Many thanks to the reviewers for your thorough and helpful responses. We have addressed the comments within the manuscript and please find our responses to your comments and suggestions below, identified by paragraphs that start with a ##.

**Anonymous Referee #1**

First, I am curious why the authors chose to use the LAI3g dataset rather than the MODIS LAI product (MCD15A2). While these datasets have similar agreement with respect to in situ LAI observations (RMSE=∼0.7; Zhu et al. 2013 RemSen), both validation exercises used a limited number of sites located in drought deciduous regions (e.g., Figure 2 in Zhu et al.) and MODIS data are known to have higher fidelity than AVHRR data (Huete et al. 2002 RemSenEnv). Moreover, while the LAI3g dataset is twice as long as MCD15A2, it is not clear how additional years of data actually benefit the model assessments made in this analysis. It would be helpful if the authors provided an explanation for why they chose to use LAI3g rather than MCD15A2. If this is not possible, then I recommend that the authors redo the analysis using MCD15A2 to ensure that the highest quality data are used.

## We agree that the MCD15A2 data set is more robust, but we felt that the length of the LAI3g data set was important, as semi-arid ecosystems have inconsistent rainfall and green-up patterns, making more data useful to cover a broader range of weather patterns. In addition, since we were focused on broad issues in CLM, not pinning the LAI values exactly to the satellite data, we expect that switching data sets would not significantly change the overall results, given that the discrepancies in the original model were so profound.

Second, while it is certainly necessary to assess model performance over the entire annual cycle, I believe it is equally important to consider how well models capture timing of seasonal metrics such as start and end of season since they largely control annual carbon uptake (e.g., Ma et al. 2007 AgForMet). For example, in Figure 2a it is apparent that the LAI3g growing season across NH C3 grasses is shifted early by an entire month. Despite the relatively high R2 and low RMSE across this region as shown in Table 2, this result suggests that the model does not incorporate the mechanistic controls for triggering leaf onset or autumn senescence. Therefore, for regions with stronger seasonality in LAI (mainly temperate grasslands), I suggest the authors perform a quantitative assessment of model bias in start and end of growing season dates.

## We agree that this would be an additional interesting assessment, however, we do not think it would change the overall results of this paper, and algorithms like those designed to assess start and end of growing season, growing season length, or amplitude do not perform well when trying to assess areas with more than one growing season per year, which are a common part of savanna-type ecosystems. We have added comments regarding this in the methods section of the manuscript (see response to reviewer 2, below). Further, our study is focused on semi-arid systems, and not on cold-deciduous phenology, which has been studied more extensively in the past (e.g. Levis et al. 2012).

Third, perhaps this is outside of the context of this study but if the authors have access to in situ observations of LAI data from a semi-arid/drought deciduous region it would be interesting and
worthwhile to validate the authors’ novel cumulative rainfall model using these data. This is briefly mentioned on Page 5821, Lines 23-24.

### We agree! There is lots of room for more assessment of drought deciduous phenology, however, there are only a small number of in situ data sets (e.g. Phenocams) in drought deciduous areas, as far as we know. Adding more data sources (both on the ground and space-borne) is a focus of ongoing proposals, etc.

Finally, while the authors describe and perform parameter sensitivity analyses in the methods and results sections (research question #2), there does not appear to be any significant discussion or formal conclusions made regarding these results later in the manuscript. Moreover, in Figures 4 and 5 the relationship between each model run and the magnitude of the varied parameter is unclear. Overall, I believe this is an important aspect of the analysis and, therefore, suggest that the authors make appropriate changes to resolve these issues.

### Latin hypercube types of analysis are difficult to illustrate, however, we agree that the discussion of the sensitivity analysis could be more detailed, and appropriate changes have been made to the manuscript explaining the results of the sensitivity analyses and what is shown in Figures 4 and 5.

We’ve added the following paragraph to explain this challenge in the methods, as well as described this issue elsewhere in the manuscript:

### To assess the performance of the different models in the Latin hypercube test we originally plotted the coefficients of determination between the different models’ LAI values and the LAI3g data at those points. However, this result did not illustrate any clear optimum in model performance either for the parameters of the existing model, nor for the rainfall threshold. We illustrate this using the time-series data in Figure 4, which highlight the unusual behavior of the model, and to assess whether the extra green-up period during the dry season had been eliminated in any of the parametric permutations. We ascribe the lack of a clear parametric signal to two effects. First, the LAI3g data were necessarily aggregated to monthly values, meaning that the primarily sub-monthly variation between ensemble members was masked. Second, the timing of the secondary leaf-on period in the dry season was the emergent property of the oscillatory (and thus somewhat chaotic) dynamics of the soil-vegetation feedback on soil moisture.

### Regarding the rainfall threshold, we’ve added the following statement in the results:

### While this new rainfall threshold improved model performance both at our points and globally (see below), we note that, except in a few exceptionally dry areas, the model did not appear to be particularly sensitive to the rainfall threshold, as long as some rain did fall, but this threshold, and the drought deciduous algorithm as a whole, deserves more research into seasonal drivers.

Minor Comments and Suggestions:

Page 5804, Line 23: Not sure that quotes are necessary here.

### Removed

Page 5807, Line 7: Please explain what BGC stands for.

### BGC stands for Biogeochemistry and designates a particular version of CLM that includes an active carbon and nitrogen cycle, but is different from the CLM-CN version (also with active C and N) which was used primarily in earlier versions of CLM. Since CLM4.5BGC is now the
default version, we have included an explanation in the beginning of the paper but changed further mentions of CLM4.5BGC to CLM for simplicity.

Page 5808, Equation 1: offset → onset

## Good catch! Looks like this error was introduced during typesetting – per reviewer 2’s comment we’ve changed all of these to psi-threshold

Page 5810, Line 6: Please explain what CRU-NCEP stands for.

## CRU-NCEP = Climate Research Unit (University of East Anglia) – National Centers for Environmental Prediction (NOAA), we did not define this following other papers published in Biogeosciences, but will defer to the Editor’s preference.

Page 5810, Line 19 (and instances afterward): gridcell → grid cell

## Changed

Page 5816, Section 3.3: The CLM naming convention (e.g., CLM4.5BGC, CLM, CLMMOD) gets a little confusing here. Perhaps it makes sense to only use CLM and CLMMOD?

## See comment above – we’ve changed CLM4.5BGC to just CLM

Page 5822, Line 12: phonological → phenological

## Changed. Whoops!

Figure 2: Need to show letters in each panel.

## Letters are in the bottom right corners of each panel – e.g. “A. NH C3 Grasses”

Figure 3: This figure is somewhat busy, although I am not sure what makes the most sense to remove. Perhaps it is okay but I suggest the authors consider alternative representations, if possible.

## We agree that these figures (and Figure 10) are challenging, however, we opted for this presentation as it shows all of the relevant details, and we felt that stacking all of the data allowed for easier comparisons than plotting things in different panels.

Figure 4: If the authors choose not to explain differences between each model run, it may also make sense to just plot a mean curve with +/- 1 standard deviation or some other indicator of variance.

## Related to the comment above, we’ve added a more clear description of the Latin hypercube analysis and discussion of these figures to better explain how to interpret them.

Figures 8/9: Perhaps it would make sense to mask out grid cells with less than some fraction of drought deciduous land cover so that it is easier to depict spatial patterns in the improvement of model performance.
## We've opted to leave all of the data clearly represented (instead of masked) with the hope that these figures will be useful to researchers interested in study regions beyond those we have focused on.

**Anonymous Referee #2**

Major comments and suggestions: The papers central result seems to be the pragmatic “fix” to stop the unrealistic simulation of multiple onsets over much of the drought deciduous parts of the world. This fix – using a cumulative rainfall threshold - is basically designed to stop anomalous onset caused by unrealistic soil water movement in CLM. While I am not completely against using these kinds of pragmatic solutions to prevent erroneous events within a model - and have had to use them myself - to base an entire paper around such a development that has no real world basis seems to be a bit overkill. The authors identify the actual cause of the models problems in simulating seasonal phenology (i.e. that a sudden drop in transpiration at the end of the growing season and a potentially unrealistic water table dynamics triggers leaf onset through a sudden increase in soil water potential), but fail to investigate potential changes in the simulation of soil water dynamics or a more real-world based soil water/phenology coupling. To me it seems that this would have been a much better starting position for model development and, based on the information that the authors (again, very clearly) explain in the paper, I can think of several avenues of investigation, provided that data is available for proper analysis:

1. Is the simulation of the “unconfined aquifer” and its interaction with the soil layers realistic? (to be fair, the authors do suggest may be tricky because of lack of data)

2. Is the sudden increase in water potential after the growing season realistic? If not, is there a way of “blocking” this sudden spike? A pragmatic solution here would probably be more appropriate than the one the authors implemented, as the fix is applied at the root of the problem.

3. Is phenology actually coupled to soil water potential? If not, is a more realistic decoupling than the one put forward possible, along with a demonstration that this is based on real world plant responses?

4. Can plants “pull out” of onset if soil water potential peaks but then suddenly drops again?

## We agree that this ‘quick fix’ is not the ideal solution to a fairly significant issue in CLM, however, our goal was to correct this one problem (dry season green-up) without significantly impacting other parts of the model. Several groups are currently working on new and better soil water solutions for the CLM, but changes to the soil water algorithms will have wide reaching impacts that will need to be assessed in their own right. The CLM4.5 is widely used by the land surface modeling community, and its development is a collaborative effort that relies on identification and publication of issues as rapidly as possible, to reduce redundant development efforts. We felt that this issue was too important to wait for those soil water parameterizations to come on line, given its impacts on the global carbon cycle in the model (new soil hydrology
representations are not finalized at this time). It is not our decision whether or not this modification is incorporated into the next release of CLM, but we wanted to make it available if deeper issues in the soil water algorithms are not resolved, or to others to use prior to the release of the new model, which will likely be up to a year from now. Regarding #3 and #4, those are great questions. KMD is currently conducting a literature review to attempt to address these questions across the semi-arid parts of the globe, but so far the answer seems to be that it depends on the plant and that there is quite a bit of variation in how plants respond to soil moisture at scales finer than the CLM plant functional type groupings. We hope that this ms stimulates more focus in the land modeling community on dry ecosystems and their processes.

With a slight shift on focus, the paper may have the potential for being an interesting case study of model benchmarking and evaluation. However, model quantitative evaluation of initial and developed model is a little spare and not systematic. A more suitable evaluation of phenology would probably need to assess the models simulation of the timing of the start, peak and end of the season, magnitude and number of onsets via quantifiable metrics, rather than just RMSE over the annual cycle and visual spatial and timeseries comparisons. Also, a proper assessment of model improvement (or degradation) over a range of model outputs outside of phenology would help guide interpretation of the results impact and guide further work (as the authors hint at in the discussion on carbon and fire on page 5820). There has been a lot of work on land surface model benchmarking over the past few years (see e.g. Randerson et al. 2009; Lou et al. 2012; Kelley et al. 2013) which could serve as a good starting point for assessment. Randerson and Kelley both design metrics for assessing simulated vs observed seasonal signals, and then relate this to vegetation cover and/or productivity (Randerson by comparing LAI, Kelley by comparing fAPAR), and both demonstrate full model assessment of both chosen area for development and full-model impacts.

## This paper did indeed begin as a benchmarking study, however, the discovery of the strange behavior of the model phenology in drought deciduous areas quickly changed the focus of the paper. As stated above, much more work is needed in this area, but we think that a description of the current state of the model is an important contribution. Benchmarking software following the concepts in Kelley et al. 2013 and Luo et al. 2012 is in development for the CLM, and will be available soon. It should be noted, however, that many benchmarking systems used monthly data as a default output. This study highlights that a great deal of information can be lost via the use of automated assessments of monthly timeseries. There are other metrics that we might have used, but identification of onset and offset from the satellite timeseries is subject to issues surrounding the subjectivity of the criteria defining onset and offset thresholds. We considered that the direct metrics of RMSE and R2 over the whole timeseries were more appropriate measures, given the complexity of the multiple growing seasons simulated in the default model. Regarding the other benchmarking metrics we could have used, we agree that this should be addressed, and we’ve added the following paragraph to the last section of the Methods and a sentence emphasizing this point in the conclusions:

## The recent focus on land model benchmarking has led to a number of additional suggested methods for assessing seasonality in models compared to data (e.g. Randerson et al. 2009, Kelley et al. 2013), however, the proposed metrics would not capture the central issue addressed in this paper – model output with two or more peaks per year, data with only one –
as they begin with the unstated assumption that seasonality is unimodal over the course of a year, as do measures of the start and end of the growing season. In Randerson et al. (2009) seasonality is assessed by identifying the month of peak LAI and comparing that to MODIS LAI (MOD15A2), and in Kelley et al. (2014) several more complicated metrics are introduced (equations 7-9) to again produce single numbers to compare a model's seasonality to a benchmark data set. In these examples, as in other benchmarking studies, the focus is on producing a single number, which, while useful, can miss important details.

Specific comments:

Region assessment – pg 5807, line 19/20. It would be useful to provide information on how PFTs are prescribed: i.e what is the dataset, what period is the prescribed cover based on etc. ## Lawrence & Chase 2007 (referenced in the ms) describes the development of CLM’s land cover data set in detail.

- pg 5807, equation 1: Should ψ offset be ψ onset?

## Good catch! Looks like this error was introduced during typesetting per your later comment we’ve changed all of these to psi-threshold

- pg 5807, line 25: Why is the number of onset days prescribed as 30?

## That’s what is prescribed in CLM currently – re-written to be more clear

- pg 5808, line 4: Why is a timestep 108000s? If I’ve got my arithmetic right, 10800s is 30 hours or 1.25 days.

## This is a typo (thanks!) – should be 30 minutes, or 1800 seconds.

- Pg 5809, line 20: why are ψ offset and ψ onset defined as different parameters despite having the same value?

## Changed to be one symbol defined as ‘soil water potential threshold’

- Pg 5809, line 22: as with the number of onset days, why is the offset period proscribed as 15 days?

## As with onset, that’s what is prescribed in CLM currently – re-written to be more clear

- 5810, lines 4-9: A little more detail on climate information would be nice: What climate variables are needed to run the model? Which of these has the largest effect on soil water potential (and therefore the phenology indices)? Are soils prescribed? Does the model require CO2 inputs? Is the 45- year run using transient or equilibrium (detrended) climate? If detrended, how is this done? What time period does the run cover? What was the climate and vegetation cover (and soil and CO2?) inputs for the equilibrium baseline state run? How was equilibrium tested for in the baseline state?

## We have included more detail in this paragraph to address these questions. It now reads as follows:
The model runs used in the global simulations described here ran for 45 years, and were started from an equilibrium baseline state generated by a standard CLM spin up run, (as described in detail by Koven et al. 2013) cycling meteorological conditions of 1948-1972. The present-day run (1965 – 2010) used CRU-NCEP meteorological reanalysis data (N. Viovy, pers. comm.; data available at: http://dods.ipsl.jussieu.fr/igcmg/IGCM/BC/OOL/OL/CRU-NCEP/) and transient CO₂ concentrations to drive the model. Soil type and land cover are prescribed in the model, and recent work has suggested that the soil resistance parameterization may be unrealistic in arid ecosystems (Swenson & Lawrence 2014). More details on CLM are available in Oleson & Lawrence 2013.

- Pg 5810, lines 18-22: I’m not sure I follow the spatial averaging and resampling procedure. Were observed cells just averaged to decide if a CLM cell should be excluded? And was averaging performed just using LAI3g cells falling entirely within a CLM cell, with resampling used to incorporate the rest?

## We’ve revised the sentences to say:

## First, the two LAI3g maps generated for each month were averaged, then the LAI3g pixels were aggregated (averaged) to match the size of a CLM grid cell (~165 pixels per grid cell). If more than 80% of the grid cell did not have values in LAI3g (mostly applicable at high latitudes), the entire grid cell was removed from further analysis. Finally, the aggregated LAI3g data was resampled using a nearest neighbor approach to align with the CLM grid for further analysis.

- Pg 5812, line 13: Does the Latin-hypercube approach test for equilibrium itself, and/or did you define an equilibrium position was?

## We defined the equilibrium state, as clarified in the next paragraph (p. 5812, line 21)

- pg 5813, line 18: Again, why 10 days?

## We investigated alternative rainfall accumulation periods between 5 and 60 days. The impact of the accumulation periods was low, (since the model is mostly sensitive to the condition that there is any rain at all, rather than to the precise definition of the threshold).

- Pg 5812 (5814, I think?), lines 3-4: I’m not sure I understand this sentence. Are the three maps of LAI for CLM4.5BGC, LAI3g and CLM-MOD?

## We’ve modified this sentence to be more clear and it is addressed in the results.

- Pg 5815, line 9: Is there an important reason for running CLM globally here?

## Not particularly – we’ve eliminated the word ‘globally’.

- Pg 5816, line 8-9: Surely a model change such as this would require a new equilibrium baseline state? If you think it doesn’t, what is the rationale?

## We discuss the equilibrium conditions on P5812 L18-L22. Spin-up of the new model to an LAI equilibrium was remarkably fast in these simulations, potentially reflecting a lack of
significant feedback between the soil nitrogen state and vegetation physiology for the semi-arid regions.

- Pg 5816, line 9-12: can this “closer match” be quantified?

## We feel that showing the maps, including the difference maps in Figure 6, particularly 6D and 6F adequately illustrate this improvement.

- Pg 5816, line14-19: Again, can this be quantified? In figure 7, it looks like some places get better (sahel, southern and western Australia etc) whilst some places get worse (i.e, Asia, north east Aus etc).

## We’ve added a confusion matrix table comparing counts of peak numbers between LAI3g and the two model runs (CLM and CLM-MOD) and this paragraph now reads as follows:

## To test whether the poor fit between CLM and LAI3g was due to multiple annual LAI peaks in CLM we counted the number of peaks per year in each data set (Fig 7). We found that in the observations, only areas in the humid tropics had multiple peaks in the LAI3g data (“peaks” in these cases being relatively small fluctuations), while CLM showed multiple peaks per year throughout many of the savanna regions of the world. CLM-MOD has more areas with only one peak, particularly in Sub-Saharan Africa. To quantify these changes to the model we constructed confusion matrices to compare the peak counts in LAI3g to those in CLM and CLM-MOD (Table 3) for grid cells with >50% drought deciduous cover (Figure 1). Overall, CLM-MOD had a slightly poorer performance, matching the number of peaks in the LAI3g dataset 42.5% of the time, while CLM matched LAI3g 43.7% of the time. However, these unweighted summary numbers mask improvements in CLM-MOD. CLM only correctly predicted a single peak 8.9% of the time, while CLM-MOD correctly predicted single peaks 59% of the time, and never did CLM-MOD predict more than two peaks in a year, matching the LAI3g data. The overall degradation in CLM-MOD is due to fewer correctly identified grid cells with zero or two peaks.

- Pg 5816, line 24 - pg 5817, line 2: Again, the improvement does not seem to be quantified. Introduction of a more systematic benchmark system (see major comments) would help with these last 3 points)

## Here, and in the above 2 comments, it would be possible to determine a single number for the change in model performance, but, as can be discerned from the global maps, the situation is actually spatially complex, and, as highlighted by Luo et al. and Randerson et al., the generation of single number benchmarking products is fraught with subjectivity. This is why we chose to present the results in their entirety, as part of the development of process representations.

- Pg 5818, line 23-pg 5819 line 20: This part seems very out of place in the discussion. The implementation of this should be in the methods, and it would be nice to see some results, even if they go in an SI. If you cannot do this, I would take this section out.

## We feel that it is important to keep this discussion in the paper as these alternative options were frequently brought up when this work was discussed among CLM users, but we’ve moved it to the methods (section 2.4).
References


Environmental drivers of drought deciduous phenology in the Community Land Model (MARKUP VERSION)

K.M. Dahlin¹,², R.A. Fisher², & P.J. Lawrence²

¹ Department of Geography, Michigan State University, East Lansing, Michigan, USA
² Climate and Global Dynamics Division, National Center for Atmospheric Research*, Boulder, Colorado, USA

* NCAR is sponsored by the National Science Foundation.

Correspondence to: K.M. Dahlin (kdahlin@msu.edu)

Abstract
Seasonal changes in plant leaf area have a substantial impact on global climate. If and when leaves are present affects surface roughness and albedo, and the gas exchange occurring between leaves and the atmosphere affects carbon dioxide concentrations and the global water system. Thus, correct predictions of plant phenological processes are important for understanding the present and future states of the Earth system. Here we compare plant phenology as estimated in the Community Land Model (CLM) to that derived from satellites in drought deciduous regions of the world. We reveal a subtle but important issue in the CLM: anomalous green-up during the dry season in many semi-arid parts of the world owing to rapid upwards water movement from wet to dry soil layers. We develop and implement a solution for this problem by introducing an additional criterion of minimum cumulative rainfall to the leaf-out trigger in the drought deciduous algorithm. We discuss some of the broader ecological impacts of this change and highlight some of the further steps that need to be taken to fully incorporate this change into the CLM framework.
1 Introduction

Ecosystems change with the seasons in response to environmental cues. Some of those cues are fixed, like day length, while others are climate-driven and therefore vary from year to year. The combination of fixed and climate-driven phenological cues poses an interesting problem in the face of climate change—climate related drivers of phenology (temperature and rainfall patterns) are likely to change (Lau et al., 2013), while fixed cues will remain unchanged. Phenological shifts due to climate change have already been identified (e.g. Parmesan & Yohe, 2003). "Phenology" can refer to a large number of patterns and behaviors in plants and animals that shift with the seasons. Here, however, because we are focused on land surface model simulations, we use phenology specifically to refer to intraannual variation in leaf area index (LAI). Leaf area can vary significantly within a year and is, therefore, a critical control on land-atmosphere feedbacks (Lawrence et al., 2012).

Recent advances have greatly improved our ability to predict seasonal patterns in northern temperate deciduous forests (Richardson et al., 2012), but our understanding of phenological patterns in stress or drought deciduous plants (also called ‘raingreen’) remains weak (Guan et al., 2014; Jenerette et al., 2010; Ma et al., 2013). The semi-arid ecosystems that host the majority of drought deciduous woody plants have relatively low biomass but make up a large fraction of global land area (~30%; Scholes & Hall, 1996). Their extensiveness alone makes them important to global radiation budgets, but additionally these systems are likely very sensitive to climate change given their apparent bistability (Scholes & Hall, 1996; Staver et al., 2011). In semi-arid ecosystems leaf-out is typically thought to be a function of water availability (Reich, 1995; White et al., 1997), however, some woody plants do leaf-out several weeks before the first rains of the season (Archibald and Scholes, 2007).
In an Earth system modeling context, the timing and magnitude of plant phenology, and how these processes may change, is critical for approximating the energy and carbon balances of the planet. Prognostic phenology has only recently been incorporated into Earth system models, however, and its fidelity, particularly in semi-arid regions, remains poorly tested (Blyth et al., 2011; Lawrence et al., 2011; Randerson et al., 2009). Lawrence et al. (2012) found that the prognostic phenology in the Community Land Model version 4 (CLM4(CN)) degraded estimates of latent heat flux and other biophysical properties in comparison to using prescribed, satellite-derived phenology (CLM4SP). Wang et al. (2013) compared intraannual variation in the fraction of absorbed photosynthetically active radiation (fAPAR) in CLM4CN to satellite-derived estimates and found substantial differences in regional averages, zonal means, and interannual trends. It is difficult, however, to isolate the impact of the drought deciduous phenology algorithm using these regional and zonal estimates.

Satellite-derived estimates of greenness, fAPAR, and LAI have greatly improved our ability to study the environmental drivers of phenology (Reed et al., 2009), however, the majority of studies have focused on northern deciduous and boreal forests (e.g. Delbart et al., 2006; White et al., 2009; Yang et al., 2012). While fewer studies have focused on remote sensing of phenology in semi-arid systems, Zhang et al. (2005) found a strong relationship between greenness onset and the start of the rainy season across the semi-arid parts of Africa. They found a weaker relationship, however, between dormancy and the end of rainy seasons, and they attribute this weakness to differences in soil properties. Similarly, Ma et al. (2013) found a strong relationship between greenness and rainfall in northern Australia in both seasonal timing and amplitude and Bradley et al. (2011) found a close relationship between rainfall and seasonality in Amazonian savannas. Interestingly, in Africa Zhang et al. (2005) also showed a strong relationship between latitude and both green-up and dormancy onset, even in the narrow band of the Sahelian and Sub-Saharan region, suggesting a possible link.
between phenology and subtle changes in photoperiod at least in northern Africa. Recently Guan et al. (2014) showed a relationship between woody plant cover and phenological timing in African savannas.

In this study we address three questions related to the representation of drought deciduous phenology in the CLM. (1) How well does the CLM capture phenological patterns of LAI among different drought deciduous plant functional types (PFTs) as compared to satellite-derived estimates?; (2) Which parameters in the current version of the CLM have the most leverage on woody plant drought deciduous phenology?; and (3) Do changes in the phenology algorithms in the CLM improve the model’s representation of seasonal cycles regionally?

2 Methods

2.1 Model Description

The CLM is the terrestrial component of the Community Earth System Model (CESM; Lawrence et al., 2011); it simulates biogeophysical and biogeochemical processes including radiation interactions with vegetation and soil, heat transfer in soil and snow, hydrology, and plant photosynthesis and respiration. In this paper we use the most recent release of the Community Land Model with active biogeochemistry, CLM4.5BGC (Oleson et al., 2013). Henceforth, references in this paper to the “CLM” will refer to CLM4.5BGC.

The CLM is run here on a 1.25° x 0.9375° grid, and each grid cell is, where applicable, divided into fractions representing vegetated land, lakes, glaciers, and urban areas. Within the vegetated fraction of a grid cell there may be multiple PFTs representing a coarse division of biodiversity along its major axes of variation: trees/shrubs/grass, broadleaf/needleleaf, C3/C4 photosynthesis mechanisms and phenological habit (evergreen, cold deciduous and stress/drought
deciduous). There are currently 15 non-crop PFTs in the CLM, four of which follow the drought deciduous phenology algorithm (Oleson et al., 2013). Figure 1 shows where these different PFTs dominate the globe. Over time the relative cover of the PFTs may shift, as may the overall fraction of vegetation, depending on shifts in land use, though these shifts are minor in recent decades. In the simulations used in this paper these shifts in PFT fractions and cover are prescribed from satellite observations (Lawrence and Chase, 2007) as opposed to emerging from vegetation competition (Bonan et al., 2003).

In the CLM, drought deciduous plants are represented by the ‘stress deciduous’ phenology type, as distinct from the evergreen or ‘seasonal’ (cold) deciduous phenology types. This designation allows for plants to lose their leaves either via the impact of cold, via the impact of drought, or via the onset of short days thus allowing the model to simulate, for example, grass vegetation growing in an environment that is both seasonally cold and seasonally dry. If the triggers for offset are not reached in a given year, drought deciduous vegetation will follow the evergreen phenology algorithm, gaining and losing fixed fractions of carbon with each time step. This stress deciduous algorithm, described in more detail below and in Oleson et al. (2013), was developed in part from White et al. (1997), though that study was particularly focused on grass phenology.

The deciduous algorithms are hierarchical, such that plants classified as ‘stress deciduous’ but growing at high latitudes or in cold climates will follow the same onset/offset rules as cold/seasonally deciduous plants. From the beginning of a dormant period a ‘freezing day accumulator’ is activated whereby time steps with a temperature below freezing (0°C) are summed and if this sum exceeds 15 days then the plants will follow both the winter deciduous and drought deciduous algorithms. Leaf onset can only be triggered if day length is greater than 6 hours, a latitude-specific sum of growing degree days has been reached (described in Oleson et al. (2013)) and the soil wetness criteria described below have been met.
In seasonally dry, warm regions (the focus of this paper) where day length is never less than 6 hours, leaf onset for the stress deciduous phenology type is determined by soil wetness. At the end of the previous offset period an accumulated soil water index (SWI) is set to zero and accumulation is calculated as:

\[
SWI^n = \begin{cases} 
SWI^{n-1} + f_{day} & \text{for } \psi_{soil3} \geq \psi_{onset\text{threshold}} \\
SWI^{n-1} & \text{for } \psi_{soil3} < \psi_{onset\text{threshold}}
\end{cases}
\]  

(1)

Where \( n \) and \( n-1 \) refer to the values in the previous and current time steps, \( \psi_{soil3} \) is the soil water potential (MPa) in the third soil layer (6.23 cm – 9.06 cm), \( \psi_{onset\text{threshold}} \) is -2 MPa, and \( f_{day} \) is a time step (30 minutes in CLM) as a fraction of a day. Onset is triggered when SWI exceeds 15 days.

The rate of leaf onset (fraction of onset per time step), which in the CLM is represented as the transfer of C and N from a storage pool to the ‘display’ leaf pool, is determined by the number of days prescribed for onset, here fixed at 30 days. The rate (\( r_{onset} \)) at each time step is defined as:

\[
r_{onset} = \begin{cases} 
\frac{2}{t_{onset}} & \text{for } t_{onset} \neq \Delta t \\
\frac{1}{\Delta t} & \text{for } t_{onset} = \Delta t
\end{cases}
\]  

(2)

where \( t_{onset} \) is time remaining in the current onset period in seconds and \( \Delta t \) is the length of a time step (108000 seconds = 1800 seconds). The flux of C out of the storage pool is then defined as the amount in the C storage pool at that time step multiplied by \( r_{onset} \). These functions result in a linearly decreasing flux out of the transfer pool, so the rate of increase in LAI over the onset period steadily decreases as C moves from the storage pool to the display pool (see Fig 14.1 in Oleson et al. (2013)).

During the onset period C and N are also transferred from storage pools for fine roots, live and dead stem, and live and dead coarse roots into these components’ respective displayed growth pools. During the growing season, C and N taken up by the plant are accumulated in transfer pools, to be used in the next growing season.
As long as the leaf onset period is complete, leaf offset can be triggered by short (<6 hr) day length, a period of cold temperatures has been reached (described in Oleson et al. (2013)) or if the soil dryness criteria described below has been met.

The offset soil wetness index (OSWI) can potentially start accumulating time steps once the previous leaf onset phase is complete. The algorithm differs slightly from the onset trigger, in that OSWI can increase or decrease as described below.

\[
OSWI^n = \begin{cases} 
    OSWI^{n-1} + f_{\text{day}} & \text{for } \psi_{\text{soil}}^3 \leq \psi_{\text{threshold offset}} \\
    \max(OSWI^{n-1} - f_{\text{day}}, 0) & \text{for } \psi_{\text{soil}}^3 > \psi_{\text{threshold offset}}
\end{cases}
\]  

(3)

where \( \psi_{\text{threshold offset}} \) is -2 MPa, and leaf offset is triggered when OSWI equals 15 days.

Similar to the rate of leaf onset, leaf offset rate is a function of the amount of time left in the offset period, here-fixed at 15 days:

\[
 r_{\text{offset}} = \frac{2\Delta t}{t_{\text{offset}}^2}
\]  

(4)

Carbon fluxes into the litter pool are only calculated for leaves and fine roots (stems and coarse roots cannot shrink). Nitrogen fluxes into the litter pool reflect retranslocation of N prior to offset. See Oleson et al. (2013)) for more details.

The model runs used in the global simulations described here ran for 45 years, and were started from an equilibrium baseline state generated by a standard CLM4.5BGC spin up run, (as described in detail by Koven et al., 2013), cycling meteorological conditions of 1948-1972. The present-day run (1965 – 2010) used as documented in (Oleson et al., 2013) and using CRU-NCEP meteorological reanalysis data (N. Vivoy, pers. comm.; data available at: http://dods.ipsl.jussieu.fr/igemg/IGCM/BC/OOL/OL/CRU-NCEP/) and transient CO\(_2\) concentrations to drive the model. Soil type and land cover are prescribed in the model, and recent
work has suggested that the soil resistance parameterization may be unrealistic in arid ecosystems (Swenson and Lawrence, 2014). More details on CLM are available in Oleson et al. (2013).

2.2 Satellite derived LAI

We compared the model-derived estimates of LAI to those estimated from the Advanced Very High Resolution Radiometer sensors (AVHRR) onboard the National Oceanic and Atmospheric Administration satellites. These data are available bimonthly and span from 1981 to 2011. They are supplied at 1/12 degree resolution. A detailed description of the development of the LAI product (hereafter LAI3g) is in Zhu et al. (2013).

To ensure the most appropriate comparison possible, the LAI3g dataset was rescaled to match the mean monthly LAI output from the CLM. First, the two LAI3g maps generated for each month were averaged, then the LAI3g pixels were aggregated (averaged) to match the size of a CLM grid cell (~165 pixels per grid cell). If more than 80% of the grid cell did not have values in LAI3g (mostly applicable at high latitudes), the entire grid cell was removed from further analysis. Finally, the aggregated LAI3g data was resampled using a nearest neighbor approach to align with the CLM grid for further analysis.

All spatial and statistical analyses were performed in R (R Core Team, 2013) using the ncdf (Pierce, 2011), raster (Hijmans and van Etten, 2013) and rgdal (Bivand et al., 2013) packages.

2.3 Comparing LAI3g to CLM LAI
We first compared LAI3g to the CLM output for 1982 (the first full year of LAI3g) to 2010 (the last available year of the CRU-NCEP forcing dataset for CLM), aggregating values by zones based on dominant PFT and hemisphere. To aggregate into PFT classes, we only considered grid cells dominated by a single drought deciduous PFT (>50% cover), permitting six possible comparisons with a sufficient number of grid cells for comparison: northern hemisphere (NH) temperate C3 grasses \((n = 180)\), NH C4 grasses \((n = 234)\), tropical deciduous trees \((n = 242)\), southern hemisphere (SH) deciduous shrubs \((n = 68)\), SH C3 grasses \((n = 160)\), and SH C4 grasses \((n = 271)\). Note that the actual number of grid cells compared from year to year varied slightly with changes in land cover and, in the case of LAI3g, available data. The counts listed here are the averages for each PFT. To visually assess the comparison between LAI3g and CLM we plotted the monthly means and standard deviations for these seven regions. We also computed the \(R^2\) and root mean squared error (RMSE) across all 29 years’ monthly values to assess CLM’s ability to fit the seasonality and the magnitude of the LAI3g values. Due to the temporal coarseness of these data and the irregular seasonal patterns found in many of our areas of interest, we did not fit a continuous function to these data.

2.4 Point Simulations and Parameter Sensitivity Tests

Given the observed mismatches between LAI3g seasonality and CLM predictions among drought deciduous woody plants (see Discussion), we conducted an analysis of how the parameters determining phenology in the model affect the model outcome. The use of global or regional simulations to assess the sensitivity of models to their structural and parametric assumptions is problematic on account of both the computational requirements to do so and the high-dimensionality of the model outputs, which can hinder understanding. To avoid these issues, we conducted a sensitivity test to the major model parameters that control seasonality in LAI at the point scale. In these point simulations we focused on low-latitude drought-deciduous ecosystems,
and selected six locations dominated by tropical deciduous trees, grasses, or a combination of the
two (Table 1).

The phenology model contains three empirical parameters that collectively describe the leaf
onset and offset algorithms.

1. Critical soil water potential ($\Psi_{\text{onset/offset threshold}}$) values for leaf onset and offset (default = -2
MPa)

2. Soil water potential days to onset/offset (SWI/OSWI) threshold (default = 15 days)

3. Quantity of carbon assimilation which is directed to current leaf growth rather than
storage ($F_{\text{current}}$) (default = 0.0)

To determine what impact the choice of these parameters has on the model outcome, we
conducted a Latin-hypercube analysis (McKay et al., 1979), beginning the model from a spun-up
state with default parameters. We then perturbed the parameters and ran the model forwards until a
new LAI equilibrium condition was detected. Because the nitrogen cycle is active in CLM, soil
biogeochemical equilibrium can in some circumstances (particularly at high latitudes) take many
decades or even centuries achieve, therefore we set a threshold for the new equilibrium state as the
absence of a trend in LAI resulting in a 2% increase over a five year period. Given the high
temperature and relatively low biomass and productivity of the ecosystems in question, in our
simulations LAI equilibrium was in practice typically reached after the first 5 year period of the
simulation. As with the global simulations, we used the CRU-NCEP reanalysis forcing data,
extracted at the six points of interest, to drive the point simulations.

The Latin hypercube methodology is a particularly efficient means of investigating a multi-
dimensional parameter space, because each run of the model perturbs every parameter, and the
algorithm ensures that the distribution of sampled points is distributed efficiently (but not
uniformly) through parameter space. For this study, since we were comparing the model output to monthly satellite observations, we did not manipulate the number of days for onset (30) or offset (15) to occur. We also did not consider the growing degree days and day length parameters because the focus of this study was on tropical and subtropical regions where these components of the algorithm are not active. To determine what impact the choice of these parameters has on the model outcome, we conducted a Latin-hypercube analysis (McKay et al., 1979), beginning the model from a spun-up state with default parameters. We then perturbed the parameters and ran the model forwards until a new LAI equilibrium condition was detected. The Latin-hypercube methodology is a particularly efficient means of investigating a multi-dimensional parameter space, because each run of the model perturbs every parameter, and the algorithm ensures that the distribution of sampled points is distributed efficiently (but not uniformly) through parameter space.

Because the nitrogen cycle is active in CLM, soil biogeochemical equilibrium can in some circumstances (particularly at high latitudes) take many decades or even centuries achieve, therefore we set a threshold for the new equilibrium state as the absence of a trend in LAI resulting in a 2% increase over a five year period. Given the high temperature and relatively low biomass and productivity of the ecosystems in question, in our simulations LAI equilibrium was in practice typically reached after the first 5 year period of the simulation. As with the global simulations, we used the CRU-NCEP reanalysis forcing data, extracted at the six points of interest, to drive the point simulations.

For this study, since we were comparing the model output to monthly satellite observations, we did not manipulate the number of days for onset (30) or offset (15) to occur. We also did not consider the growing degree days and day length parameters because the focus of this study was on tropical and subtropical regions where these components of the algorithm are not active.
For the critical soil moisture potential threshold, $\Psi_{\text{critical/threshold}}$, we investigated values from 0 to -3.5MPa (where the default value is -2 MPa). The upper end of this range is the maximum possible value for saturated soils, whereas the bottom end of the range was determined from a prior set of sensitivity tests which determined that sensitivity below this range was very low (i.e. the soil moisture potential in the third layer rarely drops below -3.5 MPa, and so leaves remain on continuously for those simulations at our locations of interest). For the number of days of onset, we followed a similar protocol, and found that the range of sensitivity was focused between 5 and 35 days (where the default is 15 days). For the fraction of displayed assimilated carbon ($F_{\text{current}}$) we varied the values between 0 and 0.5, (where the default is zero). Sensitivity of average LAI to this parameter was low in all cases, but it has an impact on the intraannual cycle, since the LAI is unchanging through a single growing season if $F_{\text{current}} = 0$.

To assess the performance of the different models in the Latin hypercube test we originally plotted the coefficients of determination between the different models’ LAI values and the LAI3g data at those points. However, this result did not illustrate any clear optimum in model performance either for the parameters of the existing model, nor for the rainfall threshold. We illustrate this using the time-series data in Fig. 4, which highlight the unusual behavior of the model and to assess whether the extra green-up period during the dry season had been eliminated in any of the parametric permutations. We ascribe the lack of a clear parametric signal to two effects. First, the LAI3g data were necessarily aggregated to monthly values, meaning that the primarily sub-monthly variation between ensemble members was masked. Second, the timing of the secondary leaf-on period in the dry season was the emergent property of the oscillatory (and thus somewhat chaotic) dynamics of the soil-vegetation feedback on soil moisture.
Once we determined that we could not eliminate the dry season green-up by changing the existing model parameters, we considered four possible additions to the model. The first three are described here but, for brevity, are not quantified in the results. First, we considered that using the third soil layer in CLM may be an arbitrary choice of soil depth, and that usage of the soil moisture potential derived drought index (‘BTRAN’, Oleson et al., 2013), which is weighted by vertical root fraction across the whole rooting depth profile, might provide a more physiologically relevant metric and be less prone to increases due to upwards moisture diffusion in the dry season. However, since the exponential root profile in the CLM weights the top soil layers (including layer 3) more strongly than the lower layers with fewer roots, this metric was just as prone to increasing water potential during the dry season as soil water potential in the third soil layer.

Second, we implemented leaf onset as function of a total column soil moisture content threshold rather than soil moisture potential. We postulated that the redistribution of water causes the erroneous behavior and that this would not impact total column moisture. However, the establishment a single global threshold for total soil moisture is challenging, as a number of different variables impact soil moisture, including the variation in soil water retention capacities between different land points, and by the interaction between leaf area, evaporation rate and deep soil moisture content. Variation in rainfall and evaporation rates affects the equilibrium water content of deep soils, which changes the total column soil moisture content between locations and years, but not the physiologically relevant upper soil moisture potential. Therefore, we abandoned this metric of phenological trigger.

Third, we considered a metric of triggering leaf flush by the rate of change of total column soil moisture, rather than soil moisture potential. However, this methodology also generates erroneous behavior, on account of the ability of the CLM hydrology model to extract water from the water table or aquifer along a water potential gradient. Thus, when water potential is low in the
bottom soil layer in the dry season, the rate of change of total soil moisture can be positive without any input from rainfall.

2.5 Rainfall Model

To correct biases uncovered in the model output (described below) we introduced a simple trigger into the model, that time-averaged 10-day precipitation must exceed a given threshold before leaf onset is triggered. This approach requires the addition of a new parameter, rain_threshold, into the model, which is the threshold over which 10-day precipitation must be for leaf-on to occur. Leaf onset is thus triggered if 10 day rain is higher than rain_threshold and if the SWI is greater than 15 days $\Psi_{\text{max}}$.

We then used a Latin hypercube approach again to determine the sensitivity of the model to rain_threshold at our six chosen geographical points. We considered a range of rainfall rates, requiring that it rain 20 mm over the course of 5 to 60 days in order for plants to begin growing leaves. Requiring that it rain 20 mm over the course of 10 days in order for plants to begin growing leaves best fit the LAI3g data in our six points. To test the global impact of these parameter changes we ran CLM4.5BGC with the new rainfall-based trigger and compared the results both at several points and globally.

2.6 Global Simulations

We used a number of different metrics to globally compare CLM4.5BGC to the LAI3g data and, later, to the modified version of the model (CLM-MOD). First we compared maps of maximum annual LAI and differences between the three maps. We also developed an algorithm to count the number of LAI peaks per year in all three data sets on grid cells with a range greater than one, by counting the number of times per year that the difference between one month’s LAI and the next was negative, then taking the mode across all 29 years. Finally, we calculated the coefficient of
determination ($R^2$) in each grid cell, comparing the monthly LAI3g data to CLM4.5BGC and CLM-MOD to identify areas with strong agreement between the remotely sensed data and the models, and areas with weak relationships.

The recent focus on land model benchmarking has led to a number of additional suggested methods for assessing seasonality in models compared to data (e.g. Randerson et al., 2009, Kelley et al., 2013), however, none of the proposed metrics would capture the central issue addressed in this paper – model output with two or more peaks per year, data with only one – as they begin with the unstated assumption that seasonality is unimodal over the course of a year, as do measures of the start and end of the growing season. In Randerson et al (2009) seasonality is assessed by identifying the month of peak LAI and comparing that to MODIS LAI (MOD15A2), and in Kelley et al (2014) several more complicated metrics are introduced (equations 7-9) to again produce single numbers to compare a model’s seasonality to a benchmark data set. In these examples, as in other benchmarking studies, the focus is on producing a single number, which, while useful, can miss important details.

3. Results

3.1 Seasonal Patterns in CLM

We found generally good agreement between LAI3g and CLM averaged across grass-dominated regions. In a comparison of monthly values from 1982 to 2010 for the single PFT dominated regions in Fig. 1, $R^2$ values ranged from 0.54 to 0.9 (Table 2) with the majority of the grass $R^2$ greater than 0.7. Figure 2 shows the monthly values across all years, and we see similar results – generally good correspondence, especially in seasonal pattern, between LAI3g and the CLM runs in the grass-dominated regions. The root mean squared error (RMSE) values in Table 2, as
well as Fig. 2B and C show that CLM does not always capture the appropriate LAI values in grasslands, but the seasonal cycle is reasonably correct.

In contrast, CLM does not successfully capture phenological patterns or values in areas dominated by woody drought deciduous vegetation. Among tropical deciduous trees CLM predicted LAI appears to be both too high and out of phase with the satellite observations (Fig. 2E) while CLM shows no apparent seasonality among deciduous shrubs in the southern hemisphere (Fig. 2F), while LAI3g shows a slight cycle ranging from 0.4 to 0.7 LAI.

3.2 Point simulations & sensitivity tests

To look more closely at seasonal patterns in drought deciduous locations we selected six points around the globe across a range of latitudes dominated by a mixture of broadleaf deciduous tropical trees, C3 and C4 grasses (Table 1) (all of which use the same stress deciduous phenology algorithm). To better understand the phenological patterns, we re-ran CLM globally using the same methods as described above but recording daily outputs of relevant parameters (including LAI, soil water potential, rainfall, and others). Plots of the seasonal cycles at these specific points using daily model output (solid green lines in Fig. 3) revealed an apparent pattern whereby CLM appears to put leaves on during the “brown season” in the LAI3g data in some of the points in addition to during the LAI3g green season. We note, however, that some areas in reality do have two separate growing seasons per year (e.g. Fig. 3E). Despite the lack of rainfall, soil water potential in the third soil layer in CLM rises during the dry season and is extremely variable in the dry season, on account of periods of high transpiration when plants leaf out (blue dot-dashed line in Fig. 3).

We used the output from the Latin hypercube approach at these six points to vary the parameters of interest (days to onset/offset, critical soil water potential, carbon assimilation) to assess whether modification of parameter values could ameliorate the problem of plants leafing out
during the dry season in CLM. We found, however, that simply varying the parameters of the existing model within the parameter space investigated (and assuming no large non-linearities in the model response surface) did not remove the dry season leaf out in the model (Fig. 4).

In order to address this issue, we considered a number of structural perturbations to the leaf-on and leaf-off algorithms (described in the discussion below), but ultimately decided on adding a new parameter, rain_threshold, to the model. We then used the same Latin hypercube approach to determine the best fitting values for this parameter (Fig. 5). This additional leaf-on criterion, set so that 20 mm of rain must accumulate over 10 days in order for leaf onset to occur, led to a removal of the “brown season” leaf out in CLM (dashed green line in Fig. 3) without preventing two green seasons per year, as is possible in some semi-arid regions (e.g. parts of Ethiopia, Fig. 3E). While this new rainfall threshold improved model performance both at our points and globally (see below), we note that the model did not appear to be particularly sensitive to the amount of rain that fell, as long as some rain did fall, but this threshold, and the drought deciduous algorithm as a whole, deserves more research into seasonal drivers.

3.3 Global simulations

To test how well the additional rainfall parameter performed globally, we ran CLM with the new rainfall parameter for 45 years (CLM-MOD) from the same equilibrium baseline state as was used in the first run described. Measures of maximum LAI (Fig. 6) and mean LAI (data not shown) in CLM-MOD showed closer matches to LAI3g than CLM4.5BGC. While CLM values remain far too high in the evergreen tropics, the maximum LAI values in deciduous savanna regions did increase appropriately in CLM-MOD to better match the LAI3g data in deciduous savanna regions.

To test whether the poor fit between CLM and LAI3g was due to multiple annual LAI peaks in CLM we counted the number of peaks per year in each data set (Fig. 7). We found that in the
observations, only areas in the humid tropics had multiple peaks in the LAI3g data (‘peaks’ in these cases being relatively small fluctuations), while CLM showed multiple peaks per year throughout many of the savanna regions of the world. CLM-MOD has more areas with only one peak, particularly in Sub-Saharan Africa. To quantify these changes to the model we constructed confusion matrices to compare the peak counts in LAI3g to those in CLM and CLM-MOD (Table 3) for grid cells with >50% drought deciduous cover (Fig. 1). Overall, CLM-MOD had a slightly poorer performance, matching the number of peaks in the LAI3g dataset 42.5% of the time, while CLM matched LAI3g 43.7% of the time. However, these unweighted summary numbers mask improvements in CLM-MOD. CLM only correctly predicted a single peak 8.9% of the time, while CLM-MOD correctly predicted single peaks 59% of the time, and never did CLM-MOD predict more than two peaks in a year, matching the LAI3g data. The overall degradation in CLM-MOD is due to fewer correctly identified grid cells with zero or two peaks.

We compared monthly data and mapped the point-wise coefficients of determination (R^2) globally to consider how well CLM LAI seasonality matched the LAI3g dataset (Fig. 8A). There were moderate to good relationships (R^2 >0.4) in the higher latitudes for the standard model, but notably poorer relationships in the lower latitudes, particularly in savanna regions. In contrast, a comparison between LAI3g and CLM-MOD showed improvements in savanna regions, with the most dramatic improvements in Sub-Saharan Africa (Fig. 8B). Predicting the phenology of the Brazilian Cerrado continues to be a challenge in CLM-MOD, shifting from two peaks in CLM4.5BGC to no peaks in CLM-MOD, but the heterogeneity revealed in the LAI3g dataset suggests that this region may need closer consideration and a separate phenology algorithm.

4. Discussion
4.1 Comparing LAI3g to CLM LAI

By comparing the satellite LAI3g data to output from CLM4.5BGC we found that while the model performed reasonably well in temperate grasslands both in seasonal cycle and in magnitude, it performed poorly in areas of tropical grasslands, mixed grassland and drought deciduous trees (savannas) and areas dominated by drought deciduous trees. Closer examination of individual grid cells in tropical regions revealed that these points often have a leaf flush during the dry season in the model, which is not the case in the satellite data, or in reality. This additional leaf flush not only impacts the phenological cycle, but also affects the overall amount of carbon stored in plants and their maximum LAI, as plants spend their stored carbon unnecessarily in the dry season, leaving less carbon available during the wet season for growing leaves. This addition of leaf carbon in the dry season also may affect the fire cycle in varying ways around the dry tropics. While these runs of the model were not coupled to a dynamic atmosphere, we expect that this dry season leaf flush could also impact the climate, potentially having an unrealistic cooling effect by moving more water in to the atmosphere during what should be a very dry time of year, but also darkening the land surface, possibly leading to a slight warming.

The mechanism behind the dry season leaf flush is an increase in soil water potential in the dry season to levels above the prescribed leaf-out threshold. These increases derive from the assumption that all of the land surface in CLM sits on top of an unconfined aquifer. In most cases this aquifer is either irrelevant because plenty of soil water is available or it is essential to plant survival in areas where aquifers do exist in the real world. In semi-arid systems, however, this extra pool of soil water becomes problematic in the dry season. The top soil layers dry out due to soil evaporation and, when plants are active, evapotranspiration, establishing a water potential gradient, which causes water to be transferred by mass flow from the aquifer up through the soil column to the shallow soil layers, until eventually the moisture potential reaches the trigger for plants to leaf
out. Per the drought deciduous phenology algorithm, once leaf out is triggered it must be completed, so plants begin to grow leaves but then the increased evapotranspiration rate quickly draws the soil moisture down below leaf off threshold levels, so leaf drop begins again, typically as soon as the leaf out period (30 days) has ended. The degree to which aquifers in reality contribute to dry season evapotranspiration is largely unconstrained because there are no global data sets for depth to water table, making it impossible to non-arbitrarily define where plants should have access to ground water and where they should not. Refinements of the soil water algorithms in CLM and access to new data sources like the NASA Soil Moisture Active Passive mission (SMAP; Entekhabi et al., 2014) will likely improve this part of the model, but like many aspects of the CLM, more global-scale data is needed.

4.2 Soil water and rainfall in CLM

To address the erroneous dry season leaf flush we tested a number of different model alterations, beginning with the least invasive – adjusting existing parameters – and ending with adding an additional rule to the existing drought deciduousness algorithm. We experimented with four alternative methodologies for triggering leaf onset, described in the methods (section 2.4) described below, but for brevity we have only shown results from the last and most effective approach. First, we considered that using the third soil layer in CLM may be an arbitrary choice of soil depth, and that usage of the soil moisture potential derived drought index (‘BTRAN’, (Oleson et al., 2013)), which is weighted by vertical root fraction across the whole rooting depth profile, might provide a more physiologically relevant metric and be less prone to increases due to upwards moisture diffusion in the dry season. However, since the exponential root profile in the CLM weights the top soil layers (including layer 3) more strongly than the lower layers with fewer roots,
this metric was just as prone to increasing water potential during the dry season as soil water
potential in the third soil layer.

Second, we implemented leaf onset as function of a total column soil moisture content
threshold rather than soil moisture potential. We postulated that the redistribution of water causes
the erroneous behavior and that this would not impact total column moisture. However, the
establishment a single global threshold for total soil moisture is challenging, as a number of different
variables impact soil moisture, including the variation in soil water retention capacities between
different land points, and by the interaction between leaf area, evaporation rate and deep soil
moisture content. Variation in rainfall and evaporation rates affects the equilibrium water content of
deep soils, which changes the total column soil moisture content between locations and years, but
not the physiologically relevant upper soil moisture potential. Therefore, we abandoned this metric
of phenological trigger.

Third, we considered a metric of triggering leaf flush by the rate of change of total column
soil moisture, rather than soil moisture potential. However, this methodology also generates
erroneous behavior, on account of the ability of the CLM hydrology model to extract water from the
water table or aquifer along a water potential gradient. Thus, when water potential is low in the
bottom soil layer in the dry season, the rate of change of total soil moisture can be positive without
any input from rainfall.

These hydrological issues in CLM are complex, and derive from the need to operate an
internally consistent global model of the water cycle in the absence of critical data at the appropriate
scale (depth to water table, the unsaturated hydraulic conductivity of deep soils, etc). In an ideal
case, improvements in hydrology might allow the existing phenological model to operate correctly.

However, here we took a more pragmatic approach and

so for a fourth method we partially
decoupled the soil hydrology and the phenology models, allowing rainfall inputs to directly impact
on leaf phenology without interacting with the assumptions of the soil-hydrology model. Leaving the
condition of soil water potential in the third soil layer in place, we then added an additional
condition which was that the rainfall accumulated over the last 10 days should be higher than a
threshold value (20 mm). Thus, if soil moisture rose above the threshold level, but little or no rain
had fallen, plants would not put on leaves. The new model performs better both for the point
simulations and in global simulations, both in terms of the seasonal cycle of LAI, where the average
point-wise coefficient of determination ($R^2$) between modelled and observed monthly satellite LAI
of drought deciduous dominated points (>50% drought deciduous cover) is significantly higher for
the new model (0.31 vs 0.13). While there was no substantial change in the overall peak count
accuracy (Table 3), CLM-MOD had zero drought deciduous dominated grid cells with >2 peaks and
a substantial improvement in the identification of single-peak grid cells (8.9 to 59%). The added
rainfall trigger did, however, reduce the number of zero peak and two peak grid cells correctly
identified. This result highlights the need for more research into the diversity of drought deciduous
phenology drivers around the world.

4.3 Impacts of modifications to the model

This relatively small change to the drought deciduous phenology algorithm had wide ranging
impacts within the CLM. Because carbon was not being unnecessarily spent to grow leaves during
the dry season, which was then not replenished since there was not enough water to maintain
photosynthesis, CLM-MOD showed substantially higher overall carbon stores in savanna regions
(Fig. 10, blue lines). Over time, this increase in vegetation carbon could lead to more realistic soil
carbon levels, which have been shown to be too low in savanna regions in CLM (Wieder et al.,
2013).
Fire is a critical component of savanna ecology and has been a focus of recent efforts to improve the CLM (Li et al., 2014). Our change to the drought deciduous phenology algorithm does have a slight impact on the fire cycle, but unfortunately, though not surprisingly, it degrades the fire model’s performance relative to a global fire data set (GFED4; Giglio et al., 2013). Comparing the average total annual fire fractions for each grid cell with drought deciduous cover greater than 50% across the time period for which we have both GFED4 data and CLM output (1996-2010) we find a correlation between GFED4 and CLM4.5BGC of 0.35 (global correlation = 0.44), and a correlation of 0.23 with CLM-MOD (global correlation = 0.33). This degradation of fire model performance is not surprising, however, given that the fire model was developed using CLM4.0CN with the erroneous dry season green up and a different forcing dataset. As shown in Fig. 9, fraction of area burned per grid cell decreases in many areas in CLM-MOD, likely due to the fact that less fuel is being produced in the dry season, and seasonality shifts as well (Fig. 10, red lines). Future work will include exploring the impacts of this change to CLM on fire and other ecosystem properties.

Two other outstanding questions about LAI in CLM remain. First, in savanna regions in the CLM LAI drops to zero during the dry season, implying that across an entire grid cell all vegetation is perfectly drought deciduous (Figs. 3-6). Reality is, of course, far more complex, as reflected in the LAI3g dataset which rarely drops below 1.0 in savanna regions when aggregated to the CLM grid. The focus of this study was on improving the timing and magnitude of peak LAI, however, improving dry season values is also a concern. This is a deeper question in the CLM, as it relates to the overall land cover data. It is possible, for example, that there is evergreen vegetation in these grid cells, while the land cover classification determines that all grid cells with a significant seasonal LAI signal are 100% ‘drought deciduous’. Even if drought deciduous and evergreen vegetation types did not co-exist in the same ecosystem type, within grid cell spatial heterogeneity might also allow their coexistence within a whole grid cell (e.g. riparian areas or areas with shallow ground water that are...
able to stay green year round). Second, though also not the focus of this study, it is clear from Fig. 6 that across the mesic regions of the terrestrial biosphere CLM is dramatically overestimating LAI. This issue, often masked when only mean annual values or zonal means are considered, deserves more attention, and it is likely that recent detailed studies of carbon allocation (e.g. Doughty et al., 2014) could improve this part of the model.

A question also remains as to whether our new representation of leaf phenology, in spite of its improved performance, constitutes a better predictive model of current and future ecosystem behavior. In general, we hope to construct ecosystem models that represent hypotheses of how plants function, that we might test against observations. In this case, we find that the existing hypothesis - that plants respond to the soil water potential of the upper soil - does not adequately represent phenological patterns. However, this explanation is complicated because the predictions depend also on the properties of soil water in the model. Given a perfect representation of soil moisture, we might find that the existing leaf-on hypothesis is a good approximation of average vegetation behavior. However, at present the coupling of these two complex systems produces unexpected results. By tying the vegetation behavior to the actual climate drivers we are reducing the complexity of the problem, however, we are also reducing the capacity of the model to be responsive to the nuances of climate drivers. For example, the same rainfall amount in high and low humidity regimes will have different impacts on net soil moisture.

Ideally, models should represent the mechanisms by which ecological processes operate in as much fidelity as we understand. The representation of drought phenology is interesting; however, as we suspect that there are many different phenological strategies in the tropics that the CLM classifies with the same algorithm (e.g. Archibald & Scholes, 2007). This means that, in the absence of the representation of these numerous phenological strategies in the model, we are really representing the net behavior of ecosystems, rather than the exact mechanisms pertaining to a single species. The
fact that CLM-MOD improved model performance most significantly in Africa and less so in Australia and South America by some metrics (Fig. 7) suggests that evolutionary differences between plants could play a significant role in determining phenological patterns between continents. In a higher-fidelity land surface model, we might ideally allow numerous phenological algorithms to compete for light and water resources, and the ecosystem LAI profile would reflect the net behavior of the successful algorithms. This type of modelling is now theoretically possible (e.g. Fisher et al., 2010, 2015), and will be investigated in future versions of the CLM.

5. Conclusions

By comparing satellite derived estimates of LAI to LAI values produced by the latest version of the CLM we revealed a small but significant issue in the CLM – the tendency for leaves to flush during the dry season in drought deciduous PFTs due to unrealistic upwards movement of water through the soil column. We tested a number of different approaches to address this issue, however we found that tying leaf flushing to rainfall directly produced results that best matched the satellite data. While this change to the drought deciduous phenology algorithm does not reflect our understanding of how plants respond to their environment in the real world, without better data on soil water movement at scales relevant to global land surface modeling it is difficult to rely on the soil water model to drive plant physiology. Changing the drought deciduous phenology algorithm to remove dry season leaf flushes improved overall LAI values in savanna systems as well as changed the amount of carbon stored in these systems and altered the fire cycle. We also emphasize that this issue would have been impossible to detect with a standard ‘benchmarking’ type of metric for measuring seasonality and was difficult to identify until daily model outputs were reported and analyzed (i.e. Fig. 3). Future work will include exploring different drought deciduous phenology
algorithms for different PFTs and testing the importance of this change in a coupled Earth system model.

Acknowledgements

The authors thank the members of the Terrestrial Sciences Section at NCAR for helpful discussions of this work, and we thank Dr. Ranga Myneni and his group for providing the LAI3g dataset. KMD was funded by an Advanced Study Program Postdoctoral Fellowship at NCAR. We would also like to acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.
REFERENCES


Table 1. List of locations for point simulations and percent cover of plant functional types (PFTs). PFTs with no coverage at any point are not listed.

<table>
<thead>
<tr>
<th>point.name</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Bareground</th>
<th>Broadleaf Evergreen Tree Tropic</th>
<th>Broadleaf Evergreen Tree Temperate</th>
<th>Broadleaf Deciduous Tree Tropic</th>
<th>C3 Grasses</th>
<th>C4 Grasses</th>
<th>Crops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brasilia</td>
<td>-15</td>
<td>-51</td>
<td>0.46</td>
<td>1.69</td>
<td>0</td>
<td>16.52</td>
<td>8.83</td>
<td>62.35</td>
<td>10.15</td>
</tr>
<tr>
<td>western Brazil</td>
<td>-6</td>
<td>-39</td>
<td>2.66</td>
<td>0</td>
<td>0</td>
<td>35.4</td>
<td>9.04</td>
<td>40.6</td>
<td>12.3</td>
</tr>
<tr>
<td>South Chad</td>
<td>11</td>
<td>18</td>
<td>1.34</td>
<td>0</td>
<td>0</td>
<td>34.26</td>
<td>0.03</td>
<td>60.22</td>
<td>4.16</td>
</tr>
<tr>
<td>eastern Zambia</td>
<td>-13</td>
<td>32</td>
<td>0.22</td>
<td>0.56</td>
<td>1.27</td>
<td>26.4</td>
<td>37.81</td>
<td>26.39</td>
<td>7.35</td>
</tr>
<tr>
<td>south Ethiopia</td>
<td>5.5</td>
<td>40</td>
<td>8.75</td>
<td>0.13</td>
<td>0.02</td>
<td>63.1</td>
<td>19.12</td>
<td>5.42</td>
<td>3.47</td>
</tr>
<tr>
<td>Darwin Australia</td>
<td>-15</td>
<td>130.5</td>
<td>15.94</td>
<td>0</td>
<td>0</td>
<td>35.73</td>
<td>0</td>
<td>48.33</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. R² and RMSE of AVHRR LAI₃g v. CLM₄.₅BGC monthly data for all 29 yrs.

<table>
<thead>
<tr>
<th>CLM₄.₅BGC</th>
<th>NH C3 grass</th>
<th>NH C4 grass</th>
<th>tropical DT</th>
<th>SH DS</th>
<th>SH C3 grass</th>
<th>SH C4 grass</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.79</td>
<td>0.72</td>
<td>0.08</td>
<td>0.05</td>
<td>0.41</td>
<td>0.85</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.15</td>
<td>0.29</td>
<td>0.18</td>
<td>0.10</td>
<td>0.27</td>
<td>0.13</td>
</tr>
</tbody>
</table>

* PFT = plant functional type, NH = northern hemisphere, DS = deciduous shrub, DT = deciduous tree, SH = southern hemisphere
Table 3. Confusion matrices comparing grid cell peak counts between LAI3g and the two model data sets. “%” rows and columns are the percent of the correct values (diagonal) compared to the sums for the respective rows and columns.

<table>
<thead>
<tr>
<th></th>
<th>LAI3g</th>
<th></th>
<th></th>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>&gt;2</td>
<td>%</td>
</tr>
<tr>
<td>CLM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>436</td>
<td>164</td>
<td>145</td>
<td>0</td>
<td>58.5</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>74</td>
<td>130</td>
<td>0</td>
<td>31.9</td>
</tr>
<tr>
<td>2</td>
<td>196</td>
<td>555</td>
<td>499</td>
<td>0</td>
<td>39.9</td>
</tr>
<tr>
<td>&gt;2</td>
<td>7</td>
<td>42</td>
<td>35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>%</td>
<td>65.4</td>
<td>8.9</td>
<td>61.7</td>
<td>NA</td>
<td>43.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LAI3g</th>
<th></th>
<th></th>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>&gt;2</td>
<td>%</td>
</tr>
<tr>
<td>CLM-MOD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>365</td>
<td>279</td>
<td>277</td>
<td>0</td>
<td>39.6</td>
</tr>
<tr>
<td>1</td>
<td>242</td>
<td>439</td>
<td>407</td>
<td>0</td>
<td>43.2</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>63</td>
<td>125</td>
<td>0</td>
<td>50.4</td>
</tr>
<tr>
<td>&gt;2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>%</td>
<td>54.7</td>
<td>59.0</td>
<td>15.5</td>
<td>NA</td>
<td>42.5</td>
</tr>
</tbody>
</table>
FIGURE CAPTIONS

Figure 1. (A) Areas of the globe dominated by a single PFT (>50%) grouped where appropriate (e.g. there are actually three grass PFTs). (B) Percent cover of drought deciduous PFTs within the natural vegetation component of each grid cell; gray areas have zero percent cover of drought deciduous vegetation. For visual clarity, grid cells with < 50% natural vegetation (e.g. grid cells that are mostly water) are not shown in both maps.

Figure 2. Annual LAI cycles for LAI3g and CLM4.5BGC averaged for 1982-2010; shaded areas represent one standard deviation. Each plot is averaged across a region as shown in Fig. 1. (A) Northern hemisphere (NH) C3 grasses; (B) NH C4 grasses; (C) Southern hemisphere (SH) C3 grasses; (D) SH C4 grasses; (E) Tropical deciduous trees; (F) SH broadleaved deciduous shrubs.

Figure 3. Seasonal cycles of rainfall (mm day⁻¹, gray bars); leaf area index (LAI, green lines and black dots) and soil water potential in the third layer (MPa, blue lines) in CLM4.5BGC and CLM-MOD for one year (2001).

Figure 4. Illustration of Latin hypercube (LH) variable exploration analysis results – here each line represents one simulation all from one year of the LH analysis without the additional rainfall trigger. Each line is from a model run with slightly different values for the variables considered. In actuality 100 simulations were performed, but for visual clarity we are showing a selection of 10 simulations.

Figure 5. Illustration of Latin hypercube (LH) variable exploration analysis results as with Fig. 4 – here each line represents one simulation all from one year of the LH analysis with the additional rainfall trigger.

Figure 6. Maximum annual LAI averaged across the 29 year time period (1982-2010) in LAI3g, CLM4.5BGC, and CLM-MOD, and the differences between these three maps.
Figure 7. Mode of annual peak count analysis for the three simulations. (A) LAI3g; (B) CLM4.5BGC; (C) CLM-MOD

Figure 8. Coefficients of determination ($R^2$) between LAI3g and the two model versions.

Figure 9. Average burned area fraction per year across the time period where data was available (1996-2010) for GFED4, CLM4.5BGC, and CLM-MOD, and the differences between these maps.

Figure 10. Seasonal cycles of rainfall (mm day$^{-1}$; gray bars), total vegetation carbon (gC m$^{-2}$; blue lines), and grid cell burned fraction (day$^{-1}$; red lines) in CLM4.5BGC and CLM-MOD for one year (2001).