Review of the manuscript “Comparison of high-resolution climate reanalysis datasets for hydro-climatic impact studies”

In this study, Wood et al. compared how four different reanalysis datasets (ERA5, ERA5-land, CERRA, and CHELSA) reproduce the observed average, extremes, and trends in precipitation and temperatures over a set of catchments located in Switzerland. These datasets differ in their spatial and temporal resolution, as well as in their calculation methods. The results show that CERRA generally outperform the other datasets, likely due to the data assimilation, which make it the most reliable dataset for hydrological impact studies.

Overall, the manuscript is well written and structured, the objectives and methods are clearly defined. It addresses relevant questions on the quality of the dataset that we use in hydrology studies (among other disciplines). Therefore, I think that the manuscript is a useful contribution to the HESS journal. Nevertheless, the comparison of some of the metrics and the consistency of the figures could be further improved before publication (see the comments below).

Specific comments

Abstract

Line 10: I suggest to rephrase the sentence about drought and floods since the datasets do not properly simulate flood and drought events, but rather the conditions (precipitation and temperature) that lead to droughts and floods.

Introduction

Lines 28-29: Could you provide a reference for the statement “Among these, the highest quality datasets are those relying on spatially interpolated observations.”?

Lines 65-66: “many applications require even higher resolution data”. Could you mention a few examples and explain why higher resolutions is needed?

Line 76: I would suggest specifying the type of drought here (I guess hydrological), since one could also be interested in understanding the drivers of atmospheric drought.

Lines 83-91: This paragraph is difficult to follow, as we move from one study to another with different objectives and spatial scales. Also, the conclusions of these studies are missing. We understand that they focused on temporal variations in precipitations (including snow) and their impacts streamflow events, but it could be more explicit.

Data and Methods section

Overall comment: the criteria for selecting the four datasets are missing. Why did you select these four datasets? As you mentioned previously, Dura et al (2024) compared seven datasets, why not including all of them in your study?

Line 182: I wonder why you did not used precipitation and temperature provided by the CAMELS-CH? Do they come from the same products?

Line 186: In my opinion, the first analyses focus more on comparing the spread of the climate metrics (figures 2 and 3), than on the absolute or relative differences between the metrics, which are presented in the supplementary material (more about that in the comments of the results section).

Lines 192-196: I didn't find this comparison between time series-based and metric-based approaches in the results. Did you observe differences between these approaches? Anyway, I would suggest to stick to the metric-based approach in the manuscript.

Line 198 and 204: Are the annual mean and other metrics such as wetdays computed over hydrological years?

Line 203: why keep both wet and dry days metrics? Aren't they 100% correlated? The wetdays is shown in Figure 2 (while drydays isn't), and reversely drydays is shown in Figure 7 (but wetdays isn't). I suggest to stick to one or the other of these metrics.

Lines 211-212: How different are the estimations of snowfall (re)computed in your study from those provided by the datasets?

Table 2: The metrics for the mean air temperature is missing in the table. Also, since you used the colddays, why not including hotdays as well (perhaps with a threshold adapted to your catchment sample)?

Results section

Overall comment: In sections 4.1, 4.2, 5.1, the analyses and the figures compare the distribution of the metrics (the boxplots), including all the catchments. From these results we can not conclude that the reanalysis products reproduce well (or not) the different metrics for each catchment. Therefore, I would suggest to plot the distribution of the bias (between obs. and reanalysis data computed for each catchment) instead. This would also help to estimate which dataset over- or under-estimate the metrics (which is difficult to assess on the figures in the manuscript). Or, even better in my opinion, to show the scatter plot of the metrics 'simulated' vs 'observed', as you did for the SPI (e.g. Figure S10). From these, we could assess the correlation between the reanalysis and the observed metrics, as well as their respective spread.

In addition, I don't understand the different types of representations of the results between:

a) the box plots in Figures 2, 3 and 5,
b) the scatter plots in Figures 4, 8 and 9,
c) and the box + violin plots in Figure 7.

a) and b) aimed at the same objective, which is to compare the datasets, why are the results presented in different ways?

c) shows the distribution of the trends, but why adding violin plot? It is redundant with the boxplot (and previous figures did not display violin plots).

Lines 278-280: "The positive reanalysis biases are particularly evident in catchments at high- and low-elevations (>2000 and < 1000 m.a.s.l.)" It seems that you have more catchments in these high and low altitudes, perhaps you could provide a histogram of the catchment altitudes?

Line 305: The differences in metrics are described in function of the altitude of the catchments, perhaps you could define the thresholds for low/mid/high altitude catchments once, and refer to these groups of catchments when describing the results?

Lines 312-315 How is this short section about the relationship between the fraction of snow and altitude relevant for the study (Figure 4e)?

Line 319: The 100% of under- and over-estimation could be related to catchment with very low fraction of snow.

Line 332: Did you mean that CHELSA overestimates solid precipitation? (the world "solid" is missing)

Figure 4: The figure would be easier to read with the name of the products on the plot (ERA5, ERA5Land …).

Line 337: I suggest to delete the parentheses about "variability from one day to the other" since this is not what the coefficient of variation (or the standard deviation) reflects. You could rank the daily precipitation from the highest to lowest precipitations during a year (which would result in low day to day variation) and get the same standard deviation value.

Figure 5: Rx1day and Rx5day are shown in the figure 5 but not discussed in the text. Moreover, the figure compares the distribution of the metrics, therefore I suggest to change the legend of the figure accordingly (perhaps to something like "Comparison of the distribution of daily to inter-annual precipitation variability"), the same apply for previous figures 2, 3, 5, 6, 7.

Line 349: I suggest to add a reference to the figure 6 and the end of this sentence.

Line 385: On figure 7d, the proportion of catchments with significant trend in Rx1day seems closer to 20% than 10%.

Line 394-396: The reasons (poor representation of dry days) for the poor trends detection in those metrics should be moved in the discussion section.

Figure 7: This figure is hard to read. Why adding violin plot on top of boxplot? The labels of plots i-o are too close to plots b-h, and can easily be confused with the x-axis of the other plot. The legend for the significance of the trends feel too complicated. The information of interest are: is the trend significant? Is it increasing or decreasing? And do we observe it in the simulated data? Wouldn't it be easier to represent using: color (e.g. from blue to red) for sig. increasing/ no-trend / sig. decreasing in observed data, and then hatch the proportion of catchment in which the simulated data failed to reproduce the observed trend?

Lines 418-419: I suggest to add a reference to the figure 8 at the end of the sentence.

Figures 8 and 9: I find the figure S10 more interesting and more straight to the point than the figures 8 and 9 in the manuscript. I would suggest replacing Figures 8 and 9 by Figure S10.

Line 469: What are the results discussed here? Figures S11/S12?

Discussion section

Overall comment: One of the objectives of the study was to investigate how the datasets perform in complex terrain, with a focus on mountain region. This should be discussed more here, since you

have low/high altitude catchments, the comparison of the behavior of the dataset in different categories of catchment could be investigated in more details. Can we use these datasets for catchments in high altitudes? Why? Why not? What are the recommendations? How could we improve the datasets for these regions?

Line 490: I think that the statement that CERRA clearly improves the representation of precipitations metrics is too strong. The analyses show that it improves how the spread is reproduced over the sample of catchment, however there is no evidence that bias is lower at the catchment scale. What about the spatial correlations of the metrics (obs. vs. reanalysis datasets)?

Lines 494-495: I suggest to move the sentence about the precipitation assimilation of ERA5 in the method section (or to remove it) as it adds nothing to the discussion here.

Lines 499-500: Is there a reference supporting the effect of bubble structures in Europe?

Line 505: The snowfalls evaluated are not those that were originally provided in the datasets, I think that it should be reminded here that it was re-calculated for this study (Lines 212-213).

Lines 525-527: If datasets and observation agree well at larger scales, should we use them (instead of more refined datasets) when studying large catchments? Should we select different dataset based on catchment size?

Lines 542-544: As stated below (lines 555-556), the isolated effect of statistical and dynamical downscaling can not be assessed with these datasets, since CERRA uses data assimilation. This sentence should be reworded.

Lines 577-580: Bruno et al. (2023) have also clearly shown how trends (in catchment evapotranspiration) can vary depending on time periods. (Figure 1 in Bruno, G., & Duethmann, D. (2024). Increases in water balance-derived catchment evapotranspiration in Germany during 1970s–2000s turning into decreases over the last two decades, despite uncertainties. *Geophysical Research Letters*, 51, e2023GL107753. https://doi.org/10.1029/2023GL107753).

Line 597: Errors of > 3 °C during winter could also explain why the snowfall metrics are not well reproduced by some datasets.

Conclusion

What the word “variability” (Lines 615, 616) refers to is not totally clear to me: is it temporal variability? spatial variability? Both?

I think that the summary of the results could be shortened, in the favor of a focus on the strength and weaknesses of the datasets in mountainous areas (which is a gap of knowledge identified Lines 96-97). More recommendations on which dataset to use, where, and why (or why not) would be appreciated.

The last sentence is vague and could be removed since all datasets have limitations (CERRA included).

Figure 12: Could this figure be introduced in the discussion section? It is very useful, although the attribution of performance grade (poorly/satisfactory/well) seems very subjective. Perhaps you could explain how these performance grades were determined in the section on methods?

Technical corrections

Line 65: the name of the product and its spatial resolution are in different orders in this sentence: CERRA at 5.5 km, vs. 6 km COSMO-REA6. I would suggest to rephrase for consistency.

Line 104: "…the representation of ~~**the**~~ all of these components…"

Line 106: "To shed light on the question **of** which reanalysis products are most suitable…"

Line 133 and table 1: "…the Integrated Forecasting System Cy41r2 and covers the period from 1940 until the present". In the text, ERA-5 started in 1940, but in table 1 it started in 1950.

Line 192: "To calculate catchment averages **metrics**, we use two complementary…".

Line 208, Table 2, Figure 3: The metric for annual min and max of temperature are written as "tg_min" and "tg_max" in the text and figure 3, vs. "tg-min" and "tg-max" in table2.

Figure 3: Labels of the metrics: rx1day in the figure vs. Rx1day in table 2 (same thing for r99ptot).

Line 396: "Figure33g" -> Figure 3g

Line 486: "…by CERRA and ~~**and**~~ overestimated..."

Line 518: "(Bandhauer et al., 2021)~~**)**~~". (there is an extra closing parenthesis)

Line 581: "or artificial trends Monteiro and Morin (2023)." I guess that the format of the citation is wrong here.

Line 634: "The CHELSA ~~**datasest**~~ **dataset** is available…"