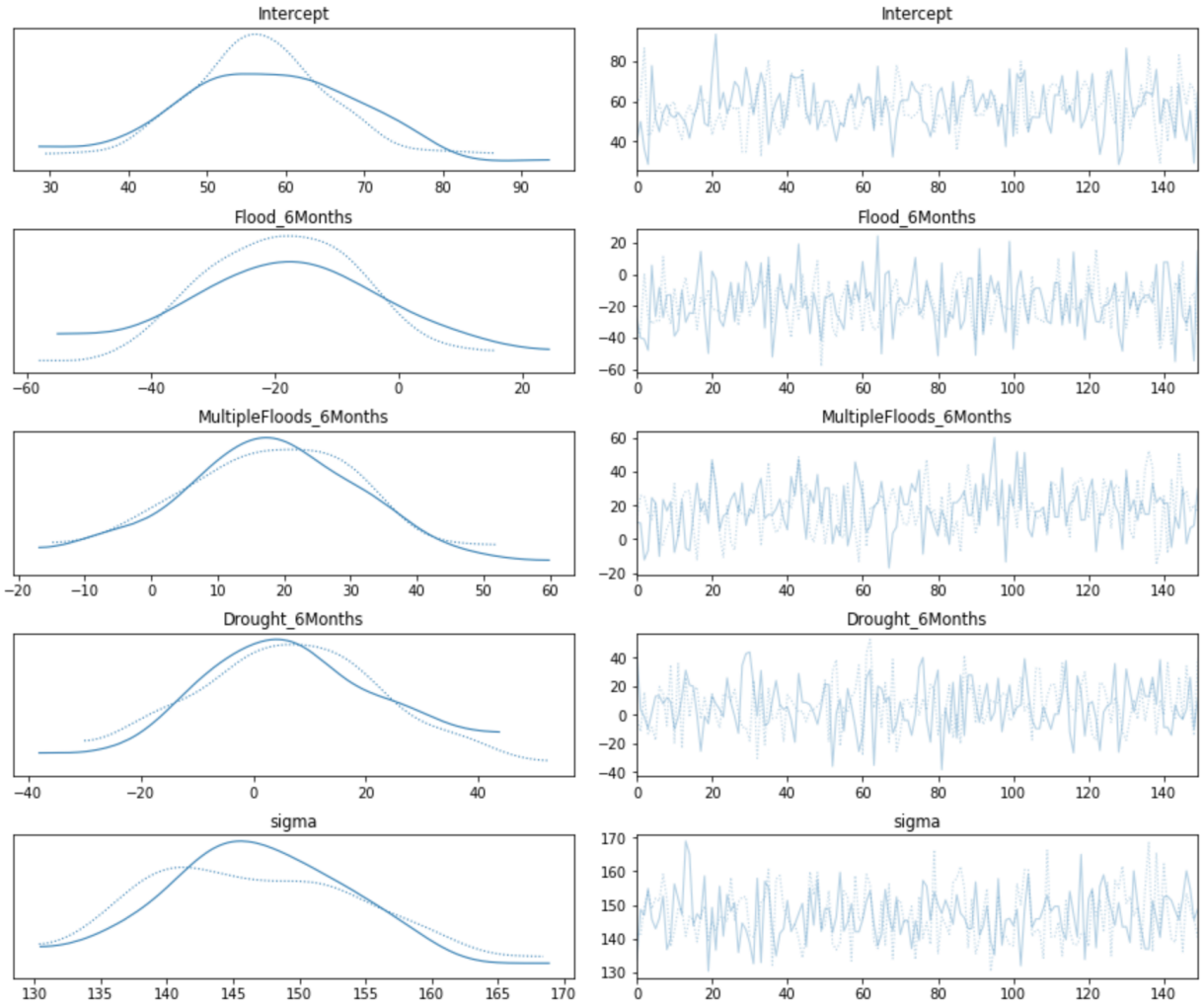
Data for Moderately Tolerant Species

```
Intercept 46.19413763532205
Flood_6Months -20.825941007892215
MultipleFloods_6Months 4.331057976519299
Drought_6Months 9.784821702141342
Flood_6Months Normalized -0.4508351508215578
MultipleFloods_6Months Normalized 0.09375774066204416
Drought_6Months Normalized 0.21181955553293932
```



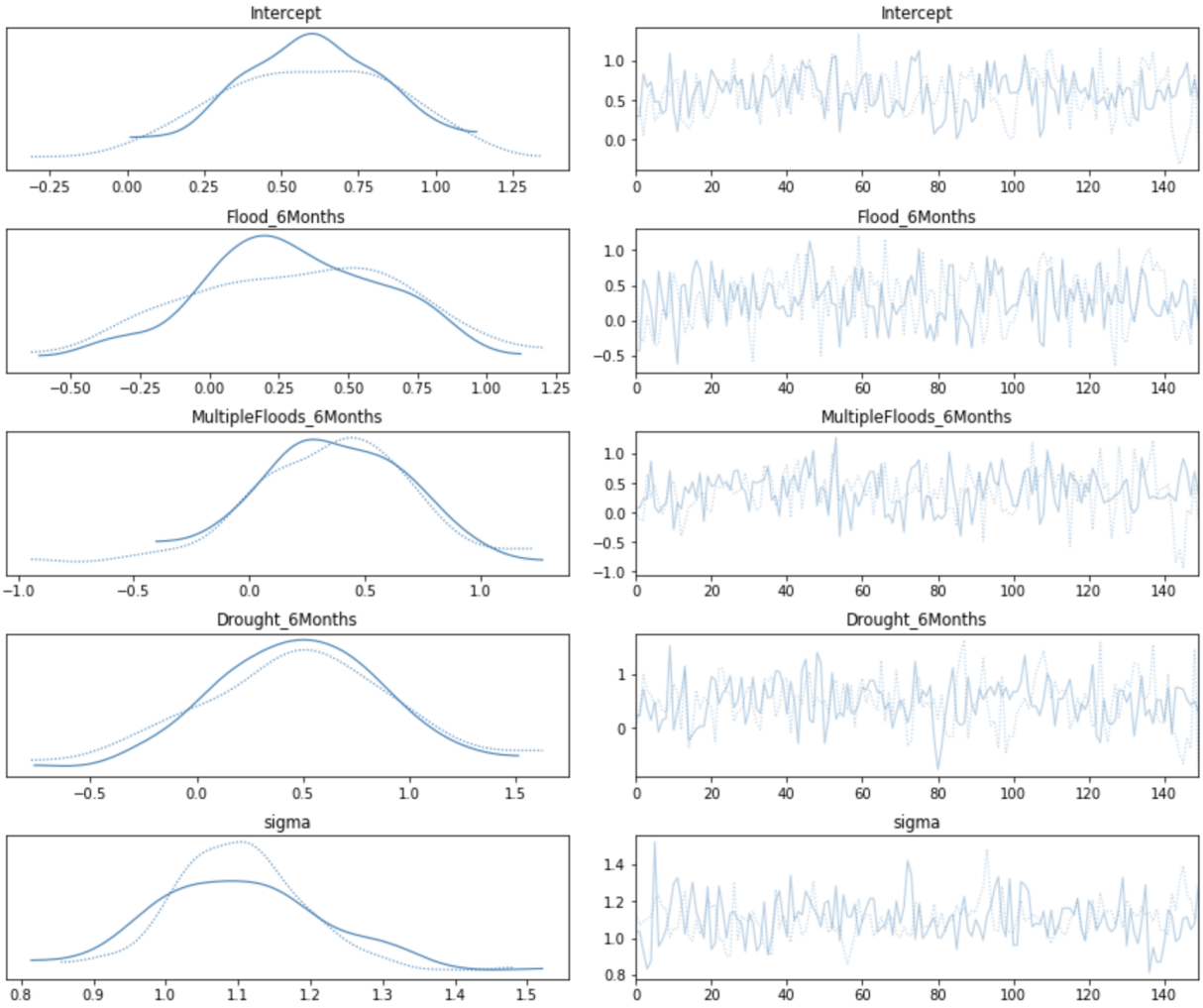
Data for Tolerant Species

```
Intercept 57.2367494043198
Flood_6Months -18.130999791576176
MultipleFloods_6Months 18.152133249336522
Drought_6Months 6.526180689205387
Flood_6Months Normalized -0.3167720036562346
MultipleFloods_6Months Normalized 0.31714123248177567
Drought_6Months Normalized 0.1140208128016585
```



Data for Unknown Tolerance Species

```
Intercept 0.5872680087324352
Flood_6Months 0.3030259390819592
MultipleFloods_6Months 0.3553500130880923
Drought_6Months 0.48044694404847416
Flood_6Months Normalized 0.5159925869893939
MultipleFloods_6Months Normalized 0.6050900233014277
Drought_6Months Normalized 0.8181050847388662
```



A2 Code

Code is included as a jupyter notebook. The sampler code seems to have an intermittent failure. I tried to figure out how to fix it, and read that changing the initialization of the sampler might fix it, but I wasn't able to get that to work.

My regression code is based on this tutorial: https://docs.pymc.io/notebooks/GLM-linear.html

The pymc3 sampler uses this algorithm: https://arxiv.org/abs/1111.4246