We would like to thank both Referees for their valuable comments and suggestions for improving our manuscript. Following Referees' comments, we carefully revised our manuscript. Please find below the point-to-point responses (in black) to referee comments (in blue). For your convenience, changes in the revised manuscript are highlighted with dark red.

Referee #1

General Comments

This paper provides a description of revisions to the ORCHIDEE-HL (high latitude) land surface model intended to improve the simulation of Northern Hemisphere vegetation cover. The results are evaluated against several fractional land cover datasets and gridded observations of GPP, biomass and soil carbon. The authors claim "significant improvements" in simulated tree distributions and this appears to be justified. A particularly strength of the paper is that simulated PFT fractions are compared with multiple observational estimates, which takes into account the combined uncertainty in the source data and in the mapping from land cover classes to model PFTs. This allows the authors to place an informed emphasis on model errors and improvements in different regions.

The manuscript is well-written throughout and the figures are clear and understandable. With only a couple of exceptions, details of the model description that were not provided explicitly in the manuscript were found easily in the references provided (e.g., Krinner et al, Gouttevin et al).

Specific Comments

1. Despite being well used, it's not clear to me whether the 6-hourly CRU-NCEP forcing resolves the diurnal cycle adequately. In particular, the simulation of photosynthesis will depend strongly on the sub-daily representation of surface insolation. How are the forcing data downscaled from 6 hours to the 30 minute model time step? If these forcing fields are valid at the same UTC time rather than the same local solar time, is there any significant longitudinal variation in how well the diurnal cycles of insolation and GPP are represented?

Response

In ORCHIDEE, the meteorological fields of climate forcing are interpolated from their original time step to the half-hourly model time step. For fields other than downward solar radiation and precipitation, the 6-hourly data in CRU-NCEP are linearly interpolated to half-hourly resolution. For the short-wave radiation in particular, it is distributed as a function of solar angle, calculated based on longitude/latitude, the day of the year and the hour, according to the method used by GSWP (Dirmeyer, 2011; ORCHIDEE code see http://dods.ipsl.jussieu.fr/orchidee/DOXYGEN/webdoc/d1/db6/solar_8f90_source.html). The forcing fields and model outputs are valid at the same UTC time, for example at each time step, only half of the earth surface has solar radiation. The diurnal cycles of insolation and GPP at different longitudes are thus corresponding to UTC time rather than their local time.

Reference:

Dirmeyer, P. A.: A history and review of the global soil wetness project (GSWP), J. Hydrometeorol., 12, 729–749, 2011.

2. The β diversity metric shows well the improvement in the high latitude tundra (Fig 5), but it doesn't highlight the greatly improved tree PFT fractions in northern Europe and eastern Canada. I would have expected this improvement between simulations to be more apparent in the metric, especially in the mean given the agreement between the observational datasets in these regions (Fig 3). It is more visible in the skill score (Fig 6) so, are there model errors and improvements that we should not expect to be able to evaluate through the use of this metric?

Response

It is true that in Fig. 5, the most highlighted regions are arctic tundra, with large β values between OLD and observational datasets and substantial improvement (β reduction) in NEW. But it can also be seen from Fig. 5 that the β metric is reduced in eastern Canada (from ~0.7 to ~0.3), and northern Europe and European Russia (from ~0.7 to ~0.4).

Tundra regions are more apparent in Fig. 5, because the OLD simulation produced very high fraction (>0.9) of needleaf deciduous trees in these regions that in reality have high fraction of bare land (PFT1); according to the definition of β diversity, this "extreme" bias of 2 PFTs, compared with evenly distributed bias among all PFTs, will more enlarge the value of β diversity. By contrast, in eastern Canada and northern Europe, besides the dominant needleleaf evergreen trees, other PFTs including broadleaf trees, grass and bare land can account for ~0.3. This relatively evenly distribution in vegetation (compared to that in tundra regions) avoids very large values of β diversity, even though the OLD simulation highly overestimated broadleaf trees in these regions. Therefore, the significant improvement in eastern Canada and northern Europe as shown in Fig. 4 did not turn into very obvious decrease of β Fig. 5.

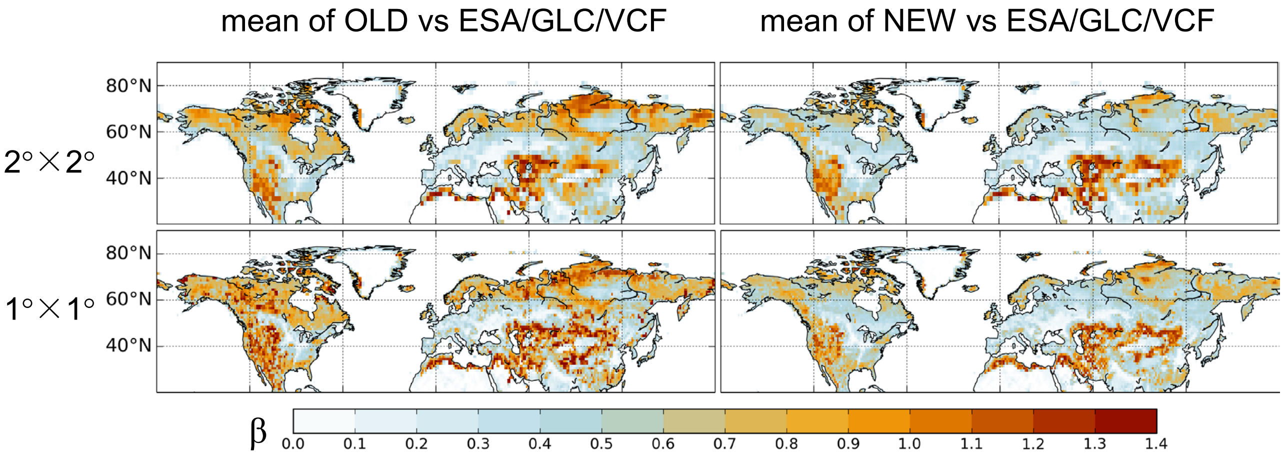
The skill score (S_V) in Fig. 6, however, presents more visible improvement in northern Europe and eastern Canada. This is consistent with the high agreement among observational datasets (i.e., small β for data vs. data) in these regions shown in Fig. 3, because S_V is defined as β (data vs. data) divided by β (model vs. data) (Eq.9), and the small S_V value for OLD, due to a small numerator, makes the difference between OLD and NEW more visible. S_V highlights the improvement that is intuitionally shown in Fig. 3; in this sense, S_V is a good metric to evaluate model performance in simulating vegetation distribution.

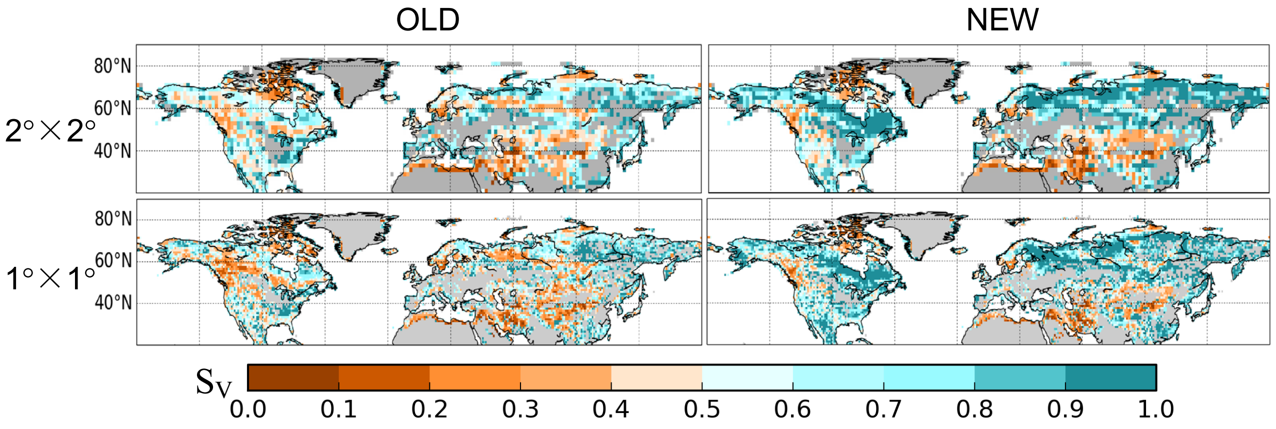
3. The authors highlight that these metrics (β, D and S) provide a framework that could be used by other models, and this type of multi-dataset analysis should undoubtedly be done in other studies. But how resolution dependent are these metrics likely to be? This would be a tradeoff between the smoothing of coarser grids making it easier for a model to match observations, but also easier for observations to match each other. So would it be reasonable

to compare models using significantly different grids? Could I calculate values for another model and compare them fairly with those in Table 3?

Response

Following this comment, we conducted two additional runs similar to OLD and NEW except for a 1°×1° resolution, in order to test the resolution dependency of these metrics. The figure below displays the new β value and skill score (S_V), compared with Fig. 5 (bottom panel) and Fig. 6:





As the figure shows, both β and S_V have similar spatial pattern in 1°×1° runs as in previous 2°×2° runs.

The β metric (Northern Hemisphere (20-90°N) mean) between models and observational datasets, and average S_V over different countries/regions are listed in the following tables:

β	2°×2°				1°×1°			
	OLD	NEW	ESA	GLC	OLD	NEW	ESA	GLC
ESA	0.58	0.56			0.70	0.62		
GLC	0.56	0.48	0.25		0.68	0.54	0.29	
VCF	0.65	0.47	0.37	0.35	0.77	0.52	0.43	0.41

S_V		Asian Russia	European Russia	Canada	USA	Europe	China	Northern Hemisphere (20°N-90°N)
2°×2°	OLD	0.68	0.63	0.53	0.66	0.62	0.57	0.60
	NEW	0.89	0.89	0.70	0.69	0.65	0.61	0.72
1°×1°	OLD	0.69	0.57	0.52	0.63	0.58	0.53	0.59
	NEW	0.87	0.91	0.71	0.73	0.67	0.66	0.74

In the coarser 2°×2° runs, due to smoothing effect, the β values for both model vs. data and data vs. data are decreased by 9~18% compared with 1°×1° runs. For S_V however, there is little difference between the two resolutions (relative differences are mostly within 5%), since the smoothing effect on both numerator and denominator partly offset each other. It indicates that the resolution at which the model runs has minor influence on the S_V metrics, and is not supposed to change the ranking of different models. Therefore, it is reasonable to calculate the skill score for other DGVMs with different grids, and compare them with the results in this study, if they adopt the same simulation protocol.

4. In the sensitivity experiments one piece of information that I couldn't glean was how do variations in the "1850" forest cover and GPP owing to spin up methodology (e.g., 1901 vs 1914, Fig 14) compare with the magnitude of 20th century change in the NEW and OLD simulations? Section 6.2 quotes 11.5% and 4.8% 20N-90N forest fraction increases with and without CO2 fertilisation, but it is difficult to compare these aggregate figures with the maps in Fig 14. This would provide some context for the warnings about spin up methodology. Also in section 6.3, the apparent motivation for the individual year simulations ("...recycled one-year climatic data are sometime used...") appears near the end after the results. It would be clearer if this was mentioned earlier in the section.

Response

Sect. 6.3 focused on the spin up methodology in terms of climate forcing, and the vegetation distribution results shown in this sector corresponded to the last year of spin up, i.e., the initial state (1850) of the transient simulation. The EXP3 experiment used the 20-year average climatology as forcing file in spin up; compared to NEW, total forest area in EXP3 increase by 5.1 Mkm^2 (22%), among which temperate trees (PFT4-6) increase by 2.7 Mkm^2, boreal needleleaf evergreen (PFT7) and broadleaf deciduous (PFT8) trees increase by 6.3 Mkm^2, and needleleaf deciduous tree (PFT9) decrease by 3.9 Mkm^2. This large variation in the initial forest cover owing to different climate forcings in spin up brings a warning on spin up methodology. Accordingly, the following sentences were added at P2242,L26: "In EXP3, temperate trees (PFT4-6) can extend northward, taking up the boreal tree positions, while the distribution of boreal needleleaf evergreen (PFT7) and broadleaf deciduous (PFT8) trees is squeezed to the climatic range of needleleaf deciduous tree (PFT9). Compared with the initial state after spin up in NEW, total forest area in the studied region (20-90°N) in EXP3 increase by 5.1 Mkm^2 (22%), among which PFT4-6 increase by 2.7 Mkm^2, PFT 7 and 8 increase by 6.3 Mkm^2, and PFT9 decrease by 3.9 Mkm^2."

To explain more clearly the motivation for the spin-up tests forced by individual year climate, the sentence "The large variance…" in P2242,L26 was replaced by "Apart from average climatology, recycled one single year climate is occasionally used in spin-up phase, which can also lead to large variance in initial vegetation distribution after spin-up due to interannual climate variability."

Technical Comments

Title: "...northern..." is a bit too vague. "...Northern Hemisphere high latitude..." would be more informative (and would reflect the model version).

Response

The title was revised accordingly: "Improving the dynamics of Northern Hemisphere high latitude vegetation in the ORCHIDEE ecosystem model".

P2219,L20: The repository that "rev1322" corresponds to isn't mentioned until Section 2.3.

Response

Since the original "Sect. 2.3 Code availability" was moved to the end (after "Sect. 7 Conclusions"), the two sentences in P2219,L20 were revised as "The basic structure of ORC-HL used in this study is shown in Fig. S1 in the Supplement, in which different processes from Krinner et al. (2005) are highlighted with red."

P2220,L13-15: It's not clear if *V* can be negative, e.g., though net biomass loss, which makes the range of possible MBG values unclear.

Response

Here *V* cannot be negative. To clarify it, the following sentence was added in the end of P2220,L15: "*V* equals to 0 in case of net annual biomass loss."

P2222,L21: "M_{SF} (t)" should be "M_{SF} (t, T_{min})", if I've interpreted the model correctly.

Response

$M_{SF}(t)$ was revised as $M_{SF}(t,T_{min})$ accordingly.

P2224,L8: Kuppel et al (2012) references a PhD thesis; can the same information be gleaned from Kuppel et al (2012), Biogeosciences, doi:10.5194/bg-9-3757-2012 ? If so, the latter reference is preferable.

Response

Kuppel et al (2012, Biogeosciences) presented a data assimilation system to optimize some ORCHIDEE parameters using measurements from temperate deciduous broadleaf forest sites, thus their results were only applied to PFT6 in ORCHIDEE; while in Kuppel's PhD thesis (Kuppel, 2012), parameters of other PFTs were optimized using the same method. So we cited the PhD thesis (accessible from Internet) rather than the paper in Biogeosciences.

P2224,L19-21: It's not clear whether the leaf age dependency was switched off entirely or whether just very long time constant (a_{crit}) was used. The values in Table 1 for evergreen needleleaf are unchanged from Krinner et al, so is a_{crit} used elsewhere in the model? If not, why quote unused acrit values at all?

Response

Apart from the v_{cmax} (or j_{max}) dependency on leaf age discussed in Sect. 2.2.3, a_{crit} is also used to calculate leaf senescence in the turnover module in ORCHIDEE, so we still listed the a_{crit} values for evergreen needleleaf (PFTs 4 and 7) in Table 1.

The leaf age dependency of v_{cmax} (or j_{max}) for PFTs 4 and 7 was switched off. This v_{cmax} (or j_{max})–leaf age relationship was introduced in Krinner et al. (2005) to account for the influence of seasonal variation in leaf age on photosynthetic activity for trees; and we removed this rule for needleleaf evergreen trees since they do not have such significant seasonal variation in leaf age as deciduous trees do. To clarify it, we added a sentence at the end of Table 1 notes: "a_{crit}: critical leaf age for leaf senescence (days); the dependence of v_{cmax} and j_{max} on leaf age for PFTs 4 and 7 was eliminated as described in Sect. 2.2.3."

P2230,L3 L9: (Equation pedantry) The sum should be from "k = 1" rather than just "k".
P2230,L19: Similarly, the sums are missing upper limits.

Response

Equation (7), (8) and (9) were revised accordingly.

P2232,L26: Should be σ_O rather than σO.
P2233,L1: Are there missing modulus symbols, i.e., $|X_{c,M} - X_{c,O}| < \sigma_O$?
P2237,L7: "SG" should be "S_G".

Response

Revised accordingly.

Fig 2: "Brighter colors..." is ambiguous wording, "Deeper colors..." would be better. Should "...relative fraction..." be just "...fraction. . .", else it's not clear what it's relative to?

Response

Fig 2 caption was revised as: "…Color indicates the fraction of three PFT groups…Deeper colors represent higher fractional covers." Similarly, the "relative" in Fig 4 caption was deleted.

Fig 2 4: I find it difficult to determine how deep or pale these maps are relative to each other (e.g., OSIB vs IIASA). A limited scale (e.g., 25%,50%,75%,100%) for the pure RGB hues would be useful.

Response

A color scale was added in Fig. 2 and 4 accordingly.