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 will include the time-series pictured below of Landsat vs MODIS anomalies in the appendix of an updated manuscript. 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.

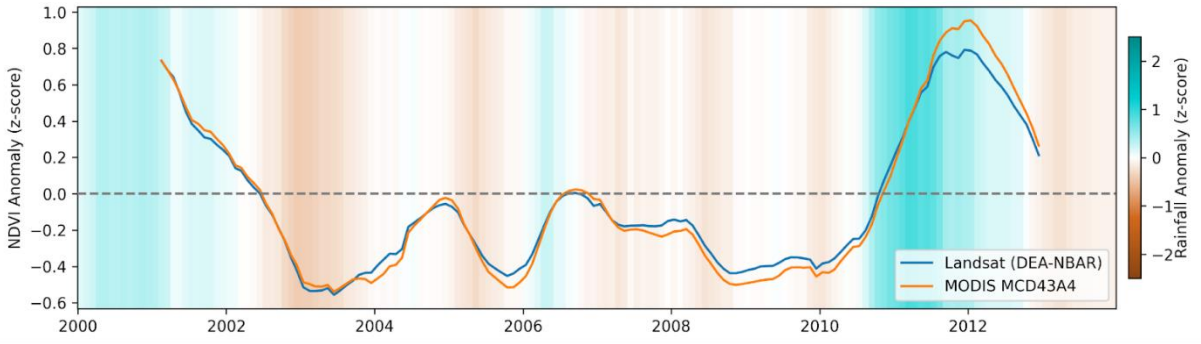


Figure: Standardised anomalies of the overlapping period between MODIS MCD43A4 NDVI and Landsat NDVI derived from the common baseline period of 2000-2012. Rainfall anomalies are derived from a longer baseline of 1982-2022.

RC3-2: 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 will highlight 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.

RC3-3: 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 can include in the appendix the below per-pixel maps of the correlation between NDVI anomalies 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.

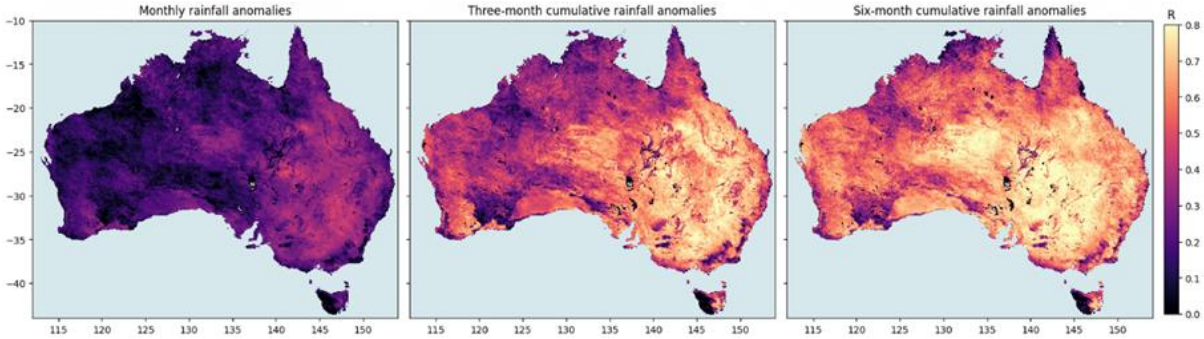


Figure: Pearson correlations between NDVI anomalies and monthly, three-monthly, and six-monthly cumulative rainfall anomalies.

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 will update Figure 8 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 (again, cumulative effects are also very important). 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 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-sourced. 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?

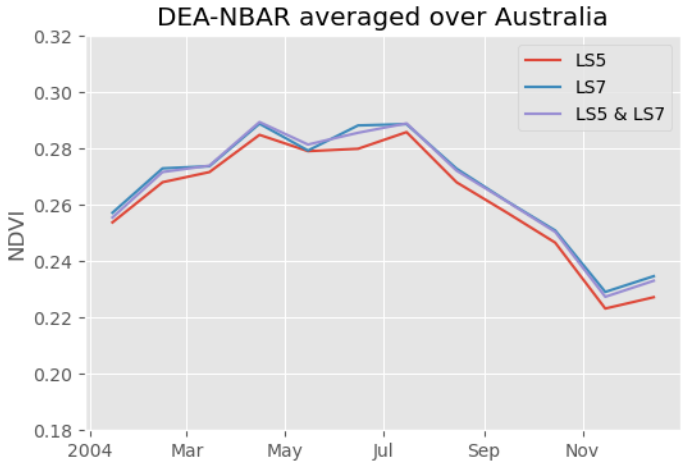
As per our last point, we have plans to do a broader analysis of phenology trends, and examine a broader range of phenometrics and their drivers in subsequent work and don't wish to overload a dataset-description paper with too much 'applications' content.

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 will include this information in an updated manuscript.

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.



RC3-8: 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.

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

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

RC3-10: 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. 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 will include 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 would prefer not to include the formula in the manuscript as it is a very common approach, but we will include a reference to a published definition.

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 will include a table in the appendix 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 will add RMSE to the statistics in the scatter plots of Figure 4 and Figure A5. 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.

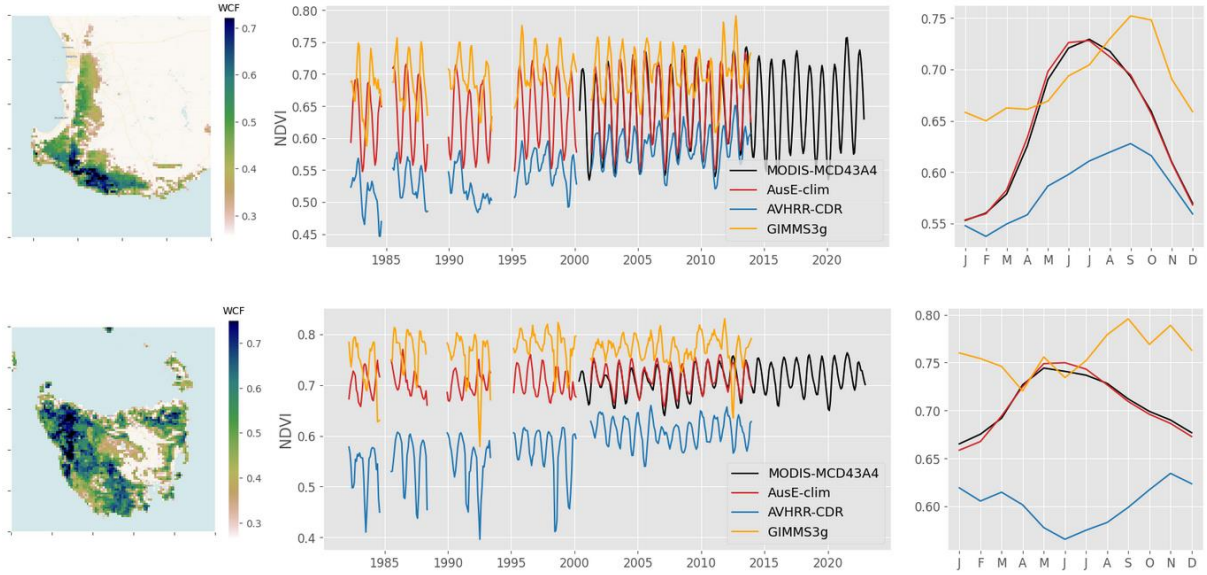
We appreciate the comment. Please refer to the discussion and figures in RC3-3 and RC2-2.

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 could include one of these time series in the appendix and make a note in the results sections that the artificial trend in CDR is removed by the GBM calibration, at the editor's discretion.



<u>RC3-31:</u> 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 an updated version of the manuscript we can update Figure 7 to include a second time-series showing the synthetic data over a densely vegetated region, similar to figure 6.

<u>RC3-32:</u> 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 will update the data section to include more information on the QC masking procedures.

<u>RC3-34:</u> 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.