Challenging terrestrial biosphere models with data from the long-term multi-factor Prairie Heating and CO_2 Enrichment experiment

Journal:	*Global Change Biology*
Manuscript ID	Draft
Wiley - Manuscript type:	Primary Research Articles
Date Submitted by the Author:	n/a
Complete List of Authors:	De Kauwe, Martin; Macquarie University, Biological Sciences Medlyn, Belinda; Western Sydney University, Hawkesbury Institute for the Environment Walker, Anthony; Oak Ridge National Laboratory, Environmental Sciences Division and Climate Change Science Institute Zaehle, Sönke; Max Planck Institute for Biogeochemistry, Biogeochemical Integration Department Asao, Shinichi; Colorado State University, Natural Resource Ecology Laboratory Guenet, Bertrand; CNRS, LSCE Harper, Anna; University of Exeter, College of Engineering, Mathematics, and Physical Sciences Hickler, Thomas; Biodiversity and Climate Research Centre (BiK-F) & Senckenberg Gesellschaft für Naturforschung, Senckenberganlage, Department of Physical Geography Jain, Atul; University of Illinois, Department of Atmospheric Sciences Luo, Yiqi; University of Oklahoma, Botany and Microbiology Lu, Xingjie; CSIRO, Oceans and Atmosphere Luus, Kristina; Max Planck Institute for Biogeochemistry, Biogeosystems Department Parton, William; Colorado State, Natural Resource Ecology Laboratory Shu, Shijie; University of Illinois, Department of Atmospheric Sciences Wang, Ying Ping; CSIRO, marine and atmospheric research Werner, Christian; Biodiversity and Climate Research Centre (BIK-F), Xia, Jianyang; East China Normal University, Tiantong National Station of Forest Ecosystem, School of Ecological and Environmental Sciences Pendall, Elise; Western Sydney University, Hawkesbury Institute for the Environment Morgan, Jack; USDA-Agricultural Research Service, Rangeland Resources Research Ryan, Edmund; Arizona State University, School of Life Sciences Carrillo, Yolima ; Western Sydney University, Hawkesbury Institute for the Environment Dijkstra, Feike; The University of Sydney, Faculty of Agriculture, Food and Natural Resources Zelikova, Tamara; University of Wyoming, Ecosystem Science and

	Management Norby, Richard; Oak Ridge National Laboratory, Environmental Sciences Division and Climate Change Science Institute
Keywords:	carbon dioxide, FACE, grassland, PHACE, temperature, soil moisture, phenology, models
Abstract:	Multi-factor experiments are often advocated as important for advancing terrestrial biosphere models (TBMs), yet to date such models have only been tested against single-factor experiments. We applied 10 TBMs to the multi-factor Prairie Heating and CO_2</sub Enrichment (PHACE) experiment in Wyoming, USA. Our goals were to investigate how multi-factor experiments can be used to constrain models, and to identify a road map for model improvement. We found models performed poorly in current ambient conditions; there was a wide spread in simulated above-ground net primary productivity (range: 31-390 g C m^{-2} yr^{-1}). Comparison with data highlighted model failures particularly in respect to carbon allocation, phenology, and the impact of water stress on phenology. Performance against observations from single-factors experiments was also relatively poor. In addition, similar responses were predicted for different reasons across models: there were large differences among models in sensitivity to water stress and, among the N cycle models, N availability during the experiment. Models were also unable to capture observed treatment effects on phenology: they over-estimated the effect of warming on leaf onset and did not allow CO2</sub-induced water savings to extend the growing season length. Observed interactive (CO2</sub x warming) treatment effects were subtle and contingent on water stress, phenology and species composition. Since the models did not correctly represent these processes under ambient and single-factor conditions, little extra information was gained by comparing model predictions against interactive responses. We outline a series of key areas in which this and future experiments could be used to improve model predictions of grassland responses to global change.

Challenging terrestrial biosphere models with data from the long-term multi-factor Prairie Heating and CO_2 Enrichment experiment

Martin G. De Kauwe[1*], Belinda E. Medlyn[2], Anthony P. Walker[3], Sönke Zaehle[4], Shinichi Asao[5], Bertrand Guenet[6], Anna B. Harper[7], Thomas Hickler[8,9], Atul Jain[10], Yiqi Luo[11], Xingjie Lu[12], Kristina Luus[4], William J. Parton[5], Shijie Shu[10], Ying-Ping Wang[12], Christian Werner[8], Jianyang Xia[13], Elise Pendall[2], Jack A. Morgan[14], Edmund M. Ryan[15], Yolima Carrillo[2], Feike A. Dijkstra[16], Tamara J. Zelikova[17], Richard J. Norby[3]

1. Department of Biological Science, Macquarie University, North Ryde NSW 2109 Australia.

2. Hawkesbury Institute for the Environment, Western Sydney University, Locked Bag 1797, Penrith NSW 2751 Australia.

3. Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

4. Max Planck Institute for Biogeochemistry, Biogeochemical Integration Department, Hans-Knöll-Str. 10, 07745 Jena, Germany.

5. Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523-1499 USA.

6. Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.

7. College of Engineering, Mathematics, and Physical Sciences, University of Exeter, Exeter, UK.

8. Senckenberg Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt, Germany.

9. Department of Physical Geography, Geosciences, Goethe-University, Altenhöferallee 1, 60438 Frankfurt, Germany.

10. Department of Atmospheric Sciences, University of Illinois, 105 South Gregory Street, Urbana, Illinois 61801-3070, USA.

11. Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019 USA.

12. CSIRO Oceans and Atmosphere, Private Bag #1, Aspendale, Victoria 3195, Australia

13. Tiantong National Forest Ecosystem Observation and Research Station, School of Ecological and Environmental Sciences, East China Normal University, Shanghai 200062, China.

14. Rangeland Resources Research Unit, Agricultural Research Service, United States Department of Agriculture, Fort Collins, CO 80526, USA.

15. Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YW, UK.

16. Centre for Carbon, Water and Food, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

17. Department of Botany, University of Wyoming, Laramie, WY 82071.

Corresponding author address: Martin De Kauwe, Macquarie University, Department of Biological Sciences, New South Wales 2109, Australia. E-mail: mdekauwe@gmail.com Phone: +61 2 9850 9256

Running head: Model-data synthesis of the PHACE experiment.

Keywords: carbon dioxide, FACE, grassland, PHACE, temperature, models, soil moisture, phenology, allocation.

Type of Paper: Primary Research Article

Word Count: 7,822

Figures: 10 (5 supplementary)

Tables: 4 (2 supplementary)

Abstract

Multi-factor experiments are often advocated as important for advancing terrestrial biosphere models (TBMs), yet to date such models have only been tested against single-factor experiments. We applied 10 TBMs to the multi-factor Prairie Heating and CO_2 Enrichment (PHACE) experiment in Wyoming, USA. Our goals were to investigate how multi-factor experiments can be used to constrain models, and to identify a road map for model improvement. We found models performed poorly in current ambient conditions; there was a wide spread in simulated above-ground net primary productivity (range: 31-390 g C m^{-2} yr^{-1}). Comparison with data highlighted model failures particularly in respect to carbon allocation, phenology, and the impact of water stress on phenology. Performance against observations from single-factors experiments was also relatively poor. In addition, similar responses were predicted for different reasons across models: there were large differences among models in sensitivity to water stress and, among the N cycle models, N availability during the experiment. Models were also unable to capture observed treatment effects on phenology: they over-estimated the effect of warming on leaf onset and did not allow CO_2-induced water savings to extend the growing season length. Observed interactive (CO_2 x warming) treatment effects were subtle and contingent on water stress, phenology and species composition. Since the models did not correctly represent these processes under ambient and single-factor conditions, little extra information was gained by comparing model predictions against interactive responses. We outline a series of key areas in which this and future experiments could be used to improve model predictions of grassland responses to global change.

Introduction

Grasslands are estimated to cover 20% of the terrestrial land surface (Lieth, 1978; Hadley, 1993) and store ~25% of the world's soil carbon (C) excluding permafrost soils (Jobbágy & Jackson, 2000; Ciais *et al.*, 2013). However, whether grasslands will be substantial C sources or sinks in the future is uncertain; estimates of future C uptake range between −2 to 2 Gt C yr^{-1} (Scurlock & Hall, 1998). Semi-arid ecosystems, including grasslands, are large contributors to both the trend and inter-annual variability in above-ground net primary production (Knapp & Smith, 2001) and net biome production (Ahlström *et al.*, 2015), over the last three decades, suggesting these ecosystems are particularly important for accurately predicting terrestrial C-cycle responses to global change.

To predict how increasing temperatures, atmospheric carbon dioxide (CO_2) and changing precipitation patterns will affect ecosystem function and species composition, multi-factor ecosystem-scale experiments have been widely advocated (Heimann & Reichstein, 2008; Luo *et al.*, 2008; Leuzinger *et al.*, 2011). Since global change factors likely cause a series of complex interactions (Fuhrer, 2003; Hovenden *et al.*, 2014), single-factor experiments may not be sufficient to investigate future ecosystem-scale responses. Further, while interactive effects are typically smaller than main effects (Shaw *et al.*, 2002; Dieleman *et al.*, 2012), they may sometimes exceed single factor effects. However, interactive effects may be contingent on environmental conditions, such as inter-annual variability in precipitation (Mueller *et al.*, 2016). As a result, multi-factor experiments can be more difficult to interpret, and underlying mechanisms harder to identify, than single factor experiments.

For example, Shaw *et al.* (2002) found contrasting results when comparing responses from single and multi-factor treatments in the Californian grasslands at the Jasper Ridge Global Change Experiment (JRGCE). In the third year of the experiment, net primary productivity (NPP) was increased in response to elevated CO_2 (eCO_2). However, the interactive effect of multi-factors suppressed the NPP response seen in the single factor response. Re-examining the responses at the JRGCE over 5 years, Dukes *et al.* (2005) concluded that NPP did not in fact respond to eCO_2. Hovenden *et al.* (2008) also found no CO_2 enhancement in ecosystem productivity in an Australian perennial grassland experiment (TasFACE). This lack of response was attributed to a reduction in soil N availability in response to eCO_2, but increasing temperature by 2°C in combination with the CO_2 treatment was found to prevent this decrease in available N. In the multi-factor Prairie Heating and CO2 Enrichment (PHACE) experiment, Mueller *et al.* (2016) found that above-ground NPP and total plant biomass both had time-dependent and interactive effects of warming and eCO_2. Above-ground NPP responses to the combination of eCO_2 and warming exceeded responses to the single factors (non-additive). Soil moisture was especially important in explaining the productivity responses to treatments as well as inter-annual precipitation variability.

Dieleman *et al.* (2012) conducted a meta-analysis using data from 150 manipulation experiments and concluded that the response of above-ground biomass to the combined treatments of CO_2 and warming was typically less than additive. These results suggest that single factor experiments, which miss the interaction, may over-estimate responses, highlighting the need to test models against multi-factor experiments. However, model comparisons to date have only explored theoretical multi-factor experiments (e.g. Melillo *et*

al., 1993; Riedo *et al.*, 1997; Pepper *et al.*, 2005; Parton *et al.*, 2007; Luo *et al.*, 2008), rather than applying models directly to experimental data.

The model-data inter-comparison approach has been useful to investigate single-factor forest experiments (De Kauwe *et al.*, 2013, 2014; Zaehle *et al.*, 2014; Medlyn *et al.*, 2015; Walker *et al.*, 2015), but it is not clear whether multi-factor experiments will be as useful to constrain models when their responses seem so diverse, and in dry environments, contingent on environmental conditions. In this paper, we applied 10 state-of-the-art terrestrial biosphere models (TBMs) to an 8-year, multi-factor (CO_2 × warming) grassland experiment. Our goals were to: (i) explore how a multi-factor experiment can be used to constrain models and (ii) identify ways to improve models based on this experiment.

Materials and methods

Site description

The PHACE experiment was located in the semi-arid grasslands of Wyoming, USA (41.18°N, 104.9°W), was established in 2006, and lasted 8 years. Mean winter and summer temperature at the site were –2.5°C and 17.5°C, respectively, with a mean annual precipitation of 403 mm (range: 224–496 mm). The site has marked variation in both annual and growing season precipitation (Fig. 1). The site was previously subject to grazing, but was fenced off in 2005. Vegetation at the site is dominated by C_3 grasses (55%), with C_4 grasses constituting 25% and the final 20% made up of sedges, forbs and small shrubs.

The experiment implemented a factorial combination of warming (ambient +1.5°C during the day and ambient +3.0°C at night) and elevated CO_2 (600 ppm; ambient = 385 ppm), with five replicates per treatment. The elevated CO_2 treatment, initiated in 2006, used Free Air CO_2 Enrichment (FACE) technology (Miglietta *et al.*, 2001). The warming treatment, initiated a year later in 2007, used infrared heaters (Kimball, 2005). In the first year (2006) an additional 160 mm of water was added (20 mm × 8 dates during the growing season) to establish growth. Further details can be found in Morgan *et al.* (2011), Pendall *et al.* (2013) Ryan *et al.* (2015) and Zelikova *et al.* (2015).

Summary of the experimental findings

Mueller *et al.* (2016) present a comprehensive summary of the ecosystem responses over the duration of the PHACE experiment. Elevated CO_2 effects on soil water content usually counteracted the desiccating effect of warmer temperatures. However, the combination of eCO_2 and elevated temperature tended to enhance soil water content early in the experiment, but reduced it after 7 years of treatment when compared to control plots under present-day CO_2 and temperature levels. Above-ground plant biomass responded positively to eCO_2 and eCO_2 combined with warming, especially in dry years when water savings were most important to growth. In contrast, while above-ground biomass did not respond to warming alone, root biomass responded positively to both warming and eCO_2, but only in wetter years, with either eCO_2 or warming enhancing production approximately 30% in wet growing seasons. As a result, total plant biomass responded consistently and positively to eCO_2 alone or combined with warming, with a 25% increase observed in the combined treatment compared to control plots. The positive effect of the combined eCO_2 and warming on above-ground plant biomass with passing years was increasingly experienced by C_3 grasses,

157 reversing biomass responses in the first few years of the experiment when C_4 grasses were
158 favoured (Morgan *et al.*, 2011). Soil nitrate availability was enhanced by warming and
159 reduced by eCO_2, although contrasting effects were observed for soil ammonium (Carrillo *et*
160 *al.*, 2012). In contrast, wetter soil conditions under eCO_2 increased phosphorus (P) availability
161 to plants and microbes relative to that of N, while drier conditions with warming reduced P
162 availability relative to N (Dijkstra *et al.*, 2012). Warming combined with eCO_2 extended the
163 seasonality of plant activity (greenness), especially because of earlier spring growth with
164 warming (Zelikova *et al.*, 2015).

165 *Experimental data*

166 To constrain the models we used five key datasets: (i) above- and below-ground biomass; (ii)
167 shoot and root N concentrations; (iii) vegetation greenness; (iv) leaf-on/off dates; (v) soil
168 water content.

169 Plant biomass (above- and below-ground) and N concentrations (elemental analyser) were
170 measured in mid-July as biomass reached its maximum (Morgan *et al.*, 2011; Dijkstra *et al.*,
171 2012; Carrillo *et al.*, 2014). Above-ground biomass measurements were obtained by clipping
172 vegetation that resided in the harvest areas (1.5 m^{-2} harvest area, but clipping 50% of this area
173 each year from alternating grids). Root-biomass measurements were obtained from cores
174 taken to a depth of 15 cm, but exclude standing crown tissues (see discussion). These data
175 exclude below-ground crown tissues estimates (see discussion). Above-ground biomass
176 estimates were corrected using pre-treatment data from 2005 to account for initial differences
177 between treatment plots and control plots (see Morgan *et al.*, 2011; also Mueller *et al.*, 2016).

178 Vegetation greenness was inferred from biweekly digital photographs taken between March
179 and October. In 2008, photographs were obtained monthly (see Zelikova *et al.* (2015) for
180 details). Phenology leaf-on and leaf-off dates for different species were obtained by direct
181 observation (Reyes-Fox *et al.*, 2014).

182 Soil moisture measurements were taken hourly using EnviroSMART probes at 10 and 20 cm
183 soil depths. These data were combined to give a total estimate of soil water content in the top
184 25 cm.

185 *Models*

186 The 10 process-based models applied to the PHACE experiment contrasted markedly in terms
187 of application, complexity and structure. Broadly, they can be considered to encompass three
188 categories: stand (DAYCENT, GDAY), land surface (CABLE, CLM4.5, ISAM, O-CN,
189 ORCHIDEE) and dynamic vegetation models (JULES, LPJ-GUESS, SDGVM). A detailed
190 overview of eight of these models and how they differ in terms of key assumptions can be
191 found in Walker *et al.* (2014), with detailed analyses of their water and N cycle responses to
192 eCO_2 found in De Kauwe *et al.* (2013) and Zaehle *et al.* (2014), respectively. The two models
193 not described in these previous analyses, JULES and ORCHIDEE, are fully documented in
194 Clark *et al.* (2011) and Krinner *et al.* (2005), respectively. Here, we provide some basic
195 assumptions in relation to growth and phenology used in each of the models that affects
196 simulations of the PHACE experiment (see Table 1).

Modelling simulations

Model participants submitted simulations covering the experimental period (2006 – 2013) for the ambient (ct), eCO_2 (Ct), warming (cT) and eCO_2 × warming (CT) experiments. Models were spun-up to equilibrium (2000 year minimum) using their standard spin-up approach accounting for site history and using a fixed CO_2 concentration of 285 μmol mol^{-1} and fixed N deposition set at the 1850 value based on Dentener *et al.* (2006). Models estimated biological N fixation (BNF) following their standard approach: CABLE uses a method based on light, N and phosphorus availability (Wang *et al.*, 2009) (BNF was estimated to be zero for the site), CLM4.5 uses an empirical relationship based on NPP (Oleson *et al.*, 2013), DAYCENT estimates N fixation as a function of climate (Parton *et al.*, 1987) and GDAY, ISAM, LPJ-GUESS and O-CN use an empirical relationship with long-term evapotranspiration (Cleveland *et al.*, 1999). Modellers were provided with stand and soil characteristics to parameterise their models so as to be representative without being "tuned" to the observations.

Experimental plots were harvested (mid-July) to simulate grazing; by contrast models did not assume any site disturbance during simulations. This choice was made because harvested plant biomass was removed from a small area of the plot only, while some of the experimental data did not come from the harvest areas (e.g., root biomass, soil moisture). Models, including dynamic vegetation models (JULES, LPJ-GUESS and SDGVM), did not simulate competition among plant functional types. Instead, models simulated the sites by weighting outputs by the average observed ambient total C_3 and C_4 above-ground biomass fractions, 0.69 and 0.31, respectively.

Data availability will be summarised and updated as appropriate at https://facedata.ornl.gov/facemds/

Results

Ambient CO_2

Fig. 2 shows the simulated above-ground net primary productivity (aNPP) in the ambient treatment plots. Whilst the models are able to capture the observed inter-annual variability (r^2 > 0.74), there is a wide spread in the magnitude of simulated values (RMSE mean = 96 g C m^{-2} yr^{-1}; range: 31-390 g C m^{-2} yr^{-1}). To explain differences among the models, we analysed aNPP by decomposing the modelled aNPP flux into its average component parts (Table 2). Each of these component terms is a simplification of how the models operate, but on an annual time-step should closely approximate simulated aNPP fluxes, allowing us to better understand causes of differences among models. aNPP can therefore be analysed as:

$$\mathrm{aNPP} = \mathrm{A_b} \cdot \mathrm{CUE} \cdot \mathrm{GPP_u} \cdot \beta \cdot \mathrm{LAI_p} \cdot \mathrm{LAI_r} \quad (1)$$

where A_b is the allocation of net primary productivity above-ground (fraction), CUE is the C-use efficiency, or the fraction of gross primary productivity (GPP) not lost as respiration (fraction), GPP_u is the unstressed GPP per unit leaf area (g C m^{-2} $leaf^{-1}$ d^{-1}), β is the water stress factor which limits productivity as water content declines (fraction), LAI_p is the peak LAI (m^{-2} leaf m^{-2} ground); and LAI_r is the integral of LAI over the year divided by the peak

LAI, and indicates LAI duration (d yr^{-1}). GPP_u is inferred from model output by dividing GPP by ($\beta \cdot LAI_p \cdot LAI_r$).

Table 2 shows a very large spread in component terms across models. The size of this variation, which is greater than the aNPP spread between models, suggests that models are arriving at the same answer for different reasons. For example, DAYCENT and GDAY predict similar average aNPP values, but to get to this prediction GDAY has a low GPP_u (4.71 g C m^{-2} $leaf^{-1}$ d^{-1}) and a high β (low water stress; 0.73). By contrast, DAYCENT has a much greater GPP_u (11.92 g C m^{-2} $leaf^{-1}$ d^{-1}) but a very low β (0.17). The most variable components among models are: (i) LAI_r (range: 77-256 days); (ii) LAI_p (range: 1.21 - 6.1 m^2 m^{-2}); (iii) A_b (range: 0.16–0.92); and (iv) β (range: 0.17–0.97). We now examine each of these components in more detail.

Observed seasonal phenology at the site, inferred from greenness estimates corresponds with measured soil water content (SWC; 5–15 cm) (Fig. 3). Drops in observed greenness agree with drops in SWC, particularly in dry years (2007, 2008), but also in a relatively wet year (2011). In wetter years (2009, 2010), greenness and SWC show little correspondence, until sufficient soil drying has occurred to drive a sudden decline in leaf greenness, around day of year (DOY) 200. Inferred vegetation greenness from digital photography does not directly correspond to leaf area index (LAI), but is well correlated with plant cover and biomass (Zelikova *et al.*, 2015), and so is a reasonable proxy against which to compare modelled LAI. With the exception of CLM4.5, modelled LAI at the site was remarkably smooth both across models and years; none of the models showed the observed strong within-season dynamics

257 seen in the observations (Fig. 4). We conclude that, in general, modelled LAI is insufficiently
258 sensitive to soil water availability in this semi-arid grassland

259 The lack of variability within the growing season is a consequence of how models determine
260 growth (Table 1). For deciduous species, DAYCENT and GDAY use the previous year's
261 stored C to grow, and in LPJ-GUESS growth is only calculated once at the end of the year,
262 based on the annually integrated NPP. These assumptions introduce a significant lag between
263 growth and meteorology and also result in very smooth growth predictions, because the sub-
264 annual scale allocation of C is not related to environmental stress. Other models (CABLE,
265 ISAM) assume specific phenological periods in which growth must occur, and end up with
266 similar smooth phenologies, which are unrelated to environmental conditions. In JULES, O-
267 CN and ORCHIDEE, the current year's growth is directly related to recently-fixed C, without
268 assumptions about specific phenological growth stages. Nevertheless, these models display
269 only marginally more within-season variability than the other models. In CLM4.5, C_3 grasses
270 were not able to grow at the site and the extremely variable LAI corresponds to the C_4 grass
271 component.

272 Table 1 summarises the key assumptions that dictate modelled leaf emergence and
273 senescence. Both CABLE and SDGVM assume that grasses do not entirely drop their leaves,
274 behaving instead like dynamic evergreen vegetation. Leaving aside these models (and
275 CLM4.5), most models predicted a later leaf onset date (mean = 40 ± 26 days, 1 standard
276 deviation) than was observed at the site. LPJ-GUESS was the exception, predicting an earlier
277 leaf onset, mean ~11 days.

Conversely, modelled leaf senescence typically occurred at or after DOY 300, which meant models were broadly consistent with the range in leaf drop dates observed at the site (Reyes-Fox *et al.*, 2014). Despite this seemingly better agreement with observed leaf senescence, the data in Fig. 2 suggest that whilst the grasses maintained standing biomass, these leaves were no longer productive. Towards the end of the growing season, there is a drop in vegetation greenness, which signifies a change in leaf chlorophyll content. By contrast, the models assume that as long as there is leaf area, sufficient soil water and radiation, leaves are actively photosynthesising. Thus, the models typically over-estimated the period that leaves were photosynthetically active by ~50-100 days, even in wet years.

Models predict LAI as a consequence of allocation of net primary productivity (NPP) and stored carbohydrates to leaves, the subsequent turnover of these tissues, and assumptions about specific leaf area. We inferred observed above- and below-ground allocation fractions from biomass data and an assumed fine-root lifespan of 5.8 years (Fig. 5). This estimate is consistent with an isotope based estimate of 6–7 years at the site (Carrillo *et al.*, 2014) and from a near-by shortgrass steppe site, which has an approximate lifespan of 5.5–7 years. As there is uncertainty about this estimated lifespan, we also show these data as above- and below-ground ratio (Fig. S1). Site data suggested that the proportion of NPP allocated above-ground (64 %) was greater than below ground (35 %). Models strongly disagreed about the proportion of C allocated above versus below-ground, and no model agreed with the observations. At the extremes, CABLE predicted that ~70% of C was sent below-ground, while ISAM, JULES and SDGVM predicted >80% was allocated above-ground (Fig. 5). Much of the details as to why these models disagree in terms of allocation have been documented previously (De Kauwe *et al.*, 2014). In agreement with these earlier findings,

301 models (GDAY, LPJ-GUESS, O-CN, ORCHIDEE) that implemented a functional balance
302 (between leaves and roots) predicted more balanced allocation fractions. Among these
303 models, higher allocation below-ground (CABLE, GDAY, LPJ-GUESS) indicated greater N
304 and/or water stress. This prediction was also in line with the DAYCENT model, which
305 allocates C to the plant tissue with the greatest resource limitation.

306 Another key explanation for model differences was related to soil water content (SWC).
307 Models were parameterised with the same soil water holding capacity, so differences in
308 predicted SWC partly relate to differences in LAI (Fig. 3), but also to soil evaporation.
309 Models disagreed on both the available SWC, as well as the sensitivity of productivity to
310 SWC. Fig. 6 shows modelled soil water time-series in a dry (2008) and a wet year (2009).
311 Despite differences in the absolute SWC, with the exception of CABLE and ISAM, most
312 models predicted consistent declines in SWC, with earlier declines in the dry year.
313 ORCHIDEE (mean = 44 mm yr^{-1}), SDGVM (mean = 62 mm yr^{-1}), O-CN (mean = 81 mm yr^{-1})
314 and LPJ-GUESS (mean = 129 mm yr^{-1}) predicted comparatively low total soil evaporation
315 fluxes across years, whereas the other models predicted ~2-3.5 times greater annual
316 evaporative fluxes. The SDGVM result is likely explained by continuous (and high) foliage
317 cover, but this does not apply to the other models which simulate lower LAI. In a semi-arid
318 system, these variations among models in predicted water losses are concerning.

319 Models also strongly disagreed on the level of water stress, shown by the growing season
320 simulated water stress factor (β; the ratio of predicted soil water content to the soil water
321 content at field capacity), which is used to limit gas exchange as water availability declines
322 (Fig. 7). β varied markedly between models. For some models (DAYCENT, JULES, LPJ-

GUESS) there is no obvious distinction between wet and dry years. This variation is caused by different assumptions among the models as to the shape of the functions used to represent the effect of water stress (Medlyn *et al.*, 2016) (Fig. S2). Notably, ORCHIDEE predicted no stress because in this version of the model (IPCC's Fifth Assessment version), the hydrological cycle is represented by a two buckets layer scheme. Using this representation, drainage or surface runoff occurs only when both buckets are full. Therefore this scheme generally underestimates runoff and consequently overestimates the soil water content and underestimates the soil water stress for plants.

Response to CO_2

We assessed modelled responses to eCO_2 by comparing results against measured above- and below-ground biomass data. We also explored modelled responses of N mineralisation, uptake and changes in N use efficiency, comparing results to summary data from the site.

To understand model predictions, we split above-ground response into C_3 and C_4 components. Fig. 8 shows marked year-to-year variability in the observed aNPP responses to CO_2 in C_3 species: observed aNPP responses were between 11% and 39%, averaging 16%. In 2009 (the wettest year), the observations showed a 6% decrease in aNPP because the ambient plots were more productive than the eCO_2 treatment plots. The modelled CO_2 effect on aNPP averaged 29% (range: -12 to 63%). However, with the exceptions of CABLE and ISAM, model responses were within the range of the observed treatment responses in most years when considering standard errors calculated across replicates. Whilst models seemingly appear unable to capture the inter-annual variability of the enhancement due to CO_2, the uncertainty on the observed responses is large, meaning most of the simulated responses are plausible.

345 Observed aNPP responses to CO_2 for C_4 species were negative for 4 of the 6 years, with aNPP
346 on average decreasing by -4%. The models predicted more modest changes in aNPP, mean =
347 5% increase, range: -27 to 16% (Fig. 9), which is within the range of observed responses
348 including the standard errors of treatment replicates.

349 The change in aNPP in response to CO_2 is itself a result of changes in GPP, autotrophic
350 respiration and allocation. To investigate these changes we separated these average responses
351 for each component for C_3 (Table 3) and C_4 (Table 4) species. We focus on differences in the
352 responses of C_3 species as this is where the models disagreed most. We examine the change in
353 autotrophic respiration by looking at the CUE, or the fraction of GPP not respired.

354 Most models predicted an increase in GPP in response to eCO_2, with the mean annual
355 increase ranging between 30-73%. JULES predicted the largest GPP response to CO_2 (mean =
356 73%) and CABLE the smallest (mean = 21%). The direct effect of CO_2 on leaf-scale
357 photosynthesis should theoretically be on the order of 25-30% (Franks *et al.*, 2013) for the
358 treatment change in CO_2 concentration. In the models the predicted effect is greater because
359 of indirect feedbacks through increased soil moisture and LAI.

360 Among the C cycle only models (JULES, ORCHIDEE, SDGVM), the mean annual response
361 of GPP to CO_2 varied strongly (range: 31 to 73%). JULES had the largest stimulation because
362 under ambient conditions, the model is particularly water stressed (Fig. 7), and eCO_2
363 alleviates this water stress, which results in large CO_2 stimulation of GPP. ORCHIDEE and
364 SDGVM predicted similar mean values (different inter-annual variability), but for different
365 reasons. At ambient CO_2, ORCHIDEE did not predict any water stress, and as a result the
366 benefit of CO_2 via water savings was negligible. In SDGVM, the GPP response to CO_2 was

low due to the high ambient LAI (Fig. 4), which meant that canopy photosynthesis was primarily light-limited. In addition, this high LAI meant that there were negligible benefits to be gained from CO_2 induced water savings, due to high transpiration.

GPP responses among the N cycle models were also not consistent (mean range: 20 to 55%), particularly evident in the year-to-year variability in the size of the enhancement. There was pronounced variability in N availability due to different levels of productivity (see Fig. 2) during model spin-up. Models could be categorised into three groups: at the low end, the mean inorganic N pool was between ~0.3–1.3 g N m^{-2} (CABLE, GDAY, LPJ-GUESS and O-CN), in the middle ~30 g N m-2 (CLM5, ISAM) and at the high end, 177 g N m^{-2} (DAYCENT). Site soil N measurements suggested an inorganic pool size (0.4 g N m^{-2}) towards the lower end of the model predictions (Dijkstra *et al.*, 2012). Most models (CABLE, DAYCENT, GDAY, LPJ-GUESS) predicted large increases (>20 %) in photosynthetic N use efficiency (GPP / canopy N; PSNUE) (Fig. S3). CLM4.5, ISAM and O-CN predicted large increases (>20 %) in N uptake (Fig. S4), which combined with increased N mineralisation (Fig. S5) in ISAM and O-CN, resulted in sustained GPP responses to CO_2 in these models. CABLE also predicted a reduction in N losses in response to CO_2, but this change was small (~0.3 g N m^{-2}) when integrated across the experiment and thus, made a negligible difference to total N availability. N losses were thought to have been low for the site (Dijkstra *et al.*, 2010).

The increases in N mineralisation (Fig. S5) in response to CO_2, particularly in the ISAM and O-CN models were at odds with the site data. Although there is no direct site evidence of N limitation, Dijkstra *et al.* (2012) showed evidence of dilution in plant N concentrations with

increasing soil water, which would suggest plant N demand increased by more than the net N mineralisation rate. The increased N mineralisation in O-CN was caused by decreased soil organic matter, whereas in ISAM, it was driven by the increased C:N ratio of the soil organic matter. Generally, these models did not predict the increased microbial N immobilisation because inorganic N pools were sufficiently saturated. Had these models started with smaller inorganic N pools (similar to that used by GDAY), then the changes in N availability in response to treatment would also have been smaller and more in line with what was observed. Models that implement a variation of the CENTURY soil model have the mechanism to predict the observed sites changes in N availability and ultimately the differences come down to the availability of N, which differed due to different end states after model spin-up.

We now examine the contribution of changes in CUE to the aNPP enhancement (Tables 3 and 4). Most models predicted modest changes although models disagreed on whether total respiration increased or decreased with CO_2 (-12 to 14%). The DAYCENT and O-CN models assume that nutrient limitation results in excess C being respired, which results in a decreased CUE at eCO_2.

Changes in allocation in response to CO_2 were low across all models, typically of the order of ±5% (Tables 3 and 4). CABLE predicted ~15% increase in the NPP allocated to the labile storage pool in both C_3 and C_4 plants, which occurs because in CABLE plants were unable to acquire sufficient N to grow tissues. This N limitation largely explains the negative response (mean = 12%) of aNPP to CO_2 despite the GPP enhancement (mean = 21%). CABLE simulated a very large labile C store: the elevated mean was 3983 g C m^{-2} yr^{-1} at eCO_2 compared to ambient, mean = 708 g C m^{-2} yr^{-1}.

411 The explanation as to why the high GPP response to CO_2 (73% enhancement for C_3 species)
412 only resulted in a more modest increase in aNPP in JULES relates to the C allocated for
413 competition (spreading). As competition is switched off, there is additional C fixed by the
414 plant that is subsequently not used during growth.

415 Shifting focus to changes in phenology, one of the principal results of the experiment was that
416 eCO_2 resulted in a longer growing season in 3 of the 5 years (Reyes-Fox *et al.*, 2014). In 2009
417 the last species to reach senescence did so 15.6 days later than in the ambient conditions.
418 However, in other years the change was smaller, 3.2 and 1.5 days in 2008 and 2011,
419 respectively (Reyes-Fox *et al.*, 2014). Notably, in 2007 (9.8 days) and 2010 (3.6) days,
420 senescence was actually earlier, shortening the growing season. These results complicate
421 drawing concrete conclusions about the effect of CO_2 treatment given the large inter-annual
422 variability, which was mediated by precipitation and soil moisture (Zelikova *et al.*, 2015).

423 Tables S1 and S2 show the change in growing season length in response to treatment in the
424 models. Leaf senescence was only delayed in the ISAM (0.8 days, range = -5 to 5 days)
425 model; however, this response did not relate to a CO_2 effect on soil water, but instead was an
426 outcome of the use of phenological phases. The senescence phase occurs only when LAI
427 declines to 95% of a prescribed upper threshold. eCO_2 results in an increase in LAI and
428 therefore LAI does not fall below this threshold, which lengthens the growing season (see De
429 Kauwe *et al.* (2014) for details). A number of models determine their leaf drop dates (Table 1)
430 based solely on air temperature (GDAY, JULES) and so miss any positive effect of any CO_2
431 induced soil water savings on growth via changes in leaf senescence. Other models (LPJ-
432 GUESS, ORCHIDEE, O-CN; see Table 1) do consider a minimum soil water status when

433 determining leaf drop, but soil water savings were not great enough to maintain the water
434 status above these thresholds.

435 Root biomass was increased on average by 11% with CO_2 treatment (Fig. 10). With the
436 exception of SDGVM, the models broadly enveloped the size of the increase, mean range: 7–
437 17%. However, models did not capture the year-to-year variability. Increased N stress
438 throughout the course of the experiment led to a greater allocation to roots in GDAY, LPJ-
439 GUESS and O-CN, as they simulate N uptake as a function of root biomass and allow
440 allocation to shift in response to resource availability. By contrast, DAYCENT predicted a
441 very small increase, because at ambient CO_2 fine root allocation was already high (Fig. 4),
442 which meant allocation to leaves was prioritised under eCO_2. SDGVM follows a leaf
443 optimisation scheme for C allocation. Responses of allocation to leaves and roots in SDGVM
444 largely matched the responses of GPP to CO_2, as grass allocation uses fixed fractions (Table,
445 1), which explains the large mean enhancement of 38%.

446 *Response to warming*

447 Observed aNPP of C_3 species only increased only in response to warming in 2011 (+53%); in
448 all other years, the warming treatment had a negative effect. However, when accounting for
449 the standard error on replicates, only one of the five years in which the response was negative,
450 did not also include the potential for a positive treatment response. CABLE apart, the models
451 generally predicted a small response of aNPP to warming, although the direction of the
452 treatment effect varied among models, plant functional groups and across years (Figs. 8 and
453 9). Among the N Cycle models, the balance between the warming-induced treatment
454 increases in N mineralisation (Fig. S5) and decreases in soil water (Fig. 7) explained

455 interannual variability in aNPP responses. Warming particularly enhanced N mineralisation in
456 GDAY and LPJ-GUESS. For C_3 species, soil water stress also increased (Fig. 7), which
457 limited responses (less mineralisation) in the O-CN and DAYCENT models. Similarly,
458 among the C-cycle models (JULES, SDGVM), the warming treatment increased water stress,
459 which reduced the aNPP response.

460 Warming consistently led to an earlier leaf expansion in the observations, mean = 5.1 days
461 (range 0.9 – 9.6 days) (Reyes-Fox *et al.*, 2014). The effect on leaf senescence was mixed:
462 shortening the growing season in 2007 (3.3 days) and 2009 (6.9 days) and lengthening it in
463 other years, 3.3 days, 0.4 and 8.5 days in 2008, 2010 and 2011, respectively. Most models did
464 predict an earlier spring growth in response to warming, as warmer temperatures meant that
465 models passed their assumed growing degree-days threshold earlier (see Table 1). However,
466 the magnitude of the change was considerably larger than observed: on average by 15.9 days
467 (range 2–24.3 days). Three of the models (CABLE, DAYCENT, SDGVM) predicted no
468 change. In DAYCENT the CO_2 effect on leaf on/off dates were prescribed, so it does not
469 capture a treatment effect. In CABLE and SDGVM, LAI is assumed not to reach zero (see
470 above). Finally, in two of the years, LPJ-GUESS predicted a delayed leaf onset (11 and 38
471 days) with warming, which was a result of limited soil water availability. The trigger for
472 growth in LPJ-GUESS is simply air temperature, which means the model attempted to grow
473 very early in some years (e.g. DOY 12 in 2010), but development is temporarily shut off
474 when soil water is below a threshold level. In the warming treatment, warmer temperatures
475 led to increased soil water depletion (via soil evaporation), which had the effect of delaying
476 leaf onset. Nevertheless, in years where soil water stores were greater (2008), the direction of
477 change in response to treatment matched the other models (not shown).

The small changes in root biomass in response to warming among the models follows the small aNPP response (Fig. 7) and, as with the response to CO_2, models again enveloped the observed change (Fig. 10).

CO_2 × warming

To examine the interactive effect, we calculated the additive response to CO_2 × warming treatment for C_3 aNPP (Fig. 8), C_4 aNPP (Fig. 9) and root biomass (Fig. 10), shown by the black horizontal lines. Observations generally show greater than additive interactions in both above- and below-ground biomass. DAYCENT is the only model to predict additive responses to the combined treatment. Models do not predict consistent interactions: responses are just less than additive, additive, or considerably greater than additive. Models that predict greater than additive interactions do so as a result of a positive effect of warming on N mineralisation (Fig. S5), combined with increased CO_2-induced water savings (Fig. 7).

In the observations from combined treatment plots, leaf expansion was earlier than in the ambient treatment, mean = 4.6 days (range 2.4 – 7 days), but the effect was smaller than in the warmed plots (Reyes-Fox *et al.*, 2014). There was a clear interaction on the leaf drop dates: the combined treatment resulted in an increased growing season length of 22.4 days in 2009 (Ct = 15.6 days), despite the warming treatment shortening the growing season by 6.9 days. Across all years, the response to the combined treatment was consistent, increasing the growing season length mean = 7.9 days (range 0.1 – 22.4 days) (Reyes-Fox *et al.*, 2014). With the exception of ISAM (not related to treatment, see above), the models did not predict the observed interaction between eCO_2 and warming on phenology.

Discussion

Evaluating models against ecosystem scale manipulation experiments has the potential to produce significant insight into model performance (De Kauwe *et al.*, 2013, 2014; Zaehle *et al.*, 2014; Medlyn *et al.*, 2015; Walker *et al.*, 2015).

Our inter-comparison has identified a number of important model failings. Several of these have been identified in previous model comparisons against FACE experiments, such as C allocation (De Kauwe *et al.*, 2014); flexibility of plant stoichiometry (Zaehle *et al.*, 2014); and sensitivity to drought stress (Medlyn *et al.*, 2016). There are however, a number of new issues identified in this study, namely: grassland phenology; link between soil water stress and growth; soil N availability; inter-annual variability; C storage / grassland physiognomy.

Soil water stress

In semi-arid ecosystems, water availability is a key determinant on productivity. The wide disagreement in the level of water stress among models (Fig. 6) is alarming, particularly given the models were all initialised with the same effective soil water bucket size. Differences in level of water stress among models drove differences in modelled productivity both in ambient conditions and in response to treatments, particularly warming. There were two main causes for these differences among the models: a large difference in simulated soil evaporation and differences in sensitivity of productivity to water availability (Figs. 7, S2).

The issue of different modelling schemes simulating sizeable differences in soil evaporation is not a new one (see Desborough *et al.* (1996)). Nevertheless, in water limited systems, it is the

519 principal control on early-growing season water in the root-zone. Data from existing eddy
520 covariance towers located at grassland sites should offer a strong constraint on modelled soil
521 evaporation fluxes.

522 Medlyn *et al.* (2016) recently questioned the empirical support for a number of the functions
523 used by the models in this study. There is therefore a clear need for models to implement
524 more evidence-based functions for the representation of drought stress (De Kauwe *et al.*,
525 2015). Considerable research is now being targeted to address this need (Zhou *et al.*, 2013,
526 2014; Verhoef & Egea, 2014). One issue is that many ecosystem manipulation experiments
527 only measured SWC in part of the root-zone profile, as at PHACE where SWC was measured
528 to 25 cm depth (Blumenthal et al. in prep). To quantify sensitivity to SWC, time courses of
529 SWC throughout the entire root-zone are required, along with information on rooting
530 distributions and regular gas-exchange measurements (e.g. Pendall *et al.* (2013)).

531 *Grassland phenology*

532 Models struggled to replicate the grassland phenology dynamics, both under ambient
533 conditions and in response to climate change treatments. With the exception of the CLM4.5
534 phenology scheme, most models predicted the growing season length in line with the
535 observed, but this blanket statement ignores some notable gross errors. A number of the
536 models were late in predicting the start of the growing season, often by as much as a month,
537 because they over-estimated the temperature required to initiate growth in this cold-temperate
538 grassland. The models that determine leaf senescence based solely on the ambient
539 temperature, did not predict the observed CO_2 effect on soil water that maintained growth in
540 some years (Reyes-Fox *et al.*, 2014). Two of the models (CABLE, SDGVM) do not simulate

true deciduous behaviour. These failures suggest that the triggers for growth and senescence in these models need to be re-examined.

In this ecosystem, vegetation greenness (a proxy for LAI) was highly dynamic in response to soil water availability (Fig. 2). The models, in contrast, are not as responsive to soil water availability and do not depict a clear threshold change in greenness with water stress. There is a clear need to improve our quantitative understanding of the mechanisms that determine the water-related dynamics of canopy greenness and senescence in grassland ecosystems.

There has been considerable work done on applying model-data fusion techniques to satellite-derived estimates of LAI, fractional cover and more recently, PhenoCams to improve predictions of LAI (Richardson *et al.*, 2009; Knorr *et al.*, 2010; Migliavacca *et al.*, 2011). For example, Hufkens *et al.* (2016) optimised a model to PhenoCam data from 14 North American grassland sites and demonstrated that a single parameterisation was able to capture the dynamics of changes in grassland fractional cover. Models could look to these studies to determine parameters constrained by data for their phenology models. However, Hufkens *et al.* (2016) did not consider the effect of eCO_2. Our results show that the models are not able to currently translate any CO_2-induced soil water savings into extended growing seasons, which has obvious consequences for predicting responses to future global change. In models that do account for soil water status when determining leaf drop (O-CN, ORCHIDEE, LPJ-GUESS), the threshold is arbitrarily defined. Phenology datasets from manipulative experiments, along with measurements of soil water status, could be used to inform this key process using similar data-model fusion approaches.

562 A further reasons for the smooth phenology simulated by models, relates to the use of a long-
563 term carbon storage pool. This pool effectively dampens day-to-day dynamics and whilst a
564 desierable process, the models currently lack fundamental controls on growth (e.g. meristems)
565 which are independent of carbon fixed through photosynthesis. The models are also unable to
566 rapidly shift allocation patterns between pools in response to changing environmental
567 conditions, such as allowing browning in dry conditions.

568 A related issue is the lack of crown biomass data. Crown biomass is a key ecosystem
569 component, acting as the principal store of reserve carbohydrates in grassland ecosystems;
570 however, it is difficult to quantify. Estimated values during the experiment ranged from < 50-
571 500 g m^{-2} and in the 2013 final harvest averaged 260 g m^{-2} (Nelson et al. in prep). Data used
572 in this study did not account for the crown biomass component, which may have biased
573 inferred allocation fractions. Assuming that including crown biomass would have doubled
574 root biomass estimates, the below- vs. above-ground allocation would be considerably
575 increased (0.52:0.48), compared to results presented in Fig. 5 (0.36:0.64).

576 *Available nitrogen*

577 Among the N cycle models, a key cause of disagreement was the simulated size of the
578 available N pools at the start of the experiment. This issue was raised previously (Zaehle *et*
579 *al.*, 2014), but the impact of model predictions is more apparent in this inter-comparison. Key
580 differences in how the N cycle is implemented, including the processes that govern the
581 amount of N fixation, the flexibility of plant stoichiometry and the ability of the models to
582 increase N uptake, affect the initial N stocks through model spin up and during the course of
583 the manipulation experiment. To constrain these differences among the models would require

584 a more complete observational record of both the N site history and the N budget. Whilst
585 there were site measurements of plant C, N, P ratios (Dijkstra *et al.*, 2012; Mueller *et al.*,
586 2016), these data are not sufficient to constrain a number of the key disagreements in the
587 change in N dynamics simulated in this study. Experimental measurements of N
588 mineralisation rates, N uptake, nitrification/denitrification rates and biological N fixation,
589 would greatly help to better constrain model uncertainties.

590 *Inter-annual variability*

591 Despite models being broadly able to capture ambient inter-annual variability (IAV) in aNPP
592 ($r^{-2} > 0.74$), they were seemingly unable to simulate observed treatment effects on IAV
593 (noting the large observed treatment uncertainties). Directly assessing the models' ability to
594 simulate observed treatment changes in IAV is not straightforward because it is not clear how
595 the timing of growth relates to the timing of photosynthetic uptake. At the extreme, a number
596 of models assume that one year's growth is entirely a product of the previous year's carbon
597 uptake and thus meteorology. Other models modulate the growth-productivity relationships
598 through the use of a labile C store. As a result, attempting to directly compare modelled time-
599 courses to growth observations is unproductive. To make progress we need more
600 experimental insight into the time lag between productivity and growth. In this experiment, as
601 is common, biomass and N concentration measurement were taken at the annual peak (mid-
602 July). These measurements do not offer a constraint as we cannot separate direct responses
603 from lagged effects.

C_3 vs C_4 competition

During the course of the experiment there were notable shifts in species dynamics. C_4 species initially prospered at the start of the experiment (Morgan et al., 2011) but did worse than C3 species in the later years (Zelikova et al., 2015; Mueller et al., 2016). This shift is an important result with implications for future predictions of species composition and ecosystem function. In this study models which had the capacity to simulate competition (JULES, LPJ-GUESS and SDGVM) did not do so they could be compared to other models without this functionality. Therefore, there remains an opportunity to further exploit the PHACE experimental data to test models that simulate C_3 vs. C_4 competition and to determine if the experimental results are predictable. However, for such a comparison to be meaningful, the key identified issues with existing models when applied to this site will need to be tackled first.

Modelling in advance of experiments

In advance of the PHACE experiment, Parton *et al.* (2007) carried out a novel study in which they used DAYCENT to predict grassland responses to treatments. Studies like this can help identify testable predictions against which hypotheses can then be compared (Norby *et al.*, 2016). Nevertheless, the Parton *et al.* (2007) study only used a single model, whereas a multi-model comparison (cf. Medlyn *et al.* (2016)) would have identified a greater range of processes in which models differed as this study demonstrates. *A priori* identification of areas where models diverge could have better helped guide experimentalists as to what key measurements would have helped constrain these model uncertainties. We strongly advocate the use of multi-model comparisons in advance of ecosystem scale experiments (Medlyn *et*

626 *al.*, 2016; Norby *et al.*, 2016); these studies need to become normal practice, rather than the
627 exception.

628 *Evaluation of models against multi-factor experiments*

629 Comparison of the models against the PHACE data has thus resulted in a clear agenda for
630 improving model predictions of grassland response to environmental change. Interestingly,
631 however, the multi-factor nature of the experiment did not add greatly to the model
632 evaluation. Global change will not affect a single factor in isolation, and thus it is widely
633 advocated that multi-factor experiments be used to probe future changes in the terrestrial
634 biosphere (Heimann & Reichstein, 2008; Luo *et al.*, 2008; Leuzinger *et al.*, 2011; Dieleman *et*
635 *al.*, 2012). In our study, however, the multi-factor comparison yielded little additional
636 constraint on model responses, for several reasons.

637 One of the main reasons that multi-factor experiments are commonly advocated is the need to
638 examine whether the main effects are additive or not when combined (Dieleman *et al.*, 2012;
639 Mueller *et al.*, 2016). However, models rarely predict additive effects; rather, they predict
640 non-linear interactions, which can sometimes be too small to be detectable. In this study,
641 models did not predict consistent interactions in response to combined treatments. Most
642 models, in line with the observations, predicted greater than additive interactions in some
643 years for both above- and below-ground biomass responses. Thus, determining whether or not
644 main effects are additive is of little help to constrain models.

645 Interactive effects in multi-factor experiments, particularly those carried out in environments
646 that experience marked inter-annual variability in precipitation, are complex to interpret and it

can be very challenging to identify the mechanisms underlying causing the observed responses. This statement is also true of the PHACE experiment, where treatment responses are overlaid on a marked year-to-year variability in responses to meteorology. Without a good causal understanding of the underlying processes, it is difficult to draw mechanistic understanding from the experiment that can be used to inform models.

However, the principal reason that the interacting responses did not help to constrain the models was because the models were unable to replicate the observed ecosystem behaviour under ambient conditions, or in response to single factor treatments. Since the interactive responses are contingent on key environmental factors such as soil water content and species composition, the models have to be able to realistically simulate these factors for their interactive effects to be comparable against data. Thus, at this stage, the most important way forwards is to use experimental data to improve model simulations of ambient conditions and responses to main effects (Norby & Luo, 2004). Future, improved, models, which are better able to simulate grassland phenology and can represent C_3 and C_4 competition, will likely find that the PHACE multi-factor dataset can provide a further constraint on our ability to predict response to global change.

Acknowledgements

Contributions from MDK, APW, XJY, KL and RJN were supported by the U.S. Department of Energy Office of Science Biological and Environmental Research program. SZ was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (QUINCY; grant no. 647204). XJL's contribution is supported by the CSIRO postdoctoral fellowship. We thank numerous individuals who

made the PHACE experiment possible, especially Dana Blumenthal, Dan LeCain, Eric Hardy and David Smith. We also thank Kevin Mueller for sharing biomass data.

References

Ahlström A, Raupach MR, Schurgers G et al. (2015) The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink. *Science*, **348**, 895–899.

Carrillo Y, Dijkstra FA, Pendall E, Morgan JA, Blumenthal DM (2012) Controls over soil nitrogen pools in a semiarid grassland under elevated CO_2 and warming. *Ecosystems*, **15**, 761–774.

Carrillo Y, Dijkstra FA, LeCain D, Morgan JA, Blumenthal D, Waldron S, Pendall E (2014) Disentangling root responses to climate change in a semiarid grassland. *Oecologia*, **175**, 699–711.

Ciais P, Sabine C, Bala G et al. (2013) Working group i contribution to the intergovernmental panel on climate change fifth assessment report climate change 2013: The physical science basis. In: *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (ed al TFS et), pp. 465–570. Cambridge University Press.

Clark DB, Mercado LM, Sitch S et al. (2011) The joint UK land environment simulator (JULES), model description - part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, **4**, 701–722.

688 Cleveland CC, Townsend AR, Schimel DS et al. (1999) Global patterns of terrestrial
689 biological nitrogen (N2) fixation in natural ecosystems. *Global Biogeochemical Cycles*, **13**,
690 623–645.

691 De Kauwe MG, Medlyn BE, Zaehle S et al. (2013) Forest water use and water use efficiency
692 at elevated CO_2: A model-data intercomparison at two contrasting temperate forest FACE
693 sites. *Global Change Biology*, **19**, 1759–1779.

694 De Kauwe MG, Medlyn BE, Zaehle S et al. (2014) Where does the carbon go? A model–data
695 intercomparison of vegetation carbon allocation and turnover processes at two temperate
696 forest free-air CO_2 enrichment sites. *New Phytologist*, **203**, 883–900.

697 De Kauwe M, Zhou S-X, Medlyn B, Pitman A, Wang Y-P, Duursma R, Prentice I (2015) Do
698 land surface models need to include differential plant species responses to drought?
699 Examining model predictions across a mesic-xeric gradient in europe. *Biogeosciences*, **12**,
700 7503–7518.

701 Dentener F, Drevet J, Lamarque J et al. (2006) Nitrogen and sulfur deposition on regional and
702 global scales: A multimodel evaluation. *Global Biogeochemical Cycles*, **20**.

703 Desborough C, Pitman A, Iranneiad P (1996) Analysis of the relationship between bare soil
704 evaporation and soil moisture simulated by 13 land surface schemes for a simple non-
705 vegetated site. *Global and Planetary Change*, **13**, 47–56.

706 Dieleman WI, Vicca S, Dijkstra FA et al. (2012) Simple additive effects are rare: A
707 quantitative review of plant biomass and soil process responses to combined manipulations of
708 CO_2 and temperature. *Global Change Biology*, **18**, 2681–2693.

709 Dijkstra FA, Blumenthal D, Morgan JA, Pendall E, Carrillo Y, Follett RF (2010) Contrasting
710 effects of elevated CO_2 and warming on nitrogen cycling in a semiarid grassland. *New*
711 *Phytologist*, **187**, 426–437.

712 Dijkstra FA, Pendall E, Morgan JA et al. (2012) Climate change alters stoichiometry of
713 phosphorus and nitrogen in a semiarid grassland. *New Phytologist*, **196**, 807–815.

714 Dukes JS, Chiariello NR, Cleland EE et al. (2005) Responses of grassland production to
715 single and multiple global environmental changes. *PLoS Biol*, **3**, e319.

716 Franks P, Adams M, Amthor J et al. (2013) Sensitivity of plants to changing atmospheric CO_2
717 concentration: From the geological past to the next century. *New Phytologist*, **197**, 1077–
718 1094.

719 Fuhrer J (2003) Agroecosystem responses to combinations of elevated co 2, ozone, and global
720 climate change. *Agriculture, Ecosystems &, Environment*, **97**, 1–20.

721 Hadley M (1993) Grasslands for our world. (ed Baker MJ). SIR Publishing: Wellington.

722 Heimann M, Reichstein M (2008) Terrestrial ecosystem carbon dynamics and climate
723 feedbacks. *Nature*, **451**, 289–292.

724 Hovenden MJ, Newton P, Carran R et al. (2008) Warming prevents the elevated CO_2-induced
725 reduction in available soil nitrogen in a temperate, perennial grassland. *Global Change*
726 *Biology*, **14**, 1018–1024.

727 Hovenden MJ, Newton PC, Wills KE (2014) Seasonal not annual rainfall determines
728 grassland biomass response to carbon dioxide. *Nature*, **511**, 583–586.

729 Hufkens K, Keenan TF, Flanagan LB et al. (2016) Productivity of north american grasslands
730 is increased under future climate scenarios despite rising aridity. *Nature Climate Change*.

731 Jobbágy EG, Jackson RB (2000) The vertical distribution of soil organic carbon and its
732 relation to climate and vegetation. *Ecological applications*, **10**, 423–436.

733 Kimball B (2005) Theory and performance of an infrared heater for ecosystem warming.
734 *Global Change Biology*, **11**, 2041–2056.

735 Knapp AK, Smith MD (2001) Variation among biomes in temporal dynamics of aboveground
736 primary production. *Science*, **291**, 481–484.

737 Knorr W, Kaminski T, Scholze M, Gobron N, Pinty B, Giering R, Mathieu P-P (2010)
738 Carbon cycle data assimilation with a generic phenology model. *Journal of Geophysical*
739 *Research: Biogeosciences*, **115**.

740 Krinner G, Viovy N, Noblet-Ducoudré N de et al. (2005) A dynamic global vegetation model
741 for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, **19**,
742 GB1015.

743 Leuzinger S, Luo Y, Beier C, Dieleman W, Vicca S, Körner C (2011) Do global change
744 experiments overestimate impacts on terrestrial ecosystems? *Trends in Ecology and*
745 *Evolution*, **26**, 236–241.

746 Lieth HFH (1978) *Patterns of primary productivity in the biosphere*. Hutchinson Ross:
747 Stroudsberg, PA, pp.

748 Luo Y, Gerten D, Le Maire G et al. (2008) Modeled interactive effects of precipitation,
749 temperature, and [CO_2] on ecosystem carbon and water dynamics in different climatic zones.
750 *Global Change Biology*, **14**, 1986–1999.

751 Medlyn BE, Zaehle S, De Kauwe MG et al. (2015) Using ecosystem experiments to improve
752 vegetation models. *Nature Climate Change*, **5**, 528–534.

753 Medlyn BE, De Kauwe MG, Zaehle S et al. (2016) Using models to guide field experiments:
754 A priori predictions for the CO_2 response of a nutrient-and water-limited native eucalypt
755 woodland. *Global Change Biology*, **22**, 2834–2851.

756 Melillo J, McGuire A, Kicklighter D, Moore B, Vorosmarty C, Schloss A (1993) Global
757 climate change and terrestrial net primary production. *Nature*, **363**, 234–240.

758 Migliavacca M, Galvagno M, Cremonese E et al. (2011) Using digital repeat photography and
759 eddy covariance data to model grassland phenology and photosynthetic CO_2 uptake.
760 *Agricultural and Forest Meteorology*, **151**, 1325–1337.

761 Miglietta F, Hoosbeek M, Foot J et al. (2001) Spatial and temporal performance of the
762 miniface (free air CO_2 enrichment) system on bog ecosystems in northern and central europe.
763 *Environmental Monitoring and Assessment*, **66**, 107–127.

764 Morgan J, Lecain D, Pendall E et al. (2011) C4 grasses prosper as carbon dioxide eliminates
765 desiccation in warmed semi-arid grasslands. *Nature*, **476**, 202–205.

766 Mueller K, Blumenthal D, Pendall E et al. (2016) Impacts of warming and elevated CO_2 on a
767 semi-arid grassland are non-additive, shift with precipitation, and reverse over time. *Ecology*
768 *Letters*, **19**, 956–966.

769 Norby RJ, Luo Y (2004) Evaluating ecosystem responses to rising atmospheric CO_2 and
770 global warming in a multi-factor world. *New Phytologist*, **162**, 281–293.

771 Norby RJ, De Kauwe MG, Domingues TF et al. (2016) Model–data synthesis for the next
772 generation of forest free-air CO_2 enrichment (FACE) experiments. *New Phytologist*, **209**, 17–
773 28.

774 Oleson KW, Lawrence DM, Bonan GB et al. (2013) *Technical description of version 4.5 of*
775 *the community land model (CLM)*. National Center for Atmospheric Research. Climate;
776 Global Dynamics Division; Citeseer, National Center for Atmospheric Research, P.O. Box
777 3000, Boulder, Colarado, pp.

778 Parton WJ, Schimel DS, Cole C, Ojima D (1987) Analysis of factors controlling soil organic
779 matter levels in great plains grasslands. *Soil Science Society of America Journal*, **51**, 1173–
780 1179.

781 Parton WJ, Morgan JA, Wang G, Del Grosso S (2007) Projected ecosystem impact of the
782 prairie heating and CO_2 enrichment experiment. *New Phytologist*, **174**, 823–834.

783 Pendall E, Heisler-White JL, Williams DG, Dijkstra FA, Carrillo Y, Morgan JA, LeCain DR
784 (2013) Warming reduces carbon losses from grassland exposed to elevated atmospheric
785 carbon dioxide. *PLOS ONE*, **8**, e71921.

786 Pepper D, Del Grosso S, McMurtrie R, Parton W (2005) Simulated carbon sink response of
787 shortgrass steppe, tallgrass prairie and forest ecosystems to rising [CO_2], temperature and
788 nitrogen input. *Global Biogeochemical Cycles*, **19**, GB1004.

789 Reyes-Fox M, Steltzer H, Trlica M, McMaster GS, Andales AA, LeCain DR, Morgan JA
790 (2014) Elevated CO_2 further lengthens growing season under warming conditions. *Nature*,
791 **510**, 259–262.

792 Richardson AD, Hollinger DY, Dail DB, Lee JT, Munger JW, O'keefe J (2009) Influence of
793 spring phenology on seasonal and annual carbon balance in two contrasting new england
794 forests. *Tree physiology*, **29**, 321–331.

795 Riedo M, Gyalistras D, Grub A, Rosset M, Fuhrer J (1997) Modelling grassland responses to
796 climate change and elevated CO_2. *Acta Oecologica*, **18**, 305–311.

797 Ryan EM, Ogle K, Zelikova TJ, LeCain DR, Williams DG, Morgan JA, Pendall E (2015)
798 Antecedent moisture and temperature conditions modulate the response of ecosystem
799 respiration to elevated CO_2 and warming. *Global change biology*.

800 Scurlock J, Hall D (1998) The global carbon sink: A grassland perspective. *Global Change*
801 *Biology*, **4**, 229–233.

802 Shaw MR, Zavaleta ES, Chiariello NR, Cleland EE, Mooney HA, Field CB (2002) Grassland
803 responses to global environmental changes suppressed by elevated CO_2. *Science*, **298**, 1987–
804 1990.

805 Verhoef A, Egea G (2014) Modeling plant transpiration under limited soil water: Comparison
806 of different plant and soil hydraulic parameterizations and preliminary implications for their
807 use in land surface models. *Agricultural and Forest Meteorology*, **191**, 22–32.

808 Walker AP, Beckerman AP, Gu L et al. (2014) The relationship of leaf photosynthetic traits –
809 V_{cmax} and J_{cmax} – to leaf nitrogen, leaf phosphorus, and specific leaf area: A meta-analysis
810 and modeling study. *Ecology and Evolution*, **4**, 3218—3235.

811 Walker AP, Zaehle S, Medlyn BE et al. (2015) Predicting long-term carbon sequestration in
812 response to CO_2 enrichment: How and why do current ecosystem models differ? *Global*
813 *Biogeochemical Cycles*, **29**, 476–495.

814 Wang Y-P, Trudinger CM, Enting IG (2009) A review of applications of model-data fusion to
815 studies of terrestrial carbon fluxes at different scales. *Agricultural and Forest Meteorology*,
816 **149**, 1829–1842.

817 Zaehle S, Medlyn BE, De Kauwe MG et al. (2014) Evaluation of 11 terrestrial carbon–
818 nitrogen cycle models against observations from two temperate free-air CO_2 enrichment
819 studies. *New Phytologist*, **202**, 803–822.

Zelikova TJ, Williams DG, Hoenigman R, Blumenthal DM, Morgan JA, Pendall E (2015) Seasonality of soil moisture mediates responses of ecosystem phenology to elevated CO_2 and warming in a semi-arid grassland. *Journal of Ecology*, **103**, 1119–1130.

Zhou S, Duursma RA, Medlyn BE, Kelly JW, Prentice IC (2013) How should we model plant responses to drought? An analysis of stomatal and non-stomatal responses to water stress. *Agricultural and Forest Meteorology*, **182-183**, 204–214.

Zhou S, Medlyn B, Sabaté S, Sperlich D, Prentice IC (2014) Short-term water stress impacts on stomatal, mesophyll and biochemical limitations to photosynthesis differ consistently among tree species from contrasting climates. *Tree physiology*, **10**, 1035–1046.

Figure Captions

Figure 1: Annual and early- to mid-growing season (day of year: 100-200) when soil water availability most limits productivity (Morgan *et al.*, 2011). In 2006 all plots were irrigated (20 mm × 8) with 160 mm of additional water. The additional water is shown by the precipitation above the black horizontal line in 2006. The annual bar shows the effect of the eight additional treatments, whereas the early- to mid-growing season bar shows the addition of the six treatments which occurred during that period.

Figure 2. Scatter plot showing the observed and modelled aNPP in the control (ct) treatment. Vertical errorbars (one standard deviation) represent cross plot (N=5) variability in observed aNPP. Note, the SDGVM model (panel j) is shown on a different x-axis range (0-700 vs. 0-350). ME is the Nash-Sutcliffe model efficiency coefficient (-∞ to 1), where 1 would indicate

perfect agreement with the observed aNPP. CI is the 95% confidence interval for the modelled values and r^2 is the coefficient of determination.

Figure 3: Greenness (number of green pixels) derived from bi-weekly digital photographs and the corresponding soil moisture content (top 20 cm) in the ambient plots. Greenness observations are shown with filled circles, with a fitted spline to aid visual interpretation. Soil moisture data represent the plot means (solid line) and minimum and maximum from the 5 ambient plots (shaded area).

Figure 4: Modelled leaf area index (LAI) from the ambient (ct) treatment, shown by sequential colours from yellow to dark green, which corresponds to years between 2007 and 2012. Grey shading indicates the range of leaf out and leaf off dates calculated from the control (ct) treatment (Reyes-Fox *et al.*, 2014).

Figure 5: Fraction of Net Primary Productivity (NPP) allocated above-, below-ground and to reproduction in the control (ct) treatment.

Figure 6: Modelled soil water profile in a dry (2008) and a wet year (2009).

Figure 7: Summer (June, July, August) soil water availability factor (β) in the control (ct), CO_2 (Ct), warming (cT) and CO_2 × warming (CT) treatments. Error bars show summer inter-annual variability across the experimental years.

Figure 8: Response of aNPP to CO_2 (Ct), warming (cT) and CO_2 × warming (CT) for C_3 species. Error bars on the Ct and cT observed treatments denote one standard error.

859 Horizontal lines on the CT treatment bars, show the estimated interactive terms if this
860 interaction was additive.

861 Figure 9: Response of aNPP to CO_2 (Ct), warming (cT) and CO_2 × warming (CT) for C_4
862 species. Error bars on the Ct and cT observed treatments denote one standard error.
863 Horizontal lines on the CT treatment bars, show the estimated interactive terms if this
864 interaction was additive.

865 Figure 10: Response of root biomass to CO_2 (Ct), warming (cT) and CO_2 × warming (CT).
866 Error bars on the Ct and cT observed treatments denote one standard error. Horizontal lines
867 on the CT treatment bars, show the estimated interactive terms if this interaction was additive.

868 Figure S1: Ratio of above– and below–ground biomass in the control (ct) treatment.

869 Figure S2: Reduction in gas exchange (β) with declining soil moisture content (θ) in 2007 and
870 2009

871 Figure S3: Response of nitrogen use efficiency to CO_2 (Ct), warming (cT) and CO_2 ×
872 warming (CT).

873 Figure S4: Response of nitrogen uptake to CO_2 (Ct), warming (cT) and CO_2 × warming (CT).

874 Figure S5: Response of nitrogen mineralisation to CO_2 (Ct), warming (cT) and CO2 ×
875 warming (CT).

876 Table 1: Summary of model phenology and growth assumptions. C is carbon, GDD is the number of growing degree-days, GDD5 is the number
877 of growing degree days above 5°C, GPP is gross primary productivity, LAI is leaf area index, maxGDD is the a maximum growing degrees day
878 threshold, N is nitrogen, NPP is net primary productivity, PAR is the photosynthetically active radiation SLA is the specific leaf area and SWI is
879 soil water index.

Models	Leaf onset	Growth	Leaf drop	References
CABLE	Leaf onset is prescribed based on a satellite climatology, i.e. no inter-annual variability. Onset dates vary as a function of latitude.	After leaf onset, 80% of NPP is allocated to leaves for a 2-week period. Following this allocation to leaves is 20% of NPP until the period 2-weeks before leaf drop, in which NPP allocation to leaves is 0%.	Leaf drop is prescribed based on a satellite climatology, i.e. no inter-annual variability. Drop dates vary as a function of latitude.	Zhang et al. 2004
CLM4.5	GDD accumulation, SWI accumulation (accumulated matric potential above a 'onset' minimum, -2MPa, in the third soil layer), and day length >6hrs. Can occur	Taken from storage pool at a linearly decreasing rate.	Sustained period of dry soil or cold temperature, or day length	Oleson et al. 2013

	multiple times in a year.		shorter than 6 hours.	
DAYCENT	Leaf onset is prescribed to occur at a fixed date.	After growth begins, leaf and root growth comes from carbon stored in previous year growing season. Peak growth is determined by temperature, water and nutrient availability, and prescribed maximum LAI that controls leaf death due to shading.	Like leaf onset, leaf drop is prescribed.	Parton et al. (1993)
GDAY	Growth begins after exceeding both a precipitation and a GDD threshold. The precipitation threshold is 15% of the annual precipitation. GDD are calculated from the sum of mean daily air temperature above 0°C for cool and for 5°C warm grasses. The thresholds are 185 and 400 days for C_3 and C_4 grasses, respectively.	For deciduous species, leaf growth comes from carbon stored in the previous year growing season. It is assumed that all growth occurs before the mid-point of the growing season, after this point senescence begins. Both growth and litterfall occur with a linearly ramping rate. These assumptions result in a symmetrical growth dynamic.	Day of year ≥ 243 and mean daily air temperature is above 0°C for cool and for 5°C warm grasses. Soil water availability has no effect on litterfall in the deciduous model.	Foley et al. (1996), White et al. (1997).
ISAM	Growth begins when: (i) daily mean root zone temperature is higher than 10 °C for 14 days and (ii) daytime length is longer than 12 hours.	There are two growth stages: (i) the maximal growth stage, where more carbon is allocated to foliage to capture PAR and (ii) the normal growth stage, where more carbon is allocated to roots/stem to acquire resources. Plant enter normal growth stage when they	Leaf drop occurs when at least one of the following four conditions below is met: (i) water stress is greater than 40% for 14 days; (ii) daily mean root zone temperature lower than 10 °C and daytime length shorter	Song et al., (2013), El-Masri et al., (2015)

		LAI exceeds half of their potential maximal LAI (set to 3). In addition, if grassland enters leaf drop stage due to water stress, but it could re-enter the growth stage, if the water stress becomes lower than 40% and other conditions for leaf onset are still satisfied.	than 12 hours; (iii) LAI higher than the potential maximal LAI or; and (iv) plant maintains normal growth for longer than 120 days.	
JULES	Growth begins when the canopy temperature (Tc) is above a threshold (5°C).	The rate of growth is $\tau_p(1-L_b)$, where τ_p is a parameter (20 yr^{-1}), and L_b is the "balanced LAI", or the LAI the plant would have in full-leaf (allometrically related to height). Growth continues as long as the plant is assimilating carbon, until leaf area index reaches $L_{b,}$, while Tc>threshold T.	When Tc drops below the threshold temperature, leaf turnover rate is modified (see eq. 47 in Clark et al.)	Clark et al. 2011 – See Section 4; Cox et al. 2001
LPJ-GUESS	Leaf onset begins after exceeding a GDD sum threshold in LPJ-GUESS. However, grasses grow with a GDD threshold of 0 by default.	Growth is calculated at the end of a year. The annually integrated NPP is then allocated to leaves and roots, with a higher fraction allocated to roots under water and/or N limitation. Grasses are inactive under cold or very dry conditions. The maximum LAI (as calculated by carbon mass for leaves at the end of the previous year divided by a SLA) is scaled with a phenology development factor (GDD5 / maxGDD; maxGDD=100). For grasses, this scalar	Once a 30-day running average temperature falls below a threshold (5°C) the cumulative GDD5 counter is reset. In the simulation we also introduced a 60-day inhibition for the GDD5 counter preventing immediate increase after the senescence event was triggered.	Smith et al. (2014)

		is also zero at any days where plant-available soil water content falls below 35% of water holding capacity.		
O-CN	Growth begins after exceeding a GDD threshold above 5°C, subject to weekly moisture above 25% of field capacity and a positive trend in weekly soil moisture. The GDD requirement adjusts to long-term annual mean temperature, and was applied here at a value of 270 and 400 days for C_3 and C_4 grasses, respectively.	Growth is modeled using a functional balance approach between leaves, tillers, and fine roots, responding to moisture and N status. Growth is fuelled from a labile carbon pool, which is filled by current photosynthetic carbon uptake and a long-term reserve (past GPP). Once the incremental net carbon gain of the canopy goes negative, most growth is allocated to seed production.	The turnover time of leaves increases once weekly temperatures drop below -2/2°C (for C_3/C_4 grasses respectively) and weekly soil moisture below 10% of field capacity. Complete abscission within 10 days commences once weekly NPP becomes negative.	Krinner et al. 2005, Zaehle & Friend 2010, with unpublished updates.
ORCHIDEE	The leaf onset scheme follows Botta et al. (2000). Leaf onset scheme for tropical grass starts after a fixed number of days after the dry season's. For boreal regions, the number of growing degree days during the past few weeks has to exceed a prescribed threshold. For temperate grass, both criteria control the leaf onset.	Leaf growth starts using C stored in reserves tissues. Once the leaf starts to grow C is fixed by photosynthesis following Farquhar et al., (1980). Once the C is fixed, it is redistributed following an allocation scheme developed by Friedlingstein et al., (1998). This allocation scheme is controlled by biophysical limitations (light, water).	Two different criteria are used separately to calculate the fraction of dying leaves at each time step. i) a meteorological criterion controlled by temperature and water stress (temperature < 4°C for C_3 and 5°C for C_4 grasses; moisture > 20% for both). ii) the leaf age itself (>120 days).	Friedlingstein et al. (1998); Botta et al. (2000)

SDGVM	For evergreen vegetation leaf onset is triggered by a GDD accumulation subject to sufficient soil water.	Leaf growth comes from stored carbon and occurs at a constant rate until the target LAI is reached.	Leaf drop is triggered when leaves reach their parameterized age. Small amounts of litterfall occur every day as a function of leaf age.	Woodward and Lomas (2004)

880

Table 2: Causes of differences in modelled aNPP. Values shown are averages across the experiment in the ambient treatment. A_b is the aboveground allocation fraction, CUE is the carbon-use efficiency, GPP_{us} is the unstressed GPP per unit leaf areas, β is the water stress factor, D is the growing season duration, LAI_p is the growing season maximum LAI, aNPPc is the inferred aNPP which is the product of A_b, CUE, GPP_u, β, D/LAI_p and LAI_p, $aNPP_a$ is the actual model output for comparison.

Model	A_b (-)	CUE (-)	GPP_u (g C m^{-2} leaf d^{-1})	β (-)	D (d yr^{-1})	LAI_p (m^2 m^{-2})	$aNPP_c$ (g C m^{-2} ground y^{-1})	$aNPP_a$ (g C m^{-2} ground y^{-1})
CABLE	0.13	0.63	8.57	0.33	249.02	1.55	54.33	54.5
CLM5	0.55	0.67	6.27	0.6	155.79	2.99	203.27	197.85
DAYCENT	0.47	0.55	11.92	0.17	126.54	1.29	63.31	64.29
GDAY	0.46	0.5	4.71	0.74	104.07	1.88	82.05	88.16
ISAM	0.85	0.53	5.3	0.82	125.53	2.98	247.15	211.89
JULES	0.82	0.32	3.6	0.2	77.96	1.38	18.86	20.02
LPJ-GUESS	0.31	0.5	4.63	0.77	218.57	2.49	122.1	129.78
O-CN	0.52	0.52	4.81	0.84	169.93	3.08	185.62	246.2
ORCHIDEE	0.47	0.53	3.3	0.97	149.91	1.21	118.13	123.31
SDGVM	0.86	0.69	4.95	0.71	256.11	6.1	542.86	526.82

Table 3: Causes of differences in the modelled aNPP response to CO_2 for C_3 species. Values shown are averages across all years. GPP is enhancement expressed as a percentage, CUE is the carbon-use efficiency, expressed as a percentage, A_b is the percentage change above-ground allocation, B_g is the percentage change below-ground allocation and S is the percentage change in allocation to labile carbon storage.

Model	GPP (%)	CUE (%)	A_b (%)	B_g (%)	S (%)
CABLE	20.65	2.86	-4.13	-11.02	15.15
CLM5	-	-	-	-	0
DAYCENT	45.45	-12.2	0.72	-0.72	0
GDAY	39.13	0	-4.55	4.55	0
ISAM	55.13	-3.07	3.74	-3.74	0
JULES	72.62	5.06	-3.57	3.57	0
LPJ-GUESS	15.44	16.62	0.64	-0.64	0
O-CN	53.66	-11.32	2.41	-2.41	0
ORCHIDEE	31.21	4.92	1.59	-1.59	0
SDGVM	33.45	-2.05	-1.73	1.73	0

Table 4: Causes of differences in the modelled aNPP response to CO_2 for C_4 species. Values shown are averages across all years. GPP is enhancement expressed as a percentage, CUE is the carbon-use efficiency, expressed as a percentage, A_b is the percentage change above-ground allocation, B_g is the percentage change below-ground allocation and S is the percentage change in allocation to labile carbon storage.

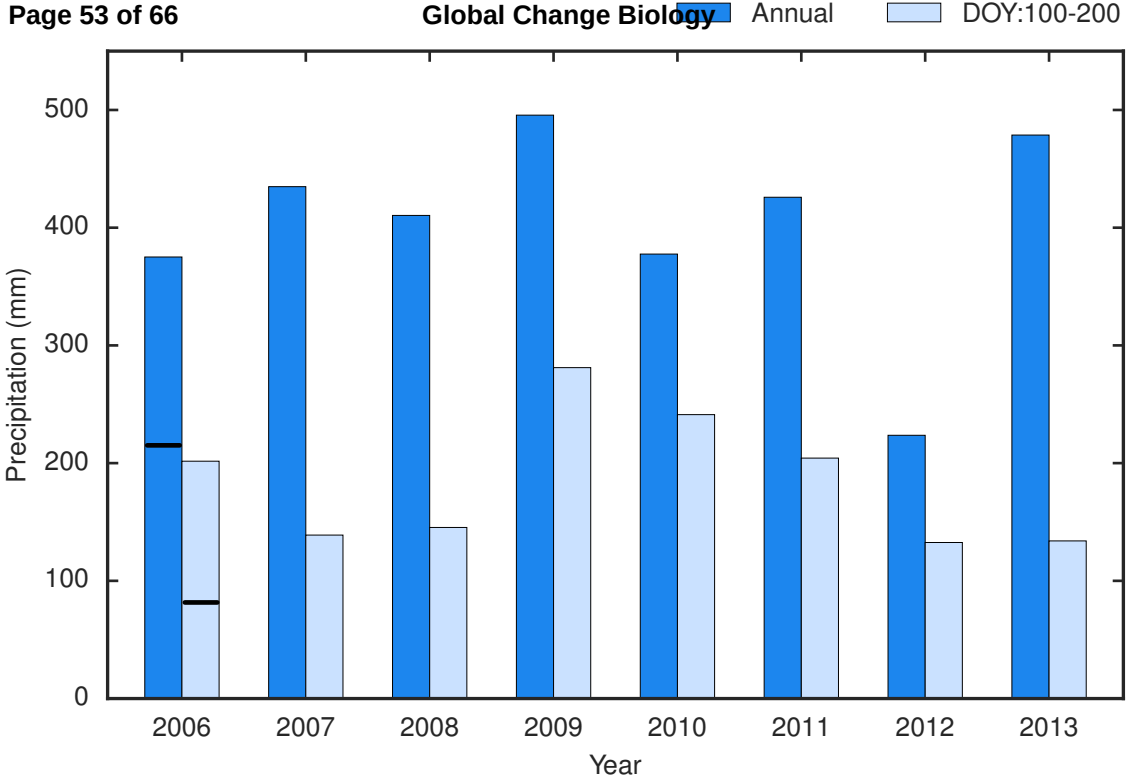
Model	GPP (%)	CUE (%)	A_b (%)	B_g (%)	S (%)
CABLE	22.42	2.98	-2.42	-11.47	13.89
CLM5	19.1	-1.72	0	0	0
DAYCENT	12.58	-4.53	0.17	-0.17	0
GDAY	16.85	0	-0.99	0.99	0
ISAM	9.43	2.7	-0.3	0.3	0
JULES	34.51	6.89	-0.87	0.87	0
LPJ-GUESS	26.37	4.69	-1.95	1.95	0
O-CN	6.8	-0.08	2.34	-2.34	0
ORCHIDEE	4.75	0.64	1.57	-1.57	0
SDGVM	10.15	-2.73	-2.38	2.38	0

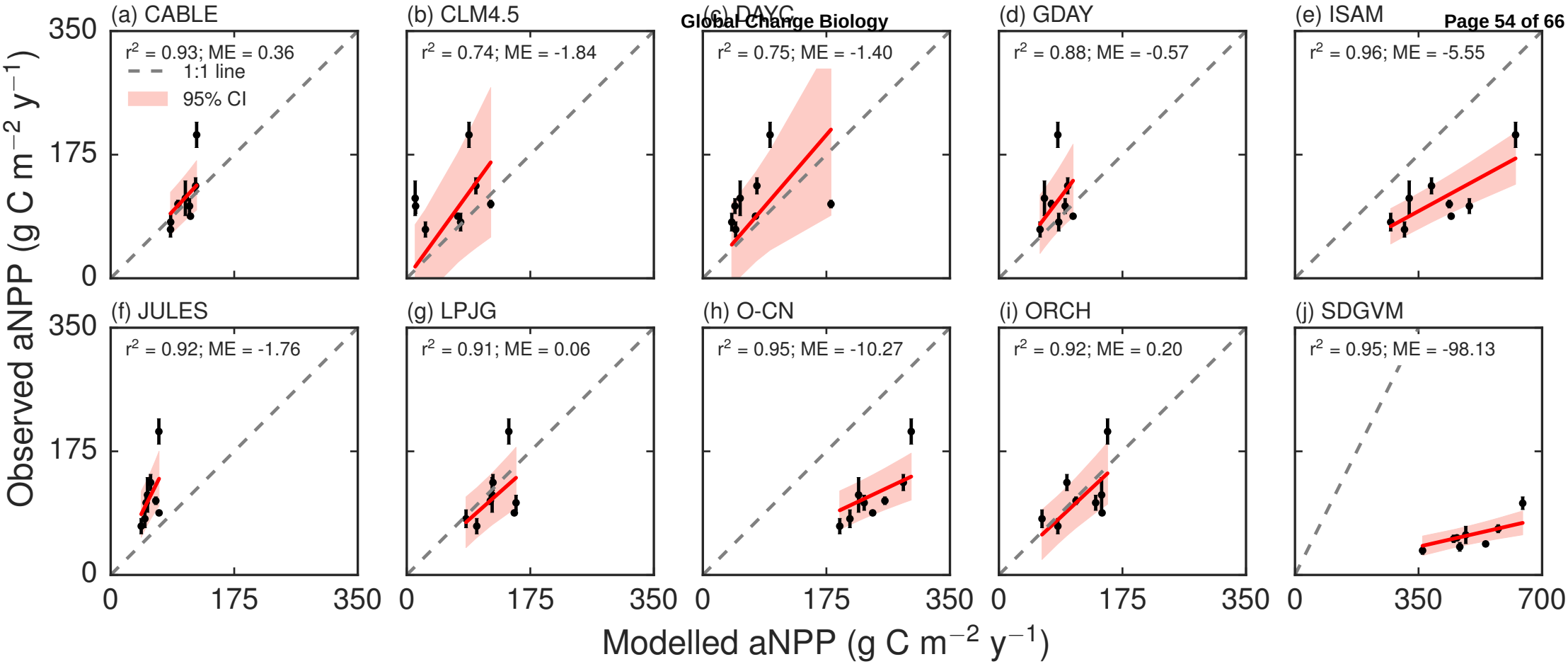
Table S1: Number of days change in leaf onset in the CO_2 (Ct), warming (cT) and CO_2 × warming treatments. Positive numbers indicate earlier onset dates. CABLE and SDGVM have been excluded, as they do not completely drop their leaves. CLM4.5 has also been excluded as the C_3 grasses did not grow and it is clear that the C_4 grass phenology does not work at this site (Fig. 3).

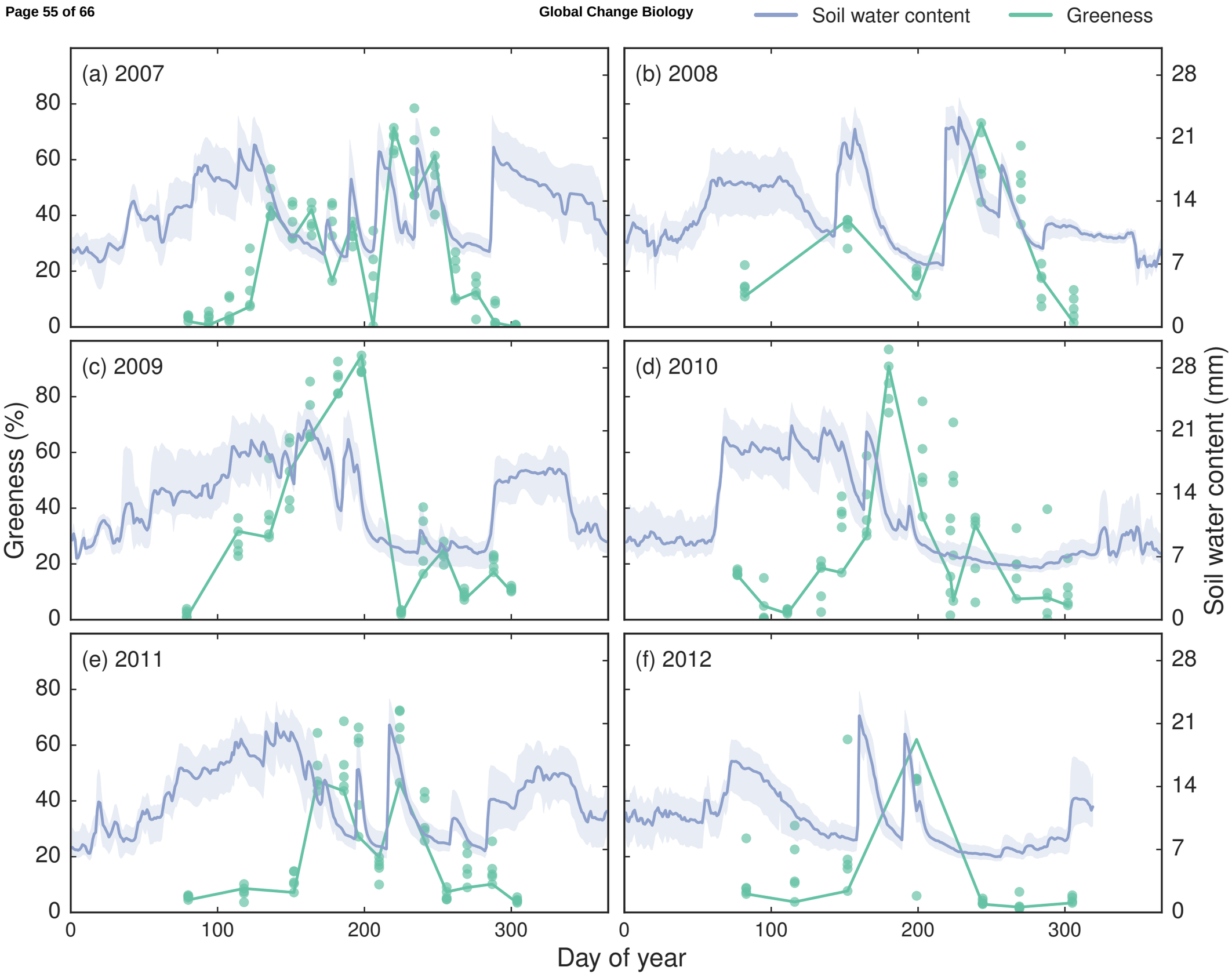
Model	Ct	cT	CT
DAYCENT	0.0	0.0	0.0
GDAY	0.0	21.7	21.7
ISAM	0.0	14.9	14.9
JULES	0.0	2.0	2.0
LPJ-GUESS	0.0	2.4	2.4
O-CN	0.0	24.3	24.3
ORCHIDEE	0.0	16.7	16.7

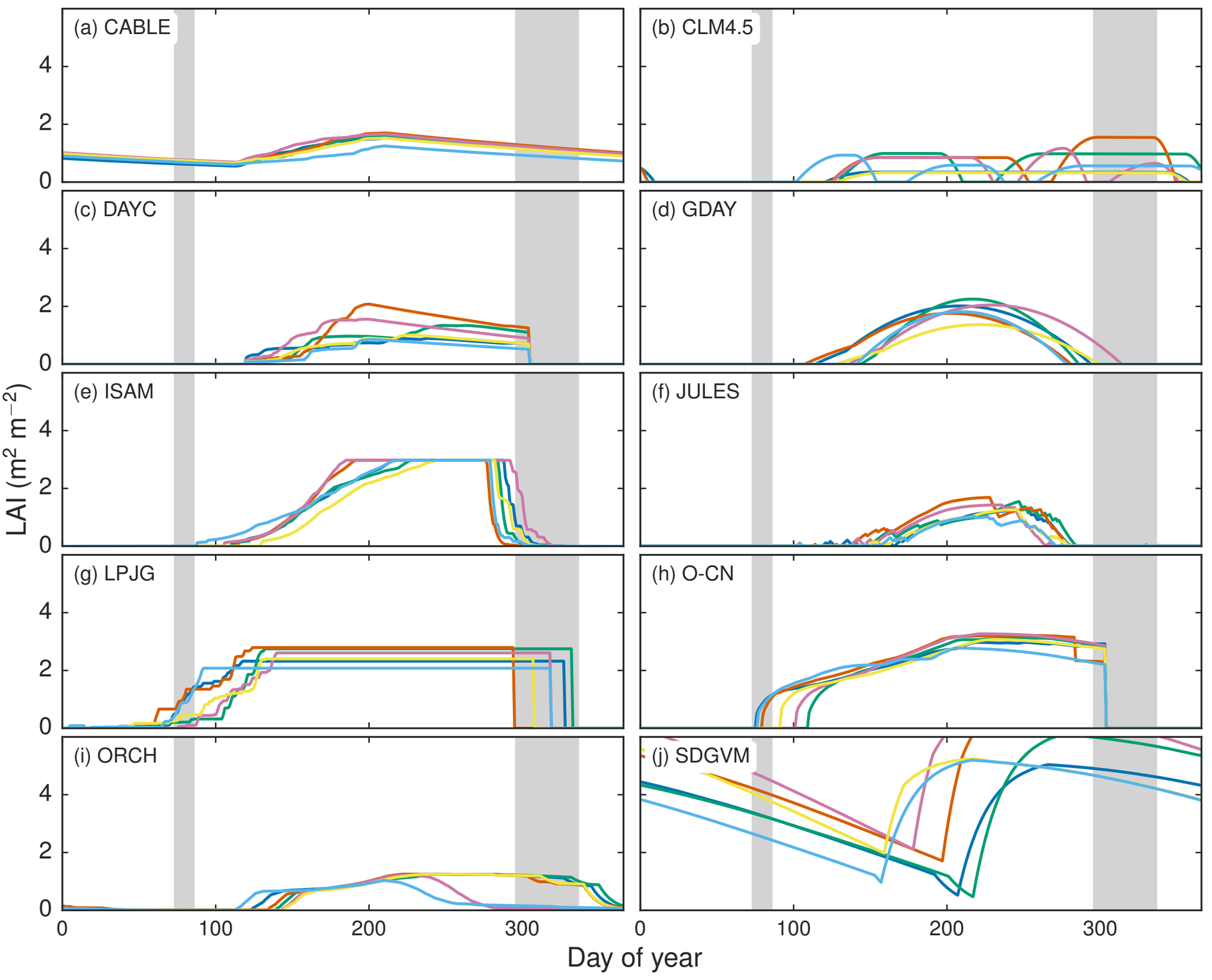
Table S2: Number of days change in leaf senescence in the CO_2 (Ct), warming (cT) and CO_2 × warming treatments. CABLE and SDGVM have been excluded, as they do not completely drop their leaves. CLM4.5 has also been excluded as the C_3 grasses did not grow and it is clear that the C_4 grass phenology does not work at this site (Fig. 3).

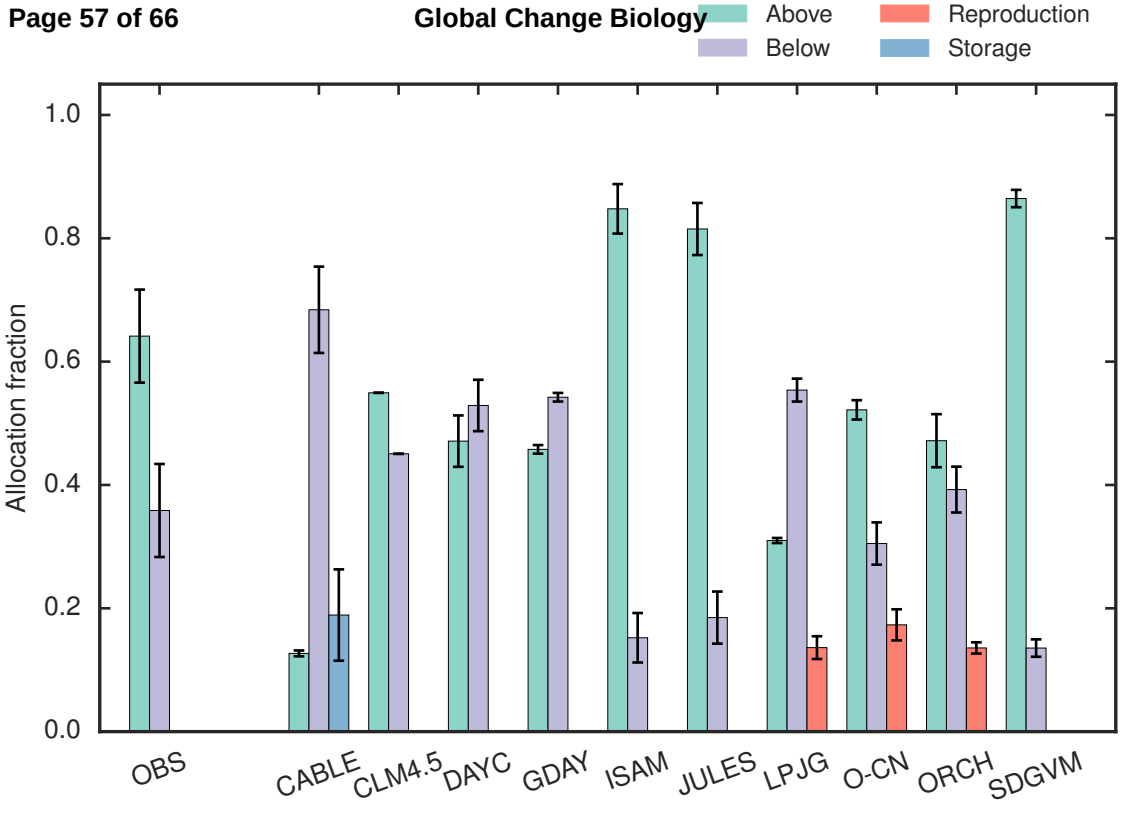
Model	Ct	cT	CT
DAYCENT	0.0	0.0	0.0
GDAY	0.0	14.8	14.8
ISAM	0.8	11.7	10.9
JULES	0.0	9.3	9.3
LPJ-GUESS	0.0	12.6	12.6
O-CN	0.0	0.0	0.0
ORCHIDEE	0.0	0.0	0.0

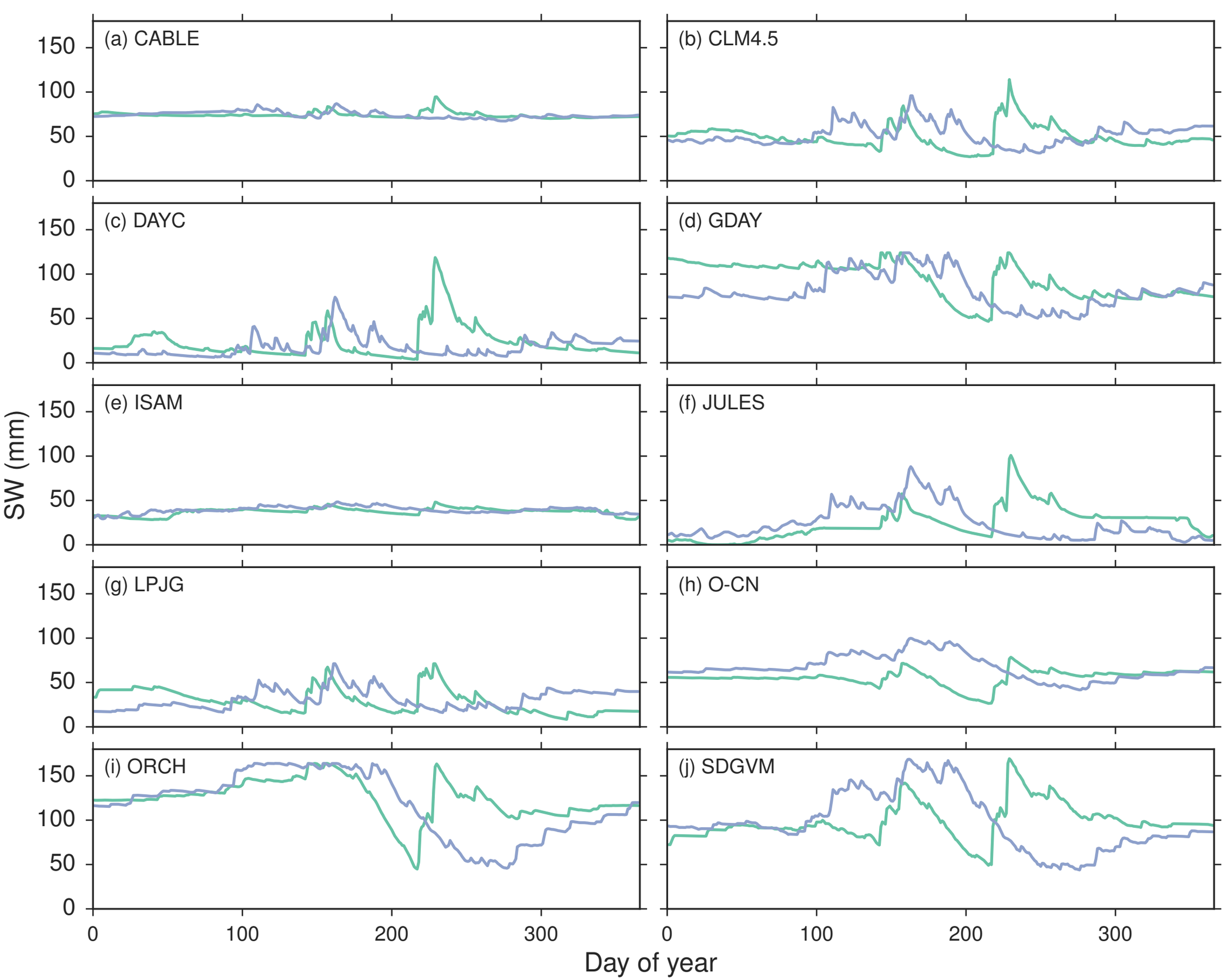

Annual
DOY:100-200
Precipitation (mm)
0
100
200
300
400
500
2006
2007
2008
2009
2010
2011
2012
2013
Year

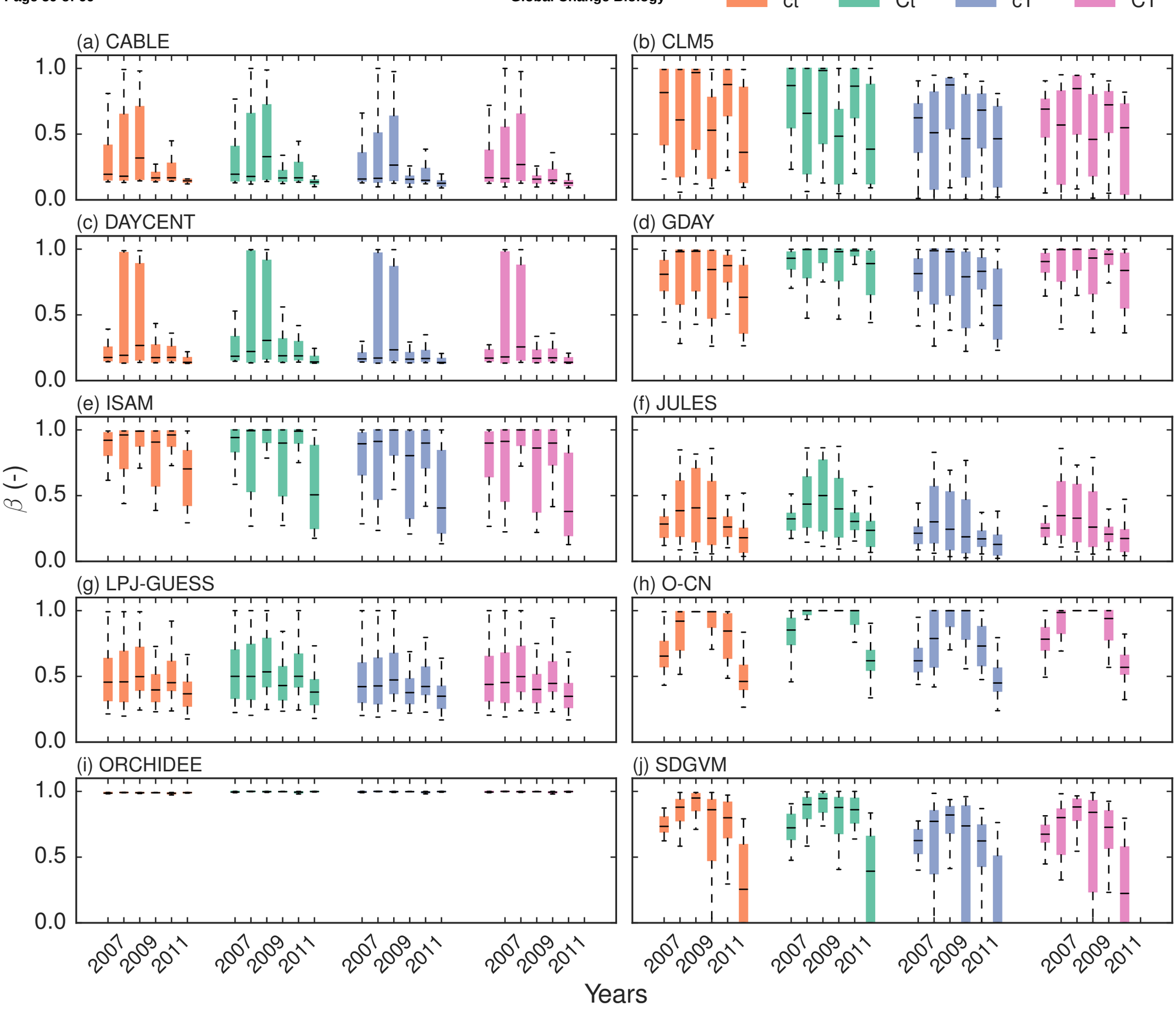

(a) CABLE
r² = 0.93; ME = 0.36
1:1 line
95% CI
(b) CLM4.5
r² = 0.74; ME = -1.84
(c) DAYC
r² = 0.75; ME = -1.40
(d) GDAY
r² = 0.88; ME = -0.57
(e) ISAM
r² = 0.96; ME = -5.55
(f) JULES
r² = 0.92; ME = -1.76
(g) LPJG
r² = 0.91; ME = 0.06
(h) O-CN
r² = 0.95; ME = -10.27
(i) ORCH
r² = 0.92; ME = 0.20
(j) SDGVM
r² = 0.95; ME = -98.13
Observed aNPP (g C m⁻² y⁻¹)
Modelled aNPP (g C m⁻² y⁻¹)
0
175
350
700

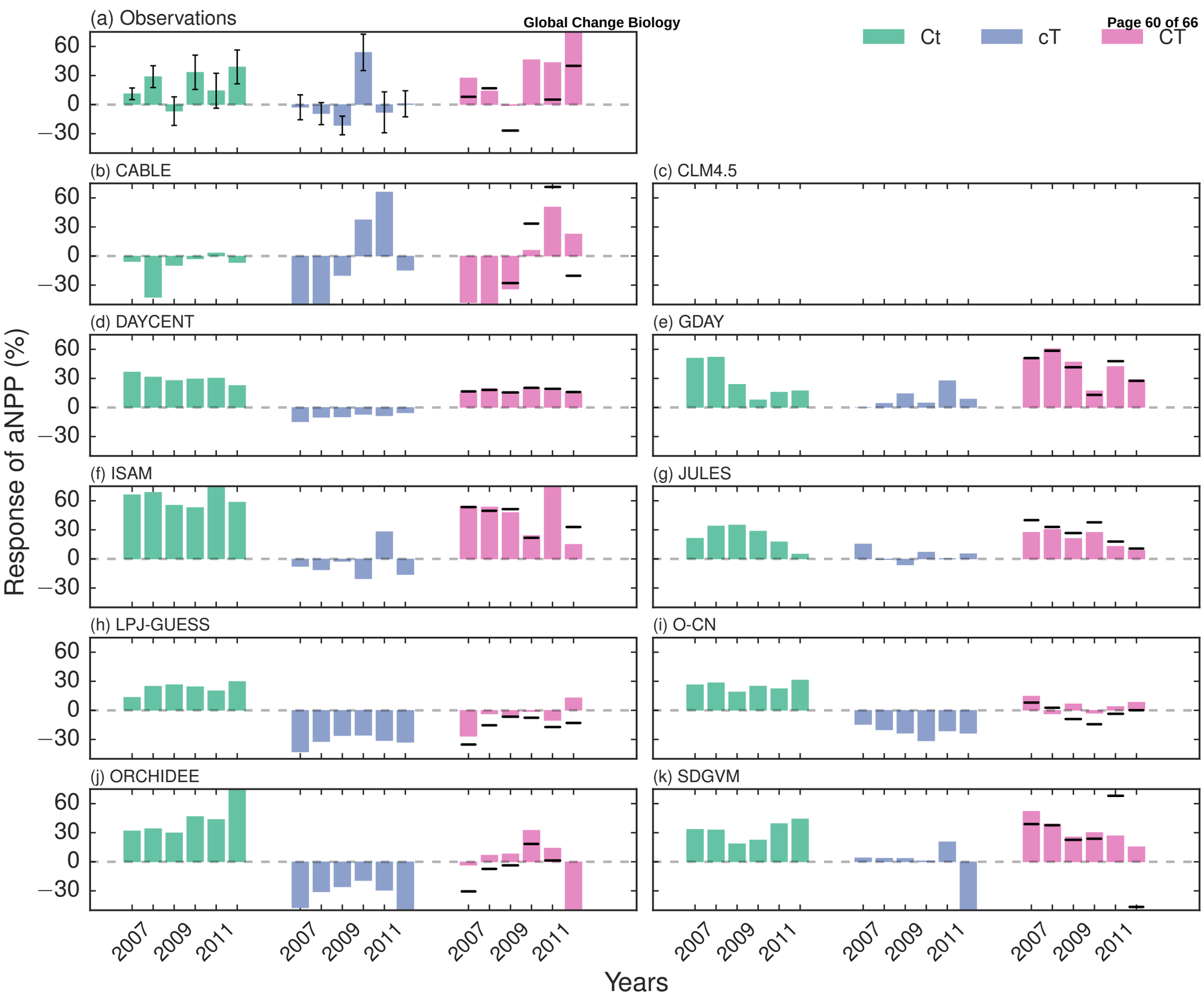

Soil water content
Greeness
(a) 2007
(b) 2008
(c) 2009
(d) 2010
(e) 2011
(f) 2012
Greeness (%)
Soil water content (mm)
Day of year
0
20
40
60
80
0
7
14
21
28
0
100
200
300

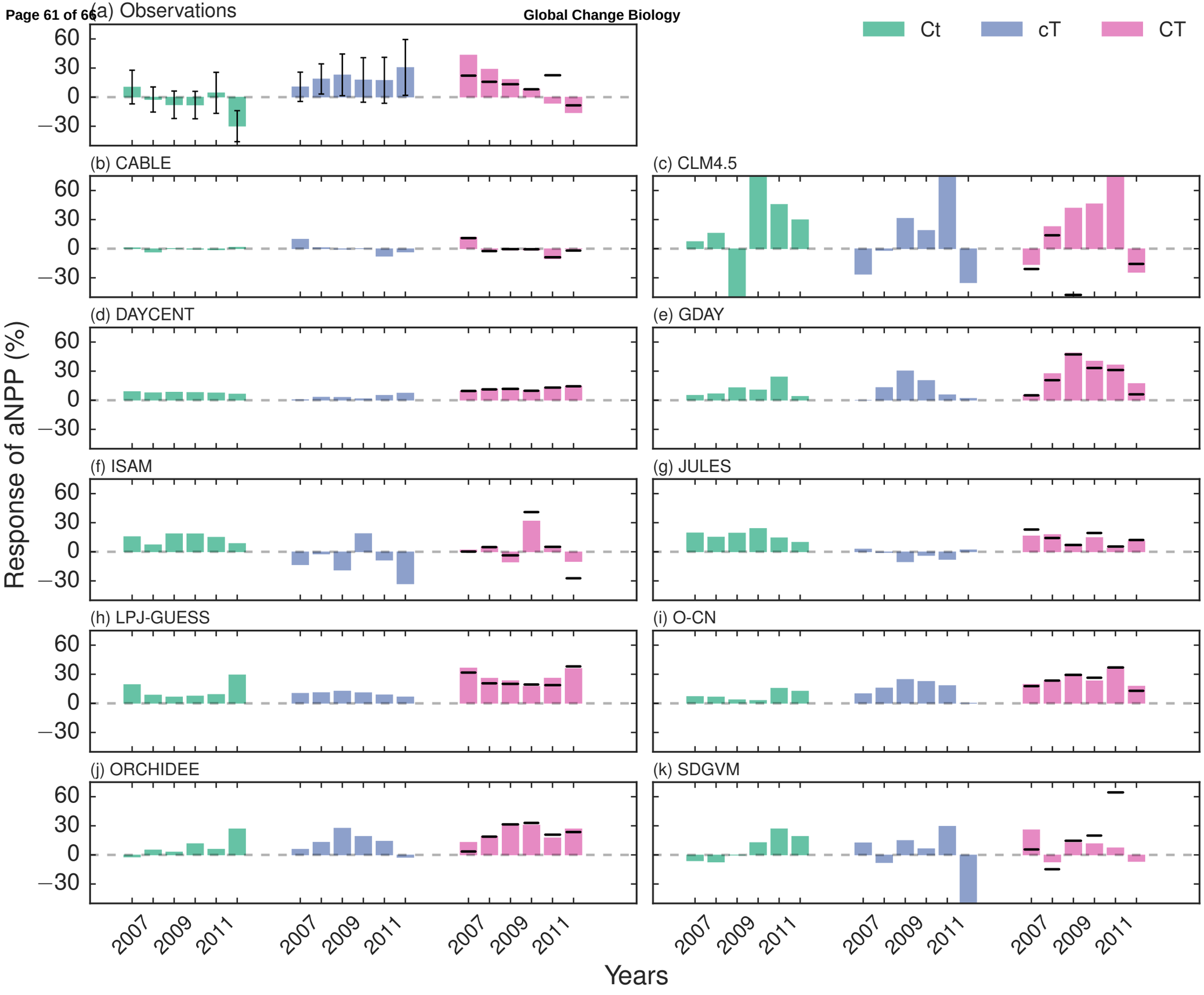

2007
2008
2009
2010
2011
2012
(a) CABLE
(b) CLM4.5
(c) DAYC
(d) GDAY
(e) ISAM
(f) JULES
(g) LPJG
(h) O-CN
(i) ORCH
(j) SDGVM
LAI (m2 m−2)
Day of year
0
2
4
100
200
300

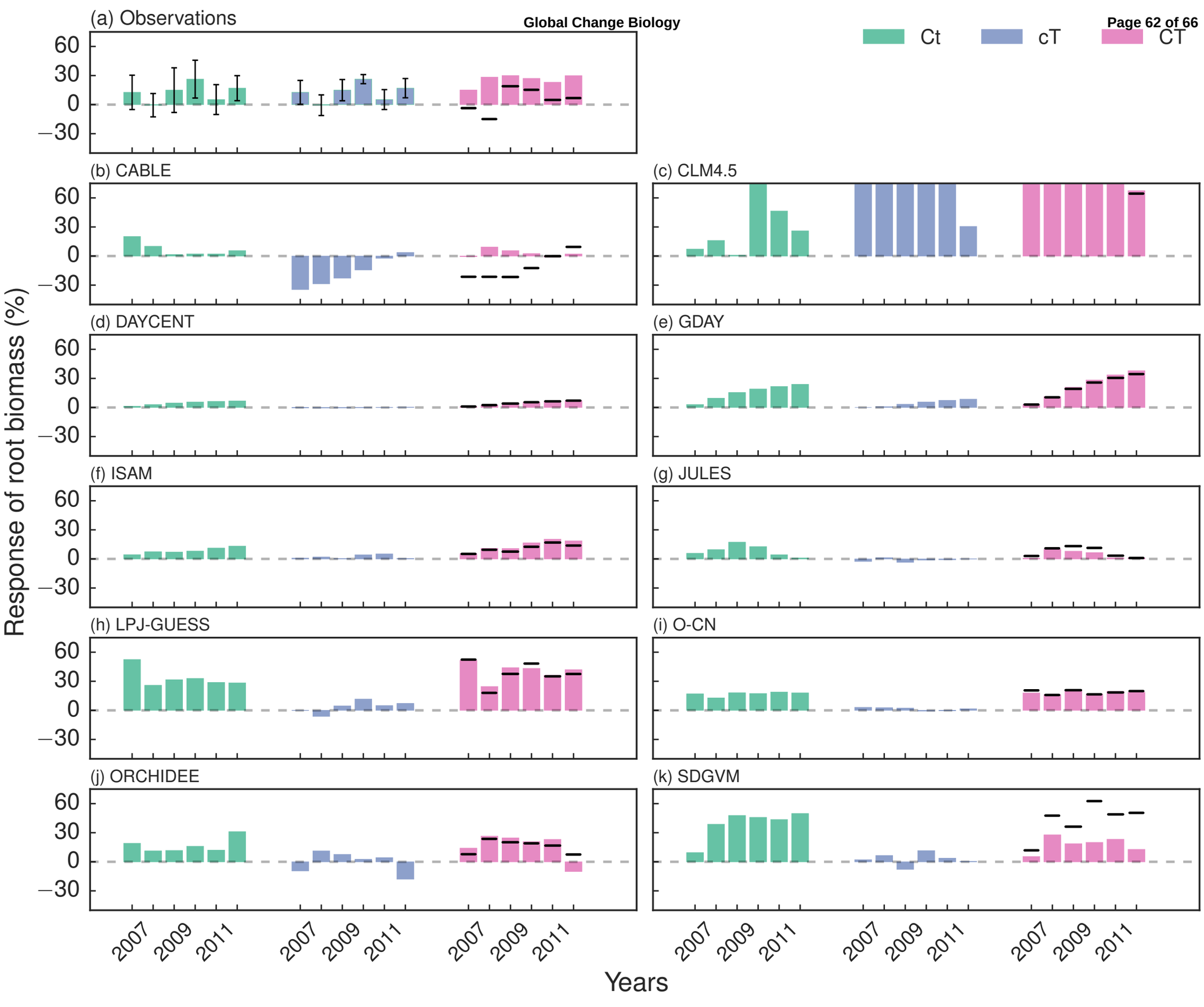

Above
Below
Reproduction
Storage
Allocation fraction
1.0
0.8
0.6
0.4
0.2
0.0
OBS
CABLE
CLM4.5
DAYC
GDAY
ISAM
JULES
LPJG
O-CN
ORCH
SDGVM

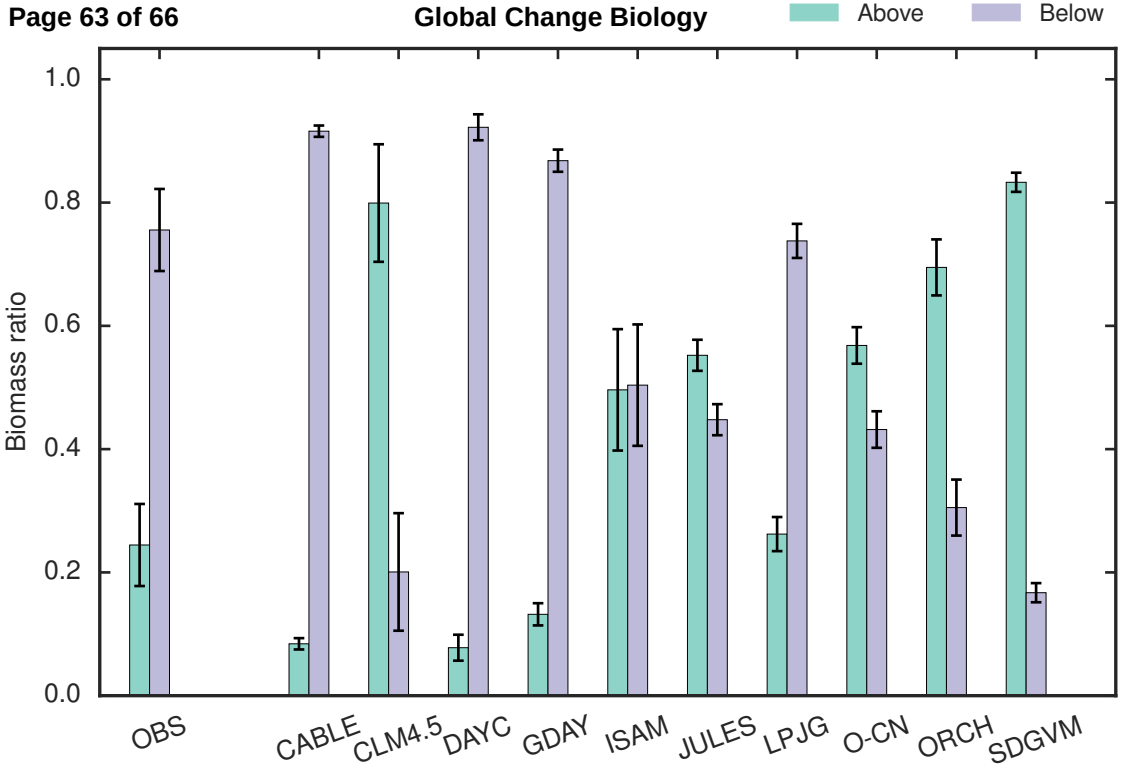

Dry year (2008)
Wet year (2009)
(a) CABLE
(b) CLM4.5
(c) DAYC
(d) GDAY
(e) ISAM
(f) JULES
(g) LPJG
(h) O-CN
(i) ORCH
(j) SDGVM
SW (mm)
Day of year
0
50
100
150
200
300

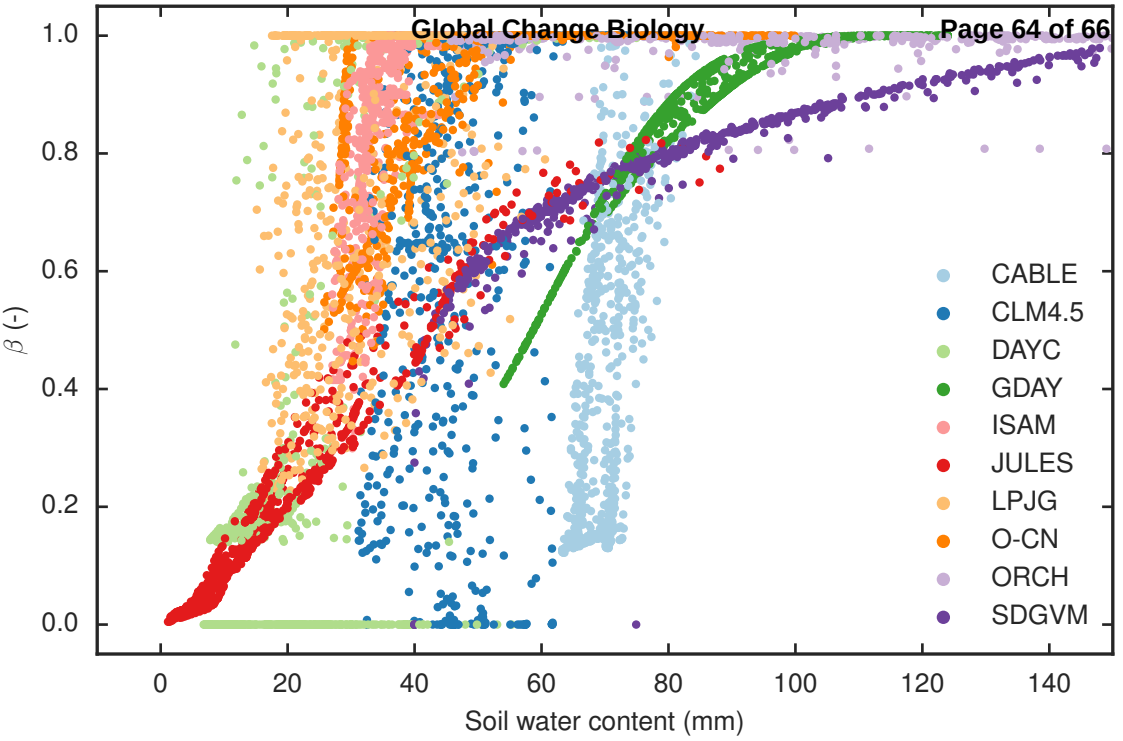

ct
Ct
cT
CT
(a) CABLE
(b) CLM5
(c) DAYCENT
(d) GDAY
(e) ISAM
(f) JULES
(g) LPJ-GUESS
(h) O-CN
(i) ORCHIDEE
(j) SDGVM
β (-)
1.0
0.5
0.0
2007
2009
2011
Years

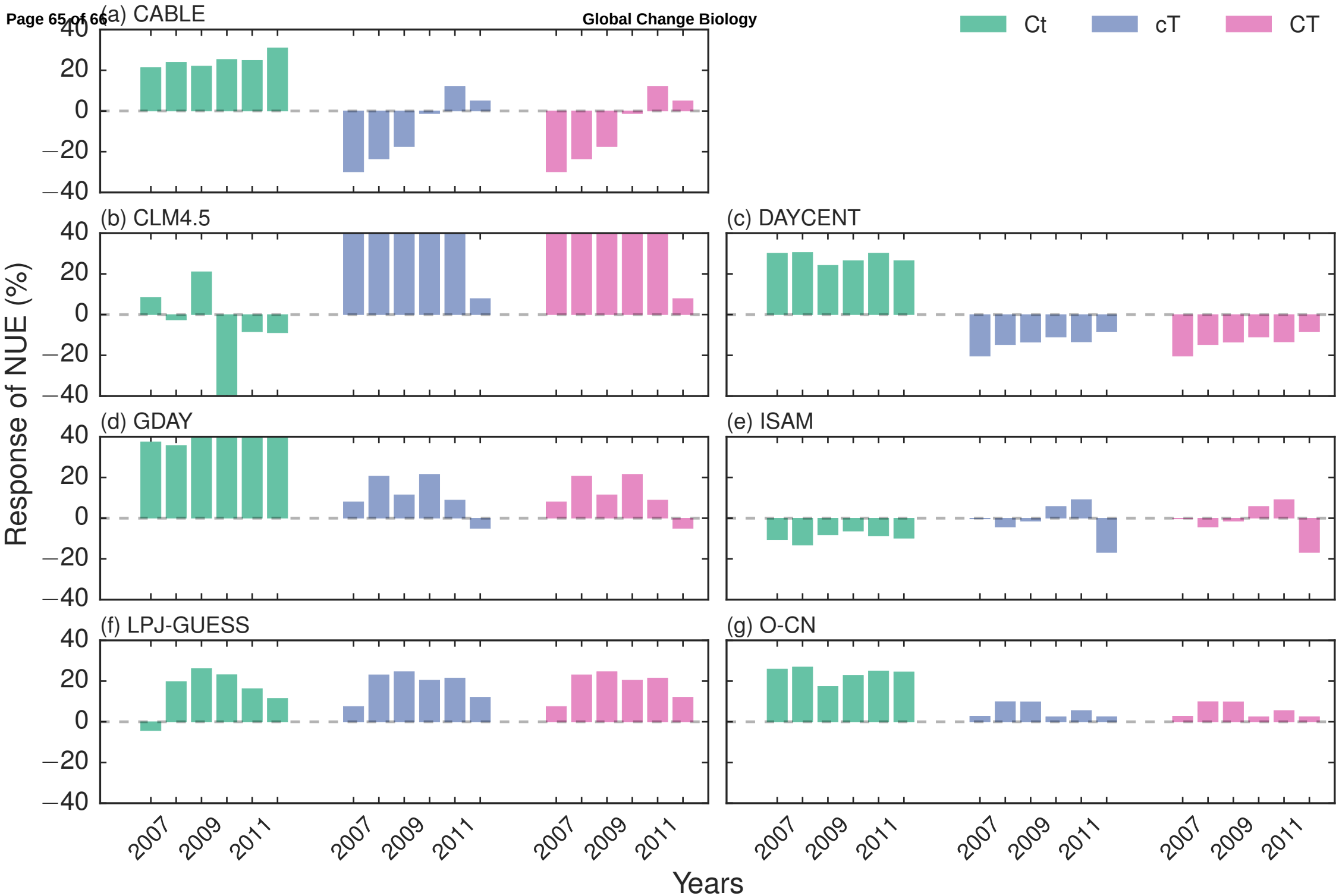

Ct
cT
CT
(a) Observations
(b) CABLE
(c) CLM4.5
(d) DAYCENT
(e) GDAY
(f) ISAM
(g) JULES
(h) LPJ-GUESS
(i) O-CN
(j) ORCHIDEE
(k) SDGVM
Response of aNPP (%)
60
30
0
−30
2007
2009
2011
Years

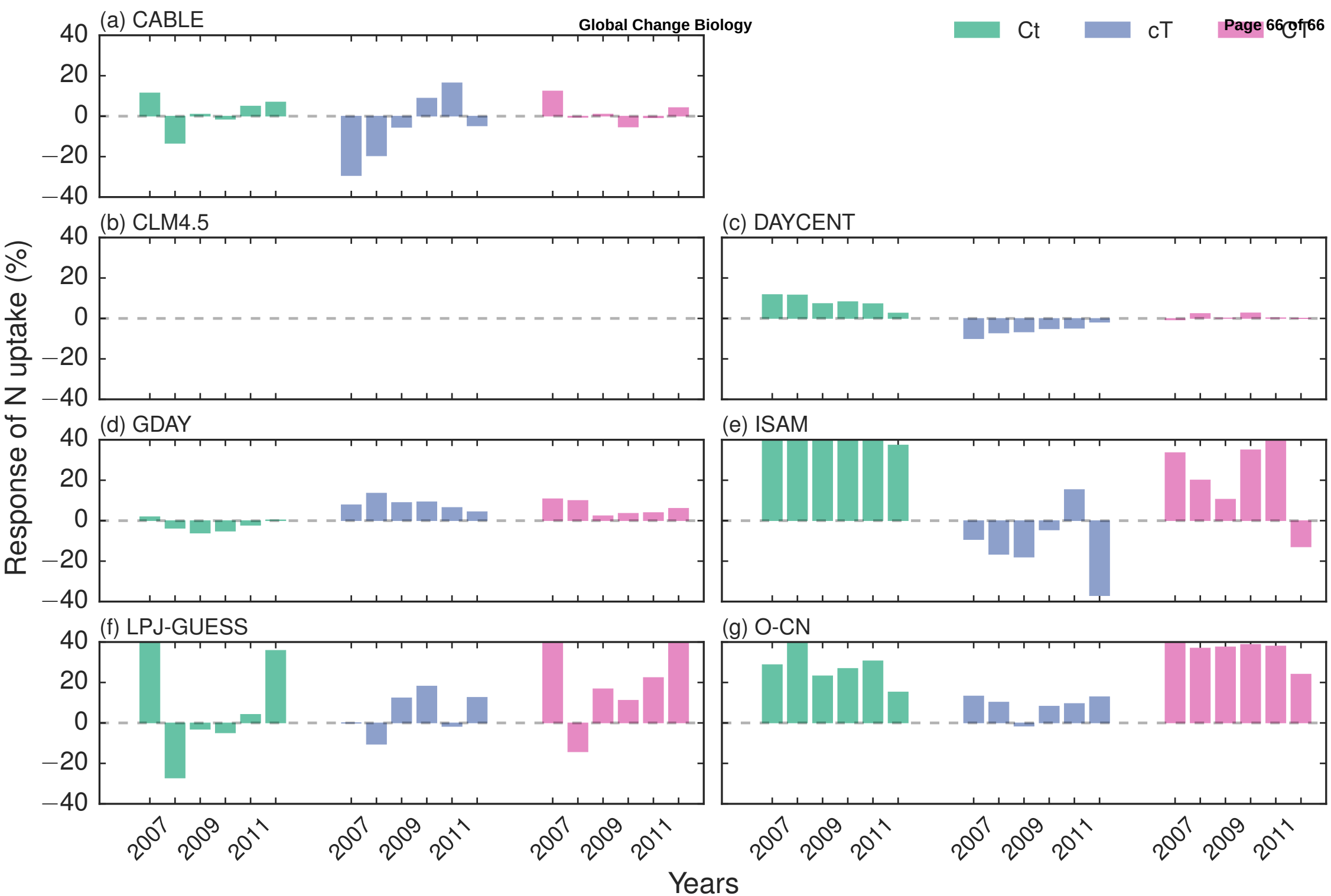

Ct
cT
CT
(a) Observations
(b) CABLE
(c) CLM4.5
(d) DAYCENT
(e) GDAY
(f) ISAM
(g) JULES
(h) LPJ-GUESS
(i) O-CN
(j) ORCHIDEE
(k) SDGVM
Response of aNPP (%)
60
30
0
−30
2007
2009
2011
Years

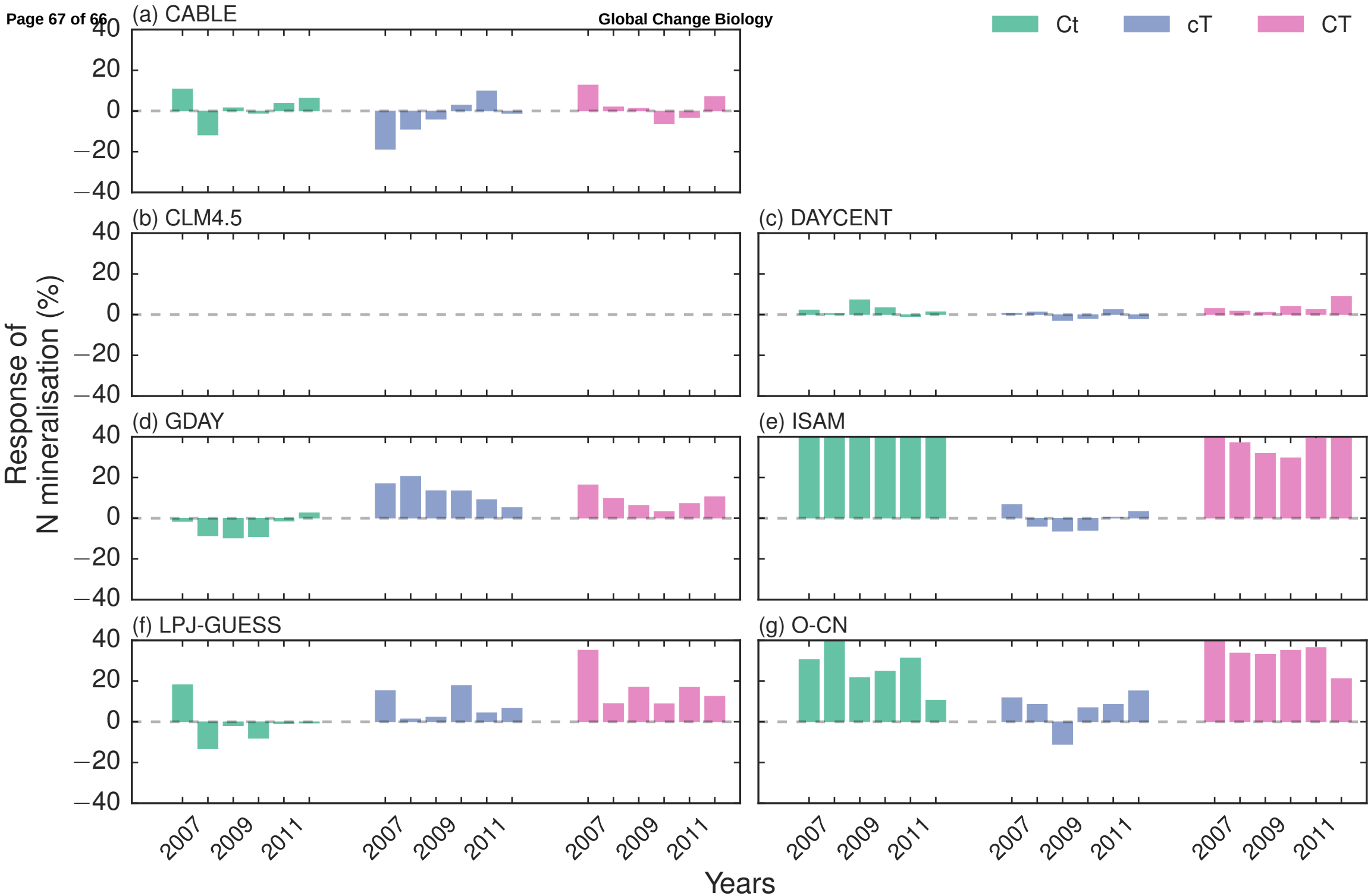

Ct
cT
CT
(a) Observations
(b) CABLE
(c) CLM4.5
(d) DAYCENT
(e) GDAY
(f) ISAM
(g) JULES
(h) LPJ-GUESS
(i) O-CN
(j) ORCHIDEE
(k) SDGVM
Response of root biomass (%)
60
30
0
−30
2007
2009
2011
Years


Above
Below
Biomass ratio
1.0
0.8
0.6
0.4
0.2
0.0
OBS
CABLE
CLM4.5
DAYC
GDAY
ISAM
JULES
LPJG
O-CN
ORCH
SDGVM


1.0
0.8
0.6
0.4
0.2
0.0
β (-)
0
20
40
60
80
100
120
140
Soil water content (mm)
CABLE
CLM4.5
DAYC
GDAY
ISAM
JULES
LPJG
O-CN
ORCH
SDGVM


Ct
cT
CT
(a) CABLE
(b) CLM4.5
(c) DAYCENT
(d) GDAY
(e) ISAM
(f) LPJ-GUESS
(g) O-CN
Response of NUE (%)
40
20
0
−20
−40
2007
2009
2011
Years


Ct
cT
CT
(a) CABLE
(b) CLM4.5
(c) DAYCENT
(d) GDAY
(e) ISAM
(f) LPJ-GUESS
(g) O-CN
Response of N uptake (%)
40
20
0
−20
−40
2007
2009
2011
Years


Ct
cT
CT
(a) CABLE
(b) CLM4.5
(c) DAYCENT
(d) GDAY
(e) ISAM
(f) LPJ-GUESS
(g) O-CN
Response of N mineralisation (%)
40
20
0
−20
−40
2007
2009
2011
Years