The authors would like to thank the Reviewer for the very helpful comments and suggestions. We will take them into consideration in the revised manuscript and the specific questions will be addressed and incorporated in the revision. Moreover, we would like to respond to some individual points:

General comments

At times it is difficult to follow what the main findings of the study are, and how the results are used and fit with the objectives (see the specific comments). This could be improved by defining the objectives of the study in the introduction more clearly, and explain how the results presented in the figures and tables contribute to the objectives.

We will clarify the objectives and questions behind this study in the introduction referring where the reader can find the results contributing to the answer of each specific question.

Simulated fluxes are compared using different phenology models, but how do these compare to the eddy covariance fluxes from the Harvard site? Are they comparable? Are your simulated fluxes within the same range? Could you add something about this to the discussion by comparing your results with that available in the literature?

As suggested by the Reviewer we compared annual gross primary productivity (GPP) estimates for the Harvard Forest reported in Urbanski et al., (2007) (Table 1) with GPP simulated with:

1) BEPS (Reference Run in the paper);
2) BEPS forced by bud-burst dates simulated by the best model formulation selected according to the AICc ($BEPS_{SW\text{-}CF2}$).

The average GPP simulated with the two runs of BEPS are comparable with the GPP estimated from eddy covariance flux measurements. GPP simulated with BEPS forced with the optimized phenology model is slightly more accurate than BEPS with the native phenology routine.

Table 1 - Annual Gross Primary Productivity reported in Urbanski et al 2007 (GPP) and simulated with BEPS and BEPSSW-CF2.

Year	GPP	GPP BEPS	GPP $BEPS_{SW-CF2}$
	MgC ha^{-1}	MgC ha^{-1}	MgC ha^{-1}
1992	11.7	14.7	14.3
1993	13.6	15.3	14.6
1994	12.4	16.6	15.8
1995	12.5	17.4	16.3
1996	13.3	17.0	14.8
1997	14.0	14.7	14.5
1998	12.1	16.4	14.6
1999	14.0	15.5	15.2
2000	14.5	15.7	14.3
2001	16.4	16.0	14.9
2002	15.1	16.7	15.0
2003	15.4	16.5	14.9
2004	17.1	15.9	15.0
MEAN	14.0	16.0	14.9
SD	1.7	0.8	0.6

GPP simulated with our runs of BEPS are also in good agreement with a recent estimates of is 14.09 $MgCha^{-1}$ of mean annual GPP estimated at the Harvard forest for the period 1992-2005 reported in Keenan et al., 2012 (Tab 1).

P886, L15: can you explain what the Sarvas function is, and how it is different from the threshold approach.

For the different model structures (Spring warming, Sequential, Alternating or Parallel) presented in the manuscript, two functional forms of the equations for the computation of rate of forcing R_f and chilling R_c have been used.
In one approach (here denoted CF1, or threshold approach), rate of chilling and forcing of the day of the year (t) are accumulated if the daily air temperature $x(t)$ is higher than a specific threshold, respectively. Considering as example the forcing, its state is specified in terms of "forcing degree-days," which are accumulated as $R_f = x(t) - T_{force}$ where $x(t) > T_{force}$ and $R_f = 0$ otherwise. T_{force} is a temperature threshold optimized as model parameters and reported in Table 1 of the submitted manuscript.
In the second approach (denoted CF2), rates of chilling and forcing are both specified as nonlinear functions of $x(t)$ according to the Sarvas model (Sarvas 1974 in Chuine 1999). More specifically, in CF2, chilling is accumulated according to the triangular function:

$$R_c = 0 \qquad \text{where } x(t) \leq -3.4 \text{ or } x(t) \geq 10.4$$

$$R_c = \frac{x(t)+3.4}{T_{chill}+3.4} \qquad \text{where } -3.4 < x(t) \leq T_{chill}$$

$$R_c = \frac{x(t)-3.4}{T_{chill}-10.4} \qquad \text{where } T_{chill} < x(t) \leq 10.4$$

In CF2, the rate of forcing is a sigmoid function of $x(t)$, and (unitless) forcing is accumulated as in follow for $x(t) > 0$:

$$R_f = \frac{28.4}{1+\exp(-0.185(x(t)-18.4))}$$

As example, we report the time series of R_f computed with CF1 (T_{force} = 0 °C) and with CF2 using the air temperature measured at the Harvard Forest during the 2009.

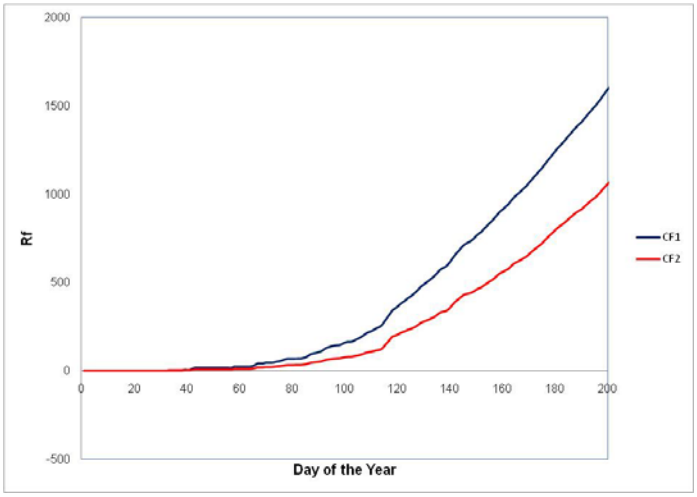

Figure 1 - Rate of forcing (R_f) computed with the threshold approach (CF1 with T_{force} = 0 °C) and with the Sarvas' function.

According to referee suggestion we will introduce a brief description of the threshold method and of the Sarvas' approach.

P889, L22: can you explain the Sen's slope estimator? Figure 4 cannot be understood now.
P895, L6: can you explain the Mann-Kendall test? Or at least how the reported numbers were derived?

The Mann-Kendall test (Mann, 1945; Kendall, 1976) and the Sen's slope estimator (Sen, 1968) are non- parametric procedures for trend testing and estimation of trend magnitude, respectively. These methods are suited for univariate time-series with monotonous trends and no seasonal or other cycles in the data (no autocorrelation in the time series).
Non-parametric methods are preferred to parametric methods (i.e. regression analysis: regression slope and test) because make no assumption for probability distribution of data.
The second advantage of non parametric methods and, in particular for the estimation of trend magnitudes, is their robustness to outliers or to abrupt breaks due to inhomogeneous time series (Hirsch et al. 1982).
The Mann-Kendall test (Mann, 1945) determines the statistical significance of trends. The test has been extensively used with environmental time series (e.g. Hipel and McLeod, 2005; Gagnon and Gough, 2005).

The Sen's Slope provides a robust estimate of the slope of a time series and it is calculated as the median of the slopes of all the pairs of ordinal time points in the time series.
Here, we used the Sen's slope estimator to assess the slope (or magnitude) of trends in bud-burst dates.
Given the wide diffusion of the methods for the analysis of environmental time series we will not provide in the manuscript a full description of the method (e.g. Hipel and McLeod, 2005; Gagnon and Gough, 2005). However, we will add a short description of both methods in the Materials and Methods section providing to the readers more references.

P892, L8-12: what can be seen/concluded from Fig. 3. Please describe. Could you also explain why species PRSE behaves very differently, while the same best model is found?

In Figure 3 we report the determination coefficient (R^2) and slope of the linear regression between observed and predicted bud-burst with the best model selected (Table 1 of the manuscript). Figure 3a shows how the best model for each species is able to explain the variability of bud-burst across species for each year. Figure 3b shows how the best models are able to explain the variability of bud-burst for each species. Figure 3c is the scatterplot of observed and modeled bud-burst for the years not used in model optimization. We will clarify the interpretation of Figure 3 both in the results and discussion section of the revised manuscript.
PRSE (Black Cherry -. *Prunus serotina*) is a species with early bud-burst. For PRSE, Spring Warming models are still selected as best model. It should be considered that model parameters are estimated for each species, therefore it can be possible that the same model can be selected for species with late budburst. Regarding the parameters estimated, we observed that PRSE requires an amount of accumulated forcing units (R_f) lower than other species while the sensitivity to temperature is comparable to other species.

P892, L18-25: can you explain how this figure should be understood and what can be concluded from it? Is it necessary to include this figure for you conclusions?

Figure 4 is a relevant figure supporting the conclusion of the manuscript. In Fig. 4 the uncertainty of the magnitude of the projected bud-burst trends simulated for each species with the best model selected as in Table 1 is represented. The width of violins represents the uncertainty in trend for each species. The differences between blue and red violins represent differences in trend and uncertainty between A1fi and B1 scenario.
Therefore, this plot summarizes the information of 1) the differences in bud-burst trends between scenarios, 2) the differences in the uncertainty of trends for different scenarios and 3) for different species. We will clarify the meaning of this plot throughout the manuscript.

P893, L8: how are the trends smoothed?

The smoothed bud-burst projection was extracted from the time series by using a local polynomial regression fitting (Cleveland and Devlin, 1988) as described in the Materials and Methods section. We will add this information to the figure caption

P894, L2: can you explain how dBB/dT is calculated in Fig. 6?

dBB/dT is computed as the ratio between the smoothed bud-burst projections for each model and the smoothed temperature. Smoothing was computed from the original time series by using a local polynomial regression fitting (Cleveland and Devlin, 1988)

P894, L16: over which period is the interannual variability calculated?

The interannual variability was computed as the standard deviation of the residual between the forecasted bud-burst and its smoothed time-series. (please see page 889 lines 17-19)

P898, L20-end: I do not understand this paragraph. What do you mean with "driver uncertainty is enhanced"? What is the message or point you ant to make here?

The driver uncertainty can be quantified as the difference between runs of the same model forced by different drivers. Changing drivers (i.e. changing scenario) we obtain larger differences for models less supported by data. As shown in Fig. 5, the spread between models at the end of simulation is larger in Fig 5a (Scenario A1fi) than in Fig. 5b (Scenario B1); in particular Spring Warming class models without photoperiod limitation (model less supported) are more sensitive to changes in driver scenario and the difference in the smoothed trend reported in Fig 5a and 5b is larger than for other models.
The impact of the use of different driver is then larger for model less supported and therefore the uncertainty due to model drivers is larger (or enhanced) if we consider models poorly supported by data. We will clarify this concept in the revised version.

General question I was curious about: is it realistic to use present day phenology parameters in predictions? Is there anything in the literature about how they change or adapt when climate changes?

We cannot address this question since we are not aware of phenological studies showing the acclimation of trees to climate change and we are therefore have no knowledge of how parameters could change or adapt to climate change.

P904, L7-8: so, these models are wrong?

Our model selection is based on the Akaike Information Criterion (AIC) that allows ranking different models considering the trade-off between the model likelihood and complexity (Anderson et al., 2000).
In our model selection (Tab 3 of the submitted manuscript), the evidence for a lead effect of models with photoperiod limitations is very strong while empirical support of models without photoperiodic limitation is weak. Models without the explicit description of the photoperiod limitation are never supported by data. This does not mean that these models are wrong but it means that, according to our dataset, these models are not able to describe accurately the bud-burst observed and that are not suited to simulate bud-burst for the Harvard Forest. We will clarify this in the revised manuscript

References:

Anderson, D.R., Burnham, K.P., Thompson,W.L., 2000. Null hypothesis testing: problems, prevalance and an alternative. J. Wildlife Manage. 64, 912–923.

Chuine, I., Cour, P., and Rousseau, D. D. (1999). Selecting models to predict the timing of flowering of temperate trees: Implications for tree phenology modelling, *Plant Cell and Environment*, 22, 1-13.

Cleveland, W., and Devlin, S. (1988). Locally-weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, 83, 596–610.

Gagnon, A.S. and Gough, W.A.(2005) Trends in the Dates of Ice Freeze-up and Breakup over Hudson Bay, Canada, ARCTIC, Volume 58, issue 4.

Hipel, K.W. and McLeod, A.I., (2005). Time Series Modelling of Water Resources and Environmental Systems. Electronic reprint of our book orginally published in 1994. http://www.stats.uwo.ca/faculty/aim/1994Book/.

Hirsch, R.M.; Slack, J.R.; Smith, R.A. (1982). Techniques of trend analysis for monthly water quality data, *Water Resources Research* **18** (1): 107–121, doi:10.1029/WR018i001p00107. 58201.

Keenan, T.F., I. Baker, A. Barr, P. Ciais, K. Davis, M. Dietze, D. Dragoni, C. M. Gough, R. Grant, D. Hollinger, K. Hufkens, B. Poulter, H. McCaughey, B. Rackza, Y. Ryu, K. Schaefer, H. Tian, H. Verbeeck, M. Zhao, A.D. Richardson. (201x) Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO2 exchange. *Global Change Biology*. In press

Kendall, M.G. (1976). Rank Correlation Methods. 4th Ed. Griffin.

Mann, H.B. (1945). Nonparametric tests against trend, *Econometrica*, 13, 245-259.

Sarvas, R.: Investigations on the annual cycle of development on forest trees active period, Communicationes Instituti Forestalis Fenniae, 76-110, 1972.

Sen, P.K. (1968), "Estimates of the regression coefficient based on Kendall's tau", *Journal of the American Statistical Association* **63**: 1379–1389, JSTOR 2285891, MR02

Urbanski, S., Barford, C., Wofsy, S., Kucharik, C., Pyle, E., Budney, J., McKain, K., Fitzjarrald, D., Czikowsky, M., and Munger, J. W.: Factors controlling co2 exchange on timescales from hourly to decadal at harvard forest, Journal of Geophysical Research-Biogeosciences, 112, 25, G02020. 10.1029/2006jg000293, 2007.