Response to interactive comment by M. G. De Kauwe (Referee) on "Water stress induced breakdown of carbon-water relations: indicators from diurnal FLUXNET patterns"

Nelson et al. present a novel exploration of FLUXNET data to derive two new data-driven drought metrics. I think their approach is very interesting and this could form part of a more nuanced way to benchmark land models in the future. Ultimately, whilst I feel that this paper should be published, I think the text requires quite a lot of clarification and redrafting.

I found the introduction & methods text a bit disjointed. There is very little in the way of text to explain or set the "why is drought a big problem" argument. To me, there was a strange focus on considering the proportion of transpiration to ET. The authors spoke a lot about uncertainty, but it wasn't very clear if their focus was data (flux measurements) or model world. In short, the authors could do a better job of framing the broader problem before they get to the hydraulic and non-stomatal limitation text. In the methods, you often jump or introduce new concepts with little back story and it becomes quite distracting.

I also think the analysis of the results could be more incisive. I strongly feel the authors are doing themselves a disservice in terms of likely citations by not picking through their results a bit further (see comments).

Finally, whilst I shouldn't have to, it still true that it isn't the norm to share code, so I applaud the authors. I suspect it is likely to lead to their work being more widely used and potentially improved upon.

> We would like to thank Dr. De Kauwe for this very in-depth and thorough review. As authors it is always welcome to get such thoughtful outside opinions to further strengthen and clarify a manuscript. We hope the revised manuscript has a much better flow which allows the reader to follow our line of thinking and limit the distractions, as well as the expanded analysis and discussion. Following are the detailed responses which we hope address these overall concerns.

Introduction ————

1. I don't immediately follow why you've introduced VPD into equation 2? Surely your estimation of GPP and ET have both already accounted explicitly for a VPD dependence? Then on line Pg 1, line 19 you say "more consistent" … more consistent with what? This is probably simply my ignorance, but I'd like to follow the logic here because you then use VPD in eqn 7 and 8.

   > We agree with the reviewer that this concept may have been inadequately introduced. It may have been rather distracting to introduce both the uWUE and iWUE equations, and as only the uWUE was use in the calculations of the DWCI and the Katul model. As such, the iWUE references and equations have been eliminated. Furthermore, we have tried to highlight that the $\sqrt{VPD}$ allows for a more stable metric which can be compared across timescales, which becomes important when comparing diurnal GPP and

ET in calculating the DWCI. Therefore we have rewritten this section, which also incorporates comment 3 from reviewer 3. This subsection now reads:

> Classically, vegetation water and carbon fluxes are linked by stomates, where an open stomate allows CO2 to enter the leaf and, consequentially, water is lost. From this, most theoretical frameworks make some form of assumption that carbon assimilation (A) and water losses (T) are both contingent primarily on leaf stomatal conductance (gs). This assumed relationship allows us to pass between the realms of carbon and water, based on the assumption that at any given time both A and T are proportional to the stomatal conductance multiplied by the difference in internal and external CO2 and water vapor concentrations. More specifically,
>
> $$A = g_s \cdot \Delta c \quad and \quad T = 1.6 \cdot g_s \cdot \Delta v$$
>
> where Δc and Δv are the differences in inner and outer stomatal cavity concentrations of CO2 and water vapor, respectively. These diffusion equations lead to the relatively consistent carbon:water ratio, generally expressed as a water use efficiency ($WUE = A/T$). At the ecosystem level where direct measurements of A and T are not available, WUE is simply calculated as the ratio of gross primary productivity (GPP) to total evapotranspiration (ET) [7]. These carbon:water links are fundamental to understanding how stomata are regulated and underly key functioning in mechanistic plant and ecosystem models. One such set of models are those based on optimality theory which posit that plants tend to optimize carbon gains to water losses, such as those described by Katul et al. [5] and Katul, Palmroth, and Oren [4]. These concepts from Katul, which carry the assumptions of RuBISCO (light) limitaion, were built upon by Zhou et al. [14] and Zhou et al. [13] to give the equation,
>
> $$uWUE = \frac{GPP \cdot \sqrt{VPD}}{ET}$$
>
> where the $\sqrt{VPD}$ accounts for the stomatal response to vapor pressure deficit (VPD). Accounting for the VPD response allows for a more stable metric of WUE that is temporally more stable and physiologically more meaningful, such as when comparing the diurnal cycles of carbon and water. As ET is the sum of both T and non-biological evaportaion (e.g. soil and intercepted evporation), often periods during and shortly after rain events are excluded from WUE estimates to minimize the influence of non-plant evaporation. Ultimately, calculations of WUE provide a simple summary of the cost in water per carbon gain and becomes an indicator for how plants have and will adapt to the physical limitations of their changing environments [6, 11].

2. Pg 2, line 2: “propagate errors …” I assume you mean in terms of a model? As actually measured fluxes would account for any drought signature? Please clarify.

> Indeed this wording was unnecessarily vague. Here we were referring to assumptions on carbon and water in the models which then affects carbon and water estimates. To clarify this point the text has been changed to:
>
> > This failure to capture the effects of drought is not only disconcerting as water limited conditions are when ecosystems are most at risk, but an incomplete framework tends to propagates errors and uncertainties from models into estimates of the water and carbon cycles.

3. Pg 2: The arguments about the uncertainty of T as a proportion of ET … do we really think that this the chief uncertainty here is drought? To me it feels like an odd framing of the argument simply because I wouldn’t expect water stress to dominate the water cycle and the uncertainty range quoted is large. I think this line of argument would be improved by simply talking about the need to understand the carbon and water cycles during water stress. I’d argue for removing all of this text.

> This section came in part from the fact that this research was motivated from a transpiration angle, and also to highlight that gaps in understanding of drought responses can be reflected in both the water and carbon cycle communities. However, as other reviewers have also noted that the introduction could be shortened, and this section has maybe become less relevant, it has been removed.

Methods ——-

4. I think it would be helpful to explain why PET was calculated. The text just jumps to we calculated PET…Also, there is a brief mention of why the approach was adopted, but it should be expanded upon. Similarly with the CSWI, you just suddenly jump to explaining it without any back story for the reader.

> We agree with the reviewer that the reasoning for calculating these parameters was not made clear, instead we simply jumped into the calculations. The PET was used in calculating evaporative fraction, which gives a consistent measure of ecosystem dryness across sites with which to compare the metrics (basically a daily aridity index). This concept is now introduced in the text as:
>
> > In order to provide a consistent measure of ecosystem dryness that can be utilized across sites, the ratio of water evaporated to potential water evaporated was calculated as evaporative fraction (EF), or the fraction of actual ET to Potential ET (PET).
>
> Likewise, the CSWI was used as a replacement to the time after rain methods for screening periods with wet surfaces and likely higher evaporation. We now mention this when introducing the concept:
>
> > In an effort to minimize contributions of evaporation, the conservative soil wetness index

(CSWI) was employed which was designed to estimate whether the ecosystem is likely to have "dry" surfaces and therefore ET is likely to be dominated by transpiration. This approach requires a certain amount of evaporation to occur after a rain event before the surface is considered to be "dry" and can be contrasted to the method of removing a set time period after rain employed in previous studies [9, 1, 6].

Hopefully by introducing these topics before jumping into the details of their calculation this section with flow better and not be so jarring.

5. The screening of data to remove contributions from the soil is potentially problematic. I've seen that other authors have used 48 time slots after rain (see Medlyn et al 2017, New Phytologist and references they cite). The authors have taken a different approach, but screening GPP < 5 g m-2 d-1 seems quite high? Presumably as you get a drought, GPP drops and this may remove some of the signal you seek to explore? Similarly an air temp of 15 deg C. Whilst admittedly not "warm", doesn't it depend where you are? There are many locations with variable day-to-day temp, even in summer. Did the authors explore any sensitivities to these assumptions?

The reviewer makes a good point that these thresholds deserve further examination. As outlined in the previous response, the CSWI was designed to be an improvement to the time after rain method, as this does not account for differences in evaporation, say if there is little radiation in the 48 hours after the rain event. However, they are difficult to compare, as they are not immediatly comparable. Upon further analyis, the median time period for the CSWI to go from fully wet (CSWI=5) to "dry" (CSWI<=0) was 3.5 days in summer (summer being periods when daily potential radiation above median daily potential radiation for that site), so more conservative than 1-2 days after rain, but on the same order of magnitude. This comparison has been added to the manuscript:

> This method was used over the more standard method of removing 1-5 days after a rain event, as it does not make the assumption that the surface will dry in a fixed amount of time, instead relying on a minimum amount of evaporation. As a comparison, the median time period for the CSWI to go from fully wet (CSWI=5) to "dry" (CSWI<=0) was 3.5 days across all sites in summer, where summer was defined as the period when daily potential radiation above median daily potential radiation for each site.

As to the GPP and air temperature limits, we have added a sensitivity analysis showing the response of DWCI and C^*_{LE} to EF within various ranges of GPP and air temperature, Figure S2 (here R1). Based on this analysis, we looked for thresholds which would show a strong signal in the metrics under increasingly dry conditions (lower EF). In regards to temperature, our initial threshold does seem to be a bit high, as temperature thresholds above 5° C showed good metric response with reasonable variability under unstressed conditions particularly with the DWCI. As to GPP, is seems we had a problem with the units

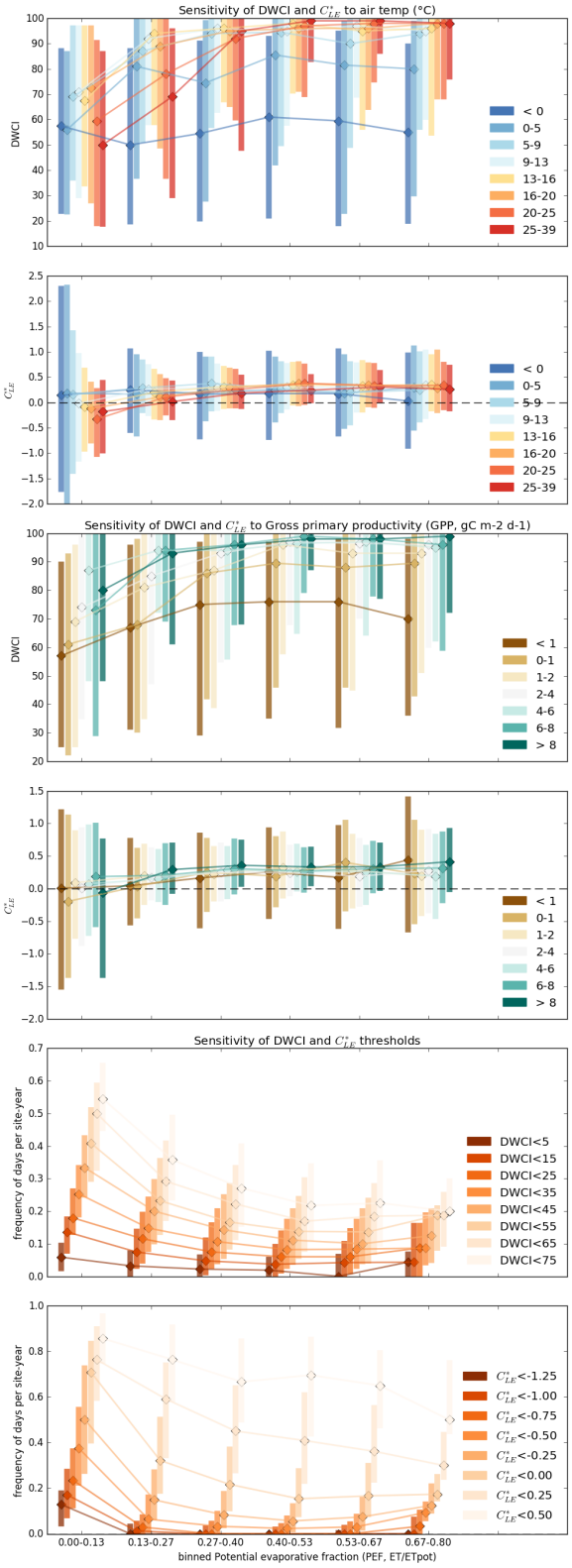

Figure R 1: Figure S1. Sensitivity analysis of DWCI and C^*_{ET} to evaporative fraction (EF=ET/potential ET) under different GPP and air temperature (Tair) values, as well as the sensitivity of frequency of decoupling (DWCI < threshold) and morning (C^*_{ET} < threshold) using various thresholds. Note DWCI of 0-100 indicate lowest-highest probability of diurnal carbon:water coupling and C^*_{ET} of -1-1 indicate one hour morning shifted-one hour afternoon shifted ET. Bins for GPP and Tair based on percentiles to give an equal number of points per bin, with the exception of GPP<1 and Tair<0, which are approximately double the size of the other bins and represent periods of low plant activity such as winter. Vertical bars represent interquartile range in all cases.

where the threshold was reported in grams $CO_2 \cdot m^{-2} \cdot d^{-1}$ instead of $gC \cdot m^{-2} \cdot d^{-1}$. With the unit conversion, this would correspond to a threshold of 1.36 $gC \cdot m^{-2} \cdot d^{-1}$. We would like to thank the reviewer for spotting the error which has been corrected in the text as well in the figures. Based on the sensitivity analysis in figure S2 (R1), we found that a threshold of 1 $gC \cdot m^{-2} \cdot d^{-1}$ works well in minimizing unnecessary variation in the metrics, as evidence by the smaller variability in the metrics during wet conditions. The issue is now discussed in the Data section in the Methods and Materials:

> Growing season was defined as all days where GPP > 1 $gC \cdot m^{-2} \cdot d^{-1}$ and daily mean air temperature > 5 °C. These threshold were shown to give good response in the proposed metrics while minimizing variability due to low diurnal signals, a sensitivity analysis of which can be found in supplementary Figure S2.

6. Similarly the assumption about precipitation and gap filling. What about filling it with reanalysis data? Assuming a gap corresponds to a 5 mm precipitation event strikes me as quite a big assumption? What happens if you simply assumed a gap = no rain? How important is this assumption for your results?

   As precipitation data can be quite hard to predict, and we did not have immediate access to the reanalysis data for this dataset, we found it prudent to simply remove those days. In this way the data screening is conservative, erring on the side of caution by removing periods if we are not sure if it has rained or not. This conservative take is possible due to the large size of the dataset. Perhaps a further analysis with the FLUXNET 2015 database would yield more data points as some of these issues have been address in the new dataset, such as inclusion of reanalysis data gap-filling. However, to make this point clear in the manuscript, we have amended the text to read:

   > Any gaps in the precipitation data were assumed to be a precipitation event of 5 mm in order to prevent any unmeasured precipitation event from biasing the results by inadvertently including rainy days.

7. Page 5, line 24. Is there any evidence of this shift? I'm not arguing it isn't true, but the authors don't cite any supporting literature. Later on in the text the authors cite Wilson, but are there any other citations? It would be good to support this point. Figure 1 is nice and useful for demonstrating the authors point.

   Indeed there has been work on shifts in ET fluxes in the literature. We have added to this section to highlight some of the relevant work, as well as further highlight the work of Wilson and it now reads:

   > As soils dry, it becomes more difficult to transport stem and root zone moisture to the leaf, potentially causing hydraulic limitations for the plant to transport water. This shift was seen in eddy covariance data in a study by Wilson et al. [12], who examined the shift of latent compared to sensible heat, which suggested that a shift in water fluxes towards dawn can be indicative

of afternoon stomatal closure. Shifts were further explored in a modeling study by Matheny et al. [8] which found that the morning shift was not well captured by models and attributed the errors to inadequate hydraulic limitations in the models. The daily cycle of wetting and drying acts as a capacitor in the hydraulic circuit, allowing water stores to be more easily transported in the morning and depleting in the afternoon. As bulk soil moisture declines, this effect may be strong enough to shift the diurnal cycle of ET significantly toward the morning. Quantifying diurnal shifts in EC data using the diurnal centroid was first explored by Wilson et al. [12]: defined as the flux weighted mean hour, or

$$C_{flux} = \frac{\sum flux_t \cdot t}{\sum flux_t}$$

where t is a regular, sub-daily time interval.

8. Page 8: similarly to where I've made this point before, you really need to introduce things better. Suddenly the text jumps to the "Katul" and then the "Boese" models, with little or no back story. To this point I've found this paper really interesting, but these jumps honestly make it hard to follow and are quite distracting, so I hope I'm being constructive here.

   Please see the response to the next comment.

- I don't really follow the benchmarking models? As to get ET, you use GPP derived from flux data and then measured VPD and Rg? Why do you need a benchmarking model? To me, you simply need to apply the method to the raw data?

  We find the reviewers comments very constructive and indicative of the high level of attention given to the review. We agree this has not been made clear and, as a consequence, has made subsequent parts of the manuscript unclear. The metrics were always calculated from the original flux data. The point of the model exercise was to benchmark whether the metrics were capturing information that the models were unable to predict, thus indicating that we are indeed providing novel information with these metrics and not something the models already capture. The idea is that if the models are unable to capture the variability of WUE, but the metrics are, then this indicates that the metrics are indeed providing some new and useful information that the models could benefit from. Here the Katul model would represent something closer to theory with some underlying assumptions, and the random forest would be at the opposite end of the spectrum, being wholly empirical with no assumptions. Therefore, if the metrics are capturing variability that the Katul model is not, it could indicate a problem with our assumptions in implementing the Katul model. However, if the metric captures variability that the random forest is not able to capture, it could indicate we are indeed adding new information that not even empirical methods are capturing. As this seems to be a key point that was causing confusion, in addition to the expanded section in the Introduction as discussed in comment 1, this section has been revised to outline

the motivation for using the models and now starts:

> In order to benchmark whether these metrics are capturing information that is possibly not being captured in modern model frameworks, three simple models were used to estimate WUE (GPP/ET) for each day at each site and compared to actual flux data. The purpose of the exercise was to evaluate if bias in the model predictions were associated with decoupled or morning shifted days, thus indicating that the metrics are corresponding to information that the models are unable to capture. Here we utilize three models to provide a spectrum of theoretical to empirical basis.

Results ——-

9. Why when DWCI < 10 is it reasonable that you have decoupling? That is stated as a fact without any support? Ditto CET < -0.5. Can you not demonstrate this for a case study i.e. the 2003 summer data from Europe, or similar?

> Based on the revieweres suggestion, we have added a case study of 6 sites during the 2003 heatwave, the results of which can be seen in the new Figure 3 (here R2). This new section shows that both monthly median values and interquartile range of DWCI and C^*_{ET} did respond during the heat wave years, although the response varied both between the sites and in timing that sites showed a response. The results of this case study have been summarize in the manuscript with the following addition:

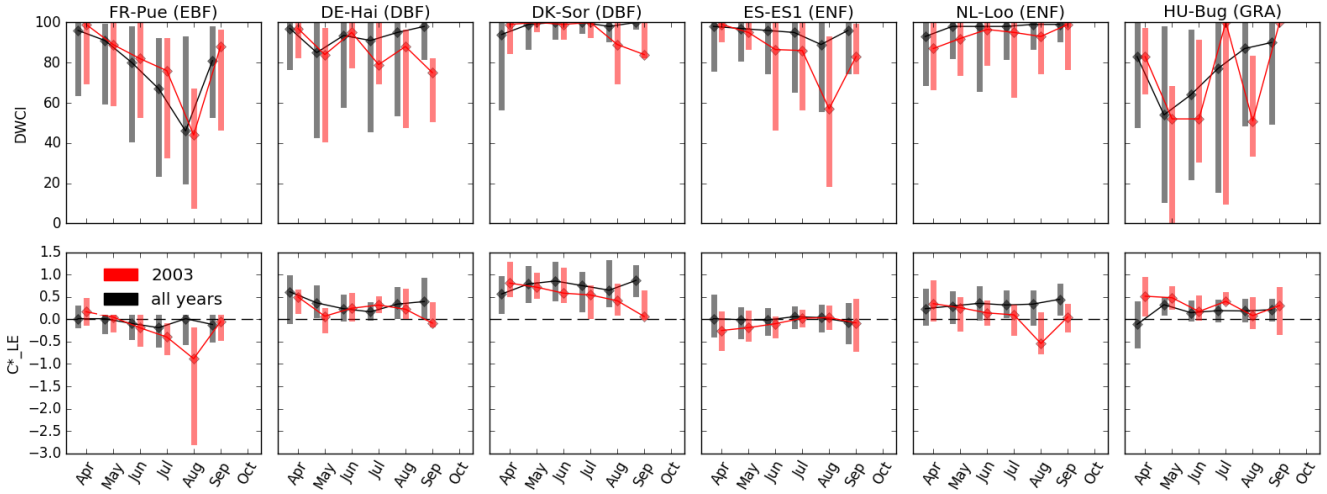

Figure R 2: Monthly median diurnal water carbon index (DWCI, lower panels) and diurnal centroids (C^*_{ET}, upper panels) for 6 sites in Europe. Data from all years available (black) is compared to 2003 (red) during which a drought event resulted in high temperatures and low precipitation throughout the summer. Note DWCI of 0-100 indicate lowest-highest probability of diurnal carbon:water coupling and C^*_{ET} of -1-1 indicate one hour morning shifted-one hour afternoon shifted ET. Vertical bars represent interquartile range. Sites from 5 plant functional types: evergreen broadleaf (EBF), deciduous broadleaf (DBF) and evergeen needleleaf (ENF) forests, as well as grasslands (GRA). Ecosystems show tendancies of morning shifts (e.g. DK-Sor and IT-Mal) and carbon:water decoupling (e.g. ES-ES1 and HU-Bug) during the drought year.

> As a case study, C^*_{ET} and DWCI time-courses for eight sites from Europe are shown in Figure 3, with an emphasis on 2003 when the continent was struck by a heatwave that was shown

to effect both the carbon and water cycles [2, 10, 3]. For DWCI, forest sties showed high water:carbon coupling throughout the growing season, with the exception of Peuchebon (FR-Pue) which showed a regular seasonal cycle of decoupling. The grassland site (HU-Bg) showed a higher variability in DWCI compared to the forest sites (all others). All sites showed either a decrease in median DWCI or an increase in variability during 2003, generally in July or August, particularly at Hainich (DE-Hai), Bugacpuszta (HU-Bug), and El Saler (ES-ES1). This increase in decoupling during 2003 is consistent with the hypothesis of non-stomatal limitations being expressed in hot, dry conditions. Median diurnal centroid values across all years varied in absolute magnitude, but were generally near or above zero, i.e. the water cycle showed no shift or an afternoon shift. One exception would be the Mediterranean oak forest of Puechabon, which shows a slight seasonal cycle of morning shifts going from a slight afternoon shift to a slight morning shift during June, July, and August. During drought years, sites that showed distinctive morning shifts were Puechabon (FR-Pue), Soroe (DK-Sor), and Loobos (NL-Loo). The framework that morning shifts are associated with water stress from soil moisture depletion would be supported by the increase in morning shifts during 2003, though factors such as species composition and access to soil water would play a significant factor and could account for the differences among sites. All sites which had significantly different ($p<0.05$, Wilcoxon rank-sum test) DWCI values between 2003 and all other years except Puechabon, whereas with C^*_{ET} only Puechabon, Soroe, and Loobos showed significant differences.

From the case study, we saw that monthly median values rarely reached below the original thresholds presented, though interquartile ranges did at some sites. Therefore, a further sensitivity of the thresholds was carried out (Figure S2 (R1)), which show that across all sites the thresholds designating decoupling or morning shifted days tended to change the magnitude of frequency, but not the response in frequency to EF. However, these values were likely low, as the response of frequency to EF was rather dampened with the DWCI<10 and C^*_{ET}<-0.5 values. Therefore, we have increase the threshold of DWCI and C^*_{ET} to 25 and -0.25 respectively, which we would stress did not fundamentally change the other figures, though it does amplify their patterns. It seems, especially in the case of this large scale study, the patterns are preserved across a large range of thresholds. This point is now discussed in the manuscript as follows:

The thresholds designating decoupling and morning shifts were 25 and -0.25 for DWCI and C^*_{ET} respectively. These thresholds were chosen to highlight frequency differences between sites and were shown to have large metric responses under dry conditions while having low frequencies under wetter conditions (see sensitivity analysis in supplementary figure S2). Furthermore, these thresholds results in a similar medain frequency of uncoupled and morning

shifted days between all site-years being 8.7% and 9.4% of days respectively. The similarity in median frequencies across site-years allowed for easier inter-comparison between the two metrics. The frequency of decoupling and morning shifts using these thresholds for each site can be found in the map found in File S1.

10. The 7 to 8% of all points being decoupled at all sites. Does that make sense? Wouldn't you hypothesize differences based on the vegetation? Rather than expect to find a universal value? I realize you have large uncertainty bounds, but I wonder what the implication of that finding is? Does it imply anything about the method at all? I don't have an immediate suggestion, I'm simply surprised.

    Indeed this statement was rather misleading, as we did not mean to indicate that 7-8% of points being decoupled at all sites. The number of points varied by site, as well as by climate and PFT as seen in Figure 3 (now Figure 4 in the updated manuscript). We have changed this section to hopefully clarify this point, as well and reference the map in File S1, which also gives information of the decoupling and morning shifts on a per site basis. However, we would again like to note that these thresholds are flexible, as the patterns are robust across thresholds from ranges of about 5-75 for DWCI and -1.0-0.0 for C^*_{ET}. As seen in the previous comment, these points have been added to the manuscript.

11. It might be interesting to see figure 3 expressed in a more informative way. Perhaps by mean annual precipitation, or spring/summer precipitation and/or an aridity index? It would also be interesting to see how variable individual years are? You clearly have this information, but it is compressed in your presentation of Fig 3 and arguably this information is very interesting and I'd argue that you're selling your paper short by not exploring it. For example, how variable was 2003/2010 vs other years for European sites?

    The reviewer makes a good point that by beginning the results with a figure that was meant to give a broad overview across all 189 sites does obscure to the reader the seasonal and interannual dynamics. Therefore we have added the case study of the European sites (now Figure 3 (R2) in the updated manuscript) which shows both the regular seasonal cycles in the dry subterranean site of Puechabon forest, and the response of the metrics in 2003 across the European sites (as discussed in the response to comment 9). Furthermore, to focus on the differences in response of the metrics between tree and grass ecosystems, we have added Figure 5 (R3) which shows that tree ecosystems show a heightened response to diurnal centroid under lower evaporative fraction levels, while grasses and savannas responded more with the DWCI. This new analysis helps support the discussion on tree and grass responses (in the Discussion section). Hopefully this added analysis, along with the original Figures 3-4 (now Figures 4 and 6) will give the reader a better idea of how the metrics respond across environmental conditions and ecosystems.

12. Similarly, do you see a shift in the centroid related to specific times in the year? Which sites shift earlier? What physically can you tie this to?

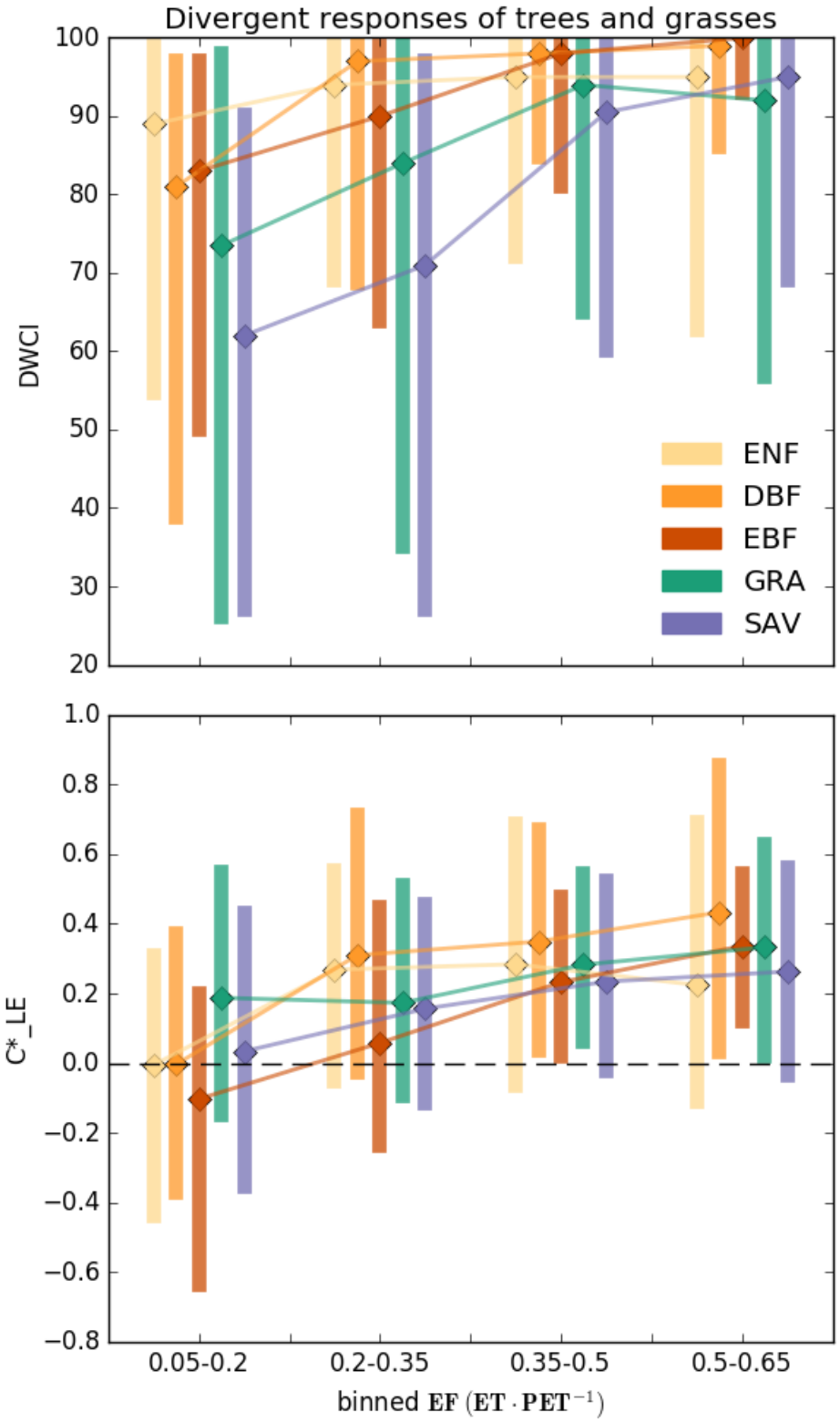

Figure R 3: Median diurnal water carbon index (DWCI, upper panel) and diurnal centroid (C^*_{ET}, lower panel) of plant flunctional types binned by evaporative Fraction (EF, low values indicate dry conditions). Note DWCI of 0-100 indicate lowest-highest probability of diurnal carbon:water coupling and C^*_{ET} of -1-1 indicate one hour morning shifted-one hour afternoon shifted ET. Evergreen needleleaf (ENF), deciduous broadleaf (DBF), and evergreen boradleaf (EBF) forests show increased morning shifts (low C^*_{ET}) with decreasing EF when compared to grassland (GRA) sites which tended to have decreased carbon:water decoupling (low DWCI) with decreasing EF. Savanna ecosystems (SAV) show a high degree of decoupling and intermediate levels of morning shifts. Vertical bars represent interquartile range.

As seen in the new Figure 3 (R2) on the European drought, some sites such as the Mediterranean oak forest Puechabon show a regular seasonal morning shift. More dramatic shifts can be seen at some sites during the 2003 drought, such as at Loobos forest in the Netherlands. The diversity of responses would hopefully indicate that the ecosystems are responding not just to atmospheric conditions but also soil water availability to the ecosystem, which we did not have the relevant data to properly analyse. However, Figure 4 (Figure 6 in the new manuscript, here R4) shows that the cleanest pattern in diurnal centroid is associated with periods of high net radiation and low latent energy, which is indicative water limited conditions. As for the DWCI, we see in the sensitivity analysis to air temperature (Figure S2 (R1)) that the decoupling tends to happen under wetter conditions at higher temperatures, so there may be a temperature component as well. We hope the added analysis focusing on particular sites during the European droughts give a better picture of how the metrics respond on different seasonal timescales.

13. I don't find figure 4 all that informative. Again I wonder if you are exploiting the interesting findings to their fullest? Which sites are most decoupled? Which vegetation types? Does it make sense to exclude the well coupled days, you're not really interested in these days?

Indeed as Figure 4 (here R4, now Figure 6 in the new manuscript) was intended to show that the response of the metrics could be seen across all sites, indicating that the metrics are indeed universally useful. However, this broad analysis then masks the dynamics seen from site to site and at different times of year. As mentioned in the previous comment, we see from the new Figure 5 (R3) that grassland and savanna ecosystems tend to decouple more and at higher evaporative fractions than tree ecosystems, possibly due to a lack of hydraulic architecture as discussed in the discussion sub-section "trees, grass, and drought stress". Regarding excluding the coupled days with regards to Figure 4 (R4, now Figure 6), as we are hoping to show the universality of the response, we want to be transparent and not mask any "false positives" one would see, as is the case of the noisiness of the response of DWCI in Figure 4a ((R4b) now Figure 6b). To highlight this point, the Figure is now referenced in the manuscript as:

> The response of both variables to drought stress is further observed in Figure 6, where low mean values of both DWCI and C^*_{ET} are associated with conditions of high net radiation and low latent energy, indicative of drought. As this figure includes all days from all sites which meet the filtering outlined in the Data subsection of the Methods, i.e. dry periods in the growing season, these figures exhibit the universality of the metrics across climates, ecosystems, and time periods.

Note the new subfigures (R4a,d), which were a response to the comments of reviewer 3 (comment 2)

14. I won't really comment on Fig 5 because I don't follow the motivation. Partly because of my question about VPD and partly because I don't see why the metrics which are data driven, require a benchmark like this? I'm not totally sold on

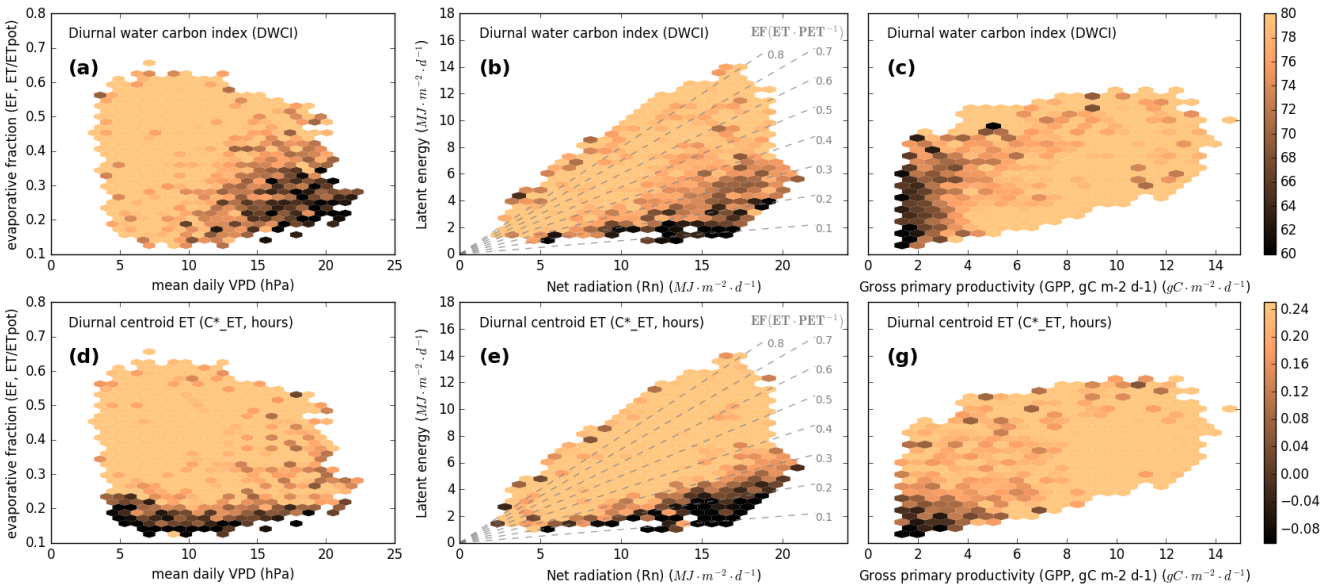

Figure R 4: Mean DWCI (upper panels) and C^*_{ET} (lower panels) with respect to evaporative fraction (EF) by vapor pressure deficit VPD (a,d), latent energy (LE) by Rn (b,e) and LE by GPP (c,g). Note DWCI of 0-100 indicate lowest-highest probability of diurnal carbon:water coupling and C^*_{ET} of -1-1 indicate one hour morning shifted-one hour afternoon shifted ET. Points with high Rn and low LE are associated with both low DWCI and C^*_{ET}, indicating that both metrics are related to water limitations. Though both metrics are associated with low EF, DWCI shows a much higher response to atmospheric demand as measured by VPD, with C^*_{ET} showing very limited response. Both metrics, and DWCI in particular, show low values with high ET and low Rn, though these points are also associated with over closed energy balances (LE+H>Rn-G). Both metrics are associated with low GPP, but the C^*_{ET} is restricted to both low GPP and ET, indicating water and carbon can decouple over a wider range of water stress. This also holds when points with energy balance over-closer are excluded (data not shown).

this being an objective means to test the approach, but appreciate why the authors have taken this approach. This is simply my opinion and I'm sure others would disagree. My first point of my discussion text below would be the way I would have been tempted to proceed.

The idea behind Figure 5 (now Figure 7) was to really get at the question of whether the metrics are giving some new information on carbon water dynamics that a model couldn't capture, as discussed in the response to comment 8. As it is clear by the responses of all the reviewers, this point was not originally made clear to the reader. In an attempt to clarify this point, we have added text to highlight this point when outlining Figure 5(now 7) which also refers back to the expanded explanation in the materials and methods:

> Figure 7 shows the difference between expected and observed WUE from the Katul, Boese, and random forest (RF) models, with respect to conditions of drought as characterized by low evaporative fraction (EF<0.2), C:W decoupling (DWCI<25), and morning shifts (C^*_{ET}<-0.25). This exercise was designed to test whether the metrics were associated with bias in the models, indicating that the metrics are able to capture information that the models are not (as further outlined in Methods and Materials subsection "models and parameter estimation").

Discussion ———-

15. How do we know the method works? What would be the best test of the method? Even if the authors don't have access to the necessary data, could they set a challenge to the community? For example, if groups had sapflux or information on non-stomatal limitations at any flux sites, do the authors have thoughts how these data could be used? How should the community push such an approach forward?

    We agree with the reviewer that though we have demonstrated the phenomenon, this broader approach would be supplemented by a site level investigation. With the added figures that show the response of the metrics during the heatwave of 2003, we hope that we give readers confidence that the metrics are indeed showing ecosystem stress. Furthermore, in an effort to make a call to the community we have added the following section to the discussion:

    > Given the broad nature of the analysis here, the metrics and hypothesis presented would benefit from site specific validations such as looking to see if the morning shits and decoupling are indeed associated with lower soil moisture levels, leaf water potentials, and/or decreases in sap flux. Sap flux in particular could give some interesting insights, as the diurnal patters in sap flux velocity will also have an offset to incoming radiation related to tree capacitance, therefore relating sap flow diurnal centroids to the ET diurnal centroid could give some information on changes in plant water recharge. Furthermore, the diurnal centroid base metrics complement

the hysteresis quantification methods such as those employed by Zhou et al. [14] and Matheny et al. [8], with the advantage of C^*_{ET} being compensation for cloudy conditions and possibly less influence of noise, though an intercomparison would be useful to explore the strengths and weaknesses of the different approaches. By providing both the equations and related code of the metrics, we the authors hope the metrics will be used by the community for both validation and to further ecophysiological understanding.

15. I think the discussion of trees vs grasses is interesting and welcomed, but I wonder if the authors looked at exploring a bit more within a functional group (i.e. by aridity etc), whether they might find something else too. Up to the authors of course.

    Based on the reviewers suggestion, we have added Figure 5 (R3) which shows a more in depth view of the tree vs grass responses based on binned evaporative fraction. The figure further shows the distinct patterns with tree and grass responses, which we have highlighted in this section of the discussion.

16. I'd argue that the authors could set aside some text to suggesting how their approach could be used in terms of benchmarking land models during drought? I'm not saying this paper has to do such a comparison, but it might be advantageous to lay the ground work. I'm guessing that the authors see modellers as potential users of their metrics? and if so, it is worth them making a case. Assessing models for responses to drought is very complicated and so their approach is welcomed.

    We agree with the reviewer that this would be a useful addition and have added the following text to the discussion:

    Though the models used here are relatively simple and lack the complexities and feedbacks found in more vigorous ecosystem models, Matheny et al. [8] also demonstrated the fundamental inability of 9 different land-surface models with 4 different stomatal conductance schemes to capture diurnal variability which the authors attributed to inadequate representation of how water gets from the soil to the leaf. Given the demonstrated phenomenon of morning shifts and decoupling across sites under dry conditions, the metrics here provide a benchmarking tool for mechanistic models to test their ability to replicate these patterns, suggesting that the models are capable of expressing hydraulic and non-stomatal limitations. Furthermore, in the case of machine learning approaches, the metrics may provide a useful input parameter which summarizes these diurnal effects, as is evidence by the bias in RF modeled WUE associated with morning shifts. As such, by demonstrating the utility of the metrics, and providing code and explanations for calculation, we hope they become useful to the community at large.

References

[1] C. Beer et al. “Temporal and among-site variability of inherent water use efficiency at the ecosystem level: VARIABILITY OF INHERENT WUE”. en. In: *Global Biogeochemical Cycles* 23.2 (June 2009), n/a–n/a. ISSN: 08866236. DOI: 10.1029/2008GB003233. URL: http://doi.wiley.com/10.1029/2008GB003233 (visited on 02/12/2016).

[2] Ph. Ciais et al. “Europe-wide reduction in primary productivity caused by the heat and drought in 2003”. en. In: *Nature* 437.7058 (Sept. 2005), pp. 529–533. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature03972. URL: http://www.nature.com/articles/nature03972 (visited on 01/16/2018).

[3] A. Granier et al. “Evidence for soil water control on carbon and water dynamics in European forests during the extremely dry year: 2003”. en. In: *Agricultural and Forest Meteorology* 143.1-2 (Mar. 2007), pp. 123–145. ISSN: 01681923. DOI: 10.1016/j.agrformet.2006.12.004. URL: http://linkinghub.elsevier.com/retrieve/pii/S0168192306003911 (visited on 01/16/2018).

[4] Gabriel G. Katul, Sari Palmroth, and Ram Oren. “Leaf stomatal responses to vapour pressure deficit under current and CO_2 -enriched atmosphere explained by the economics of gas exchange”. en. In: *Plant, Cell & Environment* 32.8 (Aug. 2009), pp. 968–979. ISSN: 01407791, 13653040. DOI: 10.1111/j.1365-3040.2009.01977.x. URL: http://doi.wiley.com/10.1111/j.1365-3040.2009.01977.x (visited on 10/24/2016).

[5] Gabriel Katul et al. “A stomatal optimization theory to describe the effects of atmospheric CO2 on leaf photosynthesis and transpiration”. en. In: *Annals of Botany* 105.3 (Mar. 2010), pp. 431–442. ISSN: 1095-8290, 0305-7364. DOI: 10.1093/aob/mcp292. URL: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcp292 (visited on 01/11/2018).

[6] Trevor F. Keenan et al. “Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise”. In: *Nature* 499.7458 (July 2013), pp. 324–327. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12291. URL: http://www.nature.com/doifinder/10.1038/nature12291 (visited on 05/22/2015).

[7] F. G. Kuglitsch et al. “Characterisation of ecosystem water-use efficiency of european forests from eddy covariance measurements”. In: *Biogeosciences Discussions* 5.6 (2008), pp. 4481–4519.

[8] Ashley M. Matheny et al. “Characterizing the diurnal patterns of errors in the prediction of evapotranspiration by several land-surface models: An NACP analysis: Error patterns in modeled transpiration”. en. In: *Journal of Geophysical Research: Biogeosciences* 119.7 (July 2014), pp. 1458–1473. ISSN: 21698953. DOI: 10.1002/2014JG002623. URL: http://doi.wiley.com/10.1002/2014JG002623 (visited on 02/15/2016).

[9] Belinda E. Medlyn et al. “How do leaf and ecosystem measures of water-use efficiency compare?” en. In: *New Phytologist* 216.3 (Nov. 2017), pp. 758–770. ISSN: 0028646X. DOI: 10.1111/nph.14626. URL: http://doi.wiley.com/10.1111/nph.14626 (visited on 01/19/2018).

[10] M. Reichstein et al. “Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: a joint flux tower, remote sensing and modelling analysis”. en. In: *Global Change Biology* 13.3 (Mar. 2007), pp. 634–651. ISSN: 1354-1013, 1365-2486. DOI: 10.1111/j.1365-2486.2006.01224.x. URL: http://doi.wiley.com/10.1111/j.1365-2486.2006.01224.x (visited on 01/16/2018).

[11] Xuguang Tang et al. “How is water-use efficiency of terrestrial ecosystems distributed and changing on Earth?” In: *Scientific Reports* 4 (Dec. 2014), p. 7483. ISSN: 2045-2322. DOI: 10.1038/srep07483. URL: http://www.nature.com/articles/srep07483 (visited on 01/12/2017).

[12] Kell B. Wilson et al. “Diurnal centroid of ecosystem energy and carbon fluxes at FLUXNET sites: DIURNAL ENERGY FLUXES AT FLUXNET SITES”. en. In: *Journal of Geophysical Research: Atmospheres* 108.D21 (Nov. 2003). ISSN: 01480227. DOI: 10.1029/2001JD001349. URL: http://doi.wiley.com/10.1029/2001JD001349 (visited on 06/13/2016).

[13] Sha Zhou et al. “Daily underlying water use efficiency for AmeriFlux sites: DAILY UNDERLYING WUE”. en. In: *Journal of Geophysical Research: Biogeosciences* 120.5 (May 2015), pp. 887–902. ISSN: 21698953. DOI: 10.1002/2015JG002947. URL: http://doi.wiley.com/10.1002/2015JG002947 (visited on 02/18/2016).

[14] Sha Zhou et al. “The effect of vapor pressure deficit on water use efficiency at the subdaily time scale: Underlying water use efficiency”. en. In: *Geophysical Research Letters* 41.14 (July 2014), pp. 5005–5013. ISSN: 00948276. DOI: 10.1002/2014GL060741. URL: http://doi.wiley.com/10.1002/2014GL060741 (visited on 02/16/2016).