The problem and promise of modeling plant leafout under warmer winters

E.M. Wolkovich[1], Justin Ngo[1], Victor van der Meersch[1], Jonathan Auerbach[2]

October 10, 2025

[1] Forest & Conservation Sciences, Faculty of Forestry, University of British Columbia, 2424 Main Mall, Vancouver, BC V6T 1Z4, Canada
[2] Department of Statistics, George Mason University, 4511 Patriot Circle, Fairfax, VA 22030, United States

Abstract

How warmer winters affect plant leafout has major implications for carbon storage and ecosystem change, but is poorly understood. Recent evidence of a slow-down in the advance of spring leafout with increasing anthropogenic climate change has reinvigorated a decades-old debate over winter 'chillling'—the concept that woody temperate plants need sufficient winter cool temperatures to budburst each spring. Yet current models of chilling make diverse predictions (JA: diverse predictions or contradictory predictions/lack of scientific consensus?), with different models given the same data often predicting everything from advanced to slowed leafout. Such variability highlights how little we understand the critical process of 'chilling,' which is never observed but often assumed. (JA: I think this is great so far. The problem from my perspective is that if this is where the field was before our work, the following sentences do not sound like much of a contribution, unless this is a review paper. Three options: 1. Change the next sentence "We show" to "We review". 2. We could replace the previous two sentences "Yet current models ... but often assumed" with "But what if anything does a chilling mechanism reveal about observable plant behavior/phenology?" or 3. We concede that the field has long recognized the non-identifiability problem, but we argue that the failure to make it precise (as we do here) has caused researcher after researcher to reinvoke chilling, which like other theoretical constructs such as the "ether", has become a placeholder that does not lead to unique falsifiable predictions and thus stalled scientific progress.) We show how current work, which routinely models 'chilling' with highly limited data has likely led to the proliferation of models—and thus forecasts—that yield little additional insight about observable plant behavior. We conclude that integrating new insights and approaches from molecular biology could better define the mechanism of chilling leading to models that make falsifiable predictions. Exposing the gaps in current mechanistic and statistical models at the same time new experiments yield richer, more informative data could greatly reduce the number of competing models of chilling that we have today—advancing our fundamental understanding of plant dormancy—and lead to much improved forecasts from crops to forest trees. (JA: I don't think the last sentence adds much. I think the point to make is that by understanding the limits of chilling as a theory, we can both 1. build better models by looking for the necessary insight from molecular biology to

fill in the details, and 2. design better experiments to test those theories.)

Plant leafout in the spring has shifted weeks earlier in many regions due to warming from anthropogenic climate change with consequences for a suite of ecosystem services, including carbon storage (**??**). What underlying processes drive this trend, however, has come under increasing debate, as recent research suggests winter warming may slow or stall this advance (**??**). Such reports focus on a two-step model of leafout where plants first require cool winter temperatures, often called 'chilling,' before they can accumulate enough warm temperatures—'forcing'—to leafout each year. But this model of leafout—especially its chilling component—is only one of many models proposed since the concept was first introduced (**??**). (JA: Is the point that the two stage chilling model is just one of many chilling models? Or are we restricting our attention to chilling models, which constitute a subset of phenology models? Either way I think the "But" is better left to the next sentence. "But while forcing ends in an observable event such as leafout, the day plants achieve their chilling requirements coincides with no known measurable event. Thus, current models of chilling are consistent with ...)

Current models of chilling predict a diverse suite of possible future leafout. Chilling in the same location under different models can easily forecast significantly increased or greatly reduced chilling and thus greatly advanced or slowed leafout (**??**). Such extreme variability of spring phenology predicted from current models suggests fundamental gaps in our understanding. While experimental evidence demonstrates winter temperatures do impact leafout (Charier et al, etc.), new research suggests major flaws in these models as currently applied, including multiple papers now suggesting estimates of 'chilling' could easily be artifacts of poor models or correlated observational climate data (**??**).

Now appears an opportune time to address the problems in models of leafout. Because accurate forecasts of spring phenology are critical for carbon storage, many crops and a number of other services important to humans, there is widespread interest in improving chilling models (**??**). At the same time, new results from molecular and cellular studies of dormancy are providing new insights into when and how chilling works (**??**), which could rule in—or out—some of the myriad current models. Here, we review the concept of chilling, its origins and potential problems, as well as new opportunities for major advances and how shifting current practices could accelerate progress. (JA: As mentioned in the abstract, I think revisiting the chilling mechanism has two benefits: 1. creating better models, and 2. designing better experiments.)

What is chilling? (JA: Great section, would not cut)

How plants in temperature-limited systems avoid leafout during warm spells in the winter has long been debated by plant biologists (e.g., **??**). Most work to date has focused on the idea that plants enter some form of dormancy in the fall, which is then released before warm temperatures in the spring begin. This idea hypothesizes that the slow accumulation of cool temperatures—or chilling—over the winter extends through periods of short warm spells in the winter and thus prevents leafout before spring.

Much of our fundamental understanding of chilling comes from studies on temperate woody fruit

crops where chilling can be critical to yield. Peach trees planted into warmer climates well outside their range often have extremely low fruitset because most flower buds do not burst (**???**). Initial studies of this phenomenon with related experiments—where cut ends of dormant branches (cuttings) exposed to cooler temperatures in chambers burst more fully and more quickly—underlies most of the models of chilling used today for crops and wild tree species (**??**).

The term 'chilling' is now used across numerous fields in plant biology to refer to a process where dormant buds exposed to cool temperatures accelerate a phenological event that later occurs after warm temperatures. Focused on how chilling can accelerate events, researchers have calculated 'chilling' required for leafout of forest trees from cutting experiments similar to those used for peaches (reviewed in **?**), ground observations of budburst (**?**), and satellite measures of greenup (**?**). These estimates rely on phenological models that have become critical across a suite of fields, from climatology where modeling how vegetation on the land surface responds to anthropogenic climate change affects carbon storage, to crop biology, where estimated chill units guide growers in which specific cultivar to plant, and has led to the cross-disciplinary field of phenology (**???**).

Alongside these more macro-scale studies of chilling, molecular approaches have also examined chilling. Many studies have focused on vernalization—cool temperatures required for flowering (**?**)—in *Arabidopsis thaliana*, with studies in woody species, especially *Populus* examining chilling before budbreak (**??**). These studies generally use controlled temperatures to vary the hypothesized amount of chilling then examine molecular and cellular responses (e.g., **???**).

Today these studies have led to over 30 basic models where accumulated chilling releases plants from dormancy and hundreds more when considering different species and cultivars (**??**). Though early debates considered whether plants were truly dormant or only growing slow ('dormancy' or 'rest' versus 'quiescent'; **?**), today most research assumes a model with two phases of dormancy (Fig. **??**).

In most models, chilling can only be accumulated under certain temperatures—traditionally above zero but below 10°C—with certain temperatures being optimal for the most rapid accumulation of 'chill units,' where some unknown sum of chill units breaks endo-dormancy. Which temperatures are most effective at providing chilling is a common question addressed in experiments, with different experiments providing different answers (**??**). This mirrors growing degree days in many ways, where a lower temperature (e.g., 5 or 10°C base temperature) is too cold for forcing units to accumulate and plants need some total sum of such units to leafout or flower, but has added complexity given the importance of both the lower and upper temperature thresholds for 'chill units' (whereas growing degree day models can often ignore the upper threshold, estimated at 25°C or above, as it is rarely reached in natural spring conditions, **??**).

Further complexity comes from the hypothesized diversity of these temperature thresholds and sums across species and populations. Most assume different species require different sums of chill units, and may have different lower, upper and optimal chill temperatures. Within species, populations may require different sums of chill units, with populations in more mild climates—where warm interruptions in the winter are more common—requiring more chill units than those in areas with cold winters, where temperatures rarely rise above zero before spring (**??**).

The problem with chilling (JA: I would cut the stuff about science being stalled. Put it in a different section if necessary, just explain the problem here.)

Chilling is a latent, unobserved process. Typically, 'chilling' describes the physiological phase (endo-dormancy) in which a plant experiences environmental temperatures that induce progress towards the next physiological phase (eco-dormancy) that ends in budburst (Fig. **??**). The problem is that the transition from endo- to eco-dormancy corresponds with no clearly measurable phenomenon, and thus the properties of each period cannot be determined from the observed leafout, even under experimental conditions in which the temperature can be manipulated. The parameters of a model governing this system are said to be underdetermined or unidentified.

For example, consider a simple model in which one would need to estimate the minimum and maximum temperatures that allow chilling to accumulate (two parameters) and the total sum of those temperature units needed to trigger a shift into the next physiological phase (often called, 'endo-dormancy break,' for one additional parameter for three total). Models then need to estimate when plants start and stop accumulating (two more parameters). An experiment that raises temperatures and observes an earlier leafout could be explained by ...

To address this, models often assign the start date of the endo-dormancy as known (e.g., starting 1 September in the northern hemisphere) and rely on assumptions to set the end date (see Box: Why has progress on modeling stalled for decades?). A common assumption, developed by early work on peaches, is that high and rapid budburst (leaf or flower) is evidence that chilling has been met (i.e. endo-dormancy has ended **?**). While this assumption is widely used, it is rarely if ever tested beyond the early work on peaches. This approach of assigning some unknown parameters as known has the benefit of avoiding adding more unknown model parameters, but it also has led researchers to be overly confident in a model of chilling where more is actually unobserved and unknown than acknowledged. (JA: I removed (without any variance) since allowing for uncertainty/randomness doesn't solve the identification problem.)

Hidden assumptions and numerous parameters can easily drive diverging models. Even if we assume high and rapid percent budburst signals sufficient chilling, most models today include parameters that cannot be uniquely identified with current data. Given experiments and models have suggested many variants on a more complicated model of chilling—for example minimum, maximum and optimal temperatures, or high temperatures that reduce previous accumulation (Fig. **??**, and see **???**)—current data are relatively uninformative to try to estimate all the parameters the models include. Further, recent models have often relied on even less data (**?**); many current methods use only observational data of the timing of leafout (or flowering) to attempt to estimate a model of chilling for different species or locations and project it forward to understand effects of anthropogenic climate change (**???**). Perhaps not surprisingly then, which model is deemed best varies strongly by method and approach (**???**), with no clear pattern.

New molecular insights could reshape the field and its models (JA: Also great, would not cut)

There is some hope that the transition from endo- to eco-dormancy is measurable. Molecular insights have long contributed to crop and forest tree models of chilling (**?**). Decades of work on vernalization have outlined the pathways—and genes—that lead to flowering only after winter's cool temperatures in biennial (herbaceous) populations of *Arabidopsis thaliana* (Fig. **??**, **??**). Research has linked some of these pathways to similar ones in woody species, and have also highlighted the sugar callose (1,3-β-D-glucan) as potentially pivotal for chilling (**??**). Multiple studies across multiple species have now shown that (1) lower temperatures appear to degrade callose and (2) the release/loss of callose appears to re-start cell-to-cell communication before budburst (**?**). Taken together, these results suggest the loss of callose—generally degraded through 1,3-β-glucanases (a group of enzymes)—may be an indicator of endo-dormancy release, though other factors, such as ABA, also often change at the same time (**??**), and may provide a similar observable signal of endo-dormancy release (**??**).

If callose is functionally a major controller of endo-dormancy and its release, then chilling models could be limited to those that match the idea of glucanase degrading callose—meaning models that include a temperature range over which the enzyme is active (Fig. **??**). In contrast, models using simple temperature thresholds (e.g., all hours below -5°C equally allow chilling) would appear less biologically accurate, as enzymes generally do not work over such a wide range of temperatures.

Other new molecular insights similarly suggest that such simplified temperature metrics used in many chilling models may not map to molecular realities. For example, new work on how slow growth itself may act a 'long-term thermosensor' (**?**) adds to an increasing number of molecular studies that suggest plants integrate long-term thermo-sensing in the winter alongside responses to short-term temperatures (**??**). The best models of leafout may thus need to integrate across multiple timescales. This could easily add complexity to models of spring phenology that are already challenged by too much complexity. But we argue that new insights from molecular biology could begin to rule out models by focusing on new experiments and modeling approaches that target the major problems facing models of chilling with richer, more informative data—if the field is open about current uncertainty in models of chilling.

Develop cross-disciplinary models with falsifiable predictions (JA: As of right now, this section seems unrelated to the unidentifiability problem. I think you have two points. 1. if chilling is real, its properties may be better understood by looking across different disciplines and combining observational/experimental evidence. 2. Is chilling really necessary for scientific progress?)

New richer data from molecular biology studies and other approaches to identify chilling (**??**) hold the promise to shift chilling from an unobserved complex process to something we understand and can robustly forecast, but will require developing models that yield clear and falsifiable predictions.

Yet progress may advance slowly given current practices in many models and experiments for estimating chilling, including the increasing the number and complexity of process-based models while fitting overly simplistic statistical models to empirical data. Below we show how these practices have limited developing models that make falsifiable predictions and outline a pathway forward.

Develop benchmarks to help discard models

A first step to improve the falsifiability of various models is to reduce the number of models to consider. Currently, crop biologists, phenological process-based modelers of forest trees, molecular biologists and hardiness modelers all develop unique and rarely compared models of dormancy and budburst (but see **?**), highlighting a major problem, but also a major opportunity. Synthesizing models—and their underlying biological understanding of chilling—across the many research fields developing chilling models today would help identify models that are equivalent. Uniting these models, first by fully defining their assumptions and conditions (e.g., what species are they designed for, what phenological or dormancy phase do they start at?), then comparing their predictions and pushing them to make different testable predictions would help build a unified model of chilling, with implications for better models of budburst, and cascading improvements in forecasts for crop yield and forest carbon sequestration.

Changing how we evaluate and compare models would also aid discarding models. To date, models are compared using different model comparison statistics and different datasets in each paper. This makes it extremely difficult to prioritize models for study and experiments, since one model may be preferred on some data for certain model statistics and another model given other data and statistics. Building standard benchmark datasets that are always tested alongside the same test statistics would alleviate some of this problem, and make it easier to identify why different datasets find different answers, thus aiding model development. Yet discarding models fully likely requires moving away from relying only on model comparisons statistics and towards models that make falsifiable predictions, which will require first embracing how poorly most models used today perform.

Highlight the unknown

Perhaps the largest problem with current process-based models of chilling is that they cannot uniquely estimate the most important aspects of chilling. Even in some of the simplest models, estimates of what temperatures accumulate 'chilling' and how much chilling is required to shift physiological stages often occupy multiple divergent options. A common example of this is one outcome where the range of temperatures that allow chilling to accumulate is wide and thus the threshold amount of chilling large, or an outcome where the range is smaller and, thus, the threshold lower. Yet researchers often only present one of these outcomes (**?**), versus showing the full range of possible values—that is, the full uncertainty. Highlighting this uncertainty instead of hiding it could quickly help compare models and guide experiments. Thus we suggest research always include routine estimates of uncertainty and clearly state any modeling choices that may have limited insights into the full uncertainty (e.g., limiting the parameter space the model can search when estimating parameter values). Given that this uncertainty only grows with complexity in

current models, focusing more on simple models in the near-term may aid progress.

Making chilling models more biologically relevant could also come from improving the statistical models of chilling that are fit to empirical data. Because process-based models are so difficult to fit, they are not routinely used to statistically fit empirical data from ground observations or new experiments.. Instead, researchers fit new empirical data with so-called 'statistical models' that often follow canonical treatment designs (e.g., ANOVA). Statistical models are usually far simpler, and make a suite of unstated assumptions that contradict the current understanding of chilling (see Fig. **??** and Box: Why has progress on modeling stalled for decades?).

Improved statistical models could challenge some of these assumptions by inching closer to the biology. Simple log transformations better match the non-linear accumulation model of chilling and forcing (**?**). Models could also relax the assumption that 'chilling' and 'forcing' treatments correspond to different physiological states by testing whether the start dates may vary with the treatments, as the actual endo-dormancy break could equally occur in either cool or warm treatments. Both these alternatives are simple to implement and thus could be included as alternative models that test alternative hypotheses (Fig. **??**). Integrating other statistical approaches could also relax additional assumptions to provide insights into how chilling works biologically. For example, instrument variable approaches could help in studies attempting to manipulate chilling and forcing where the underlying state is not known. All these models, however, are likely to be limited in the inference they provide without substantial increases in data (**?**).

Redefining chilling through new experiments

One important way to leverage new molecular insights for modeling is through new experiments designed to identify novel ways to more directly observe and measure chilling. Experiments testing for evidence of callose loss using the temperature treatments commonly applied in past studies (**?**) could be complemented by studies with other cellular and molecular markers (**?**). Testing these methods together alongside previous-used markers of dormancy shifts—including the often-used bioassay of high and rapid budburst at higher temperatures (indicating endo-dormancy release), and additional methods, such as weighing flower buds (**?**) or tracing water reactivation into cells (**???**)—could help align both new and old methods.

Because chilling is an unobserved process we argue that comparing methods to measure chilling should be a major priority for the field. This comparison will need to allow for the reality that different methods may measure different processes and, thus, terminology may need to adapt as well. As a first step, research could stop referring to treatments in experiments as 'chilling' or 'forcing,' or other terms that assume an underlying physiological state, and instead focus on the actual treatments (e.g., 'cool temperatures before warm'). Currently, many experiments use the term 'chilling' to refer to a treatment where researchers do not know the physiological phase (**??**); for example, cuttings or buds from woody plants are often chilled at 5°C for 6 weeks in the dark in a 'chilling' treatment, then transferred to warming 'forcing' conditions.

Current experimental designs are unlikely to radically challenge models of chilling, but small tweaks may still offer important insights. In particular, experiments considering multiple cool and warm temperature treatments could measure hardiness to improve the model of how hardiness and dor-

mancy interact (**?**). Cool treatments may also be useful if they tested for the effect of light regimes given the prevalence of cool temperature ('chilling') treatments in the dark to date (**?**).

Model experimental and observational data together

One way to increase the amount of data used to estimate chilling models is to include both experimental and observational data together in one model. This is rarely (if ever) done, in part because of how differently they may be observed, but also because of the challenging diversity of environmental conditions across these two data types. For example, many experiments apply cold temperatures in the dark, while photoperiod shifts each day in observational data, or many experiments achieve extreme differences in cool and warm temperatures (e.g., very minimal cool temperatures, or very warm temperatures) while natural climate data lacks such extremes. Many of the original studies that led to the concept of chilling, however, were developed from datasets that created greater extremes in observational data—focusing on crops planted well outside their natural range (e.g., peaches in Florida and Israel) and bridged across observational and experimental studies more often (**??**).

Experiments bridging across methods may have the greatest opportunity to provide data that would truly challenge current models of chilling. Testing models across large environmental gradients in the field is one of the best ways to find out where models work—and where they fail. For example, molecular biologists tested vernalization models by through comparing predicted to observed flowering times in a common garden study across Europe (**?**), and supported the temperature-dependent growth model by testing its predictions of what happens when growth is altered but temperature is held constant (**?**). Similar examples for challenging other models of chilling date back over 40 years to when many of the models used today were developed (**???**), but could take place now. Models of chilling can make predictions under lower field chilling then test them using individuals planted beyond the range (either planting those individuals now or identifying such cases in forestry provenance trials or similar). Process-based modelers could also challenge their models more through more dramatic variation in biology, via experiments that include mutants or similar variants.

Fitting experimental and observational data should drive coherency in what chilling models perform best and reduce support for some models. This would limit the growing number of studies that have compared different process-based models on observational phenology data—where 'natural' field conditions likely often satisfy chilling requirements for wild species (**??**)—to find the simplest models perform best. By adding more extreme experimental conditions we expect more complex models will perform well. Even if models cannot fit both data types together we suggest that new research should should include comparisons of the model performance on observational data and experimental data—models that cannot fit both data types should be flagged as indicating a potential problem with the model. At the same time, datasets must provide the necessary information to fit chilling models (e.g., for experiments—what was the dormancy induction temperature, what as the thermo and photo-periodicity of each stage of the experiment).

We argue that one of the main reasons for stalled progress on modeling chilling is that most models—from the old to new ones—cannot actually estimate the parameters in them. Taking a

simple example of a chilling model with three parameters—the minimum temperature for chilling, the maximum temperature, and the accumulation needed—shows that there are multiple solutions. Considering just two of these possible solutions highlights how the temperature range trade-offs with the accumulation: if the temperature range is wide (lower minimum, higher maximum) then the accumulation required will be higher, while if the range is smaller, then the accumulation needed is lower. The full suite of possible solutions is effectively endless (and the trade-off between range and accumulation is not linear, as it depends on the full width of the range, but also its placement relative to 0). Further, this model is not actually one of only three parameters as two additional parameters were set as known (start day of accumulation was set at 1 September, and the endodormancy break date at 30 January) so that the model could even be fit using common algorithms. This reality is present in every model of chilling, but it is rarely presented clearly.

Researchers tacit approach to these major issues likely has contributed to the expansion of models over the last few decades without any clear advances. While various models have added complexity via the shape of the optimal temperature range for chilling, allowing accumulated chilling to be reduced, shifting the start date of chilling, and/or allowing chilling and forcing to act at once (**????**), none of these have swept through the field. These new models of chilling have all added parameters, but none of the parameters added to chilling models in 40 years that have been successful enough to be added to all models of chilling.

Being clear about model uncertainty, the full number of parameters and how well they fit would advance progress on chilling through multiple routes. First, it would help all researchers in the field recognize what is fundamentally unknown and thus focus more research in these areas. Second, it would highlight which parameters are most often fixed (effectively assumptions in the model) versus fit to data, and to which type of data. With this, more research could easily compare across models and datasets to give better overviews of what is known, assumed, or most often studied (i.e., what parameter model studies try to fit). Given the extensive list of proposed complexities to chilling models (**???????**), having simpler models (fewer parameters) that are routinely used to compare to more complex models, would likely help the field advance. Highlighting uncertainty in findings from experiments would also aid modeling studies to be more upfront about assumptions and limitations (Fig. **??**).

1 References