Review for "Upscaling dryland carbon and water fluxes with artificial neural networks of optical, thermal, and microwave satellite remote sensing" by Dannenberg et al.

Dannenberg et al. present an approach for estimating dryland GPP, NEE, and ET by training an artificial neural network (ANN) with remote sensing signals (optical vegetation indices, thermal observations, and microwave soil moisture/temperature). The study is novel, scientifically sound, well written and within the scope of Biogeosciences. I would recommend this paper for publication but have a few revisions I think should be addressed, mainly around paper presentation and clarification on methodology.

> We thank the reviewer for their helpful and supportive comments, and we are glad that they found our manuscript interesting and useful. We respond to each comment individually below.

Minor Concerns:

- *The structure of the introduction and methods have some overlapping material. For example, the fourth paragraph of the introduction in lines 63-72 mentions that plant physiological responses are not necessarily reflected in optical signals, but this paragraph doesn't make the connection between optical VI's that are sensitive to greenness specifically. Discussion of 'greeness'-based metrics failing comes later in the methods section in lines 141-155 but I think it would be useful to draw the connection earlier in the introduction. In addition, the same paragraph in lines 63-72 says "microwave, thermal, and visible wavelengths can capture complementary information about plant and ecosystem stress that is unattainable from optical VIs alone". An explanation as to WHY these indices are useful is available in the methods but could be moved further to the introduction.*

> These are excellent suggestions. As suggested, we have moved much of this information from the methods to the suggested places in the Introduction (section 1, paragraph #4), which now reads (new text in **boldface** and moved text in red): "**The normalized difference vegetation index (NDVI), for example,** is the most widely used vegetation index, but it sometimes fails to capture temporal dynamics of carbon and water fluxes in drylands (Yan et al., 2019; Smith et al., 2019; Wang et al., 2022). **While other optical vegetation indices overcome some of the weaknesses of NDVI,** combining different types of remotely sensed observations—such as those from microwave, thermal, and visible wavelengths—can capture complementary information about plant and ecosystem stress that is unattainable from optical VIs alone (Smith et al., 2019; Stavros et al., 2017; Guan et al., 2017). **Land surface temperature from thermal imaging, for example,** is an important determinant of carbon and water fluxes because, among other reasons, both photosynthesis and respiration involve temperature-dependent enzymatic reactions (Farquhar et al., 1980; Atkin and Tjoelker, 2003) and because it is a key indicator of latent heat flux, which cools leaves and land surfaces (Bateni and Entekhabi, 2012)."

- *The final paragraph of the introduction could be rephrased to make the hypothesis/study aim clearer. Specifically, the first sentence states, "Here, we develop and test an*

approach for data-driven prediction of a full suite of carbon and water fluxes that are specially adapted for drylands using…" but I think this can be much stronger to highlight the value of the study. Something along the lines of, "We aim to improve the prediction of GPP, NEE, and ET based on remotely sensed metrics by using…"

We like this suggestion and have adopted the language suggested by the reviewer. The first sentence of this paragraph now reads: "Here, we aim to improve estimation of dryland GPP, net ecosystem exchange (NEE), and evapotranspiration using an extensive network of eddy covariance observations and multi-source satellite remote sensing."

- *Somewhere in the methods should include the number of test/train data points used.*

This is a good suggestion, though the answer is a bit complicated. Because each member of the DrylANNd ensemble is trained with one site withheld, and since the period of record varies among the different eddy covariance sites, there is not a fixed number of test/train data points; the exact number will vary depending on which site was withheld from that particular model. For example, the site US-Hn3 has only two years of available records (2017-2018) and thus only 24 monthly flux observations. The 20 ensemble members from which US-Hn3 was withheld would therefore have more available data points in the training and validation sets than would the 20 ensemble members from which the sites with complete records during the study period (e.g., US-SRM, US-Mpj) were withheld. However, in the Methods, we do state the percentages that were used for training (75%) and validation (25%) in the development of each individual ANN; the exact numbers of observations, however, would vary. We now acknowledge this complexity in section 2.3, paragraph #2 (changes in boldface): "Each ANN in the ensemble (§2.4 below) was initiated with randomly assigned weights and biases based on the Nguyen-Widrow method (Nguyen and Widrow, 1990) and with different random subsets of observations for model training (75%) and validation (25%)**, with the precise number of data points used for each individual ANN varying slightly depending on the length of the withheld site's data record**."

- *The final paragraph of the methods discusses the authors approach for testing the importance of predictor variables. Has this approach been used in other studies? Some validation of this approach or references for more information would be useful.*

This is a good point. The methods used to test variable importance in this manuscript are novel but grounded in prior work, such as the stepwise selection approaches that have proved suitable for recognizing the most influential variables in artificial neural networks (Gevrey et al. 2003). To address this in the manuscript, we have made the following change to section 2.4, paragraph #3 (changes in boldface): "Second, we tested the leverage of each time-varying predictor variable by repeatedly (100 times) randomly permuting each variable (thus destroying its information content) and re-running model predictions**, similar to established perturbation and stepwise methods for uncovering the most critical variables in ANNs (Gevrey et al., 2003).**"

- *The color palette of figures could be adjusted to follow more a 'intuitive' color scheme e.g. dark green for ENF – this is not critical but might help with figure readability.*

  This is a fair point. We played around with some different color schemes early in the manuscript development process, but we wanted to avoid anything that had a combination of greens and reds to make sure that it's color-blind friendly. Ultimately, we prefer to stick with the existing color scheme since we find it visually appealing and are reasonably confident that the color gradients will be distinguishable by anyone with red-green colorblindness.

Line edits:

Line 37: intensity of water limitation feels like awkward phrasing

We have changed this to just "water limitation."

Line 53: It might make more sense to move this like to the end of the last paragraph so someone scanning the paper could easily find "First, Second, Third" in the three paragraphs talking about the unique nature of drylands.

While we see the reviewer's point, we think that the current placement of the "Several issues…" sentence fits best thematically in its current paragraph.

Line 54: It might be nice to define mesic

We have modified this to read (changes in boldface): "…in **wetter,** more mesic systems **where moisture tends to be more plentiful**…"

Line 59: "the effects of soil moisture stress…" but it's the effects of ALL soil moisture right?

We have changed this to just say "Soil Moisture…" instead of "The effects of soil moisture stress."

Lines 53-60: I found this paragraph a little difficult to follow as several sentences are quite long. I think it would be worth revisiting for clarity.

We have revised this paragraph (including changes made in response to the previous two comments and splitting one of the longer sentences into two shorter sentences).

Line 67: Satellite-based estimates of fPAR should still be fine, it's just that the plants aren't responding to the increase in light by being more photosynthetically active. I would rephrase this.

This is a good point. We have removed the part of the sentence that refers to fPAR.

Line 88: 'however' is unnecessary

We have made this change in the manuscript.

Line 90: can be more specific with 'uniqueness'

We have modified this to read: "… 'uniqueness' of dryland fluxes **to their specific location (i.e., low predictive power of models for sites on which they were not trained)**…" [Our use of the term "unique" in this case is referring to the Haughton et al., 2018 study, but we have tried to be clearer and more explicit here about what we mean by that term.]

Line 91: 'other places and other types of ecosystems' seems redundant

We have changed this to just say "…other regions."

Line 94: 'for example' is unnecessary

We have made this change.

Line 97-100: I would rephrase to put the emphasis on the finding of the study, not the author, and just present the citation at the end.

We have made this change in the revision.

Lines 113-117: References to sections might be useful

We have added these throughout this sentence

Line 117: 'global-scale estimates' – of ecosystem fluxes?

We have revised this to read "…global-scale carbon and water flux estimates."

Line 185: 'compositing' is confusing and maybe incorrect?

We are reasonably confident that we are using this term correctly as maximum value compositing is long-standing technique for aggregating vegetation indices and minimizing noise (e.g., Holben, 1986; Townsend & Justice, 1986). We now include the classic Holben reference directly following "compositing" to make it clear that this is referring to a specific, long-established technique.

Line 192: this statement deserves a citation

We have modified this sentence to read (changes in boldface): "ANNs are effective at finding underlying relationships within multidimensional and multisource datasets**, including nonlinear relationships and interactions among predictor variables (Olden et al., 2008)**."

Line 194: '... predictions of multiple variables.' Deserves a citation

We have added a reference to Atkinson & Tatnall (1997). Full reference is listed below.

Line 210: here could be a good place to include the number of test/train data points

As discussed above in response to Minor Comments, the exact numbers will vary depending on which site was withheld from any given ANN. We therefore think that it makes most sense to just state the percentages of data points that were used in the ANN training and validation sets, though as stated above, we now explicitly state in the revised manuscript that the exact numbers will vary among the ANNs depending on which site was withheld from training.

Line 328: 'Interestingly' is unnecessary

Good point! We have removed it.

Line 333: 'However' is unnecessary

Deleted.

Line 340: 'modeling' feels like the wrong term to use here – I think predicting or estimating would be more accurate since modeling implies process based (to me).

We changed this to "estimating."

Line 403: 'thermal data' – it might be better to say LST here?

We have made this change in the revised manuscript.

Figure 2: I think it would be useful to say what the input variables are in the figure (not just the outputs)

Great suggestion! We have modified Fig. 2 (below) to be less of a general conceptual figure and to more explicitly show our specific structure and input variables.

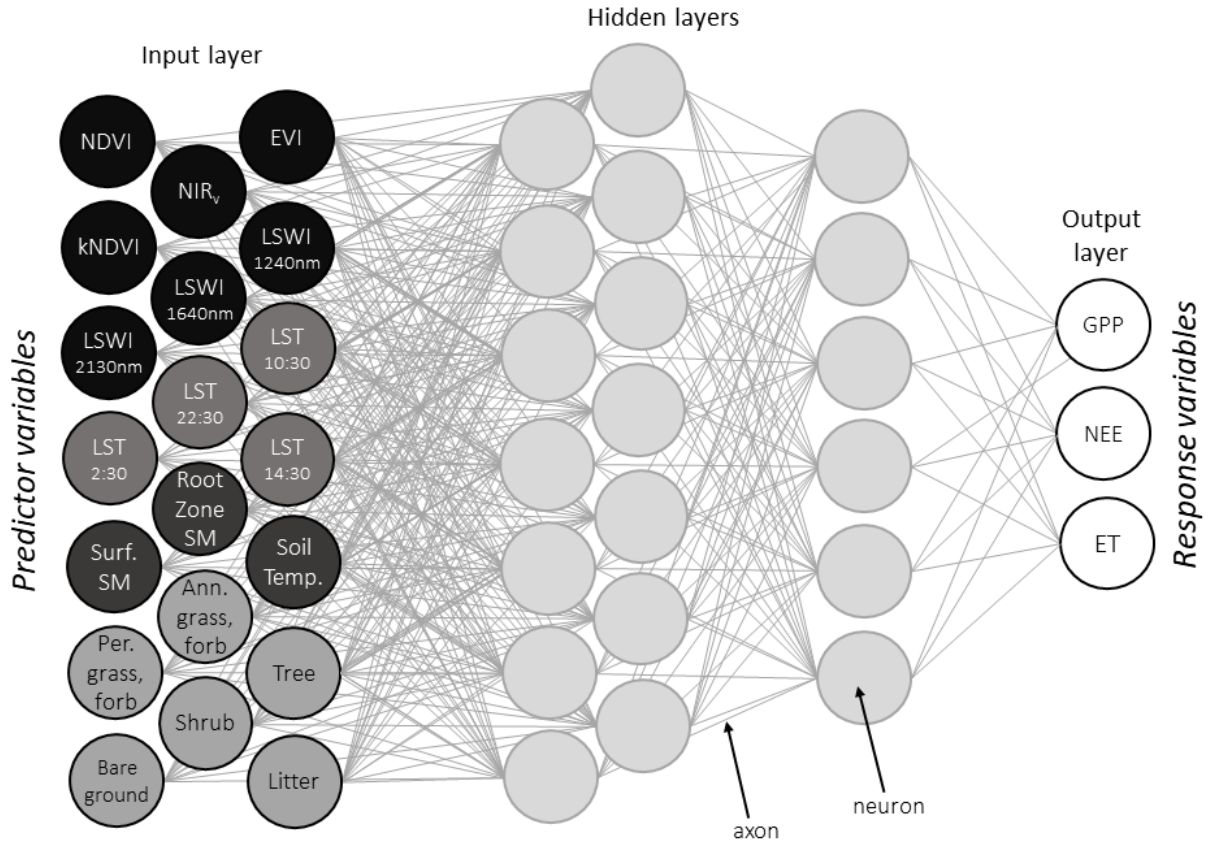
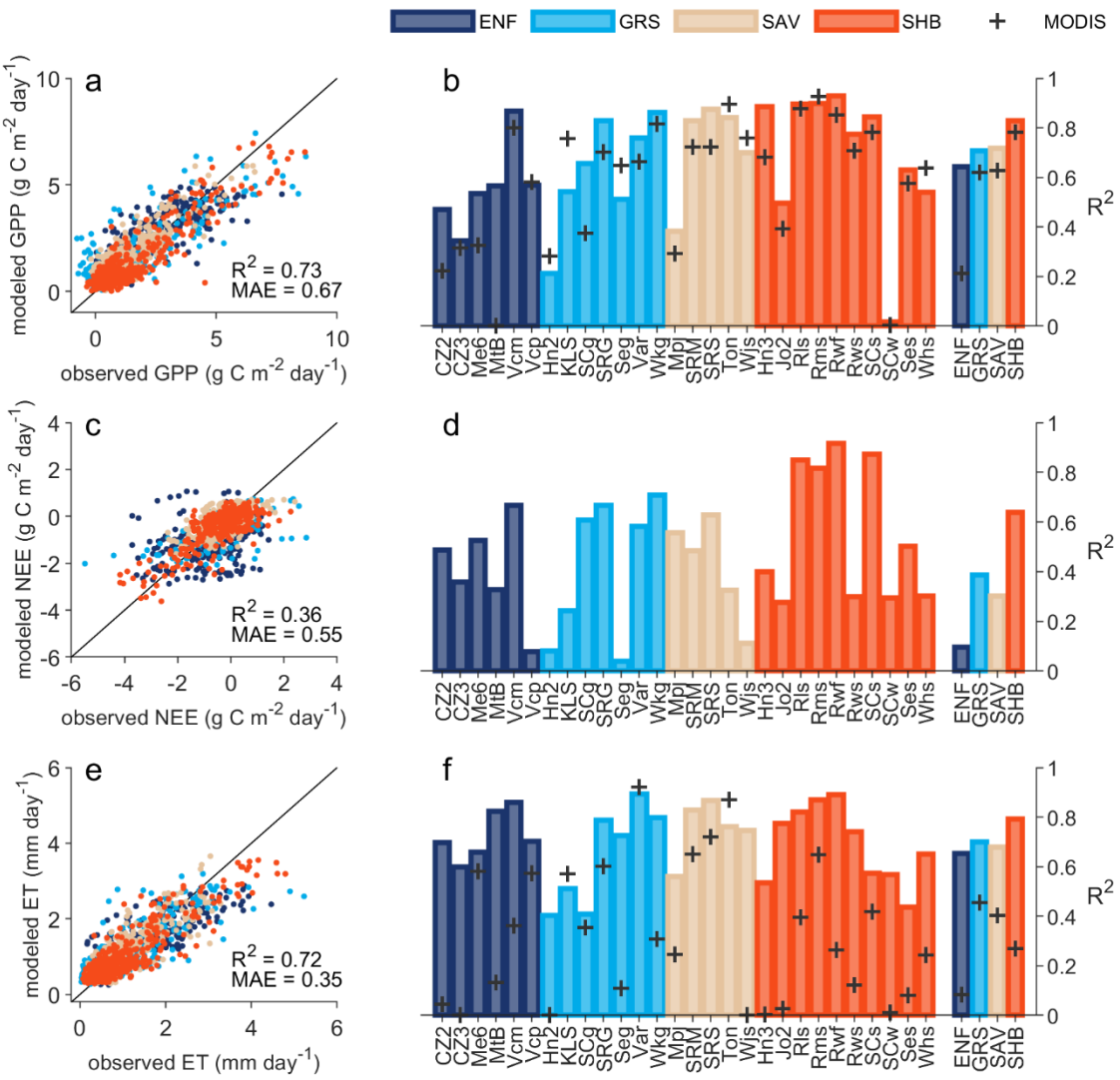


Figure 3: the + indicator is a bit difficult to see/compare with the bars – it might be easier to see in black or a different shape.

As suggested, we have made the + signs darker (below). We agree with the reviewer that this does indeed make them considerably easier to see.



Figures 5, 6: I think it would be useful to indicate on the figures somewhere which sites fall under which land cover classification category

We have now added the three-letter land cover code (ENF, SAV, etc) to the top-right of each subplot to indicate which sites belong to which cover class. (See below for the new version of Fig. 5 with the class labels. The new Figs. 6 and S1-S4 are also now labeled the same way.)

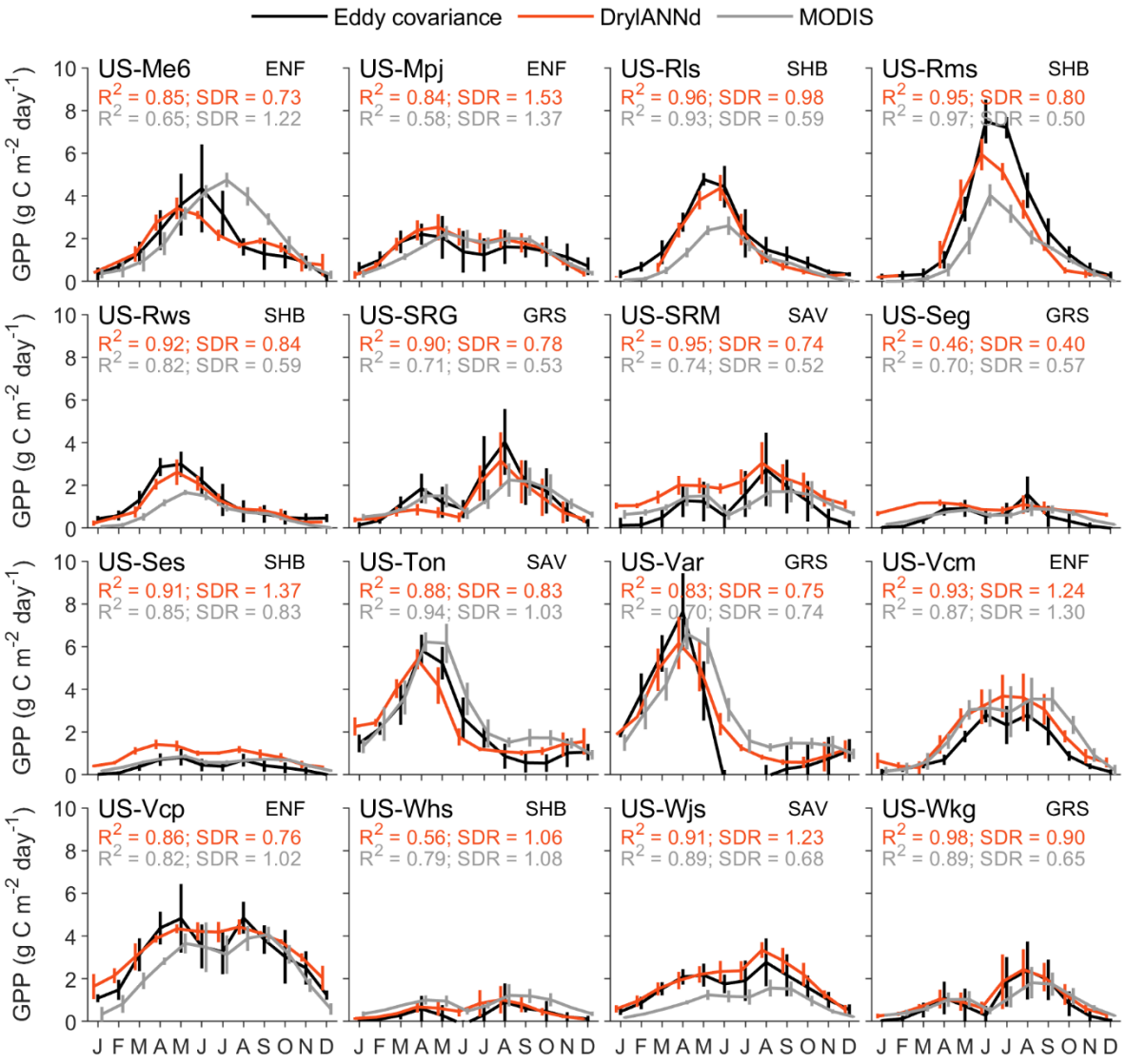


Figure 7: It's unclear to me what the lines in a and c are

The lines are showing the linear relationship between predicted and observed GPP and ET for each individual site in the flux network, the idea being to show not just how the model performs at capturing across-site spatial variation but also within-site temporal variation. We have further clarified this in the figure caption (changes in boldface): "**The orange and gray** lines in (a) and (c) show the linear relationship between estimated and observed GPP and ET for each **individual** site **during the six-year training and evaluation period**…"

Figure 8: Do the lines connecting the scatter points represent anything? If not I would remove

We had intended the lines to be a helpful visual cue, allowing readers to easily see how the variable importance changed for a particular plant functional type (as well as making

it easier to see where the points overlapped in some cases). However, we can see how this may be either confusing or add extra visual clutter, so we have removed the lines as suggested by the reviewer.

References (excluding those that were already included in the original manuscript):

Atkinson, P. M. and Tatnall, A. R. L.: Introduction neural networks in remote sensing, Int. J. Remote Sens., 18, 699–709, https://doi.org/10.1080/014311697218700, 1997.

Gevrey, M., Dimopoulos, I., and Lek, S.: Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecol. Modell., 160, 249–264, https://doi.org/10.1016/S0304-3800(02)00257-0, 2003.

Holben, B. N.: Characteristics of maximum-value composite images from temporal AVHRR data, Int. J. Remote Sens., 7, 1417–1434, https://doi.org/10.1080/01431168608948945, 1986.

Olden, J. D., Lawler, J. J., and Poff, N. L.: Machine learning methods without tears: a primer for ecologists., Q. Rev. Biol., 83, 171–93, 2008.

Townsend, J. R. G., and Justice, C. O.: Analysis of the dynamics of African vegetation using the normalized difference vegetation index, Int. J. Remote Sens., 7, 1435-1445, 1986.