We thank the reviewer for their constructive comments and we address their various concerns below.

In revision, we have now fixed the WUE bug (see below) and this now allows us to investigate both the carbon and the water simulations by CABLE. We now very clearly demonstrate that models need to implement different sensitivities to water stress at xeric sites, otherwise models will underestimate carbon and water fluxes during drought. Finally, following the reviewer's suggestion we have restructured the results to make the text clearer for the reader.

The focus of the paper in on modelling drought impacts on ecosystem gas exchanges, with the hypothesis that species respond differently to drought. While interesting for the LSM community, the hypothesis is somewhat trivial to ecologists, who know species behave differently in respect of drought.

It is one thing to 'know' that there are differences among species in response to drought; it is another thing entirely to quantify these differences and use them to attempt to predict drought impacts on forests at landscape scales. The question addressed in our work is whether or not these differences among species in sensitivity actually matter when simulating land surface fluxes. This is a far from a trivial question.

There are two important points to make here:

1. The current state-of-the-art in terms of drought modelling (not just LSMs), assumes PFTs have similar responses to drought. For a given location, differences in wilting point and field capacity are a function of soil texture and as such, different PFT occurring on the same soil type would have an identical limitation to gas exchange during drought (assuming both PFTs transpired the same amount of water). Other models, CABLE included, also assume differences among PFTs in terms of rooting depth; however, these differences are often very small. Thus, the approach taken in this paper, testing empirical drought sensitivities into CABLE, is highly novel.

2. Implementing a species-dependent response to drought is a non-trivial problem. As we demonstrate in figure 2, the relationship between soil water content and soil water potential is highly non-linear. The consequence of this

nonlinearity is that whether a model assumes gas exchange is limited by soil texture, or a species-dependent water potential, can sometimes have very little effect on modelled fluxes. The importance of species-level differences in response to drought for vegetation fluxes depends on the water potential a plant "sees". In this paper we tested three alternative approaches to determining this water potential. Our results have implications for the wider modelling community. For example, we demonstrated that the approach implemented within the CLM (one of our tested approaches), likely fails to capture drought responses despite depending on water potential, rather than soil texture.

The approach is to try three different parameterisations, and also three different root uptake models, and evaluate model outputs against 5 flux sites over a European drought. The core output of the paper is table 4, where a range of statistics are applied to the comparison of observation and fluxes.

I remain to be convinced of a main conclusion – that there is high drought sensitivity at northern sites. The hypothesis testing is not robust. I worry that there are a range of alternative model tweaks that could get similar improvements in the flux comparison. We really need further independent checks on model outputs using other data streams, for example local LAI data, biomass increments, soil moisture time series etc.

As the reviewer does not elaborate specifically on why they are not convinced the northern (or most mesic) sites can be characterised as having high drought sensitivity, it is difficult to address this point. Nevertheless, we have improved our justification of the rationale behind this hypothesis in the introduction: "*We hypothesised that drought sensitivity would increase with latitude, as sites transitioned from xeric to mesic. We hypothesised that trees at more mesic sites, with a greater abundance of available water than at xeric sites, would be more vulnerable to shorter duration droughts and thus have higher drought sensitivity (or lower resistance to drought). Therefore, accounting for this latitudinal gradient in drought sensitivity would improve the performance of CABLE.*"

We agree with the reviewer that it is certainly possible that "alternative model tweaks" may result in improved model-data agreement. Such tweaks may include adjusting the site texture or root depth (both which would change access to soil water), but we would suggest these kinds of tweaks would need to be carried out on a site-to-site basis. It would be (highly) unlikely that a single adjustment would improve the model across all 5 sites (as our changes do). In fact, Ukkola et al. 2015 (Hydrol. Earth Syst. Sci. Discuss., 12, 10789–10825) demonstrate this very point using the CABLE model across 20 flux tower sites. They found that the effect of adjusting soil properties had a limited ability to improve model-data mismatch, with improvements limited to individual sites. In the paper we are instead taking an evidence-based approach to model development, rather than tweaking individual parameters to improve data-model mismatch and then attributing improvement to this property.

Consequently, we disagree that the hypothesis testing is not robust, we have tested the new model simulations with the most direct data stream available – fluxes of latent heat and flux-derived GPP from the 5 flux tower sites. Ideally, we would of course have liked to use additional data streams to test model improvement. In fact, we make this very point in the discussion: "*Access to water by deep roots could be a potential alternative explanation for the low drought sensitivity that we inferred at the southernmost (xeric) site, Espirra. Here the dominant species is not native to the region, but rather a plantation of blue gum (Eucalyptus globulus), a species that is generally found to have high, not low, drought sensitivity (White 1996; Mitchell et al. 2014). Many eucalypts have a deep rooting strategy (Fabiao et al. 1987), suggesting a possible alternative explanation for drought tolerance at this site. More in-depth study of fluxes and soil moisture patterns at this site would be needed to determine the role of rooting depth.*"

Unfortunately none of the reviewer's suggested alternative data streams are appropriate for model testing at this scale. With respect to LAI, see our extended comment below. Changes in biomass increment are a function of drought impacts on growth and carbon allocation as well as gas exchange, and are only available at longer time scales (e.g. annually) giving relatively little information to constrain the models. With respect to soil moisture data: datasets of deep soil water layers (CABLE simulates depths up to 4.6 m) are rarely measured, and if they are, are not freely

available. Satellite derived estimates of soil moisture content only extend to the very top few centimetres (<10 cm) and so do not allow us to test deep soil water access appropriately. Hence, currently available soil moisture information is of limited use in constraining drought responses.

We know that the PFT approach is a weakness due to its one-size-fits-all approach, and therefore finer scales of parameterisation will help. The problem is to figure out how to make that happen in a tractable and robust manner, and this paper is not written in a way to tackle that problem.

We agree that our paper does not solve the problem, but we do believe that we take a very important first step towards tackling it, which is to explore whether or not finer scales of parameterisation would actually make a difference to land surface simulations. As discussed above, this is not a given. We demonstrate that differential sensitivity to drought can be important in determining drought responses. In the discussion we propose some potential ways forward for implementing within-PFT variability in modelled responses to drought: "*Global vegetation models would require a more sophisticated approach that relates drought sensitivity to the climate of each pixel. One potential solution would be to develop an empirical correlation between drought sensitivity and a long-term moisture index (e.g. the ratio of mean precipitation to the equilibrium evapotranspiration; Cramer and Prentice, 1988; Gallego-Sala et al. 2010). Previous studies have demonstrated the feasibility of linking model parameters that determine plant water use strategy to such a moisture index in global simulations (Wang et al. 2014; De Kauwe et al. 2015). Such an approach would require a concerted effort to collate appropriate data, as there are few compilations to date of traits related to drought sensitivity (but see Manzoni et al. 2011; Zhou et al. 2013). Another, more challenging, alternative, would be to develop optimization hypotheses that can predict vegetation drought sensitivity from climate (e.g. Manzoni et al. 2014).*"

The existence of a model bug is another major concern that undermines confidence.

Both reviewers have raised an issue with respect to our section outlining the water use efficiency (WUE) bug and the ensuing analysis. We acknowledge therefore that we were not clear enough in our original text.

The bug *only* affected the modelled fluxes of gross primary productivity during periods of drought: *all* water fluxes were correctly simulated (at all times). During periods where there was an inadequate soil water supply to meet the atmospheric demand for simulated transpiration, transpiration was reduced to the available amount of water in the root zone. To maintain a coupled carbon and water cycle, photosynthesis should also be reduced, but wasn't. This has the effect of producing erroneously high WUE during periods of extreme water limitation, i.e. free carbon in exchange for water might be another way to interpret this. It was for this reason we chose to focus *all* of our analysis on latent heat and transpiration and *not* on gross primary productivity. Nevertheless, we did originally show the gross primary productivity fluxes so that the readers could judge for themselves the erroneously increased water use efficiency during drought (i.e. the effect of the bug).

We have now fixed the bug (here we acknowledge assistance of Ying-Ping Wang and Vanessa Haverd of CSIRO), and so we now remove all text that referred to this from the manuscript. Hopefully this should allay any concerns the reviewers have about the impact of the bug on any of our results. We do note that part of the fix requires changing the method of root extraction in the standard version of the model. Consequently, as well as reducing erroneously high GPP during drought, there are also small changes in the latent heat fluxes.

Abstract The text is not clear about what drought response is analysed – is it C cycle, water cycle, energy balance?

We agree with the reviewer, this was not clear. We have amended the text to say: "*We tested whether variable drought sensitivities are needed to explain the observed large-scale patterns of drought impact on the carbon, water and energy fluxes*"

Introduction:

We respectfully disagree. The point of this paragraph was to set out exactly what was done in terms of this paper. It is important therefore to be explicit and describe the changes that were made to CABLE. We do broaden the implications of the results in what we consider to be the appropriate place, which is the discussion text.

The stomatal model used here should not be regarded as a 'modification' of an existing empirical model. The model, as derived by Medlyn et al. (2011), follows the theory of optimal stomatal behaviour and is functionally equivalent to the Ball-Berry model, but with the advantage that biological meaning can be associated with the model parameters.

We defined what we mean by "optimally":

"*We build on the work by De Kauwe et al. (2015), who introduced a new g_s scheme into CABLE. In this scheme, stomata are assumed to behave optimally; that is, when plants maximise carbon gain, whilst simultaneously minimising water loss over short time periods (i.e. a day) (Cowan and Farquhar, 1977) leading to the following formulation of g_s (Medlyn et al. 2011)*"

We tend to disagree that this is premature. The evidence would appear to us to be reasonably clear that there are changes in apparent Vcmax in response to drought in the majority of cases where this has been examined (Cornic et al. 1989; Epron and Dreyer 1992; Medrano et al. 1997; Parry et al. 2002 Keenan et al. 2009; Egea et al. 2011; Zhou et al. 2013). The Wright et al. study cited would appear to be an exception rather than the rule. Furthermore, we highlight to the reviewer that there are a number of existing models that *only* limit gas exchange via reducing the V_{cmax} (J_{max}), for example CLM4.

p. 12357 l. 5. Constant Jmax/Vcmax ratio is assumed Misson et al. (2006) hypothesized that Jmax is more sensitive to low water availability than Vcmax, so drought conditions may also lead to a decrease in the Jmax/Vcmax ratio. Data from Wright et al. (2013) support this hypothesis.

We agree with the reviewer that such a hypothesis, i.e. that J_{max} is more sensitive than V_{cmax} is potentially an interesting hypothesis to test in a model. We should clarify, the model assumes a constant J_{max}/V_{cmax} ratio; we did not specifically add this for the purposes of this study. As such, in this paper we were only testing an equal reduction in both properties. A logical follow-on to this work would be to examine the impacts of such a hypothesis, but this is outside the scope of this current study and we leave this to other interested researchers.

l. 8. Refers to eqn 6, but is this an error?

We thank the reviewer for spotting this mistake. It has now been corrected.

Model Simulations:

The models are run with MODIS LAI as a driver – but there is a problem in that MODIS LAI contains significant biases when used at site (flux tower) scale. Thus the LAI drivers used are unlikely to be correct, and this will lead to model biases. This issue needs to be addressed.

We agree with the reviewer that the prescribed MODIS LAI may not be a true reflection of observed site dynamics. However, there is unfortunately no practical

alternative. Measured site LAI would need be collected at the necessary temporal and/or spatial resolution to force a LSM (i.e. daily to monthly and 1 km^2). Such measurements may have been made at a small subset of sites but are not freely available. Hence, the most practical solution is to use MODIS data, as is commonly done when simulating flux sites with land surface models.

CABLE, similar to other models, does have the means to dynamically simulate LAI, but there is no reason to expect this would out-perform MODIS. We address the issue that because the model prescribes LAI (as is standard among LSMs), it likely misses the effect of canopy defoliation in the discussion text: "*During droughts, plants are often observed to shed their leaves as a self-regulatory mechanism to reduce water losses (Tyree et al., 1993; Jonasson et al., 1997; Bréda 20 et al., 2006). During the 2003 heatwave, at Hesse an early reduction of approximately 1.7 m^2 m^{-2} was observed, similarly at Brasschaat there was a observed reduction of 0.8 m^2 m^{-2} and at Tharandt needle-litter was increased during September until November, with LAI estimated to be 0.9 m^2 m^{-2} lower (Bréda et al., 2006; Granier et al., 2007). In contrast, models typically fix turnover rates for leaves and as such this feedback is largely absent from models. During periods of water stress, models do simulate an indirect reduction in LAI via down-regulated net primary productivity; however this feedback is much slower than is commonly observed. Not accounting for the canopy scale feedback will result in models over-estimating carbon and water fluxes and thus losses in θ during drought.*"

Ultimately the aim of this study is not to address how well CABLE simulates LAI and we do acknowledge that errors here will impact on simulations during drought (see above); however, it is by no means the main cause for data-model mismatch during drought periods.

Why are these species (Quercus, Cedrus) chosen? Why not use the species that are found at the flux sites (Table 2)? A consistent approach would be more valuable.

The simple answer is that the data required to parameterise the model for the actual site species are not available. This issue is directly related to the reviewer's original point that "*the hypothesis is somewhat trivial to ecologists, who know species behave*

differently in respect of drought" – and yet there are relatively few data currently available with which one could parameterise such a model. Instead, the approach we have taken is a parsimonious one: we have taken a range of different species sensitivities to soil moisture from a recent meta-analysis. We explain in the discussion that in the absence of more appropriate site level data: "*One potential solution would be to develop an empirical correlation between drought sensitivity and a long-term moisture index (e.g. the ratio of mean precipitation to the equilibrium evapotranspiration; Cramer and Prentice, 1988; Gallego-Sala et al., 2010). Previous studies have demonstrated the feasibility of linking model parameters that determine plant water use strategy to such a moisture index in global simulations (Wang et al., 2014; De Kauwe et al., 2015). Such an approach would require a concerted effort to collate appropriate data, as there are few compilations to date of traits related to drought sensitivity (but see Manzoni et al., 2011; Zhou et al., 2013). Another, more challenging, alternative, would be to develop optimization hypotheses that can predict vegetation drought sensitivity from climate (e.g. Manzoni et al., 2014).*"

Water use efficiency bug I appreciate the openness of the authors on this issue. But I remain unclear on the implications of the bug and to what degree it invalidates the conclusions of the paper. Is photosynthesis over-estimated during drought? – this would seem to indicate that the paper must only focus on water and energy responses to drought. The authors seem to suggest that root water supply is always sufficient to meet demand, and so transpiration is never down-regulated – but I am confused as I would suggest that water limitation is a definition of drought, and that water limitation must have occurred in Europe in 2003 due to high demand by plants and low rainfall. This issue needs much more clarity if the paper is to be useful.

See earlier response.

Results:

The opening of the results should be targeted towards key knowledge, rather than a bland listing of tables and figures.

This section needs sub-headings to provide structure. It is hard to see what has been learned. We need clear statements.

We thank the reviewer for these suggestions. We have restructured the results accordingly.

When referring to GPP, be clear whether this is flux-derived or model-derived.

We agree with the reviewer this information was unclear. In the methods we now clarify this: "*Model simulations were compared to measured latent heat flux at each of the FLUXNET sites. In addition, simulated GPP was compared to flux derived (site modelled) GPP. These flux GPP estimates are calculated from the measured net ecosystem exchange (NEE) of carbon between the atmosphere and the vegetation/soil, and the modelled ecosystem respiration (R_{eco}), where GPP is calculated as NEE + R_{eco}.*" Throughout the test we refer to "observed" GPP as flux-derived GPP.

P 12361. "CTRL simulation" repeated

We thank the reviewer for spotting this mistake; it has now been corrected.

The discussion here on the WUE bug just confused me further. I don't know what we can learn from these simulations when a bug is complicating matters so much.

We have addressed this issue above.

There are simulations for sand and clay soils. Why not use an appropriate soil parameterisation for the site in question? This would target the analysis more effectively. At present the comparison across soil texture is confusing.

It is possible that the reviewer has confused the results of figure 2, with subsequent model simulations carried out with CABLE? In figure 2, we have picked two extremes out of the soil texture continuum simulated by CABLE and shown, in an idealised way, how the model simulates the reduction of photosynthesis with decreasing soil water content. In the remaining figures, where CABLE is run at the different flux tower sites we do not assume a sand or clay soil texture, rather, we use the representative soil texture class as given by Zobler (1999) for the corresponding site pixel.

We provided a range of statistical characteristics of model-data performance for the readers benefit. Our feeling was that if we had only provided RMSE, it is likely another reviewer may have requested additional statistics. We disagree that the statistics contradict the RMSE; the NSE values unsurprisingly show the same mesic-xeric transition as the RMSE. The Pearson's correlation coefficient does show a different story, but this is unsurprising as this is showing the linear correlation between the model and the observations. As such, as long as the model and observed change in the same direction a high correlation will be obtained, but this does not strictly tell us whether the error is smaller or larger, hence the focus on the more meaningful RMSE statistic.

We have an entire section of the methods dedicated to testing root zone weighting, page 123570-123571. In the results section, figure 2 explores the effect of root zone effects on drought. Given the reviewers comment above in regards to texture (which we address), we think it is possible that the reviewer misunderstood the purpose of this figure. Hopefully it will now be clearer that we link to these issues earlier in the manuscript.

The discussion text in section 4.3 is relevant to the wider issues of modelling drought, many of the topics explored in this section link back to other areas of the manuscript. Section 4.3 addresses four areas: (i) rooting depth, (ii) leaf shedding, (iii) stomatal vs mesophyll responses and (iv) soil heterogeneity. The assumption of rooting depth is an important issue, as it will of course impact upon the incurred drought stress of the model and is not a uniform assumption across models. We also make the point that this could be an alternative explanation for the results at the Espirra site (see above). The reviewer had asked previously about errors that may come from using MODIS LAI; the leaf shedding section addresses this issue. In response to the reviewer's comment, these paragraphs have been shortened and re-focused.

The final paragraph is a summary of the key learnings of the paper.

We have added this information.

In short, because 2002 did not experience a drought and this paper is focussed on the drought of 2003. We show the departure from 2002, which characterises the severity of the drought in Table 2. Overall, we would argue little would be gained by adding additional figures showing responses in a non-drought year. Furthermore, on all plots we show a line representing "no drought", which depicts an approximation to the size of the drought in 2003.

We have improved the explanation of this figure.

We disagree the figures were unstructured: they are ordered to show a mesic-xeric transition between sites, detailing the impact of the different drought sensitivities and different methods to obtain soil water potential (SWP) at each site. On each panel we show the observed, control simulations and three different drought sensitivities. We do not see any other way to succinctly display this information. Hopefully with the new restructuring of the text, it will be clearer to the reviewer.