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

Referee comments are highlighted in black, with our response below in red in each case.

In this manuscript, Page et al. apply the stochastic antecedent modeling framework to predict fluxes of carbon and water across a number of Australian flux tower sites. They found that consideration of lagged meteorological effects significantly improved the prediction of fluxes, and this improvement was greatest at arid sites. This result was replicated across a number of different modeling approaches that consider different ways of accounting for these lags.

The use of methods to understand the influence of temporal lags on ecosystem fluxes is a very important way to improve our understanding of ecosystems and guide the development of broader models such as TBMs, and the authors do an excellent job of highlighting this. The writing throughout the manuscript is incredibly clear, as are the analyses. As a result, I have very few technical comments. Great job!

Thank you very much for your positive summary of our manuscript.

I do have a few broader concerns, largely surrounding the precise contributions this manuscript is trying to make to the field.

- As far as I can tell, the methods and results are very similar to that of Liu et al. 2019, which the authors cite extensively. I think that a stronger case needs to be made for what new contribution is being made here. I agree that the paper makes a compelling case that lagged effects matter for improving flux predictions, but does not make much of a distinction for how this manuscript adds on to the Liu et al. work, or Samuels-Crow et al. 2020, or the other cited papers using the SAM model (or other lagged models) to predict fluxes. The primary novelty seems to be that they are using sites that are more closely distributed than Liu et al., which the authors claim can help bypass some of the confounding factors brought up in the Liu et al. paper. This may be true to some extent, but the sites in this manuscript are also quite spaced apart, and in my view is it just speculation that having a different geographic distribution of sites improves the interpretability of the results.

Thank you for highlighting this issue. The study certainly builds upon Liu et al. (2019) and we agree that a stronger argument for the manuscript's novelty is required.

We make four novel contributions in this manuscript:

1. We evaluate our approach on latent heat and evaluate how legacy affects both NEE and latent heat fluxes. Liu et al. only explored NEE. As our results show, the interpretation for both predictability and ecosystem legacy differ markedly by flux, with latent heat considerably more predictable (Figure 1)
2. We compare the SAM model to another empirical approach, which constitutes a robust, independent assessment of SAM (not in either Liu or Samuels-Crow et al.).
3. We demonstrate how to normalise the sensitivity coefficients, and hence allow comparison between environmental drivers. This is a significant step forward from Liu et al. (2019).

4. Finally, we make an important framing argument around the need to first understand “ecosystem baseline predictability” for TBM assessment, a point rarely considered in most flux model intercomparisons.

We tend to disagree with the reviewer’s point about site selection not affecting the interpretability of results. Liu et al. (2019) reviewed 42 sites located across the US, Europe, Australia, Russia and the Amazon, which cover 9 IGBP classifications (CSH, DBF, EBF, ENF, GRA, MF, OSH, SAV, WSA). Our sites cover only 4 IGBP classifications (EBF, GRA, SAV, WSA), on the same continent, while keeping a similar rainfall gradient (Liu et al, 320 – 1651 mm y^{-1} with an outlier at 3041 mm y^{-1}; our study 256 – 1491 mm y^{-1}). As such, we have reduced the confounding impact of vegetation type (and continent, which includes various environmental factors) while maintaining the aridity scale of the study. This is further reduced by focussing predominately on a single species – for instance, the woody vegetation at 9 of our sites is dominated by Eucalypt species. The NATT in particular was specifically deployed to allow for examination of savannah functioning over an aridity gradient, and it has been suggested that spatial patterns of flux here are dominated by structural changes as opposed to vegetation species differences (Hutley et al., 2011). As a consequence, we are able for example to hypothesise reasons for the low memory impact at AU-Cum and the inverse response to VPD at AU-Wom. However, we accept both the benefit of this could be better explained, and that further exploration can be performed here. We have hopefully addressed these shortcomings below.

Changes include:

L9 – We added a sentence to the abstract that highlights the novelty of the study:

> “By focussing our analysis on a single continent (and predominately on single genus), we reduced the degrees of variation between each site, providing a novel chance to explore the unique characteristics that might drive the importance of memory.”

L85 – We added a sentence to explicitly mention the NATT sites as designed for this kind of study. This highlights the novelty of applying an existing method to these specific sites, insofar as the reduction in confounding factors:

> “We include the sites of the North Australian Tropical Transect (NATT) as an explicit case study, since this "living laboratory" covers a steep rainfall gradient without a correspondingly strong change in vegetation (Hutley et al., 2011).”

L86 – We included an additional sentence to highlight the inclusion of latent heat and the purpose of this:

> “By applying the SAM framework to λE in addition to NEE, we explore how the timescales of response vary between these coupled fluxes which can improve our understanding of the processes involved in environmental memory.”

L95-100 – We have added the IGBP classifications of the sites so that the increase in homogeneity in this study is provided further evidence.

> “These vary from tropical grasslands to semi-arid shrublands and savannahs **(IGBP classifications of grassland, savannah and woody savannah)** along a steep rainfall

> gradient (312 to 1486 mm annual precipitation) running from north to south in the Northern Territory, Australia. Secondly, we grouped the Southern Australian Woodland Sites (SAWS). These were selected as sites with a greater proportion of woody vegetation than the NATT sites **(IGBP classifications of savannah, woody savannah and evergreen broadleaf forest)**…"

L376 – This has been split into two sentences that should reinforce the benefit of the k-means modelling in improving confidence in the SAM results:

> "The k-means modelling in this study has provided a novel, independent check on the suitability and performance of the SAM approach. The consistent results increase our confidence in the findings from the SAM model **and reduces the likelihood that our findings are influenced by the structural assumptions of the SAM model.**"

L381 – We have included an additional sentence here which helps to explain the importance of the k-means approach we used in comparison to SAM:

> "Since the k-means clustering plus regression often outperformed the SAM model, we have identified that the SAM approach does not provide an upper bound on the information available from the flux data. As such, our results highlight the need to explore the role of environmental memory using different approaches, including use of alternative machine learning techniques."

L388 – We also changed "the SAM approach is well-positioned…" to "the SAM approach, **together with other machine learning techniques,** is well-positioned…" which further reinforces our suggestion that multiple empirical models should be used in parallel to explore memory effects.

- While I do think that methods such as these are important tools for showing that lagged processes matter in TBMs, the discussion does not seem to offer much targeted recommendation besides claiming lagged processes matter (which is not novel). Do the results suggest any specific ways that lags could or should be incorporated? Which processes/drivers could feasibly be incorporated into TBMs? On L354 there is text about how the results could guide model development… how?

This is a legitimate comment, which we appreciate you highlighting. Unfortunately, identifying all of the precise processes involved in our results was beyond the scope of this current piece of work. We have suggested some important features that may explain the observed behaviour (e.g. VPD at AU-Wom, (xylem embolism resistance at AU-Cum), but the truth is that the current observations do not allow us to further distinguish why behaviour differs across sites. Instead, we see our work as being motivation, it should facilitate further studies to explain the behaviour we see, likely requiring further site instrumentation e.g. data from the new critical zone observatory network (Monitoring Australia's life-sustaining 'Critical Zone' resources, 2021).

In revisions, we have taken the sentence mentioned from L354 and used this as the introduction instead for a further paragraph at the end of section 4.1. This highlights some proposed next steps in identifying the processes to be included in TBMs:

> “By characterising the extent of individual site memory statistically, we hope to stimulate future site measurement campaigns, hypothesis development that examines what drives memory variability, and ultimately guide model development. As for which TBM modules would need to be adjusted to fully capture environmental memory, this approach needs to be applied at many more individual sites. This would allow us to identify functional relationships to a greater extent. However, such application needs to carefully pursued, using not just SAM but other machine learning approaches (such as the k-means clustering plus regression as we have demonstrated), to ensure that any results are process-based and not just structural assumptions from the use of a single modelling approach. By combining multiple empirical studies of environmental memory, we can understand the key lags that aid prediction of ecosystem fluxes and how these vary across site characteristics.”

- I completely see the idea behind including CABLE results, but in my opinion they do not offer much by themselves. It seems obvious that these complex lagged effect models can predict fluxes better than CABLE. For one, they are statistical models run on the actual data and not process-based models. CABLE was also not calibrated, so the actual model performance is unknown. Is there any idea how a fully calibrated CABLE might perform compared to the lagged effects models?

Yes, there is, and indeed we could motive this better. To address this, we have amended the paragraph beginning on L216 to:

> “For further comparison, we also consider the performance of an uncalibrated TBM in simulating site NEE. The TBM used was the CSIRO Atmosphere Biosphere Land Exchange (CABLE) model (Kowalczyk et al., 2006), a land surface scheme that can be run offline with prescribed meteorological forcing (De Kauwe et al., 2015b; Decker et al., 2017; Haverd et al., 2018; Ukkola et al., 2016b; Wang et al., 2011), or fully coupled (Lorenz et al., 2014; Pitman et al., 2011) within the Australian Community Climate Earth System Simulator (ACCESS; Kowalczyk et al. 2013v). CABLE models the exchange of carbon, energy and water fluxes at the land surface, representing the vegetation with a single layer, two- leaf (sunlit/shaded) canopy model (Wang and Leuning, 1998) and a detailed treatment of within-canopy turbulence (Raupach, 1994; Raupach et al., 1997). Soil water and heat conduction are numerically integrated over six soil layers (to 4.6 m depth) following the Richards equation. CABLE can be run with interactive biogeochemistry (Wang et al., 2011) and vegetation demography (Haverd et al., 2014), but both were switched off as leaf area index was prescribed on a per site basis. **CABLE is a state of the art TBM that performs similarly to other TBMs used in global coupled modelling (Best et al., 2015).** We applied CABLE to the sites uncalibrated, meaning that there was no optimisation of parameters to improve the performance at each individual site. Instead, default parameters were taken from the assumed dominant plant functional type (i.e. for savannah ecosystems, CABLE was either run as a grass or an evergreen broadleaf tree) at each flux site location. **CABLE’s reported performance at the 13 sites in this study is then essentially the performance one might expect if CABLE were run in a global coupled model – unlike the empirical models it is being compared to, it is not calibrated with site data, so in some sense this is not a fair comparison. Nevertheless, there are strong indicators that local calibration of TBMs offers relatively minor performance increases (i.e. that structural inadequacies remain), and that empirical approaches benefit to a much greater**

> **degree by the inclusion of local calibration information (Abramowitz et al., 2007). There is also compelling evidence that TBMs share biases (Haughton et al., 2016). We suggest therefore that this comparison should highlight how much more appropriate empirical approaches are for investigating ecosystem memory effects than TBMs with additional parametrisations, where existing structural inadequacies in TBMs could cloud the interpretation of the inclusion of lagged effects. Effectively,** these **CABLE** model runs represent a lower bound on the possible performance of TBMs at each of these sites **and so the** comparison between the statistical approaches and CABLE provides insight into the role of underlying site predictability (including environmental memory) in model-observation evaluations."

Line by line comments

L2-3: This sentence is a bit confusing, are direct versus long-term physiological responses supposed to be in contrast to each other?

We have changed the sentence to:

> "Constraining climate-carbon cycle feedbacks requires improving our understanding of **both the immediate** and long-term plant physiological responses to climate."

This clarifies that we are talking about the different timescales at which plants react to environmental conditions.

L11-14: Would there be an easy way to quickly describe what type of memory effects helped improve model performance? Specificity could help here.

We agree with this comment and have modified the sentence to specify that the memory is represented as the lagged antecedent drivers. As such we changed the sentence from "both fluxes were more predictable when memory effects were included in the model" to "both fluxes were more predictable when memory effects (expressed as lagged climate predictors) were included in the model."

L57: Probably most accurate to just say non-structural carbohydrates here, since 'storage' implies longer time scales than the likely lag between photosynthesis and synthesis of structural carbon.

The reference to "storage" was removed and replaced by "non-structural carbohydrates" as suggested.

L109: How were they predicted? Is this a typo? It is confusing to mention anything regarding prediction here.

Here we had used "predicted" to refer to the modelling of fluxes using the climate predictors. We have changed this to "modelled" for clarity.

L121: Where were PET data from?

The PET data is calculated by Trabucco and Zomer (2018) from the WorldClim dataset. We have added two further citations here (Zomer et al., 2007; Zomer et al., 2008) to clarify the source of this data.

L146: Are these short-term predictors half-hourly? Hourly? What time scale?

Line 116 has been amended to make it clear that the observations are collected at a half-hourly resolution and were aggregated to daily measurements. It has become:

> "All OzFlux data were extracted at a **half-hourly** timestep, **aggregated to daily data,** screened to only include complete calendar years and then mean-centred."

We also have modified L146 to

> "CLIMATE_n(t)is a weighted sum of **daily** climate **measurements**".

L151: Similarly, is long-term precipitation always 1 year? It is unclear from the text, which says "up to a year prior". Is this summed precipitation, or averaged?

We have changed this sentence to

> "Long-term precipitation (**mean rainfall calculated over varying periods**, up to 365 days prior, **Table 3**) is included…".

This should draw the reader's attention to Table 3 which provides the timeblocks that prior year precipitation was split into it.

L390-411: I think the other possibility – that the trend just doesn't hold when considering smaller sample sizes – should be given equal value here. This text could also be shortened since it is a fairly minor result in my opinion, but I will just mention that as a suggestion.

As suggested, we have slightly condensed this paragraph and included further reference to the potential impacts of a smaller sample size (where previously, this was only mentioned in the Results section). We believe that the initial section of this paragraph is important, since a key aim of the paper was to reduce site variance by considering just a single continent. As such, this was kept but the additional two examples of "lost correlation" were removed and additional sentences added to discuss sample size. Combined, the paragraph has changed to:

> "One of the key conclusions from Liu et al. (2019) was that as sites become more arid, the importance of antecedent effects increases. However, Liu et al. (2019) considered 42 sites from across the globe, incorporating a wide range of biomes and species. As such, there is potential for confounding factors to be influencing the importance of environmental memory at each site. This study reduces some of this uncertainty by limiting its scope to 13 sites, all located within Australia. This means that, as well as limiting the diversity of species and climates studied, a greater understanding of each individual site is possible. Similarly to Liu et al. (2019), these sites were grouped by biome, although we only had two groups - savannahs/grasslands within the NATT and woodlands in the SAWS group. Each biome group contains sites with a range of MAP and WI values. When these sites are viewed together, the importance of memory is strongly correlated with site aridity

> (improvement in R2 between CC and EM models, ρ= -0.72, p-value < 0.01), consistent with the conclusions of Liu et al. (2019). In contrast, when the sites are split by our vegetation groupings, this significant correlation is only seen for the SAWS group (ρ= -0.86, p-value< 0.05). The NATT sites had no correlation between site memory and aridity (ρ= -0.49, p-value = 0.36), despite having a very strong rainfall gradient. Note that both groups have ranges of aridity that include sites spanning from "arid" to "humid" with the WI at NATT sites between 0.12 and0.75 and at SAWS sites between 0.11 and 1.2 (Trabucco and Zomer, 2018). This result **potentially** indicates that grouping many sites together to explore relationships based on a single metric can obscure more nuanced understanding of the processes involved, or the key site characteristics driving such relationships. **However, this loss of correlation when grouping by vegetation type could simply be an artefact of the smaller sample sizes. By decreasing from a sample of 13 sites to 6 or 7, it is more likely for erroneous relationships to be identified (or not).”**

L401: Typo, “groupd”.

Corrected.

L440-442: This is a pretty big claim, especially considering that the basis for it is just from a few arid sites in one region. Is there any reason to believe that model performance in arid regions is better globally (or even outside of Australia)? The improvement you see could be related to model structure or data quality or other factors, as opposed to just aridity.

This is a very valid criticism and we have taken it on board. Since we argue that aridity is not a strong indicator of site predictability, this statement was at odds with the rest of the manuscript. As such, the sentence has been changed as follows to better reflect our intentions regarding how studies such as ours should guide TBM development and evaluation:

> “We suggest that models tested at more arid sites, where we report higher predictability both with and without memory taken into account, should be expected to perform better than models tested at those sites which exhibit a lower baseline predictability of fluxes.”

has become

> “Our results indicate that process-based TBMs tested at sites that exhibit greater predictability from simple empirical models, such as SAM or k-means as used in this paper, might be expected to perform better than TBMs tested at those sites which exhibit a lower baseline predictability of fluxes.”

Figure 5. The caption makes it seem like the “+ all climate lags” is equivalent to the EM model, which in my understanding, it is not. Plus, the red crosses differ from the “+ all climate lags” boxplots, which makes that wording confusing.

You are correct and this was poorly worded. The caption has been amended to more accurately reflect that the EM model and “+ climate lags” k-means model are equivalent only insofar as they are driven by exactly the same predictors.

> “"+climate lags" is the model where every lagged environmental variable is included, which corresponds to the EM SAM model.” has been changed to “"+ all climate lags" is the model where every lagged environmental variable is included, and hence utilises exactly the same predictors as the EM SAM model.”

References

Hutley, L.B., Beringer, J., Isaac, P.R., Hacker, J.M., Cernusak, L.A., 2011. A sub-continental scale living laboratory: Spatial patterns of savanna vegetation over a rainfall gradient in northern Australia. Agricultural and Forest Meteorology, Savanna Patterns of Energy and Carbon Integrated Across the Landscape (SPECIAL) 151, 1417–1428. https://doi.org/10.1016/j.agrformet.2011.03.002

Monitoring Australia’s life-sustaining ‘Critical Zone’ resources, 2021. TERN Australia, viewed 8 October 2021, <https://www.tern.org.au/news-ozczo-announcement>

Trabucco, A., Zomer, R.J., 2018. Global Aridity Index and Potential Evapo-Transpiration (ET0) Climate Database v2. CGIAR Consortium for Spatial Information (CGIAR-CSI).

Zomer, R., Bossio, D., Trabucco, A., Yuanjie, L., Gupta, D., Singh, V., 2007. Trees and water: smallholder agroforestry on irrigated lands in Northern India, IWMI research report. Colombo, Sri Lanka.

Zomer, R., Trabucco, A., Bossio, D., Verchot, L., 2008. Climate Change Mitigation: A Spatial Analysis of Global Land Suitability for Clean Development Mechanism Afforestation and Reforestation. Agriculture Ecosystems & Environment 126, 67–80. https://doi.org/10.1016/j.agee.2008.01.014