We thank the reviewer for their positive and constructive comments and we address the various concerns below. Referee comments are highlighted in red, with our response below in each case.

Following the helpful reviewer suggestions we have made a significant effort at reshaping the manuscript to aid its interpretability along the suggested lines. In particular we have:

I. Clarified what we mean by "optimal" behaviour and clarified what the MED-C simulations add.
II. Restructured the entire results and discussion sections, adding new sub-sections.
III. We have added two additional figures: figure 5 to deal with the issue of boundary layer conductance and figure S2 to deal with the issue of our g_0 assumption.
IV. Removed results information previously presented in the method section (which involved reordering figures).
V. More clearly addressed the question of why one would want to move away from an empirical Ball-Berry approach in the discussion.
VI. In figure 9 we have now masked missing data area in the data products, for example over the Sahara desert. These areas are not masked in CABLE by default, but have very small fluxes and as such, the previous figure 9 showed an erroneous bias between model and data which related to missing data.

In combination we hope this redrafting will help more clearly define how we have advanced modelling with this manuscript.

This paper compares various stomatal schemes within the CABLE land surface model and undertakes an evaluation at site and global scale. The goal of this paper is to compare a well known empirical model of stomatal conductance with a slightly modified version. The original model is the Ball-Berry-Leuning model, and has global parameters from unknown derivation. The novel model is from Medlyn et al, which is presented with some different parameterisations. In both models conductance is a function of the same soil moisture stress function, VPD, CO2 concentration and gross assimilation rate. The key change is that one of the model parameters, g1, is now

proportional to a marginal carbon cost of water, and varies with climate. The gs models also have different sensitivity to VPD.

The conclusion of this paper is "This work paves the way for broader implementations of optimisation theory in LSMs and other large-scale vegetation models ". I remain unconvinced for a number of reasons –

1. Model testing needs more effort. Testing at flux sites is limited to comparison with LE data. Comparison with GPP and H should also be included, to show that C-water interactions are effectively coupled and build trust in the model(s). Detailed statistics are required, and discussion about model validity. I have more detailed comments below in the section on single site results.

We address the reviewers concerns on these issues in answer to their detailed suggestions below.

2. Optimal is arguable. I am unclear why a gs model that is described as "optimal" needs to be calibrated with empirical gs data. My understanding of an optimization model is that it predicts optimal behaviour (gs), without direct calibration, and is then verified against independent data (of gs). If a model has to be calibrated, it can't really be revealing some fundamental biological property. Can the authors clarify what they mean by 'optimal', and perhaps be more cautious in their claims in this regard?

We improve the clarity of what we mean by "optimal" in the main body of the text by adding a new subsection 4.1 to the discussion, but we also address the reviewer's query here.

The reviewer's main question is how the approach we have taken can be called an optimisation model when it needs to be calibrated. An optimisation model needs to include both *structure* and *parameterisation*. Most optimisation models predict some behaviour based on maximising revenue and/or minimising costs. The model structure is the solution to the optimisation problem for given revenue / cost values; the parameterisation is the revenue / cost values themselves. These revenue / cost values are sometimes able to be estimated from down-scale physiological measurements but often they represent concepts that are difficult to measure, in which case they may be assigned notional values or calibrated against data (e.g. see Bonan et al. 2014,

Geoscientific Model Development, 7, 2193-2222 Figure 5, or Franklin 2007 New Phytologist, 174, 811-822, for an example using a model of optimal allocation). Although the parameter values are calibrated, the model is nonetheless an optimisation model because of its structure. Also, the parameter values have meaning, and we can predict the direction of their variation among PFTs or with environmental conditions.

In our case, the stomatal conductance model is derived from optimal stomatal theory following Cowan and Farquhar (1977), who posited that stomata are operating *optimally* when they act to maximise carbon gain, whilst simultaneously minimizing water loss. There is one input required, which is the marginal cost of water. Given a value for this cost, the model uses the optimisation principle to predict the time course of stomatal behaviour and its responses to environmental variability. Thus, this optimisation model differs from empirical models because the model *structure* is derived from theory, not based on observation.

We do not yet have a method to predict exact values for the marginal water cost from first principles, although Prentice et al. (2014), Ecology Letters, 17, 82-91 has taken some steps in this direction. Instead, Lin et al. (2015) predicted qualitatively how the parameter should vary among PFTs, with environmental conditions, and with other plant properties such as wood density. For example, from first principles, it was predicted that g_1 should increase with mean annual temperature and should be higher in angiosperms than gymnosperms. These predictions were tested against the synthesis dataset and found largely to hold (Lin et al. 2015). Thus, although the parameters may be calibrated against data, there are predictions underlying these calibrations.

Furthermore, there is clear scope for extending these predictions in the future. For example, we discuss ways this parameter could be linked to biological traits, such as wood density, or linked to modelling behaviour under water stress. Empirical models do not offer many ways forward. In contrast, we argue that the use of an optimisation scheme opens up new avenues for theoretical advances in land surface modelling (see discussion and Zhou et al. 2013; 2014).

3. Global evaluation lacks conviction. The biome differences between the models are regarded as significant – for instance a 30% reduction in evapotranspiration for evergreen needle-leaf forests and tundra PFTs. These differences are highlighted in figures 6 and 7. But the latitudinal outputs of all the models are similar in figure 8. These results seem contradictory. By their own admission the impacts on the model outputs of the new scheme are negligible. The outstanding mismatches in the GPP signal (Fig. 8) are not solved by the new scheme. So I am left to conclude – why not just stick with the BB model, and its basic parameters? The new approach has not moved the modelling forward, even after a lot of extra effort.

We disagree with the reviewer on their conclusion that we have not "*moved the modelling forward*" and would be better "*just [to] stick with the BB model, and its basic parameters*" for the following reasons:

1. The original BB/Leuning parameter used by CABLE (and within most LSMs) is only loosely based on empirical data. In many cases the actual original source of the calibration data has not been documented, or cites the Sellers et al. 1996 paper (which does not document the original source of the g_s data). From a reproducible science perspective that is unsatisfactory. To that end, we derived a parameterisation for the g_1 parameter on a PFT-basis for CABLE and implemented it in CABLE. We also note that this scheme and parameterisation may be applicable more widely in other LSMs, or alternatively a new parameterisation could be obtained by utilising the freely available global synthesis of stomatal behaviour (Lin et al. 2015). To the reviewers point about "*why not just stick with the BB model, and its basic parameters?*". What we have achieved here is to directly link model parameters to ecological data in a transparent fashion, which crucially, from a climate modelling perspective does not degrade current model performance.
2. Whilst the changes in fluxes for CABLE for many PFTs are small relative to the control, this result may be somewhat specific to CABLE due to assumptions to do with boundary layer conductance; see new results (Line 374-392: Decoupling factor), figure 5 and discussion section (lines 600-638). By comparison, the JULES or O-CN LSMs (and perhaps others the authors are not aware of) would (broadly) predict that vegetation was far more coupled than CABLE and thus the role of g_s would be more important. Whilst

we do find evidence from the literature to support the decoupling assumed by CABLE, these are from relative few studies and the suggested ranges in decoupling are large such that it is still an area of considerable model uncertainty.

3. The results shown here have the potential to evolve as CABLE (or other models) introduce more representative PFTs. For example, if CABLE were to introduce a Eucalyptus PFT instead of assuming the vegetation was evergreen broadleaf forest, the g_1 values would change from 4.1 to anywhere from 1.4 to 8.4 depending on assumption made in regard to representative species. Similarly if the model explicitly represented a Savannah PFT instead of assuming a shrub PFT, g_1 would change from 4.7 to 3.0 or 7.2, depending on whether the model used deciduous or evergreen savannah values.
4. Finally, as the new g_1 parameter has biological meaning it has the potential to be linked to other plant traits such as wood density, or be predicted from them. Similarly, g_1 can be hypothesised to behave differently during drought for different species as has been shown recently by Zhou et al. (2013; 2014). In this paper we demonstrate that one could derive g_1 values based on bioclimatic indices rather than fit values based on measured data. Thus, introducing this new model opens new avenues for model development not currently available to a model using a BB/Leuning model.

We have clarified these points more clearly with a new revised text in the discussion section.

So, I am not led towards the authors' conclusions – which seem to lack foundations in the content of the paper. The structure of the paper is also problematical, with results and methods inter-mixed, key details left out, and long paragraphs that lack clear topics. The paper needs to build on the model outputs, model-data comparisons, towards clear and universal conclusions.

We have restructured the paper as the reviewer suggests removing the results section from the methods. The key details the reviewer cites as being left out (discussed below) are in fact in the methods section. To address the issue of "long paragraphs that lack clear topics" we have restructured the text and adopted more focussed sub-headings to aid the reader.

More detailed comments:

Single site results We are pointed towards figures and tables recording the results of this model experiment, but the text is not helpful in guiding the reader towards the critical outputs. My inspection of table 5 shows a mixed set of results for each model and no clear patterns. What has this exercise shown us?

To address the reviewers concerns we have redrafted this entire section as previously described and this should aid the reader in interpreting the presented results.

Figure 3: Why are observed GPP and H not added to the panels? It would help to have a clear evaluation of the model-data mismatch across all sites and variables.

We do not have access to the *measured* GPP data to add these to the panels as the reviewer suggested. The PALS dataset that we used in this study was traditionally not focused on the biophysics and model comparisons such as PILPS ignored fluxes such as GPP as few LSM simulated them. Furthermore, we do not feel it would be appropriate to show such a comparison as the GPP data would not be a directly observed property, but rather represents a series of assumptions about ecosystem respiration (as GPP is estimated from Net Ecosystem Exchange and total respiration). We choose to compare to the more directly observable property, LE (for which we did have observations).

We did not originally add H panels because we felt there was already a sufficiently complicated story considering multiple flux sites (6), multiple flux variables (3) and multiple simulations (3). We feel that adding H would add little to the current narrative, given the small changes in currently presented fluxes. We hope that the text revisions we help make the key messages more apparent to the reader.

Fig. 4. I am confused why MED-P shows a pronounced dip in mid-day E, but not in LE for DJF. Is this an evaporation/transpiration issue? This figure is only referenced once in the paper, with little detail provided in section 3.1. Surely more focus in required, and this figure should be referenced from the discussion.

We agree with the reviewer more focus should be given to this figure and acknowledge that the original text was confusing and have revised the text both in the results and discussion sections (4.4 Minimum stomatal conductance, g_0).

The reason the MED-P shows a clear dip during the midday due to high VPD driving stomatal closure in the E flux but not the LE flux relates to what these properties represent in the model. LE is a combination of the water flux from the canopy as well as the soil flux from the understorey. This explains how the LE can show no discernible dip and adds to the complication of attempting to use the observations to interrogate the data. What we have highlighted here is a key model failing due to a widely assumed model parameter (in this and other LSMs). The result of this assumption would likely be overlooked a model-data intercomparison because they would focus on the LE data-model mismatch, and thus not observe the midday dip in E, which could result in wrongly attributing the error to an incorrect process.

p. 6858 l.20. Is there any confirmation that this boundary layer hypothesis is correct? Can the values of boundary layer conductance be provided in support? Are these values defensible?

We now add a new figure, 5 and supporting text both in the new results (Line 374-392: Decoupling factor), figure 5 and discussion section (lines 600-638 to demonstrate this boundary layer issue. We also discuss CABLE's assumption in comparison to the wider literature and assumptions in other LSMs.

Global results I would suggest a restructuring of this section. We are presented with many tables and figures, but without topic sentences to highlight the critical results. It would help the reader to have the salient points of this comparison presented step by step, with reference to specific figures and tables to provide direct support. Currently the reader is referred to 3 figures and 2 tables in first few lines, without guidance as to the key points.

We have restructured the global results section, breaking up the text and separating the text related to GPP from E. We have also added sub-sections similar to the results to help guide the reader

I notice that MED-L differs by seemingly similar magnitudes to MED-P and MED-C from the LEU model – it would be useful to have a statistical analysis presented in the text (i.e. Mean % differences in each case). It is helpful that errors are provided for the mean PFT analyses with each model (although exactly what these errors are is not explained in the table captions). I would like more discussion of what these errors mean and how they affect the interpretation. For instance, if the errors are larger than the differences, I suspect we assume there is no significant difference. If this approach were used, it would be possible to highlight in the tables which differences we should take notice of as important.

We have added text to both tables 6 and 7 to clarify this in line with the reviewer's suggestion. We have also added mean % difference to each case as suggested in the main body of the text.

I also wonder why the error on deciduous needleleaf is the smallest in table 6, and yet this PFT lacked calibration data, so one would expect a large error.

Not necessarily so. Whilst we did not have appropriate data to derive a PFT parameter we did assume that the evergreen needleleaf forest parameter would be an appropriate parameter to use for the PFT. It does not follow that this PFT should have the largest error. In fact if the error is actually small, one interpretation might be that our assumption was a valid one and that the g_1 parameter for evergreen needleleaf forests is likely a reasonable representation of deciduous needleleaf forest stomatal behaviour.

Despite the authors' assertion in section 4.1, GPP and ET are not principally controlled by gs (and climate) – soil moisture and LAI are other (and often more) important variables in reality (and in LSMs). The short term response to a change in climate (e.g. more drought) may be an adjustment in gs, but the long term response is an adjustment in LAI. Do these variables (LAI, soil moisture) differ among any of the simulations with the various calibrations? i.e. We need to know whether it is just gs variation that is generating the model differences. . .. [you do confirm prescribed LAI later in the discussion I see, but this really needs to be set out in the Methods].

We did clearly indicate our assumptions relating to both soil moisture and LAI in the methods section. In regards to soil moisture in section 2.2, we state: "*β represents an empirical soil moisture stress factor. For these simulations we used the standard*

CABLE implementation throughout". In section 2.3 we state: "*For both the site-scale and global simulations, LAI was prescribed using CABLE's gridded monthly LAI climatology derived from Moderate-resolution Imaging Spectroradiometer (MODIS) LAI data*". To improve clarity we have now moved the text relating to the soil moisture function to section 2.3 and the text now reads: "*For both the site-scale and global simulations, LAI was prescribed using CABLE's gridded monthly LAI climatology derived from Moderate-resolution Imaging Spectroradiometer (MODIS) LAI data. In all simulations, we used the standard soil moisture stress function, β, defined in Equation 3*.".

We have also added text to the discussion highlighting that the assumption to do with using a MODIS LAI climatology could be a potential cause of model-data bias (lines 582-590).

It is also important to register that model-data mismatches for GPP/ET may be significantly affected by LAI and soil moisture uncertainties, i.e. better gs predictions may not be the answer to a perceived problem. I would like the authors to discuss this issue.

We agree with the reviewer on this point and have added text in the discussion. In relation to soil moisture we state: "*Inadequate simulation of soil moisture availability by LSMs is often identified as a key weakness in surface flux prediction (Gedney et al. 2000; Dirmeryer et al. 2006; Lorenz et al. 2012; De Kauwe et al. 2013b). In LSMs, as soil moisture declines, gas exchange is typically reduced through an empirical scalar (Wang et al. 2011) accounting for change in soil water content, but not plant behaviour (isohydric vs. anisohydric) (Egea et al. 2011). Bonan et al. 2014 recently showed that during drought periods, the formulation of the soil moisture stress scalar was likely to be the cause of error in g_s calculations, rather than the g_s scheme itself. Zhou et al. (2013, 2014) demonstrated that the g_1 parameter could be linked to a more theoretical approach to limit gas exchange during water-limited periods, by considering differences in species water use strategies.*" And LAI: "*Another avenue of potential bias may relate to the use of a prescribed (as is typical in LSMs) MODIS LAI climatology, which has been reported to be inaccurate over forested regions (Shabanov et al. 2005; De Kauwe et al. 2011; Sea et al. 2011; Serbin et al. 2013). It is important to note that the sensitivity to stomatal parameterisation may be larger*

when using prognostic LAI. In prognostic LAI simulations there may be feedbacks from changes in g_s to LAI that could cause larger differences between the Medlyn and the standard Leuning model, both in terms of the different timings of predicted flux maximums and associated feedbacks on carbon and water fluxes"

The paragraphs in the discussion are long and hard to follow. Paragraphs have multiple topics, switching from ET to GPP without warning. Please restructure, improve the topic sentences, and refer back to figures and tables consistently.
We have revised the discussion text as the reviewer suggested, clearly separating discussion topics.

Also, I get lost among the model calibrations – when the test reports a parameter value was "used in CABLE", which model run is being referred to?
We have clarified the text so that it is clear which model we are referring to.

p. 6861. The discussion on boreal forests lead to a warning about ET modelling, but I don't understand it. The recalibrated g1 parameter in the MED model reduced boreal ET. I suspect this tells us something specific about the CABLE model, rather than some general result about gs in boreal regions. ET in boreal latitudes is a complex outcome of moss, understorey and forest canopy interactions with snow and permafrost. Are these processes included in CABLE?
We have removed this paragraph in line with both reviewers' suggestions.

Section 4.2 on the g1 parameter seems to reprise the results of previous papers. There are no references to figures or tables produced in this paper. I would suggest this entire section be removed, or significantly rewritten to link to the current work.
This section attempted to address the issue the reviewer raised of what makes this approach "optimal", highlighting the potential avenues in which further model development could proceed. We have now revised the text more fully integrating it in the new section 4.1 in the discussion.

Section 4.3 What is the relevance of the first paragraph? Some potential values of g0 are mentioned, but there is no conclusion.

We have redrafted all the text in relation to g_0 in a new subsection 4.4. In addition we have added a new supplementary figure (S2) supporting the choice of g0 values used in this paper.

Section 4.4. It is good to read here about other components of the land-atmosphere exchange pathway in CABLE, and issues with boundary layer conductance. I would expect that comparison with eddy flux data, including LE, H and Rnet would allow testing of these problems. In fact I expected this would be the role of figures 3 and 4, although the paper pays little attention to these figures and the model-data mismatch. It is good also to read here about compensatory effects that can minimise the role of gs on ET. Can the authors Bonan et al (2014) found improved simulation with their optimal scheme during drought periods – that should be noted here.

We do highlight the Bonan result in our new discussion section of 4.1: "*This result is similar to that of Bonan et al. (2014), who implemented the optimal stomatal conductance scheme the CLM LSM, following Williams et al. (1996). In their implementation they solve the optimisation problem numerically (Eq. 1), with the additional assumption that leaf water potential cannot fall below a minimum value, effectively replacing the empirical soil water scalar used here (Eq. 3). As we did, Bonan et al. (2014) found that model performance using the optimisation scheme was not degraded when compared to the original empirical stomatal conductance (Ball et al. 1987) scheme.*" And we discuss the findings of Bonan in reference to our drought text in the same section: "*Bonan et al. 2014 recently showed that during drought periods, the formulation of the soil moisture stress scalar was likely to be the cause of error in g_s calculations, rather than the g_s scheme itself*". Beyond this, we do not discuss this further as these improvements are outside the scope of this paper and relates to further assumptions they make to do with leaf water potential (see above), which are not relevant to our changes in the stomatal conductance scheme in this paper. We have highlighted a potential avenue to improve drought modelling, linking the new gs scheme with the previous work by Zhou et al. (2013; 2014), however this is not explored here, but is the subject of ongoing work.

p. 6846, l. 26: vegetated surface evapotranspire – not all losses are through stomata.

In the original manuscript we did not state "all" fluxes are via the stomata, but rather that most occur via the stomata. We have added additional text to clarify interception losses as well: "*This latent heat exchange involves a transfer of water vapour to the atmosphere; for vegetated surfaces this transfer (i.e. transpiration) occurs mostly through the stomatal cells on the leaves as they open to uptake CO_2 for photosynthesis, but also includes interception losses from the canopy.*"

p. 6847, l. 19. There are examples of more mechanistic stomatal models that have been widely tested, e.g. SPA model of Williams et al. Please make this clear.
We do highlight two such examples in our original text: "*Whilst more mechanistic g_s models have been proposed (e.g. Buckley et al. 2003; Wang et al. 2012)…*". We disagree with the reviewer that SPA would be an example of such a mechanistic model. The SPA model is similar to the model implemented here, but has additional assumptions related to leaf water potential that effectively replaces the requirement for a soil moisture stress function.

p. 6848, l. 5. I can't find any reference in the Bonan et al. paper that their calculations are "highly computationally expensive".
This point is not directly raised in their text, but it is nonetheless true. They solve the optimisation problem by iteration, which adds an additionally looping constraint to any model implementation. We have removed any reference to this in our new version of the manuscript.

p. 6849, l.. 17. "excessive evaporation". This would suggest the problem with CABLE relates to soil or wet leaf evaporation, and not the stomatal modelling of transpiration.
We have amended the text as follows: "*Similarly, Lorenz et al. (2014) showed that CABLE when coupled to ACCESS, predicted excessive ET across much of the northern hemisphere, leading to unrealistically small diurnal temperature ranges. The new stomatal parameterisation predicts reduced transpiration across northern latitudes (Figs. 8d and 9d), however this only results in a small improvement in the spatial agreement when compared with the GLEAM product (Table 9), suggesting that there are other causes not related to g_s for the model-data bias.*"

p. 6866, l.5. It is not correct that numerical solutions behave incorrectly, compared to analytical solutions, for simulating optimal stomatal responses to increase CO2.

We disagree with the reviewer on this point. We have added text to the main document clarifying this: "*In this instance, the analytical solution is preferable to the numerical optimisation because it correctly captures stomatal responses to rising atmospheric CO_2 concentration, whereas the full numerical solution does not. In the full numerical solution, optimal stomatal behaviour differs depending on whether RuBP regeneration or Rubisco activity is limiting photosynthesis, and the predicted CO_2 response is incorrect when Rubisco activity is limiting, unless the stomatal slope g_1 is assumed to vary with atmospheric CO_2 (Katul et al. 2010; Medlyn et al. 2013). The analytical solution, in contrast, assumes that stomatal behaviour is regulated as if photosynthesis were always RuBP-regeneration-limited, which yields the correct CO_2 response.*"

p. 6853, l. 11. Do you mean you fitted eq 7?

Yes, we have corrected this.

p. 6853. I am confused as to why results are presented in the Methods section.

We have moved this text to the results section.

References.

Please check the reference list, there are several citations that are missing.

We have fixed the missing references.

Table 6 and 7 captions. Please explain what the +/- means.

We have fixed the captions.