About this file: It's Lizzie's attempt to keep notes (from emails etc.) and thoughts organized.

Why use survival models for plant phenology? This comes from Mike Betancourt's email on 9 June 2020:

The Wang-Engel model (and its ilk) is a threshold model — it presumes that phenological transitions are deterministic. Once the latent forcing

$$f(t) = \int_0^t C(T(t))dt$$

where C is the Wang-Engels function, T is temperature, and t is time, passes a threshold the transition always happens immediately. The statistical model assumes that variation comes from the unmodeled heterogeneity in the threshold for each plot/vine/tree/unit/etc.

The deterministic nature of the threshold is what makes the model so tricky to implement generatively. We need to compare the observed transition time to the expected transition time; the expected transition time is deterministic conditioned on the parameters and the temperature time series, but requires the integration (or summation) and inversion.

Another approach is to consider the phenological transition as not deterministic but rather *stochastic*. In other words as the latent forcing increases the *probability* of transitioning increases but there's no finite forcing for which the transition is guaranteed to always happen. In practice, however, a probability sufficiently close to one will be a practical guarantee for all intents and purposes.

In this case the probability of not transitioning by time t is given by the survival function

$$S(t) = exp(-\int_0^t C(T(t))dt)$$

where C could be the same Wang-Engels function only without a normalization restricting it to be less than or equal to one. At t = 0 the survival function is 1 and as time, and the latent forcing, increases the survival function decreases monotonically.
The likelihood is then just

$$\pi(y|\theta) = P[\text{not transitioning until time}t] \times \pi(\text{transitioning at exactly time}t) = S(t) \times C(T(t))$$

In other words because we only have to compute a latent probability and not a deterministic threshold we don't need the expensive inversion step needed in the previous model.

This is otherwise known as a survival model—a pretty well-studied approach in epidemiology and related fields—and can also be interpreted as the outcome of a continuous time hidden Markov model, connecting it to the discrete time hidden Markov models we were discussing.

The survival model approach still requires interpolating temperatures to compute the integral needed for the survival function but because it doesn't require an inversion it's straightforward to implement in Stan. In terms of predictions there shouldn't be much difference between the two, modulo issues of unmodeled censoring in the observed data that would have different consequences for the two approaches. The biggest question for considering the survival model approach is whether or not the phenology field would be okay with the change from a deterministic threshold transition to a probabilistic one.

Below from 15 July 2020 email:

In the original Wang and Engel model a phenology transition is a deterministic event that occurs when the integrated Wang an Engels units pass some threshold,

$$\int WE(T(t))dt = \text{threshold}$$

where t is time and T is temperature. This can be also be interpreted as

$$\frac{\int WE(T(t))dt}{\text{threshold}} = 1.$$

In a survival model a phenology transition is not deterministic. Instead as the integrated Wang and Engel units increase the probability of not transitioning decreases towards zero. The probability of not transitioning up to some time is given by the survival function

$$exp(-\gamma \int WE(T(t)dt))$$

gamma plays a similar role as the threshold in the deterministic model, scaling the integrated Wang and Engels units to the right magnitude.

In fact if one approximates the stochastic survival process as being deterministic after the survival probability decreases to some p_{thresh} then the transition occurs when

$$exp(-\gamma \int WE(T(t))dt) = p_{thresh}\gamma \int WE(T(t))dt = -log(p_{thresh})$$

or

$$(\frac{\gamma}{-log(p_{thresh})}) \int WE(T(t))dt = 1,$$

from which we can identify

$$\text{threshold} = \frac{-log(p_{thresh})}{\gamma}.$$

Incidentally this is how I converted the prior on threshold to a prior on gamma.

Mike sent a first version of this model in mid June. But the numerical integrator was slow, "the numerical integrator seems to be quite slow which limits the scaling of the model. I think that to speed things up we'll have to play with the numerical integrator to optimize performance. Another possibly would be to model the aggregated energy directly with something like an I-spline." Mike fixed this issue by early July though:

The reason why the numerical integrator was being so problematic was that the linear interpolation let to cusps in the integrand that messed things up.

As a proof of principle I assumed that the temperatures followed a sinusoidal behavior in which case things behave much better. With Lizzie's priors everything behaves reasonably well, although the numerical integration is still pretty expensive—it takes about an hour to run 10 observations. Possible speeds ups would come from computing the integrations in parallel, not integrating from day 1, and reusing the same intergral (i.e. integrating from day 1 to first observed transition then integrating from that transition to the next one, etc, and adding things up).

The biggest limitation right now is the temperature model—I'm guessing that the sinusoid is drastically oversmoothing. Based on your all domain expertise we could build up a better model with additional Fourier components, splines, or the like.

This means we needed to think on the temperature model! You can check out tempmodels.pdf (in temperature folder) for more on that.

Issues with the Wang & Engel: Mike says, "One thing that I've learned so far is that the WE model is pretty non identified. By changing the WE threshold (which sort of becomes gamma in my model) and T_{min}, T_{opt}, and T_{max} in the right way you can explain the same transition time multiple ways. The question is how much we can actually constrain those parameters scientifically to well-pose the model!"

Right now (September 2020), we're surviving on priors, Mike wrote, it fits "largely because of the relatively strong prior densities. If those were loosened up at all then the model would fit very differently."

We should think hard if we can get at these values from other angles (again, Mike on 12 July 2020), "The other thing to consider is how T_{min}, T_{opt}, and T_{max} and gamma might manifest in other observable phenomena that would allow us to pin them down without just resorting to prior information that may be limited due to long running assumptions in the literature. Even it's just something like T_{min}, T_{opt}, and T_{max} t vary from plant to plant but gamma doesn't, in which case the heterogeneity might be enough to separate out the various effects."