

The author present a nice study, evaluating existing metrics and a new metric for the onset and end of the rainy season with respect to vegetation data. One major result is that threshold based metrics needed to be calibrated region-specifically as they have done with the NDVI. Furthermore, they test sensitivities of the metrics, and they apply them to climate projections, showing that despite projected increase in precipitation, the rainy season onset and end do not show significant trends.

The study is interesting and relevant, however, the study could profit from an improved structure and some clarifications in the section Material & Methods as stated below in the specific comments.

Dear Reviewer, thanks a lot for your assessment and your helpful comments. Below, we addressed them point by point. All references to line number refer to the revised manuscript **without** track changes.

Line 30: “hereafter named metrics” -> this seems a bit unnecessary. Maybe, it’s fine if you just start the next sentence with: Broadly, these rainy season metrics...

Following your suggestion, we changed the sentence removing the unnecessary part.

Line 37: It’s called the Standardized Precipitation Index (SPI), maybe denote it like this. Also, it is a bit a confusing citation. You are talking about RSO and RSE metrics, but the SPI is used to separate the rain season into wet and dry periods? Consider removing this citation.

We agree that this might have caused some confusion. The cited paper does use the SPI to derive interannual variability or wet and dry anomalies for pre-defined periods where and onset and end would typically occur but as you point out the authors do not directly use the SPI to derive RSO and RSE. We therefore followed your recommendation and removed the citation.

Line 46: it would be nice if you mention some examples of uncertainties here.

We revised the sentence, now mentioning examples of sources of uncertainties (l.43f.):

“However, observational and gridded precipitation time series are typically subject to significant uncertainties, such as spatial representativeness issues, measurement errors (such as undercatch in windy conditions), temporal inconsistencies, and biases in retrieval algorithms (e.g. Kidd et al. 2017; Pollock et al. 2018).”

Line 61: How do you deal with irrigation when using a vegetation metric for independent validation?

We use the LSP data from a previous paper by our team (Hänchen et al. 2022). To exclude pixels where the seasonality is evidently decoupled from the rainy season, the NDVI data was filtered based on an autocorrelation analysis approach put forward by Verstraete et al. (2008): For each pixel, the timeseries was split into a 14-month (growing season plus/minus 1 month) timeseries and using a 3-week rolling window the autocorrelation was calculated. Then, all pixels with more than one local maximum were removed. Irrigated agriculture typically has a second “green peak” during the dry season and is therefore excluded by this method. For our data specifically, on average 7.23% of all pixels were removed per growing season, most of which are located either at the valley floor (where larger-scale irrigated agriculture exists) or at the highest altitudes (where there is no or little vegetation cover). While not specifically related to irrigation, we would like to mention that we additionally masked the NDVI data based on a land cover classification, where we removed all pixels which intersected with the classes flooded vegetation, urban areas, bare areas, water, snow and ice.

However, as you made this important comment, we believe that our manuscript was not clearly conveying this information and readers could ask themselves the same question. We therefore now extended the text of the “Data” section (l.133f.):

The data were i) filtered on quality assurance criteria, ii) gap-filled and smoothed using a Gaussian process regression algorithm (Belda et al. 2020), iii) masked based on unimodal seasonal vegetation development and land-cover data to exclude pixels which are evidently decoupled from the rainy season.

Verstraete, M. M., Gobron, N., Aussenat, O., Robustelli, M., Pinty, B., Widlowski, J. L., and Taberner, M.: An automatic procedure to identify key vegetation phenology events using the JRC-FAPAR products, *Adv. Space Res.*, 41, 1773–1783, <https://doi.org/10.1016/j.asr.2007.05.066>, 2008.

Hänchen, L., Klein, C., Maussion, F., Gurgiser, W., Calanca, P., and Wohlfahrt, G.: Widespread greening suggests increased dry-season plant water availability in the Rio Santa valley, Peruvian Andes, *Earth System Dynamics*, 13, 595–611, <https://doi.org/10.5194/esd-13-595-2022>, 2022.

Line 65: I think the sentence is nicer, if you remove “regarding such independent data”

Indeed. We removed it. Now the changed sentence reads (l.64f.): *“Spectral vegetation indices, which serve as proxies for land surface greenness, are a promising candidate for calibrating rainy season metrics in semi-arid regions due to their high spatio-temporal resolution and availability from satellite data.”*

Fig. 1: The topography colorscale here is not very intuitive. It could be useful to reduce the colorscale of topography to something simple, i.e. green to brown in order to give an intuitive idea of the topography

We changed the colormap to a more classic topography style. One of the reasons we used the cubehelix colormap initially was to allow a contrast between the sub-map of the Rio Santa basin which shows NDVI in a greenish colormap. With the topographic colormap most of the amazon lowlands would be green which is visually distracting against the other panel. We now used a more typical colormap, but also now masked all areas below 500m to ensure the contrast. This has also the benefit that the attention goes to the high-elevation Andes area on which this paper focusses.

Line 145-162: I would suggest to first mention the three observational datasets (WRF, CHIRPS, AWS) and describe them shortly, and only then the scenarios. I.e. Move everything after “In addition, Potter et al. (2023)” to a later position in the paragraph.

Rearranged according to your suggestion.

It could be nice to mention the abbreviation that you use in the figures already here (WRF, CHIRPS, AWS). For example, it's quite obvious that AWS means automatic weather stations, but it's never mentioned, I think. For reproducibility, could you state which three stations you used?

Thanks for noticing, we introduced the AWS abbreviation, added the 3 station locations (Yungay, Recuay and Santiago) including their coordinates (l.148f.). Additionally, we plotted them on Figure 1 to allow readers to see where they are located.

With respect to CMIP 5 downscaling, maybe it's good to mention briefly, what type of emission scenarios RCP4.5 and RCP 8,5 are.

Agreed, we now wrote (l.149): *In addition, Potter et al. (2023) produced statistically downscaled projections based on a 30-member CMIP5 ensemble from 2019 to 2100 using quantile delta mapping for both the Representative Concentration Pathways (RCP) 4.5 and 8.5 scenarios. These scenarios represent different greenhouse gas concentration trajectories, where the number indicates the associated radiative forcing in 2100 (in Wm^{-2}). RCP4.5 is a stabilization scenario with moderate mitigation efforts, while RCP8.5 represents a high-emission, business-as-usual trajectory.*

Line 196: and the our -> “the” not needed?

Thanks for noticing, removed.

Line 196-205: this small section is quite relevant. Maybe you could consider making an own small chapter out of it in a new chapter, for example: 2.2 Rainy season metrics, 2.3 The new “bucket” metric, 2.4. Calibration of threshold-based metrics, 2.5 Sensitivity analysis 2.6 Future projections.

We agree and now moved this small section below the previous section 2.3, gave it its own header and overall followed your suggestion of the order of chapters.

Also, in the Figures, you compare the thresholds provided by the authors to your calibrated thresholds. Maybe, you can mention that you will be comparing them as well in the section on the calibration?

Good idea, we now changed the start of the new section to (l.224f.): *Using the NDVI-derived SOS_{NDVI} and EOS_{NDVI} as targets, we first tested the initial parameters provided by the corresponding authors. We then calibrated each threshold-based metric, along with our novel metric (defined below), by adjusting their parameters (c.f. Table 1) for each of the three precipitation time series.*

Are all the necessary information given, on how you apply the Differential Evolution optimization for reproducibility of your study?

We believe they are. Under the section *Code and data availability* at the bottom of the manuscript (l.474f.), we provide a link to a repository containing all the code and data necessary to reproduce our results or to adapt it for other assessments.

Line 232: what is an almost large number of rainy seasons? Maybe consider deleting the “almost”?

Apologies for this oversight, this was a typo. Removed.

Line 247: individual models were excluded individually -> maybe one individual is enough?

Yes, removed.

Sect. 2.4: It could also be helpful to separate the sensitivity analyses and projections in to separate chapter, because you will be doing two separate things with the climate projections? Do you state anywhere how you calculated the trends for the past and the future (maybe I just didn't see it)? That could go into such a short section?

Thanks for the suggestion, we briefly describe that we did a linear regression in the caption of Figs. 6 & 7 but following your comment, we now split the sections and extended the previous brief part describing the trend analysis (l.253f.): *To reliably determine trends in future CMIP5 projections, we first*

excluded individual models for each rainy season metric if they produced five or more invalid values out of 81 seasons (2019–2100). Invalid values occurred when the conditions for RSO or RSE were not met within a given hydrological year. For the remaining data, we calculated linear trends separately for the historical WRF and CHIRPS datasets, as well as for both RCP scenarios of the CMIP5 ensemble, using linear regression. Trend significance was assessed using a Wald Test, with the null hypothesis that the slope is zero.

Line 266: always add the unit after the RMSE, i.e. 8.8 and 14.4 days.

Of course, sorry for the oversight, we added the units throughout the manuscript.

Fig 2 and 3: I find the figures nice, but very full. Could you consider either moving the calibration data or the RMSE (or both?) out of the Figure into a separate table?

We agree that the figures are quite full; however, we have discussed this extensively and believe that the current version has the advantage of allowing everything to be viewed at once, rather than requiring readers to go back and forth between a table and the figure. Therefore, we would prefer not to remove any information from the figure, although we acknowledge the trade-off between keeping the figures busy and making them informative.

Please mention more clearly that INIT WRF are the thresholds provided by the authors, maybe even mention this in the section very you describe the calibrations instead of just mentioning it in the caption?

Following one of your previous comments that is now more clearly described in the new Section 2.4. *Calibration of threshold-based metrics* (from I.223). Please refer to our answer above. Furthermore, we revised the caption of Figure 2 where now $INIT_{WRF}$ is added for more clarity.

Line 278ff: Here, mention that you talk about the INIT WRF shown in the figure, otherwise it is not so clear. Also, specifically state the very high RMSE, when the parameters of the rainy season metrics are not tuned. I think this is a very relevant result of your study. For the agricultural perspective, a RMSE of 34 days (e.g. Gurgiser) is very different from an RMSE of 12 days. Also, could you state why you looked at the original threshold only using WRF, and not for example the AWS?

We agree that this is one of our key results, and we have already addressed it thoroughly starting from I. 434 in the revised manuscript. However, our aim in this section is to make a different point: we want

to highlight that each precipitation dataset yields different optimization results, which suggests that it is unlikely for a single parameter set to work universally across all precipitation data. This underscores the need for calibration on a case-by-case basis.

Regarding the WRF data, we focused only on the initial results because, as you pointed out, the presentation of the results is already quite dense. Applying the initial setup to the other datasets, in our view, would not provide additional insights. Moreover, the WRF data is by far the most reliable dataset for the region and the primary focus of our study. The other two precipitation datasets serve mainly as a comparison to demonstrate that our framework functions independently of the input precipitation data and to emphasize the necessity of calibration.

Fig 4: Please add inside the Figure a legend with the colors of the precipitation data sets. This can increase the readability a lot.

Thanks for the suggestion, we now revised accordingly. It now matches the tabular style of the two previous figures.

Lines 308 ff: You could also call Sect. 32. Sensitivity analysis of rain season metrics?

Yes, changed.

Fig. 5 This figure is quite small. Maybe you can enlarge the figures a bit by adding the rainy season metric title at the top of the figure (i.e horizontally). Like this you have more space in the horizontal dimension.

Apologies if we misunderstand your comment but moving the metric names to the top does not work as they represent the y-axis labels. Besides, we think that the font sizes in the plot are quite similar to the ones in the main text and this might also improve once the typesetting by the journal is done.

Conclusion: Could you add a sentence summarizing the results of the sensitivity tests?

The first paragraph of the conclusion (l.434f.) already briefly mentions these results but we revised it to make this a bit clearer. Changes are marked in bold letters: "*Based on several precipitation and remote-sensing derived land surface phenology data, we introduced a novel calibration strategy for rainy season metrics applied in semi-arid regions. For all three considered precipitation datasets, we find that the threshold-based rainy season metrics, once calibrated, are able to capture the interannual variability found in a vegetation greenness proxy in the Rio Santa basin and exhibit sensible sensitivities to potential hydroclimatic changes. More objective and flexible metrics on the other*

hand have comparably low skill regarding this task. These objective metrics seem to exhibit implausible sensitivities that can potentially render them uninformative or even misleading under certain conditions of rainy season change. We therefore recommend that the usage of such methods should be at least critically reviewed on a case-by-case basis to ensure that no false conclusions are drawn or misleading practical recommendations are made."

Line 456: I don't think you can say it is an "unprecedented" number of future projections since these have been published by Potter already?

Here we meant that so far, no other assessments of rainy season change in the broader region has used such a large number of future projections not that we introduced something new. However, to avoid confusion, we revised the statement to (l.464f.): "*Using the bucket metric together with other calibrated and sensitivity-tested rainy season metrics and a large number of future projections [...]*"

Also, we had a very similar statement in the previous chapter at l.409 of the previous manuscript: "*Our results use an unprecedented number of calibrated and sensitivity-tested rainy season metrics combined with a large-ensemble of high-resolution, bias-corrected future precipitation data.*"

We changed that as well to: "*Our results incorporate an exceptionally large number of calibrated and sensitivity-tested rainy season metrics, combined with a high-resolution, bias-corrected large-ensemble of future precipitation data.*"