Modeling masting

From individual tree behavior to population-level synchrony

This document focuses on a project Lizzie and I have been working on as a way to model masting for individual trees for one species across sites and we think could be extended with Eléonore's help to work well for stand-level seed trap data across species and sites. I start by outlining the background and approach for this particular model, then get into the math! (The math may look complicated if it's been a few years since your last math class, but it will definitely come back with more immersion in it in January)

1 The motivations behind the project

The ecological motivation

Population level patterns such as masting emerge from individual trees' responses. Seed production results from each individual tree's reproductive biology (i.e. biological **constraints**) and response to the environment (**cues** and 'vetos'). Working at the individual-level may allow us to better understand the drivers of the variability and synchronicity of reproduction at the *population* scale. Our final goal is to understand to what extent climate change could really disrupt forest regeneration—and provide better forecasts!

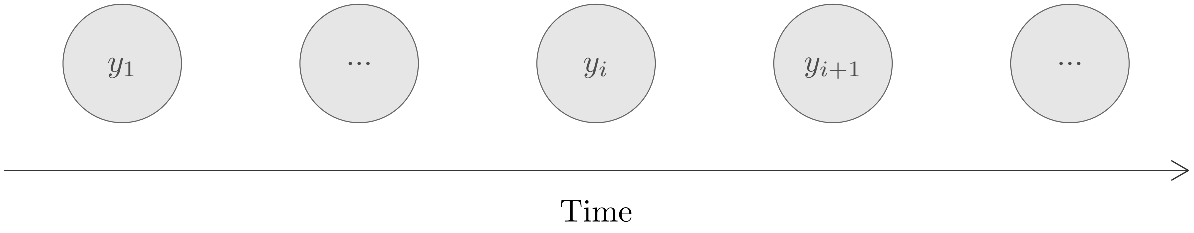
The technical motivation

Bayesian magic! Our approach models the (latent) reproductive states of each *individual* tree, and the consequent amount of seed production. The states explicitly encode the **constraints** that shape tree reproduction – in particular, the fact that most trees need at least two years between flower bud differentiation and fruit maturation. We do not standardize away the complexities of individual-level reproduction (e.g. with normalized stand-level indices between 0 and 1...). On the contrary, modeling tree-level reproduction allows us to obtain *population* estimates — and thus direct inferences on the population-level variability and synchronicity! And the cherry on top: we incorporate climatic **cues** at different key moments of the reproductive cycle.

2 A generative model for masting

2.1 Previous approaches and limitations

The observations we have—whether for an individual tree or a seed trap—are a time series of yearly seed counts y_i.



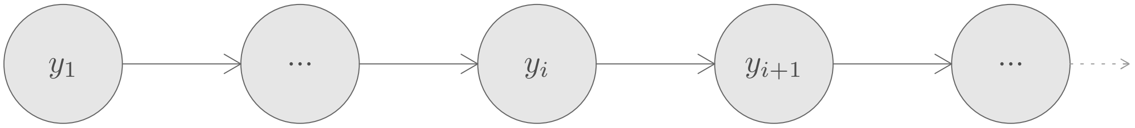
The following quote illustrates one approach of statistical model of masting:

> “We fitted a zero-inflated, negative binomial mixed model to the annual number of initiated seeds in each tree, with fixed factors that included summer temperatures in 1 and 2 years before seedfall, [. . .] and seed production in the previous year to account for possible resource depletion. [. . .] We included [. . .] a first-order temporal autocorrelation structure.”

If we ignore the zero-inflated part, we could write this model as something like:

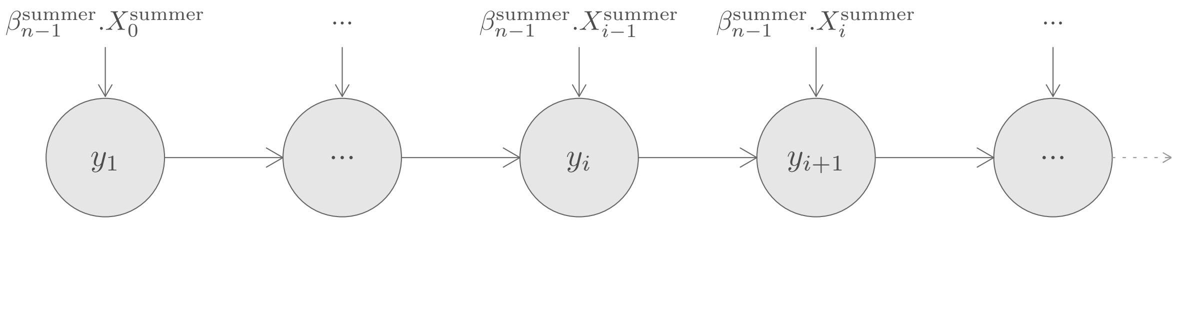
$$y_i \sim \text{NegBin}(\mu_i, \phi)$$
$$\mu_i = \alpha + \beta_{n-1}^{\text{summer}}.X_{i-1}^{\text{summer}} + \beta_{n-2}^{\text{summer}}.X_{i-2}^{\text{summer}} + \beta_{n-1}^{\text{seed}}.y_{i-1} + \rho.\epsilon_{i-1} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, 1)$$

The idea is to introduce some dependency between the seed counts:



The inclusion of both a lagged response of $\beta_{n-1}^{\text{seed}}.y_{i-1}$ and a first-order autocorrelation structure on the residuals $\rho.\epsilon_{i-1}$ seems. . . quite tricky.

But more importantly, the effect of summer temperature does not directly depends on the previous state. In other words, whatever the value of $\beta_{n-1}^{\text{seed}}.y_{i-1}$, a tree that experiences a warm summers would be predicted to have an increasing seed production—regardless of the previous reproductive state.

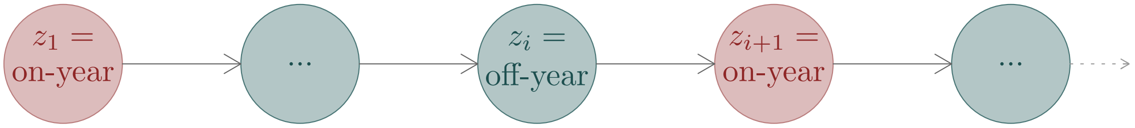

$\beta_{n-1}^{\text{summer}}.X_0^{\text{summer}}$
$\cdots$
$\beta_{n-1}^{\text{summer}}.X_{i-1}^{\text{summer}}$
$\beta_{n-1}^{\text{summer}}.X_i^{\text{summer}}$
$\cdots$
y_1
$\cdots$
y_i
y_{i+1}
$\cdots$

It's hard to include the effect of the previous summer temperature depending on a tree's previous state because each individual tree could be in a different reproductive state, but the model treats all years for all trees as the same. The model does not allow to differentiate effects of climate during a non-masting year or a masting year. We may be able to address this, however, if we start with a model at the individual tree level.
From there, we can infer reproductive behavior at the population scale—including the degree of potential synchrony.

2.2 A new approach grounded in biology

Researchers have worked on the reproductive biology of trees for decades—and in particular fruit trees. In parallel, people working on masting have defined characteristics of masting and developed clear hypotheses about the evolutionary mechanisms that could cause masting. There is thus a great opportunity to develop more generative models, that explicitly incorporate biological knowledge of flower and fruit development in order to better understand how masting works.

To understand the reproductive behavior that arises at the population level, we need to model individual trees' reproduction. At the tree level, floral buds are initiated the year before flowering—in the same time as fruits of the current year start developing. During a large crop year (an "on"-year), the presence of many fruits depress flower initiation because of hormonal inhibition and resource trade-off. Thus, the next year will likely be an "off"-year. This behavior is called **alternate bearing**. It is not always a 2-year cycle, as an "on"-year can be followed by several "off"-years. This clearly defines two states for a tree.



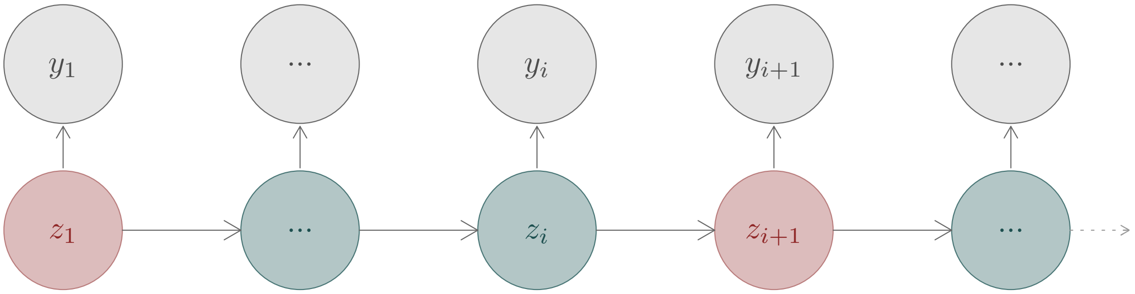
Each state z_{i+1} is dependent **only on the preceding state** z_i (which is called the *Markovian* assumption). These states are latent, because we observe only the seed counts. The model I will describe briefly in the next section is thus called a Hidden Markov Model (HMM): *hidden* because the states are latent, and *Markov* because of the Markovian assumption.
With some parameters θ (parameters are things like the slope and intercept in your model, here θ stands in for one or more of them and effectively represents—with just one Greek letter—a lot of the rest of your model), the probability of observing some sequences of the two first states is:

$$p(z_1, z_2 \mid \theta) = p(z_1) \times p(z_2 \mid z_1, \theta)$$

And then, adding the next state:

$$p(z_1, z_2, z_3 \mid \theta) = p(z_1) \times p(z_2 \mid z_1, \theta) \times p(z_3 \mid z_2, \theta)$$

And so on ...
Each observed seed count at each step arises from the state of the tree.



This gives us this likelihood for the two first seed counts:

$$p(y_1, y_2, z_1, z_2 \mid \theta) = p(z_1)\, p(y_1 \mid z_1, \theta)\, p(z_2 \mid z_1, \theta)\, p(y_2 \mid z_2, \theta)$$

And for all observations until $i+1$:

$$p(y_1, \dots, y_{i+1}, z_1, \dots, z_{i+1} \mid \theta) = p(z_1)\, p(y_1 \mid z_1, \theta) \prod_{k=2}^{i+1} p(z_k \mid z_{k-1}, \theta)\, p(y_k \mid z_k, \theta).$$

So, what parameters do wo have to estimate?

2.2.1 Transition between states

Let's look at the part which concerns the latent states.
First, we need to infer the initial state probability $p(z_1)$, i.e. the probability that the first year is an on-year or an off-year. Because we have only two states, $p(z_1 = \text{on}) = 1 - p(z_1 = \text{off})$—i.e. we have only one parameter to estimate.
We also have to infer the probabilities of transition $p(z_k \mid z_{k-1}, \theta)$. With two states, this corresponds to this transition matrix:

$$M = \begin{bmatrix} p(z_k = \text{on} \mid z_{k-1} = \text{on}, \theta) & p(z_k = \text{off} \mid z_{k-1} = \text{on}, \theta) \\ p(z_k = \text{on} \mid z_{k-1} = \text{off}, \theta) & p(z_k = \text{off} \mid z_{k-1} = \text{off}, \theta) \end{bmatrix}$$

That we could write as:

$$M = \begin{bmatrix} \tau_{\text{on}\to\text{on}} & \tau_{\text{on}\to\text{off}} \\ \tau_{\text{off}\to\text{on}} & \tau_{\text{off}\to\text{off}} \end{bmatrix}$$

The probability of assignment of z_2 to one of the two states is then obtain with the product:

$$\begin{bmatrix} p(z_2 = \text{on}) & p(z_2 = \text{off}) \end{bmatrix} = \begin{bmatrix} p(z_1 = \text{on}) & p(z_1 = \text{off}) \end{bmatrix} \cdot M$$

Again, because we have only two states, the transition matrix corresponds to only two parameters, for example:

$$M = \begin{bmatrix} \tau_{\text{on}\to\text{on}} & 1 - \tau_{\text{on}\to\text{on}} \\ \tau_{\text{off}\to\text{on}} & 1 - \tau_{\text{off}\to\text{on}} \end{bmatrix}$$

In summary, we have 3 parameters to infer: the initial probability of masting $\tau_0 = p(z_1 = \text{on})$, the probability of transitioning from an off-year to an on-year $\tau_{\text{off}\to\text{on}}$, and the probability of staying into an on-year state $\tau_{\text{on}\to\text{on}}$.

2.2.2 Seed production parameters

Now, what about the parameters of the observational model that links the latent state to the observed seed count? The seed production depends on the state of the tree. Basically, it's a mixture model between two probability distributions.

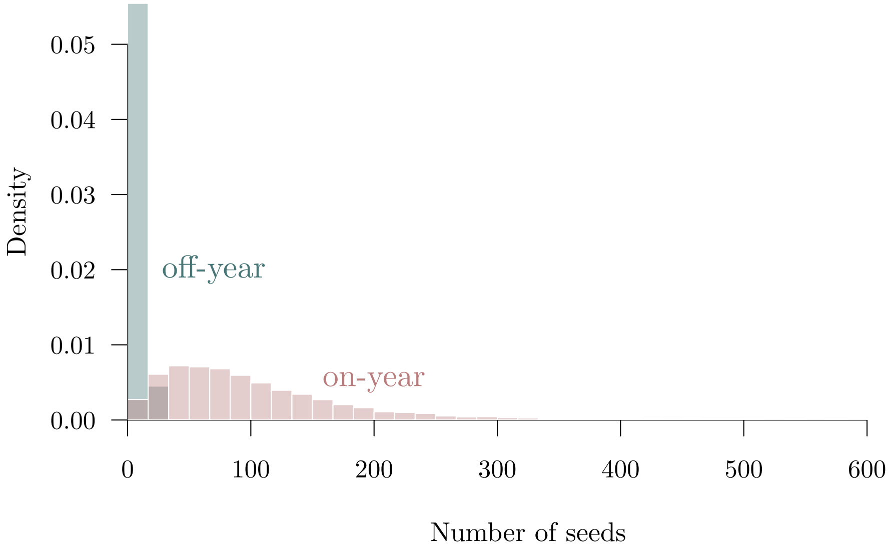
If the tree is in an off-year, then we model the (likely low) seed count with a zero-inflated negative binomial distribution:

$$y_i \begin{cases} = 0, & \text{with probability } \theta \\ \sim \text{NegBin}(\mu_{\text{off}}, \phi_{\text{off}}), & \text{with probability } 1-\theta \end{cases} .$$

If the tree is in an on-year, then we model the seed count with a negative binomial distribution:

$$y_i \sim \text{NegBin}(\mu_{\text{on}}, \phi_{\text{on}})$$

This mixture of two probability distributions would look like this:



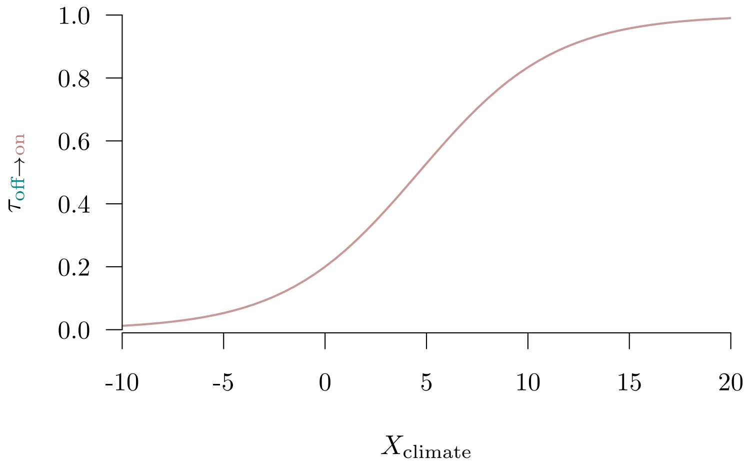
2.2.3 Adding climate!

Explicitly modeling reproductive states and the subsequent seed production allows to think more deeply where the different climatic cues should impact the individual reproductive cycles and the population-level masting behavior.

The states explicitly encode the developmental constraints, i.e. the temporal overlap between floral bud initiation and fruit development, that explain the alternate bearing at the individual level. The transition between the states are controlled by the matrix transition. If the transition matrix stays constant across time, we would have an **homogeneous** HMM. But what if climate controls the transition between the states?
It's quite easy to include some variations in the transition matrix M, to have an **heterogeneous** HMM. For example, rather than having a constant probability $\tau_{\text{off}\to\text{on}}$, this probability could vary with the local climate:

$$\text{logit}(\ \tau_{\text{off}\to\text{on}}\{t\}\ ) = \text{logit}(\tau_0) + \beta \ \cdot \ X_{\text{climate}}\{t\}$$

This would give us this kind of functional form, here with $\beta > 0$:

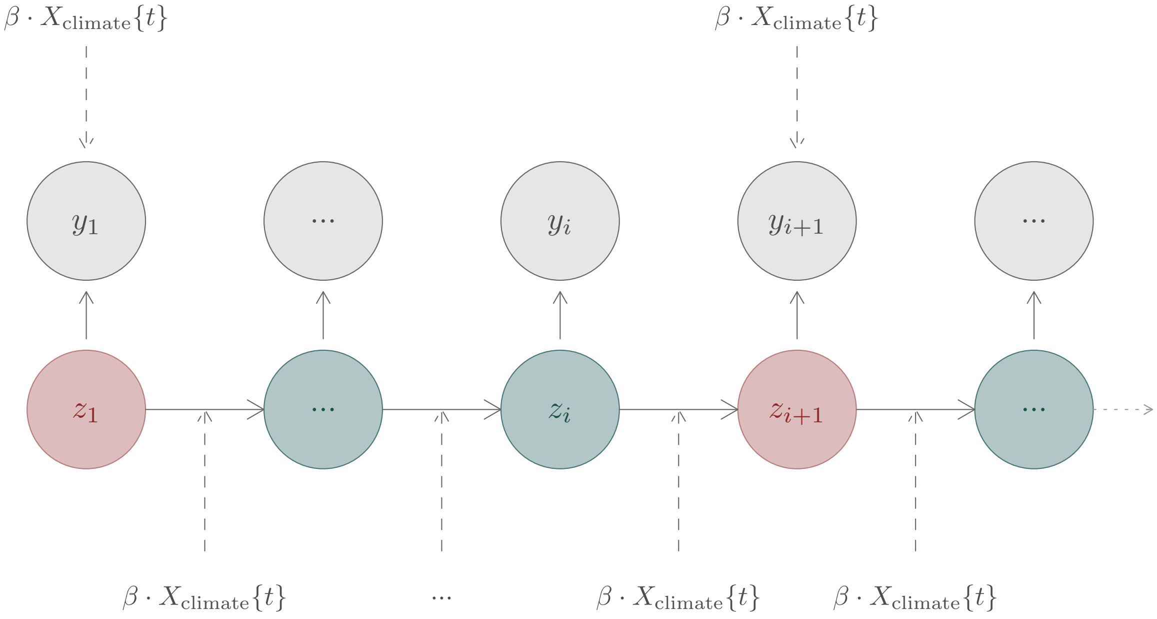


Once the tree is assigned to a state, climate may also impact seed production. The number of seeds in a off-year could be seen as a "background noise", and we would thus prioritize adding some climatic predictors on the seed production of an on-year. A simple approach would be to modify the location parameter μ_{on} of the negative binomial like:

$$\mu_{\text{on}}\{t\} = \mu_0 + \beta \ \cdot \ X_{\text{climate}}\{t\}$$

Obviously, the covariates X_{climate} could be different for the transition probabilities and the seed production.

In summary, we could schematize the model as follow:

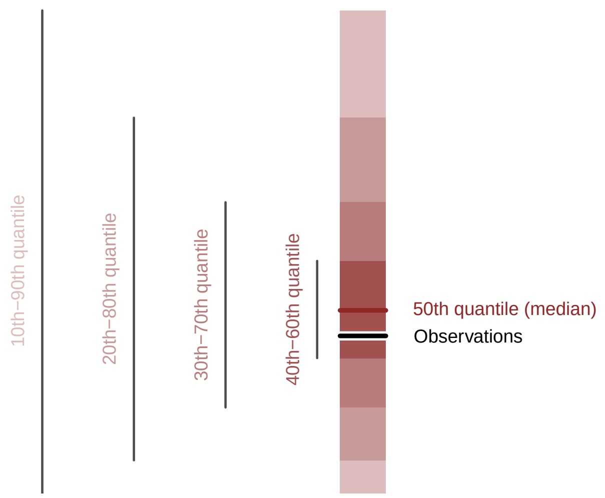


$\beta \cdot X_{\text{climate}}\{t\}$
$\beta \cdot X_{\text{climate}}\{t\}$
y_1
...
y_i
y_{i+1}
...
z_1
...
z_i
z_{i+1}
...
$\beta \cdot X_{\text{climate}}\{t\}$
...
$\beta \cdot X_{\text{climate}}\{t\}$
$\beta \cdot X_{\text{climate}}\{t\}$

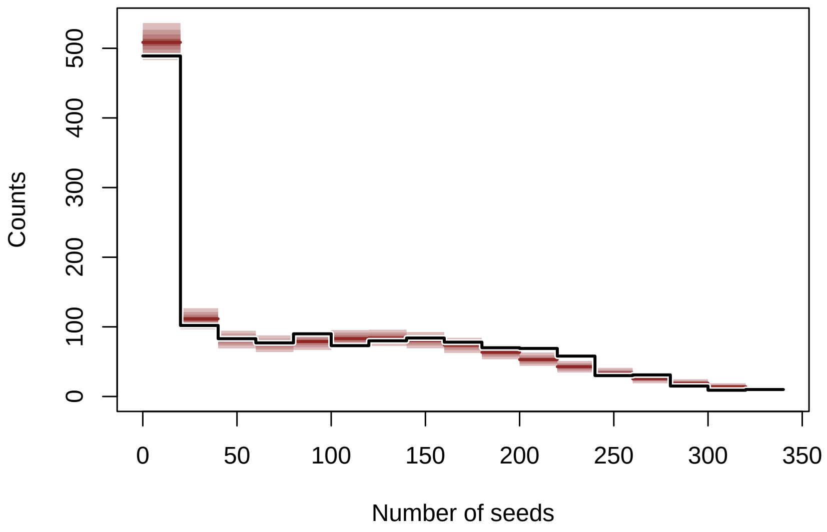
3 The case of beech masting dynamics

We applied our model to data collected in England. We have annual seed production time series for 57 trees across 7 sites. The maximum duration is 43 years, but most series are shorter (and contained some missing values). For each observed tree and each year, the model estimates the reproductive state—given the previous state—and the subsequent seed production. Climate is included as an explicit driver of both state transitions (probability matrix), and seed production (number of seeds). From these individual reproductive dynamics, the model allows to scale up to population-level behavior and investigate how climate interacts with biological constraints to impact masting.

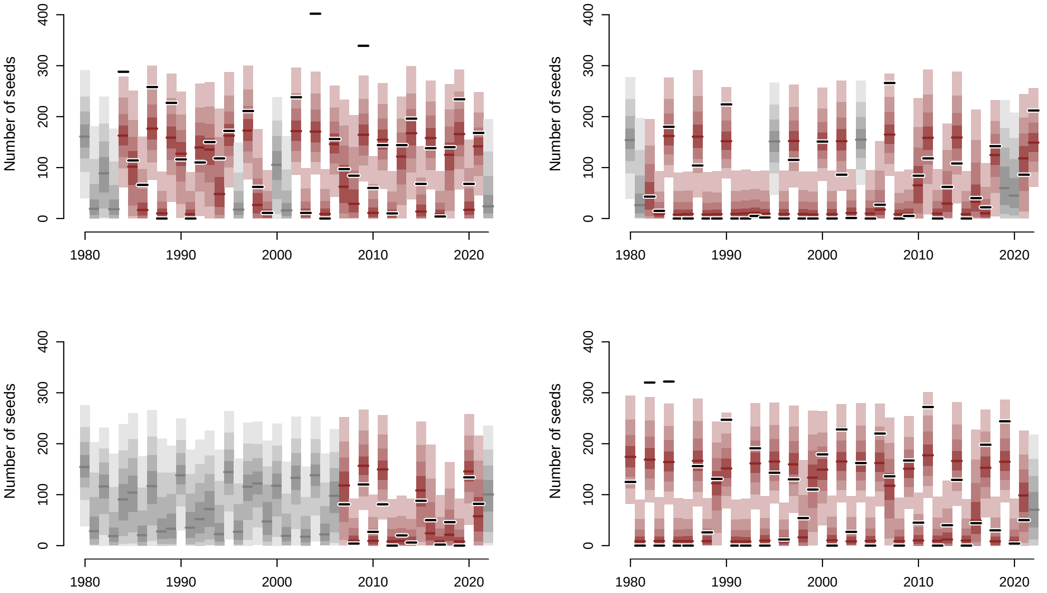
The diagram below illustrates how to interpret the following plots. The red bars correspond to the model predictions.



First, we can check for any inconsistencies between the observed data and the behavior of the posterior predictive distribution (what we call “retrodictive checks”). The histogram below compares the distribution of all observations (black line) to the posterior predictive distribution (the predictions by the model, in red).

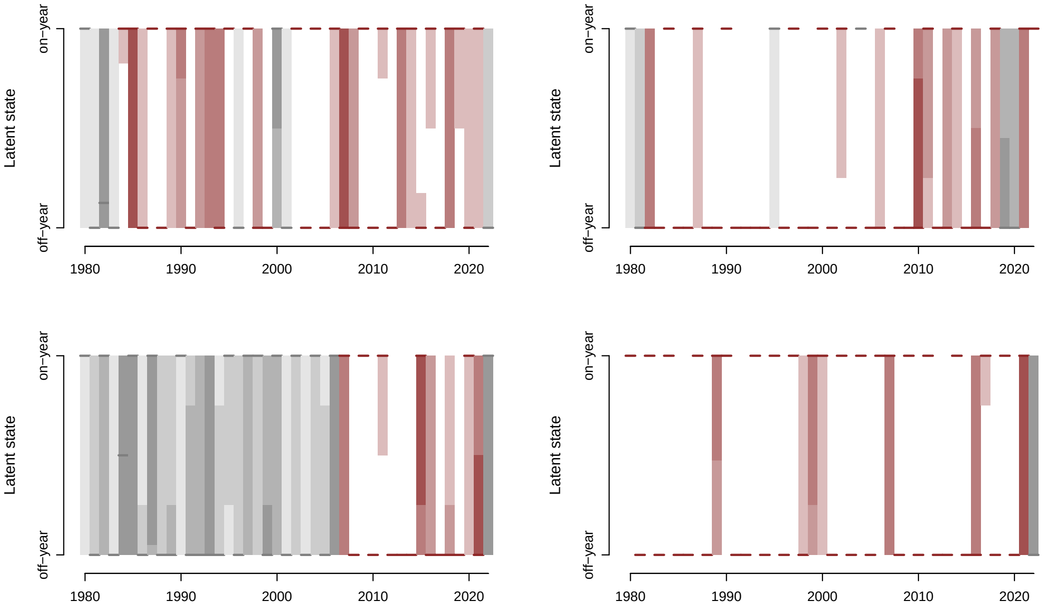


Everything looks consistent. Next we can explore how consistent the observations and the posterior predictions are for some random trees.



The grey bars indicate years with missing observations. A strength of the model is that it still allows to estimate the reproductive state and seed production during those years with missing data.

We can also look at the latent states that are estimated by the model. Remember: these are *latent*, so we don't have observations (no black lines).


Latent state
on−year
off−year