Response to reviewers 03/06/2024

We thank the reviewers and editors for taking the time to review our manuscript and provide constructive comments, we believe their comments have substantially improved the overall manuscript. Accordingly, we have resubmitted a revised manuscript.

Below we provide point-by-point responses to each reviewer where reviewer comments are in bold and our responses in italics. We have coded each reviewer comment in the format [reviewer]-[comment number] (e.g RC1-1, or RC2-1 etc.). This reduces repetition as we often refer to comments/responses from other reviewers.

These responses are in large part the same as those provided during the interactive discussion as we provided detailed point-by-point responses at that stage. However, minor revisions have been made to the responses to align the responses with the actual edits made in the revised manuscript, and line-numbers have also been added where appropriate.

Kind regards,

Chad Burton

RC1

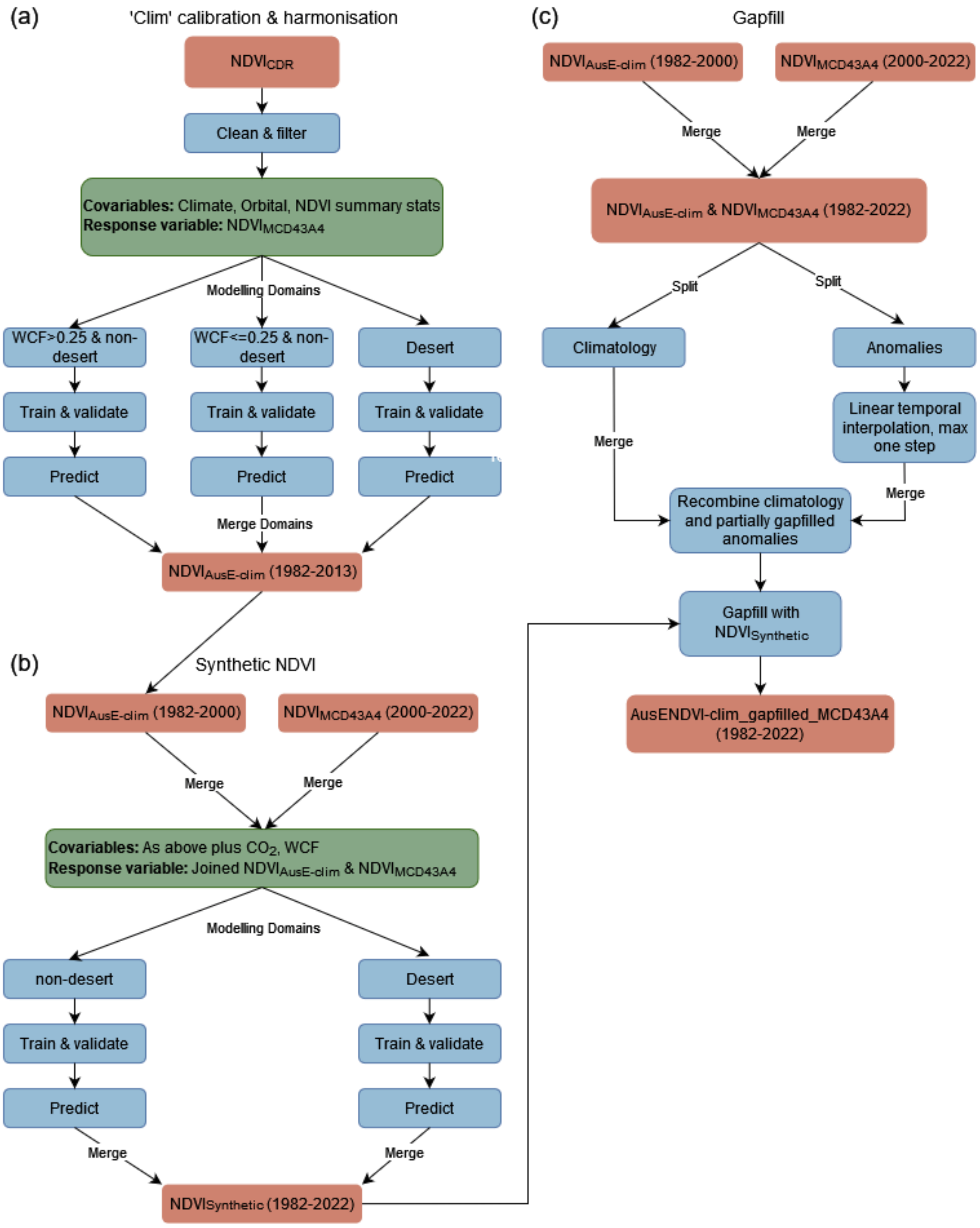
This study proposed several versions of AusENDVI and these NDVIs can be used for studying Australia's changing vegetation dynamics and carbon, and water cycles. The paper is generally organized. The new data set would be useful for the Earth system science studies. However, I still have questions about the structure of the article. Considering these and due to the following major concerns and suggestions, I would recommend it with major revision and to determine whether to accept a revised version.

Major concern:

RC1-1: I think the article needs a flowchart to show each step, which helps the reader understand the importance of the data processing process. So far I found in the section 'Data and Code Availability' that the author lists each version of AusENDVI, but in fact, I am confused about which step each version of the data is obtained through.

We thank the reviewer for noting the need for a flowchart to clarify the methods. We agree that a flowchart would increase understanding, especially since the modelling is broken up into several domains that can get confusing to follow. We had neglected to include one originally in the interests of keeping the number of figures to a minimum but are now convinced, as per your suggestion, that there is a need for it. The flowchart below will be edited into a revised manuscript with the following caption:

Figure 1: Flowchart describing the methods of calibration and harmonisation (a), and the development of a synthetic NDVI (b) for gap filling (c). a) Shows the method for the 'clim' model type, the methods for 'noclim' are the same but climate variables are removed from the covariables and 'noclim' is not gap filled. Red coloured boxes denote datasets, blue boxes denote processing steps, and green boxes describe the response variables and covariables used for modelling.


(a)
'Clim' calibration & harmonisation
NDVI$_{CDR}$
Clean & filter
Covariables: Climate, Orbital, NDVI summary stats
Response variable: NDVI$_{MCD43A4}$
Modelling Domains
WCF>0.25 & non-desert
WCF<=0.25 & non-desert
Desert
Train & validate
Train & validate
Train & validate
Predict
Predict
Predict
Merge Domains
NDVI$_{AusE-clim}$ (1982-2013)
(b)
Synthetic NDVI
NDVI$_{AusE-clim}$ (1982-2000)
NDVI$_{MCD43A4}$ (2000-2022)
Merge
Merge
Covariables: As above plus CO$_2$, WCF
Response variable: Joined NDVI$_{AusE-clim}$ & NDVI$_{MCD43A4}$
Modelling Domains
non-desert
Desert
Train & validate
Train & validate
Predict
Predict
Merge
Merge
NDVI$_{Synthetic}$ (1982-2022)
(c)
Gapfill
NDVI$_{AusE-clim}$ (1982-2000)
NDVI$_{MCD43A4}$ (2000-2022)
Merge
Merge
NDVI$_{AusE-clim}$ & NDVI$_{MCD43A4}$ (1982-2022)
Split
Split
Climatology
Anomalies
Linear temporal interpolation, max one step
Merge
Merge
Recombine climatology and partially gapfilled anomalies
Gapfill with NDVI$_{Synthetic}$
AusENDVI-clim_gapfilled_MCD43A4 (1982-2022)

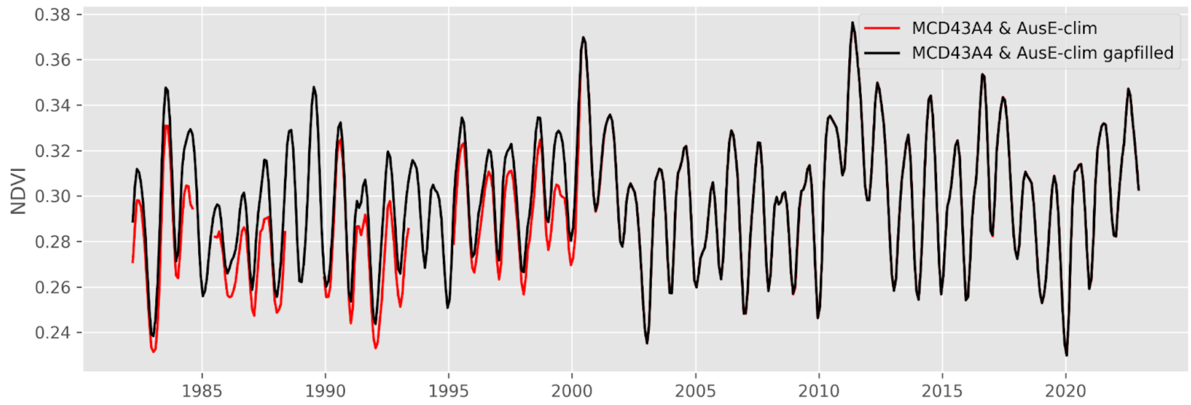
<u>RC1-2:</u> I don't think 'Quality of existing NDVIs' is the key part of the article, this part of the results could be replaced by comparing the performance of AusENDVI with other NDVIs, e.g. by adding on the performance of AusENDVI in Figure 2, and then transforming this part into the second part of the results.

We respectfully disagree that the section examining the quality of existing NDVI products over Australia is not crucial to the article. We feel it is necessary firstly to help establish the underlying scientific need to develop an Australian-specific long-term NDVI, and secondly to help educate potential users of both AusENDVI and other global NDVI products on the limitations and advantages of these datasets. Of course, there are many ways to structure the article, and your suggestion may increase the overall efficiency of the article by including more results in the same figure set. However, we argue this would reduce the emphasis on the intercomparison and thereby lessen one of the objectives of the study i.e., to determine the suitability of existing NDVI products for long-term vegetation monitoring in Australia.

<u>RC1-3:</u> I think the first part of the results could be to highlight the results of each step, especially 'before and after the calibration and harmonization' and 'before and after gaping fill'. Of course, these are already in the results, but they should be in the same section to highlight the results of each step of the enhancement.

We believe it is necessary for the results of the calibration and gap filling to be in different subsections as there are two distinct modelling efforts occurring here. In the first instance (section 3.2), we report the results of harmonising AVHRR to MODIS, and in section 3.3 we report the results from the creation of a synthetic NDVI for gap filling. The creation of a synthetic NDVI is a different enough process from the harmonisation that we argue it requires its own subsection (different models, modelling domains, and input data). Note also that the two sub-sections immediately follow each other so narratively the current structure of the article is similar to your suggestion. We hope the inclusion of a flowchart will also help clarify this and make more obvious the need to separate the results of the two different steps.

We also aim to include the plot below into Figure 7 (now Figure 8 in the updated manuscript) to show the 'before and after' results of gap-filling. As missing data tends to be in the higher NDVI regions (wet, cloudy, forested regions), gap filling has the tendency of increasing NDVI when averaging over the continent. Figure A5 in the updated manuscript also shows the time series of AVHRR-CDR before and after the calibration/harmonisation, averaged across all of Australia and broken down by bioclimatic region. We are open to including this in the main part of the manuscript at the editor's discretion but for now have left it in the appendix.



<u>RC1-4:</u> Is it possible to find field measurements of NDVIs in Australia to provide absolute accuracies for individual NDVIs, and if so, this would be an important support for demonstrating the accuracy of AusENDVIs.

In short, no. There are no in-situ field measurements of NDVI that are comparable to the spatial and temporal scales of AusENDVI (the area of pixels in AVHRR are ~25 km^2). However, note that MODIS MCD43A4 surface reflectance data (from which we calculate NDVI as the response variable for the harmonisation) is a well calibrated and validated remote sensing product, and the validation performed in our study is based on random pixels selected from MODIS. We also included a comparison with the Digital Earth Australia Landsat surface reflectance product as this product has all the same types of corrections (atmospheric, BRDF etc.) (Byne et al. 2024) as MODIS MCD43A4 and is therefore a fair and independent inter-comparison dataset.

<u>RC1-5:</u> The discussion is too lengthy, my suggestion is that it could be broken up into subsections.

We will revise the manuscript to include subtitled sections in the Discussion, and where possible we will edit for clarity and brevity.

<u>RC1-6:</u> As with 'Trends in peak-of-season phenology', I would suggest that the authors do the same study again, using the available NDVIs, do a trend analysis of annual averages, and then compare it to the results in the literature, to sidely bolster the credibility of these data.

We appreciate the reviewer's suggestion here and below we have assessed the annual-average NDVI trends across Australia for the different NDVI products to see how they differ. AusENDVI closely reproduces the observable trends in GIMMS3g (coefficients: AusENDVI-clim=0.00056 NDVI yr-1, AusENDVI-noclim=0.00049 NDVI yr-1, GIMMS3g=0.00062 NDVI yr-1; Fig. 10 in the revised manuscript). Trends in MODIS MCD43A4 over the shorter interval from 2000-2013 displayed a similar slope to AusENDVI and GIMMS3g (0.00051 NDVI yr-1). Trends in the two GIMMS-PKU products are approximately half those of the other products. This result reinforces our previous assertions that no pre-existing AVHRR-based NDVI product both reproduces close agreement with the MODIS record while simultaneously reproducing satisfactory results in the pre-MODIS era. We have included in the revised manuscript a section on this analysis (Section 3.4, lines 449-460)

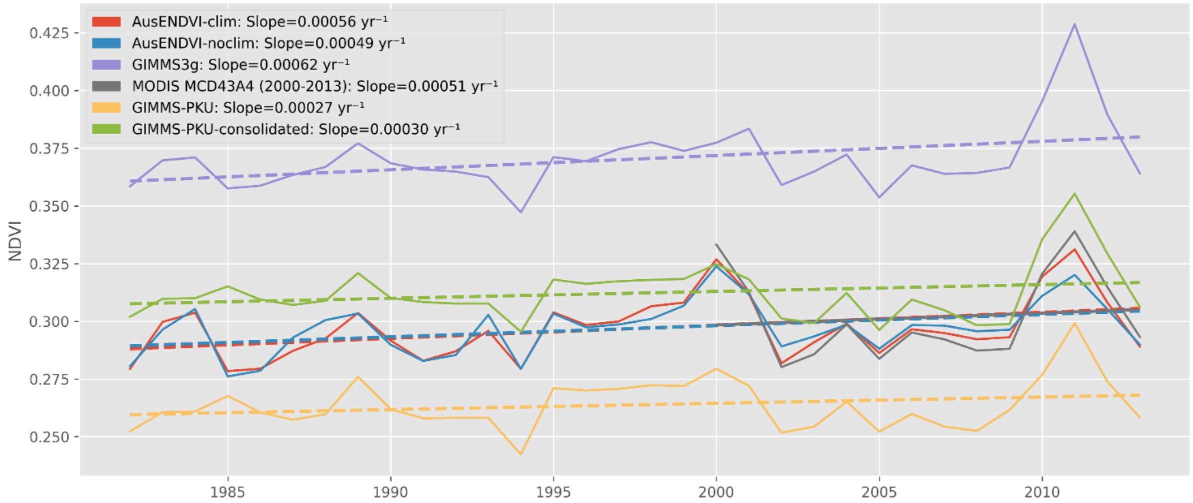


Figure: Annual average NDVI trends summarised over Australia for the overlapping period of 1982-2013. All data gaps have been matched between datasets, and datasets have been reprojected to match the resolution of GIMMS3g. Trend lines have been fitted using ordinary least-squares regression and coefficients are expressed in terms of NDVI yr^{-1}.

References

Byrne, G., Broomhall, M., Walsh, A. J., Thankappan, M., Hay, E., Li, F., ... & Denham, R. (2024). Validating Digital Earth Australia NBART for the Landsat 9 Underfly of Landsat 8. Remote Sensing, 16(7), 1233.

RC2

In the manuscript titled "Enhancing Long-Term Vegetation Monitoring in Australia: A New Approach for Harmonising and Gap-Filling AVHRR and MODIS NDVI", Burton et al. reconstructed new harmonised NDVI datasets in Australia using the GBM method. The manuscript and figures are well prepared. I appreciate the extensive work conducted in this study, like, comparing existing datasets, producing new datasets and applications. However, from my perspective, this paper may still lack sufficient novelty to warrant publication in ESSD. Below, I outline my main concerns and provide point-to-point comments.

Main concerns:

RC2-1: In the context of the existing abundance of NDVI datasets such as VIP15 NDVI, GIMMS NDVI3g and the latest PKU NDVI, authors still aim to produce new NDVI datasets, which is challenged. I encourage this work, but authors fail to show strong motivations for doing so (like, data unavailability or any issues present in existing datasets).

We wholeheartedly agree that there is an abundance of existing global NDVI datasets, and we have gone to considerable effort to include many of the most prominent datasets in a detailed intercomparison. In the introduction, we list several well-known discrepancies with existing NDVI products (lines 66-70), and also make note that the recent PKU-GIMMS product has yet to be widely assessed by the community owing to its recent release. This is why we set our first objective of the study to assess the pre-existing datasets to determine if they are suitable for studying the long term biogeophysical impacts of global change on Australia's terrestrial vegetation. Note that while there are many studies at the global scale that assess existing NDVI products, none have focused on Australia, and we see this inter-comparison as itself a valuable contribution to the Australian research community.

Ultimately, we conclude that GIMMS3g, CDR, and GIMMS-PKU have significant deficiencies (sensor transition issues, poor correlation, and/or high error with MODIS). GIMMS-PKU-consolidated offers a real improvement over other products, however, GIMMS-PKU-consolidated still has shortcomings, primarily that it does not display realistic inter-annual variability in the 1982-2001 period, and displays a lower trend in annual average NDVI from 1982-2013 than GIMMS3g and AusENDVI (figure and comments on annual average trends are in the next response, RC2-2). Hence, we argue there are further advances that can be made by optimising to the regional scale, by including a range of new features such as climate variables in the calibration, and by developing a more robust gap-filling technique. In short, our aim is to develop the best possible NDVI dataset optimised for the needs of the Australian research community, that iteratively improves on previous datasets, just as GIMMS-PKU iteratively improved on GIMMS3g.

RC2-2: According to the results (like, figures 2 & 8), I think PKU-consolidated dataset has been produced well, and compared to PKU data, your dataset does not show any significant and necessary improvements. Therefore, I would suggest highlighting clear improvements than other existing datasets.

The recent release of the GIMMS-PKU-consolidated dataset showed significant improvements over previously existing global NDVI datasets as it effectively remediated some sensor transition issues, aligns well with MODIS, and, at the global scale, better reproduced the greening trend observable in MODIS. However, over Australia, it is our contention that it fails to reproduce realistic inter-annual variability in the pre-MODIS era as indicated by its lack of agreement with the Landsat record in Figure 3a (figure 4a in the updated manuscript), and the distinct lack of *rainfall-driven inter-annual variability as shown in Figure 3b (Figure 4b in the updated manuscript) and Figure 8b (Figure 9 in the updated manuscript), respectively. This is important as the terrestrial biosphere's response to climate extremes (droughts, heavy rainfall) is of paramount importance to study given the changing frequency of climate extremes in Australia (Lewis et al. 2017). How Australia's ecosystems are responding to these changes may depend on the shifting seasonality of rainfall, warming air temperatures, and increasing atmospheric CO_2 concentrations which all affect plant physiology. We cannot effectively study these impacts and mechanisms (at the continental scale) if vegetation variability from 1982-2000 is artificially subdued.*

In the figure below we develop the statistical relationships between twelve-month rolling mean standardised rainfall and NDVI anomalies, averaged across Australia for different periods and different products. If we consider the slope of the linear relationship between rainfall and NDVI to be a reasonable approximation of the sensitivity of NDVI to water supply (and we assume there should be approximate stationarity in these relationships), then AusENDVI-clim in the 1982-2000 period (c) displays a similar sensitivity and correlation as MODIS does in the 2000-2022 period (b). Contrast this with GIMMS-PKU-consolidated which has a substantially lower sensitivity in the 1982-2000 period (d) than it does in the 2000-2022 period (e) (approximately half the sensitivity). While we may expect some changes in water-supply sensitivity over the decades due to effects such as CO_2 fertilisation, a doubling of water-supply sensitivity is highly unlikely. It is clear that AusENDVI is responding more realistically to rainfall-driven interannual variability than GIMMS-PKU-consolidated, which we consider an iterative advancement. We have included these scatter plots in an updated manuscript, along with the time series of AusENDVI-clim and GIMMS-PKU-consolidated anomalies (figure 9 in the updated manuscript). We have also adjusted the results and discussion accordingly (Section 3.3 in the revised manuscript).

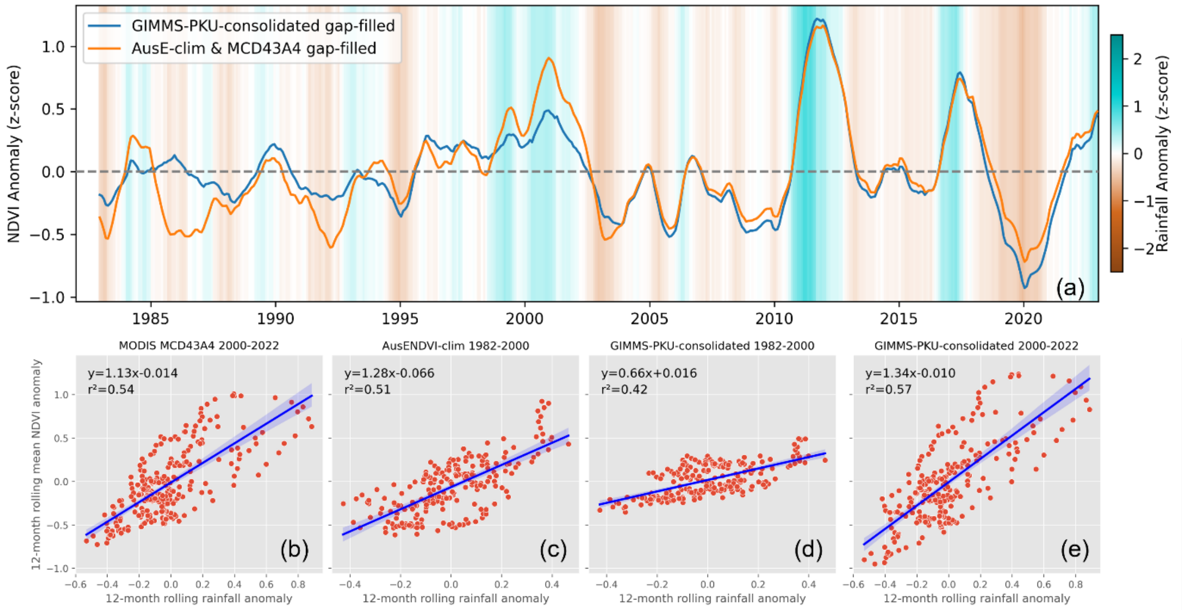


Figure: a) Standardised NDVI anomalies of AusENDVI-clim (1982-2000) merged with MODIS MCD43A4 (2000-2022), and GIMMS-PKU-consolidated. Both datasets have been gap-filled identically following the methods described in section 2.3. b-d) Relationships between twelve-month rolling mean standardised rainfall and NDVI anomalies averaged across Australia for different periods. Rainfall, AusENDVI and GIMMS-PKU-consolidated anomalies have been calculated against a 1982-2022 baseline. MODIS NDVI anomalies have been calculated against a 2000-2022 baseline. The relationship y=mx+c denotes the linear regression slope between rainfall and NDVI anomalies where y is NDVI anomalies, x is rainfall anomalies, and m is the slope coefficient. The slope coefficient can be considered an approximation of the sensitivity of NDVI to anomalous water supply.

Additionally, in the second figure below we show the annual average NDVI trends across Australia for the assessed NDVI products. Trends in the two GIMMS-PKU products are less than half those of MODIS, GIMMS3g, and AusENDVI. This result reinforces our assertion that no pre-existing AVHRR-based NDVI product can both reproduce close agreement with the MODIS record while simultaneously reproducing satisfactory results in the pre-MODIS era. We aim to include the annual average trend analysis in a revised manuscript.

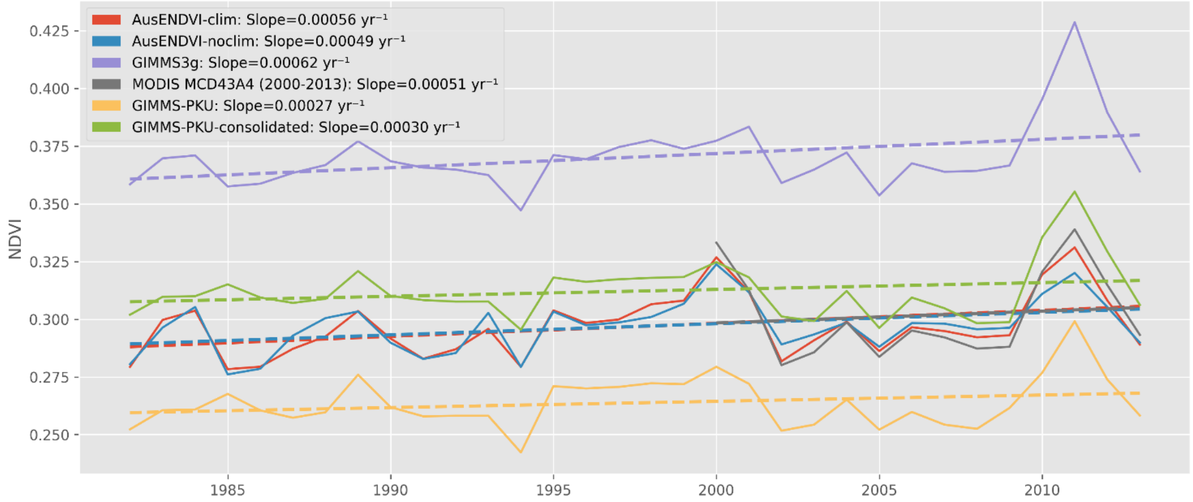


Figure: Annual average NDVI trends summarised over Australia for the overlapping period of 1982-2013. All data gaps have been matched between datasets and datasets have been reprojected to match the resolution of GIMMS3g. Trend lines have been fitted using ordinary least-squares regression and coefficients are expressed in terms of NDVI yr^{-1}.

To summarise, the advantages of AusENDVI are that: 1) it closely reproduces the MODIS record in terms of seasonality, interannual variability, and trends in annual-average NDVI, 2) it reproduces anomalies in the Landsat NDVI record in the pre-MODIS era (back to 1988), and shows realistic rainfall-driven interannual variability back to 1982, 3) gap-filling in AusENDVI does not rely on methods such as filling with a climatology, spatial interpolation methods, or lengthy temporal interpolation methods that are unreliable where wide-spread and lengthy data-gaps occur, 4) it has a higher spatial resolution than any of the GIMMS datasets and is built using inputs that apply the full suite of atmospheric and BRDF corrections, and 5) the methods and code for its development are entirely open-source. No other existing product can lay claim to all these attributes which is why we argue AusENDVI is a worthwhile addition to the suite of NDVI products available. We have edited this last paragraph into the conclusion of the updated manuscript.

Other comments:

<u>RC2-3:</u> No ground observations (like, Flux or PhenoCam sites) to validate your data?

It is unlikely that eddy-covariance flux tower GPP would have a proportional relationship with NDVI at the 5 km scale, and across the many different land covers (Camps-Valls et al 2021). Likewise, the small phenocam network in Australia does not record NDVI values. Instead, they record RGB images that can be converted to 'green chromatic coordinate' values but GCC values are not directly comparable to NDVI (Hufkens et al. 2018, St Peter et al. 2018). Regardless, there still exists a large mismatch in spatial and temporal scales between phenocams and AusENDVI (or any other AVHRR NDVI dataset, the area of pixels in CDR-AVHRR are ~25 km^{2}). Hence, there is no ground validation data for an independent assessment of our data. However, note that MODIS MCD43A4 surface reflectance data (from which we calculate NDVI *as the response variable for the harmonisation) is a well-calibrated and validated remote sensing product, and the validation performed in our study is based on random pixels selected from MODIS. Likewise, we also include a comparison with the Digital Earth Australia Landsat surface reflectance product as this product has all of the same types of corrections (atmospheric, BRDF etc.) (Byne et al. 2024) as MODIS MCD43A4 and is therefore a fair and independent inter-comparison dataset.*

<u>RC2-4:</u> For any designed steps (e.g., gap filling), it is expected to see the comparison of results for before and after processing (can refer to the guide: <u>https://lpdaac.usgs.gov/documents/1328/VIP_User_Guide_ATBD_V4.pdf</u>).

For the gap-filling, we will insert the figure shown in our response to RC1-3. Figure A5 in the revised manuscript shows the time-series of CDR-AVHRR before and after the calibration/harmonisation, averaged across all of Australia and broken down by bioclimatic region. We are open to including this in the main part of the manuscript at the editor's discretion.

<u>RC2-4:</u> Add a flowchart to summarize each step and processing.

We thank the reviewer for this suggestion and we will include in the revised manuscript the flow-chart shown in our response to RC1-1.

<u>RC2-5:</u> Add some quantified results in the abstract to show the reliability/enhancement of your datasets.

We have added the statistics to the abstract that show the model agreements with observation, along with the statistics that shows the agreement between the synthetic NDVI and observations.

<u>RC2-6:</u> Lines 30-35, provide spatial and temporal resolutions information for your 41-year dataset.

We have included the spatial and temporal resolution in the abstract in the revised manuscript.

References

Lewis, S. C., Karoly, D. J., King, A. D., Perkins, S. E., & Donat, M. G. (2017). Mechanisms explaining recent changes in Australian climate extremes. Climate Extremes: Patterns and Mechanisms, 249-263.

Camps-Valls, G., Campos-Taberner, M., Moreno-Martínez, Á., Walther, S., Duveiller, G., Cescatti, A., ... & Running, S. W. (2021). A unified vegetation index for quantifying the terrestrial biosphere. Science Advances, 7(9), eabc7447.

Hufkens, K., Filippa, G., Cremonese, E., Migliavacca, M., D'Odorico, P., Peichl, M., ... & Wingate, L. (2018). Assimilating phenology datasets automatically across ICOS ecosystem stations. International agrophysics/International Advertising Association.-New york, 32(4), 677-687.

St. Peter, J., Hogland, J., Hebblewhite, M., Hurley, M. A., Hupp, N., & Proffitt, K. (2018). Linking phenological indices from digital cameras in Idaho and Montana to MODIS NDVI. Remote Sensing, 10(10), 1612.

Byrne, G., Broomhall, M., Walsh, A. J., Thankappan, M., Hay, E., Li, F., ... & Denham, R. (2024). Validating Digital Earth Australia NBART for the Landsat 9 Underfly of Landsat 8. Remote Sensing, 16(7), 1233.

RC3

This manuscript by Burton et al. proposes a new long-term NDVI dataset specifically for Australia (AusENDVI) by harmonizing and gap-filling AVHRR and MODIS data. Compared to global NDVI datasets, localized AusENDVI could provide optimized NDVI observation with the aid of prior knowledge. To this end, I agree that the AusENDVI could be a promising dataset for better understanding long-term vegetation dynamics in Australia. However, the current manuscript faces many major issues and lacks essential information that shows the superiority of AusENDVI. My overall attitude is somewhere between a severely major revision and rejection. That's dependent on how the authors respond to the following comments.

Major comments:

RC3-1: First, NDVI is a spectral index calculated from red and near-infrared reflectance. Discrepancies of band settings (spectral range, FWHM, etc.) between sensors could be an important driver of the NDVI difference. This is the case for the three types of sensors involved in the manuscript, i.e., Landsat TM/ETM+, AVHRR, and MODIS. However, this source of NDVI differences in band setting has been completely ignored in the evaluation of current global NDVI datasets and generation of AusENDVI. For example, the authors failed to compare the two reference datasets, Landsat TM/ETM+ and MODIS in the manuscript.

We wholly agree with the reviewer that spectral sampling is different between sensor specific time-series, and we ought to have raised this point in the introduction and discussion sections, which we will do in an updated manuscript.

However, the main reason for calibrating and harmonising AVHRR to MODIS is largely to ameliorate the differences in spectral sampling between the sensors so they can be combined to produce a consistent time-series. Spectral sampling differences also cannot explain the fairly large inconsistencies between AVHRR-based products such as CDR, GIMMS3g, and GIMMS-PKU, for example in the seasonal cycles shown in Figure 2e. These differences must be due to other aspects of their processing such as those we listed in the discussion (different atmospheric corrections, cloud contamination, gap-filling procedures, etc.).

We thank the reviewer for pointing out that we had neglected to include a comparison between MODIS and Landsat. We have included the time-series pictured below of Landsat vs MODIS anomalies in the appendix of an updated manuscript (Fig. A2). As you can see, in terms of inter-annual variability, there is very good agreement between the sensors. The differences in spectral sampling between MODIS and Landsat are why we use Landsat only for validating inter-annual variability in the pre-MODIS era (using annual anomalies aggregated over the continent), since mean differences in NDVI between sensors are removed by anomalies.

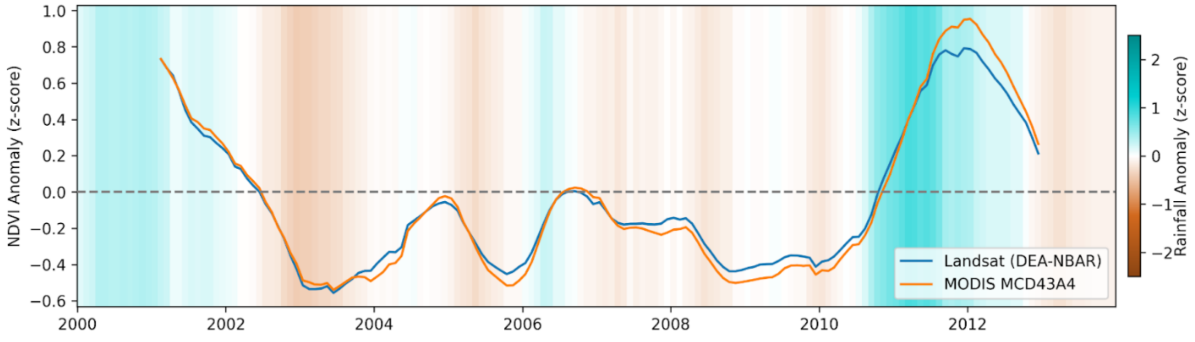


Figure: Standardised anomalies of the overlapping period between MODIS MCD43A4 NDVI and Landsat NDVI derived from the common baseline period of 2000-2012. Rainfall anomalies are derived from a longer baseline of 1982-2022.

<u>RC3-2:</u> Second, for some reason, the temporal resolution of the AusENDVI has been missing in the Abstract and Conclusion section of the manuscript. For a long-term dataset, the temporal resolution is a critical attribute that determines how well the AusENDVI could capture the abrupt vegetation changes due to climate or anthropogenic disturbances. As far as I could find in the manuscript and the data repository, AusENDVI provides monthly data records. It could be disappointing because the temporal resolution of current global NDVI datasets such as NDVI3g and NDVIpku is half a month. This issue is related to another one in that AusENDVI uses median composites while NDVI3g, NDVIpku, and MODIS NDVI use maximum composites. Why is the median? Will that underestimate vegetation growth such as vPOS?

We apologise for not making explicit in the abstract and conclusion that it is a monthly product, we have highlighted the temporal resolution in an updated manuscript.

There are quite a few advantages to aggregating to the monthly scale. Firstly, it reduces the concerns of matching overlap times between all the differently sourced datasets and thereby increases comparability between datasets. Secondly, and most importantly, it helps lessen the impact of noisy sub-monthly signals that arise from unmasked residual clouds etc. that imperfect QC bands miss. And lastly, it makes deriving relationships between covariables like climate simpler as a number of these variables come as monthly aggregates. Moreover, we argue that a monthly product is sufficient for all the likely use cases of this dataset: monitoring long-term changes to vegetation due to global environmental change (CO_2, warming, rainfall changes), and for use in driving or validating land surface models. AusENDVI has a coarser temporal resolution than the GIMMS products, true, but it also has the not-insignificant advantage of a higher spatial resolution.

Both maximum and median compositing techniques are robust to outliers, which is the principal reason for using them. It's true that GIMMS3g and GIMMS-PKU use maximum compositing techniques in their development of a 16-day product, but when we loaded these datasets (and the MODIS 16-day product) we took monthly medians so the differences between reanalysing the data using max instead of median is likely to be small (i.e. the difference between the median of two values in a month or the highest of two values in a month), and differences should only occur in the overall mean NDVI response not temporal dynamics. However, to test this, we would be quite willing to rerun the analysis with maximum-value compositing if so requested.

There is no reason to expect that trend values in vPOS will change substantially by switching from median to maximum compositing, though the actual values of vPOS may increase marginally. We argue such a change would not be of material consequence.

<u>RC3-3:</u> Third, the most impressive feature of AusENDVI is that it accounts for the dominant role of precipitation in Australia. However, the strong relationship between precipitation and NDVI has been an unproved precondition in the manuscript. The authors must demonstrate pixel-wise precipitation-NDVI relation before the relationship is used to evaluate NDVI products and generate AusENDVI. For example, in Figure 8b, the abrupt increase of NDVI in 1984 does not seem to follow the precipitation anomalies (Note the authors use the precipitation anomalies to argue the deficiency of other NDVI products). A literature review without a pixel-wise relation map is not enough.

We apologise for not including any figures demonstrating the strongly water-limited nature of vegetation in most of Australia - this is an oversight that stems from our familiarity with the landscape. We have included in Figure 4 of an updated manuscript a per-pixel map of the correlation between annual NDVI and precipitation anomalies to demonstrate this relationship. The figure highlights that cumulative rainfall anomalies are strongly correlated with NDVI anomalies over the majority of Australia's land mass.

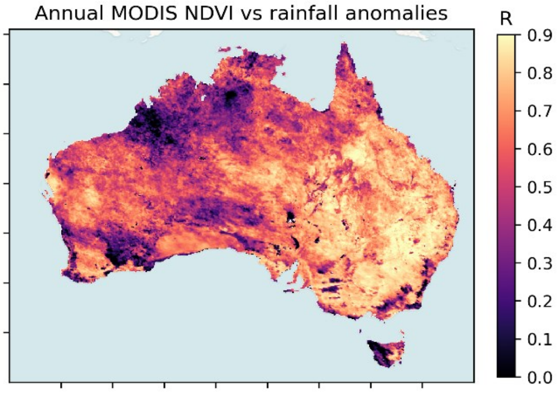


Figure: Pearson correlations between annual standardised NDVI anomalies and annual rainfall anomalies.

The apparent large peak in NDVI in 1983-1984 in figure 8b is accentuated by being bracketed by a severe drought in 1982, and modest droughts in 1985-86 (both 1983 and 84 were above-average rainfall years though - interactive maps of annual rainfall anomalies can be examined [here](). When shown as a standardised anomaly against a 40-year baseline it is in fact a fairly modest positive anomaly (see figures in RC2-2). We have updated Figure 9 in the revised manuscript to include the figure shown in our response to RC2-2 so anomalies are easier to deduce. Furthermore, although the rainfall stripes are a nice visualisation of aggregate rainfall patterns, they conceal much about the seasonal timing and spatial allocations of rainfall within a given year that can matter for the vegetation response. In the figure shown in our response to RC2-2, we also include the statistical relationships between twelve-month rolling mean standardised rainfall and NDVI anomalies, averaged across Australia for different periods and different products. If we consider the slope of the linear relationship between rainfall and NDVI to be a reasonable approximation of the sensitivity of NDVI to water supply (and we assume there should be approximate stationarity in these relationships), then AusENDVI-clim in the 1982-2000 period (c) displays a similar sensitivity and correlation as MODIS does in the 2000-2022 period (b). Contrast this with GIMMS-PKU-consolidated which has a substantially lower sensitivity in the 1982-2000 period (d) than it does in the 2000-2022 period (e) (approximately half the sensitivity). While we may expect some changes in water-supply sensitivity over the decades due to effects such as CO2 fertilisation, a doubling of water-supply sensitivity is highly *unlikely. It is clear that AusENDVI is responding more realistically to rainfall-driven interannual variability than GIMMS-PKU-consolidated, which we consider an iterative advancement.*

RC3-4: Last, the authors failed to demonstrate the improvements of AusENDVI in critical aspects such as long-term trends of vegetation and SOS.

In terms of long-term trends, please refer to the comments and figures made in our response to RC1-6.

On the trends in phenology (we assume the reviewer meant 'POS'), it would be our preference not to include another large figure and discussion intercomparing phenology trends between datasets. The phenology analysis was included as a short use case of AusENDVI to demonstrate its capability, but otherwise, the paper is intended to focus on the derivation of the data. We have plans to do a broader analysis of phenology trends (and examine a broader range of phenometrics) in subsequent work and it is our contention that including such an analysis here is unnecessary as we already demonstrate the merit of AusENDVI through a number of figures. Also, note that figure A2 Fig. A3 in the revised manuscript) does show some of the differences between products for the average month-of-maximum NDVI (averaged over the years 2000-2013).

To summarise, the advantages of AusENDVI are that: 1) it closely reproduces the MODIS record in terms of seasonality, interannual variability, and trends in annual-average NDVI, 2) it reproduces anomalies in the Landsat NDVI record in the pre-MODIS era (back to 1988), and appears to show realistic rainfall-driven interannual variability back to 1982, 3) gap-filling in AusENDVI does not rely on methods such as filling with a climatology, spatial interpolation methods, or lengthy temporal interpolation methods that are unreliable where wide-spread and lengthy data-gaps occur, 4) it has a higher spatial resolution than any of the GIMMS datasets and is built using inputs that apply the full suite of atmospheric and BRDF corrections, and 5) the methods and code for its development are entirely open-source. No other existing product can lay claim to all of these attributes which is why we argue AusENDVI is a worthwhile addition to the suite of NDVI products available.

Some minor but still important comments:

RC3-5: Line 96-97. Why are SOS and EOS not included?

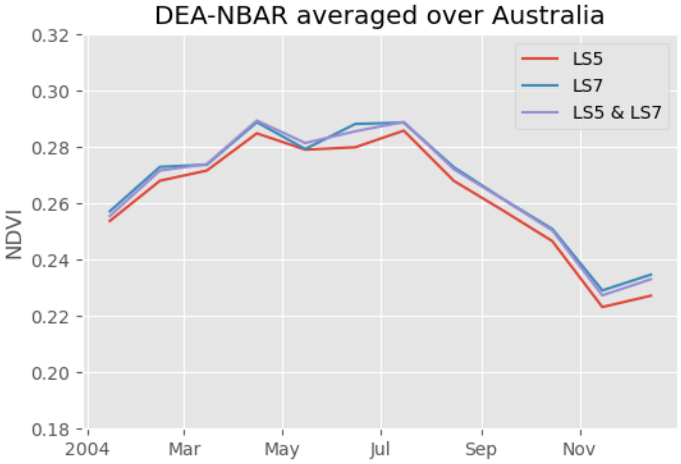
As per our last point, we have plans to do a broader analysis of phenology trends, and examine a broader range of phenometrics and their drivers in subsequent work and don't wish to overload a dataset-description paper with too much 'applications' content.

RC3-6: Line 104. When is averaging used and when is nearest-neighboring used?

Averaging is used when a finer-resolution dataset is downsampled to coarser resolution (e.g. 5km → 8km), and nearest-neighbour sampling is used when two datasets have the same/similar spatial resolution but either different projections or slightly different grid extents. We have included this information in an updated manuscript (line 105).

RC3-7: Line 105. How to deal with the radiometric difference between Landsat TM and ETM+ (Berner et al., 2020; https://doi.org/10.1038/s41467-020-18479-5)?

Firstly, the specific Landsat datasets we used, Digital Earth Australia's surface reflectance NBAR product - Collection 3, is a very high quality and consistent surface reflectance product that is calibrated and validated to the Australian continent (Byrne et al. 2024). These corrections *minimise the differences between sensors (the paper you linked uses USGS C1 Landsat which is less processed/corrected than DEA-NBAR). To demonstrate this, the figure below summarises NDVI over Australia for the year 2004 (the first year when both sensors are running for the full length of the year). These time series broadly match in terms of variability and magnitude, though LS7 has a slightly higher mean NDVI. Note that when acquiring the Landsat data, all LS5 and LS7 data are averaged together for a given month, so the actual values used in the manuscript sit between the lines (purple line in the plot below). Secondly, for the years 1988-2000 the time series only consists of Landsat 5 which is the key period we are assessing. Lastly, we only use Landsat for assessing annual anomalies so any differences in mean responses between sensors are minimised.*



<u>RC3-8:</u> Line 212. Why is the median rather than the maximum value?

We are unsure what the reviewer refers to here as line 212 does not discuss medians, but in general, we use medians for temporal compositing to reduce the influence of outlier values.

<u>RC3-9:</u> Line 122. Please provide more information on the use of the quality assurance band.

We have included more specifics in the text (line 128), but the notebooks in the linked github repository also detail how data was loaded and masked (see here).

<u>RC3-10:</u> Line 128. Simply removing data in sensor transition would not only eliminate the gradual effect of sensor degradation but also the valuable information of NDVI anomaly. Note the eruption of Pinatubo (1991) and the transition of AVHRR2 and AVHRR3 (around 2000) are not accounted for.

We remove data at the sensor transitions of the CDR product because the large anomalies associated with some of these transitions are unrealistic and do not reflect conditions on the ground. For example, see Figure 2 in Tian et al. (2015), where transitions between AVHRR sensors N9 and N11 result in an extremely large negative anomaly.

We do not remove data during or immediately after the Pinatubo eruptions in 1991. We do remove data associated with the transition from sensors N11 to N14 during the second half of 1993 through 1994, though there is little-to-no data recorded in the CDR product over Australia in these years anyway. It is our understanding that after ~3 years the impact of aerosols from Mt Pinatubo on the NDVI signal waned. It is important to note that the CDR product includes an aerosol correction, something the GIMMS products generally lack; we understand that GIMMS3g does include a correction specifically for the Pinatubo eruption but does not have aerosol corrections otherwise.

We did not remove data at the year 2000 transition from sensors N14 to N16, although a lot of data is missing in this period so its influence on the time-series is probably limited.

RC3-11: Line 131 & Figure A1. Explain the reason why some regions experience lower data availability. How does the data availability affect the evaluation of NDVI products and AusENDVI accuracies?

Coastal, alpine, and tropical regions experience fewer 'good quality' observations principally due to the greater abundance of clouds in these regions from prevailing weather systems. For example, western Tasmania experiences >1500 mm of rainfall per year which means the chances of acquiring a clear satellite image are lower. A map of annual mean rainfall over Australia can be viewed here. The patterns of high rainfall largely coincide with regions of lower data quality. Exceptions to this 'rule' sometimes occur where bright objects (e.g. salt lakes) can be misidentified as cloud by quality-assurance bands, such as is the case in parts of central arid Australia.

Whenever products are compared in the manuscript, all datasets are reprojected onto a common grid and data gaps are matched between all datasets. Basically, a mask is created that identifies all missing pixels in all datasets, and then that common mask is applied to every dataset. This ensures a fair and valid comparison (we have edited in these two statements to line 108 in the revised manuscript). Even in those areas with a lower data volume, there are still dozens of valid pixels in the time series so statistics are fairly robust. One caveat to this might be that, in those very cloudy regions (e.g. tropical forests along northern Queensland), the statistics probably relate the agreement mostly during the dry seasons as the wet seasons will have fewer observations.

It is also reiterated that the differing volumes of data across the continent are partly why we implemented a stratified, equalised random sampling approach for the training and validation samples. Providing the same number of samples for each bioclimatic region regardless of data availability or area reduces bias in validation statistics.

RC3-12: Table 1. Please provide the temporal resolution of the datasets.

We have included the 'native' temporal resolutions in an updated version of the manuscript.

RC3-13: Line 137. Why not use existing MODIS NDVI products (MOD13Q1, MOD13C1, etc.)? It looks like AusENDVI and NDVIpku are based on different MODIS products. Will be the difference reflected in the evaluation of NDVIpku?

We used MODIS MCD43A4 as we argue it is the highest quality MODIS dataset that been released owing to: 1) its full suite of atmospheric corrections, 2) BRDF corrections, and 3) inclusion of both the AQUA and TERRA satellites which extends its time-range back to the year 2000. Also, the inclusion of BRDF corrections aligns MODIS with both the AVHRR-CDR product and the Landsat product. It is true that the GIMMS-PKU-consolidated product was trained on a different version of MODIS, but we make no contentions in the paper that GIMMS-PKU-consolidated does not align well with MODIS - it does. Presumably comparing it with the exact version it was trained on would show further agreement, but given we ourselves did not wish to use that same version of MODIS we felt it was better and simpler just to train and compare on the best version of MODIS available.

RC3-14: Line 141. How are standardized anomalies calculated?

“Z-score” standardised anomalies are calculated as (x - m) / s where x is a monthly NDVI observation, m is the long-term mean NDVI for the given month, and s is the long-term standard deviation in NDVI for the given month. It is a standard way to track inter-annual variability. We have included this formula in the revised manuscript (line 150).

RC3-15: Line 146. More details are needed for the outperformance of GBM. For example, are all the models optimized in parameters?

Yes, all models were optimised and the GBM model outperformed them in terms of fitting accuracy and speed. We did not include further information on discarded methods as they are not critical to the paper, but for those interested we have made the scripts for the GAM method available here along with the other code). We only wished to include this sentence so the reader would be aware that we had done our due diligence by testing on a few different methods before selecting gradient-boosting.

RC3-16: Line 152-153. “…in the heavily forested regions where there was little to no agreement between NDVIMCD43A4 and NDVIAVHRR…”. How was pixel quality considered in calculating agreement?

All AVHRR and MODIS products were masked with their corresponding pixel-quality layers during loading and temporal compositing so only good-quality observations are retained. Then, when comparing one product to another, data gaps between all products were matched, as detailed in the comment above.

RC3-17: Line 155. Why is longitude not included? Give more details on NDVIMCD43A4 summary percentiles.

We had two reasons for this. Firstly, it proved not to be a particularly useful feature in the predictions (as evidence by low feature importance). Secondly, in an early version of the product, including both latitude and longitude introduced some artefacts into the predicted values. Ultimately, longitude is prone to overfitting. We did not feel this was important information to share with readers (since we also tried other features that did not make it into the final model configurations).

MODIS summary percentiles were calculated per pixel over the 2000-2022 period. So over the 22 year time-span, we extracted the 5th, 50th (median), and 95th percentile values. For the training and predictions, these values are simply replicated at each time step, so they are effectively static layers (i.e., not varying through time). We will include an extra note in an updated version of the manuscript describing the time-range over which the percentiles are calculated.

RC3-18: Line 178. Please list the hyperparameter values used.

Thank you for the suggestion. We have included a table in the appendix (Table A1) with the hyperparameters used for fitting the harmonisation models and the synthetic NDVI models.

RC3-19: Line 180. In addition to absolute error, a measure of error that reflects the relative error is also needed. Such a measure is particularly important for dense vegetation.

We have added RMSE to the statistics in the scatter plots of Figure 5 and Figure A8 in the revised mansucript. However, arguably this information is already in the manuscript as the per-pixel coefficient-of-variation plots in figure 5b and 5d show relative error by dividing RMSE by the long-term mean NDVI.

RC3-20: Line 185. How are the long gaps spatially and temporally distributed, particularly for dense vegetation?

We argue that Figure A1 does a reasonable job of showing how data gaps are distributed spatially. We could potentially devise an additional figure that shows how gaps are temporally distributed, but we are not sure how much value this will add to the manuscript and are somewhat wary of adding too many figures. The time series in Figure 6 does show data gaps for two forested regions.

RC3-21: Line 191-192. What do you mean by methods in the bracket?

We are not entirely sure what the reviewer is referring to here, but the methods in the brackets refer to common spatial interpolation methods that can be used to fill gaps in data by extending values from nearby data points. There are quite a few methods available for this, some of which we listed in the brackets.

RC3-22: Line 198-199. Linear temporal interpolation may under or over-estimate values for seasonal peaks or valleys or other abrupt signals.

We completely agree with the reviewer, which is why we limited linear temporal interpolation to a single time step. This limits over or under estimation of temporal dynamics to a minimum.

RC3-23: Line 206-207. Why is not WCF used as a feature in data harmonization but in synthesis?

We did not wish to use WCF in the harmonisation as that product is partly built with the use of annual Landsat composites. Since we wanted to use Landsat as a validation of inter-annual variability in the pre-MODIS era, we felt using WCF might bias the results.

RC3-24: Line 219. Will there be any issue related to the calculation of phenology when up-sampling from monthly to two-week intervals?

The day-of-year values for POS are sensitive to temporal resolution since, at the monthly time scale, the POS value is really 'month-of-year' rather than 'day-of-year'. Upsampling is thus required to increase the temporal resolution and resolve any inter-month shifts in phenology. Upsampling of this magnitude is fairly common practice in remote sensing land-surface-phenology studies, and given this applications section is not the major focus of the study, we would suggest that a sensitivity analysis may overburden the manuscript. Note that in section 2.4 we highlight that DOY values are only an approximation.

RC3-25: Line 238-239. How was the comparison made if there are data gaps brought by, for example, clouds? What if there are insufficient valid data between 2000 and 2013 for the calculation of CV and R?

*This point was addressed in RC3-11(data gaps are matched between all datasets before comparisons are made). Even in relatively low data volume areas, there are still enough good quality observations to perform statistical calculations. For example: 14 years * 12-months/per year * 0.35 (a low fraction of data) = 59 good observations. The exact procedures we used can be found in this jupyter notebook, in the github repository referred to in the manuscript.*

RC3-26: Line 242. R2 (in the text) or R (in the figure)?

Thanks for pointing out this mistake, we will correct it.

RC3-27: **Line 256-257. Present the length of the growing season please.**

We are not sure how informative this would be but can add it.

RC3-28: **Line 279-280. Solid evidence is required.**

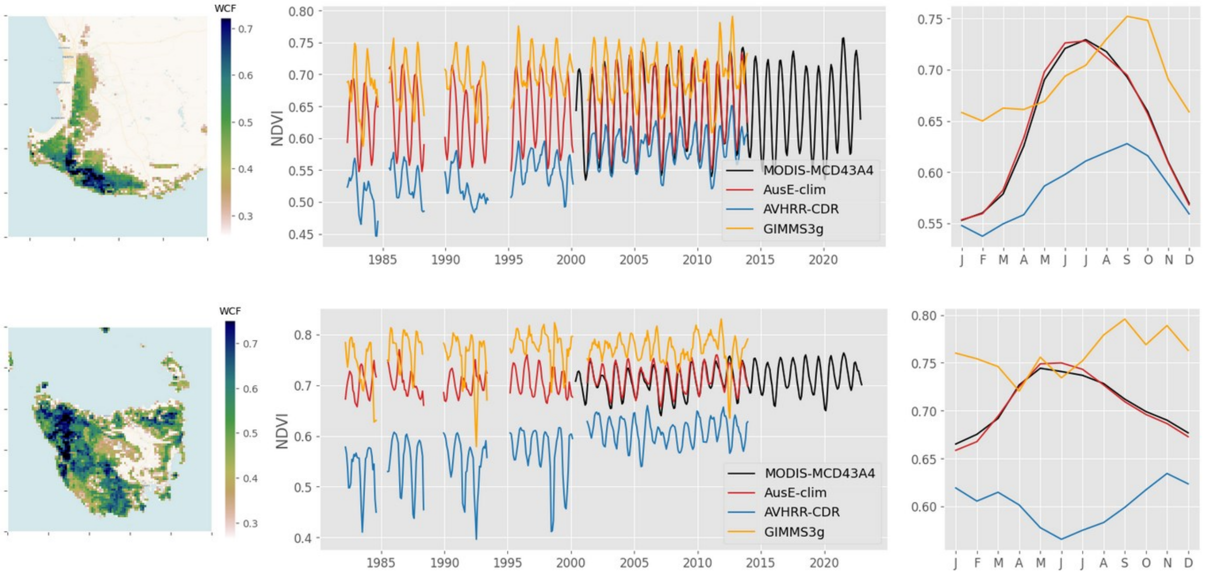
We appreciate the comment. Please refer to the discussion and figures in RC3-3 and RC2-2.

RC3-29: **Figure 5. It would be interesting to see a similar residual NDVI map for NDVIpku.**

There is a low residual signal between MODIS and GIMMS-PKU-consolidated, as one would expect based on the agreement maps in figure 2d and 2h. We argue that including this in the manuscript would not add much value - we did not contest that GIMMS-PKU-consolidated agrees well with MODIS. Figure 5 is intended to show the results of our calibration and harmonisation, not GIMMS-PKU's.

RC3-30: **Figure 6. Notice that the increased trend of NDVI before 2000 in AVHRR-CDR disappears in AusE-clim.**

The trend in CDR is almost certainly an artificial artefact of step changes between sensor transitions and poor calibration over these regions. Below we replot the same figures but including GIMMS3g which has had sensor transitions ameliorated and the trend slope is much less than for CDR in either of the two regions plotted. Similarly, MODIS over the 22-year period does not show trends like those of CDR, yet the likely drivers of greening (CO_2 and warming) all continue to increase from 2000-2022. As these plots are intended to show 'before-and-after' calibration results, we would prefer not to muddle them by including additional time series. However, we have included one of these time series in the appendix (Figure A6) and make a note in the results sections that the artificial trend in CDR is removed by the GBM calibration (line 384- 389).



RC3-31: **Figure 7. Focus needs to be placed on vegetated, particularly densely vegetated areas. Also, in Figure 7e, is the red dot line calculated without any observation data?**

Yes, the synthetic NDVI data plotted in Figure 7e is created using only climate data, MODIS summary percentiles, and annual WCF - averaged over Australia the synthetic data does a great job of replicating observations.

In the updated version of the manuscript we include a figure in the appendix (Fig A7) that shows the synthetic NDVI timeseries disaggregated by low and high woody cover fraction, and in the results section we describe these results (line 398-401).

RC3-32: Line 370. What do you mean by 'gaps in the NDVIPKU-consolidated dataset'? Non-data or data with poor quality?

We mean some pixels have no data because the QC layers that come with GIMMS-PKU labelled these pixels as poor-quality observations. Specifically, for GIMMS-PKU we kept only those pixels labelled as 'good-quality AVHRR'. And for GIMMS-PKU-consolidated we kept only those pixels labelled as 'good-quality AVHRR' and 'good-quality MODIS' and where the harmonisation was run by the random-forest model. We have included the PKU QC masking procedures in Table 1

RC3-34: Figure 8. Note that NDVIpku is generated from a different MODIS NDVI product. A comparison between MODIS NDVI products may be beneficial.

Noted, but we do not wish to overwhelm readers by adding further figures and discussion of inter-comparing versions of MODIS. We argue this would not add much value to the manuscript. As stated previously, we agree that GIMMS-PKU-consolidated agrees well with MODIS. We will also update Figure 8 to include the anomaly time-series shown in our response to RC3-3 and RC2-2. As anomalies remove the mean value, the agreement between GIMMS-PKU and MODIS will narrow, and the focus instead will be on differences in inter-annual variability.

References

Byrne, G., Broomhall, M., Walsh, A. J., Thankappan, M., Hay, E., Li, F., ... & Denham, R. (2024). Validating Digital Earth Australia NBART for the Landsat 9 Underfly of Landsat 8. Remote Sensing, 16(7), 1233.