# Update 2024-07-10

## Likelihood

The number of cones $c_i$ produced by a stand measured on a given day $i$ is Poisson distributed:

$$P(c_i \mid \overline{c}) = \frac{\overline{c}^{c_i} e^{-\overline{c}}}{c_i!} \tag{1}$$

The number of number of cones $\overline{c}$ that we expect to see is given by the energy-conserving equation that we've discussed before,

$$\overline{c}_i = c_0 + \alpha \langle T \rangle_{i-l_0, w_0} + \beta \langle T \rangle_{i-l_1, w_1} - c_{i-l_2} \tag{2}$$

where e.g. $\langle T \rangle_{i-l_k, w_j}$ denotes the moving average of the temperature $T$ over a window of size $2w_j + 1$ days surrounding the day $i - l_k$. Here, $\alpha$ and $\beta$ are fit parameters which determine the relative importance of each year's sunlight contribution to the stand's energy reserves. $c_0$ is the initial energy reserves of the stand at the beginning of our observations.

The likelihood of observing the data $\{T_i, c_i\}$ from our dataset is just the product of the probabilities of each observation:

$$P(\{c_i, T_i\} \mid \overline{c}_i) = \prod_i \frac{\overline{c}_i^c e^{-\overline{c}_i}}{c!} \tag{3}$$

where $\overline{c}_i$ is the expected number of cones on day $i$, given by Equation 2. This is the **likelihood** distribution; it is the probability of observing our data given our model.

## Priors

I chose some prior probability distributions based on what I know about cone production. These characterize the epistemic uncertainty about our system:

| Parameter | Prior | Unit of measure | Comment |
|---|---|---|---|
| $c_0$ | $\mathrm{Uniform}(0, 1000)$ | # of cones | Initial energy reserves (number of cones) at start of dataset; can be between 0-1000 cones |
| $\alpha$ | $\mathrm{HalfNorm}(10)$ | cones/°F | Weakly informative choice of half-normal distribution, since this is probably a small number |
| $\beta$ | $\mathrm{HalfNorm}(10)$ | cones/°F | Weakly informative choice of half-normal distribution, since this is probably a small number |
| $w_0$ | $\mathrm{Uniform}(1, 100)$ | days | Window size used to calculate the average temperature in the first year. Probably in the range of 1-100 days long |
| $w_1$ | $\mathrm{Uniform}(1, 100)$ | days | Window size used to calculate the average temperature in the second year. Probably in the range of 1-100 days long |
| $l_0$ | $\mathrm{Uniform}(180, 545)$ | days | Lag time of the moving average of the temperature in the first year; constrained to be 0.5 to 1.5 years before the measured crop |
| $l_1$ | $\mathrm{Uniform}(550, 910)$ | days | Lag time of the moving average of the temperature in the second year; constrained to be 1.5 to 2.5 years before the measured crop |
| $l_2$ | $\mathrm{Uniform}(915, 1275)$ | days | Lag time used to get the last cone crop, constrained to be 2.5 to 3.5 years before the measured crop |

## Posterior

Using the likelihood (Equation 3) and the priors (Table 1), we can construct the **posterior** distribution using Bayes' theorem:

$$P(\bar{c}_i \mid \{c_i, T_i\}) \propto P(\bar{c}_i) P(\{c_i, T_i\} \mid \bar{c}_i) \tag{4}$$

Using MCMC, we can sample from this distribution to get an idea of what it looks like.

## MCMC

## Next Steps

After some debugging it looks like the sampler is working reasonably well, but it clearly hasn't converged. The Markov chains for the lag and window size in the first year vary wildly, but we have to pay attention to the fact that the coefficient of the first year moving average term *did* converge to zero, which is why the lag and window size were able to vary so erratically - no matter their values, they had no impact on the cone count. In any case, we probably need to reparameterize in order for the model to converge.

If we can get a converged model post-reparameterization, the next thing to do will be to carry out some posterior predictive checks, i.e. generate fake data using these probability distributions to see if it looks like the data we measured. If they look similar, we'll know we've captured the important parts of the generating process that led to these datasets, and we'll actually be able to start connecting these parameter values with what we know about reproductive processes.

## Modeling

Assume

$$c_{\text{obs}} \sim P(c_\mu) \tag{5}$$

Consider various models for $c_\mu$:

### $n$-Years Preceeding Model

$$c_{\mu,i} = c_0 + \sum_j \alpha_j \langle T \rangle_{\gamma,i-j} - \beta c_{i-k} \tag{6}$$

Where $j$ runs over a few years preceeding the cone crop in year $i$. Here, $\langle T \rangle_{\gamma,i-j}$ means an average of the temperature for $\gamma$ days starting on day $i-j$, and $c_0$, $\{\alpha_j\}$, $\beta$, and $\gamma$ are fit parameters.

Generally these models are sort of unmotivated in the sense that the number of years included in the sum is arbitrarily chosen, although they are motivated by literature suggesting that the important reproductive processes leading up to cone production occur in either two or three years preceeding the cone crop - some species have a year of reproductive “dormancy”, where immature cones remain on the tree for a period of time.

### Resource-Accumulation Model (RAM)

$$c_{\mu,i} = c_0 + \underbrace{\alpha \int_0^{t_i} T(t)\,\mathrm{d}t}_{\substack{\text{Photosynthetically} \\ \text{Active Radiation}}} - \int_0^{t_i} c(t)\,\mathrm{d}t \tag{7}$$

Here the resources accumulated by the tree over time are considered: the Photosynthetically Active Radiation (PAR) received each day is approximated as being proportional to the temperature that day; a potentially dubious approximation. The available resources of the stand include all the PAR absorbed since the beginning of the dataset less any spent on cone production.

#### Other resource expenditure

Leaves, wood, and roots cost a lot of energy. One important nuisance parameter is the energy expenditure on wood/leaf/root growth. We can modify the RAM to include seasonal changes in non-cone resource expenditure:

$$c_{\mu,i} = c_0 + \underbrace{\alpha \int_0^{t_i} T(t)\,\mathrm{d}t}_{\substack{\text{Photosynthetically} \\ \text{Active Radiation}}} - \int_0^{t_i} c(t)\,\mathrm{d}t - \int_0^{t_i} R(t)\,\mathrm{d}t \tag{8}$$

The instantaneous resources available are thus

$$\alpha T(t) - c(t) - R(t) \tag{9}$$

and the change in expected cone crop from year i to year j is

$$\Delta c_{\mu,i \to i+1} = \alpha \int_{t_i}^{t_{i+1}} T(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} c(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} R(t)\,\mathrm{d}t \tag{10}$$

## Transformations

Monte Carlo samplers are sensitive the data fed into them; generally they sample efficiently when data is distributed $\sim N(0,1)$.