Response to Reviewer 1

Dear Laurent,
We thank you very much for your positive and very constructive review. We appreciate the time you took to review our manuscript in such detail. We agree with most of your comments and decided to follow your recommendation to change some of the figures and give more detail to difference between elevation bins (especially figure 2, 3, 5 and 6). In that regard we switched from presenting the distribution of the metrics to focusing on the biases on the catchment level.

In the following your comments are in black and our response in green.

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).

Thank you very much for your positive assessment!

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.
Reviewer 2 had the same comment. We will rephrase this to “meteorological drought” and “extreme precipitation events”.

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?
We will provide additional references in the locations that you suggest.

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.
Yes, this should say “hydrological drought”. We will clarify this.

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.
We tried to be concise here, but we might have been too minimalistic with the information. We will provide more details on the studies.

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?
Response to the dataset selection: Our interest are the newest generation reanalysis products at higher resolution. Hence, our choice of the ERA5 suite and its derivates. Compared to the older generation reanalysis (ERA-interim), ERA5 has not only been improved in spatial resolution but also

in physical representation. The focus on the higher resolution reanalysis meant that we a priori exclude the other widely used global reanalysis (i.e., MERRA2 and JRA-55) which only provide data at very coarse resolution (>50 km). Despite the short simulation period, the datasets CERRA and Chelsa were chosen as they are expected to be extended back in time, soon offering a valuable tool for long term analysis. Further, there is currently only a limited number of studies available that evaluated CERRA or Chelsa, which in our opinion meant that these datasets required a thorough comparison. We will add a sentence on the reasoning of data choice.

Response to the datasets in Dura et al (2024): The focus of Dura et al (2024) was on France which meant that their data selection was tailored to the availability over France. Three of the datasets in Dura et al (2024) are only available over France and one dataset is a CPM run that dynamically downscaled another RCM which in turn has been driven by ERA-interim. Therefore, these datasets either didn't provide any data over the catchments or it is a second order derivative of an old reanalysis product.

Line 182: I wonder why you did not used precipitation and temperature provided by the CAMELS-CH? Do they come from the same products?
The meteorological data in CAMELS-CH are based on the same data from MeteoSwiss as we used in our study. Other available meteorological data within the CAMELS-CH dataset are post-processed MeteoSwiss data from the PREVAH hydrological model. We chose to use the original MeteoSwiss products instead to have full flexibility on the analysis. Especially for the climate indicators, which we calculated on the grid cell first and then averaged over the catchments, required us to use the gridded data and not the timeseries data from CAMELS-CH.

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).
We agree with you. We took your advice and adapted the figures to show differences rather than the spread of the climate metrics. Please see the response to the other comments.

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.
We didn´t plan to show a comparison between the two approaches. We decided to go with the two approaches as we think that the different analyses require different approaches. For the metrics mean precipitation and mean temperature, annual or seasonal, the difference between the approaches is only marginal. For the other metrics the different approaches lead to different results. We will extent the justification of the two approaches in the text.

Line 198 and 204: Are the annual mean and other metrics such as wetdays computed over hydrological years?
Only the snow related analysis is based on the hydrological year, other metrics are based on the calendar year. For the snow metrics we explicitly mentioned this in the text, but for the other metrics we didn´t. We will clarify this.

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.
As reviewer 2 also mentioned this, we will streamline the presentation of metrics and will select consistent metrics across figures. This means that we will likely remove dry days as a metric.

Lines 211-212: How different are the estimations of snowfall (re)computed in your study from those provided by the datasets?
We did not quantify this. We chose to use a consistent approach to separate liquid and solid precipitation as not all datasets provide snow height or snow water equivalent or the separation of liquid and solid precipitation. Further, in many applications (e.g., hydrological models or land surface models) snow is calculated within the model itself often following a similar approach using a temperature index approach.

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)?
We will adapt the table after we streamline the set of metrics that we want to show in the manuscript. This means we will include the missing mean air temperature metric and might remove some other metrics. We chose to show cold days, as this is the threshold for the separation of liquid and solid precipitation, and therefore is a universal threshold. We could have shown hot days as well, but as you say we would have had to specify a catchment specific threshold. As we cover high elevated and low elevated catchments a "hot day" will require very different thresholds, which to some extent would be an arbitrary choice. Further, as we already compare many different metrics, we choose to not add more metrics.

Results section
As your next comment covers several topics, we will split your comment and reply to the individual topics.
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).
We fully agree with you. We will follow your recommendation and will adapt the figures accordingly to show biases instead of the distribution of the metric.
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.
This would be informative, however by displaying the results in the style of Figure S10 would increase the number of figures dramatically. Each dataset would require its own panel (i.e., as in Figure S10), which would mean that if we want to retain all seasons in figure 2, we will require 40 panels, and 24 panels for figure 3. Therefore, we will keep the boxplots, but agree with showing the distribution of biases instead of climatological means.
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,
To show multiple metrics in a concise way.
b) the scatter plots in Figures 4, 8 and 9,
Here, we compare individual metrics, hence we can allow for a more detailed presentation of the results.
c) and the box + violin plots in Figure 7.
Needed a way to compare (a) the magnitude in trends and (b) the consistency in trends. Both of this information need different ways of conveying the information.

a) and b) aimed at the same objective, which is to compare the datasets, why are the results

presented in different ways?
See our reply above. Figures 2, 3, and 5 include multiple metrics and therefore need a concise way of presentation. Figures 4, 8, and 9 show individual metrics per figure.
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).
We will likely change the presentation of results in figure 7. We will certainly streamline the presentation and will change the violin plots to boxplots. See also the response to your specific comment on figure 7.

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?
The three elevation bins (high: >2000, mid: 1000-2000, and low: <1000 m) are almost equal in the number of catchments. We have adapted the figures to now also include the information on the three elevation bins. The boxplots are now overlaid with the median bias of catchments within each elevation bins (three different markers).

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

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

Line 332: Did you mean that CHELSA overestimates solid precipitation? (the world "solid" is missing)
Yes, thank you for noticing. We added "solid" for clarification.

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

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.
We will remove the misleading information.

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.
We will add the information on the variability of Rx1d and Rx5d in the text.
Further, following your comments and recommendation we changed figures 2, 3, 5 and 6 to show biases instead of the distribution of the climate metrics. However, in figure 7, we would like to keep the comparison of the distribution of trends rather than their biases, because due to the short time period the trend estimation itself might not be as reliable, as we also mention in the discussion, and therefore we would like to relax the comparison a bit and not show biases.

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

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

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.
Yes, this could fit into the discussion section. We will see whether this fits better in the discussion or not.

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?
All your points are valid. For consistency we will change the violinplots to boxplots. We will try your suggestion on the visualization of the trend matching. We are further thinking about splitting the current figure into two figures: a) trend magnitudes (boxplots) and b) trend consistency (barplots).

Lines 418-419: I suggest to add a reference to the figure 8 at the end of the sentence.
Good idea! We now include a reference to figure 8 and 9.

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.
We appreciate your suggestion! While the figures in the supplementary material give the overall behaviour, the figures 8 and 9 give a closer look into two explicit extreme events. In many studies only the overall behaviour is compared and then from this "general" statements on individual events are drawn. Therefore, we here decided to compare individual events that could be discussed in some more detail and give the overall picture in the supplementary material. Your comment is nevertheless useful as we had to critically reflect on the presentation of results again.

Line 469: What are the results discussed here? Figures S11/S12?
No, we refer to Figure 11. We added the reference to the respective figure & panels at the end of the sentence.

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?
Thank you for comment. All your points are valid. In response to your comment, we chose to include the elevation information in a more explicit manner in the figures. We now include the median biases of the three elevation bins in various figures.

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)?
We still think that this information holds true. As we will switch to showing biases instead of the distribution of the metrics this will become more apparent.

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.
We think that this information should remain here. It is indeed not directly relevant for our results, however, in other regions the advantage of data assimilation might be smaller as ERA5 includes data assimilation of precipitation data itself. Hence, this information is required for consistency reasons and limit the transferability of the results to regions without data assimilation in other reanalysis products.

Lines 499-500: Is there a reference supporting the effect of bubble structures in Europe?
We will do another screening of the literature, however, not many studies have used the CERRA dataset yet. Therefore, this will likely remain our own hypothesis.

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).
We will clarify this.

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?
Not necessarily. Here, we talk about precipitation variability. Indeed, the results by Monteiro and Morin 2023 suggest that on larger scales the datasets agree well, however, we interpret this that all datasets can represent large scale variability, however, as our results suggest not all datasets can represent smaller scale variability. The "good" representation of the large scale variability is to some extend explainable because of the data assimilation that constrains all reanalysis datasets to represent the large scale climate states. However, as we argue here the smaller scale variability is modulated by topographic features, which are resolution dependent.

If we move away from the specific case of "variability" we could argue that if we only study large catchments, then the choice of the dataset is not as relevant, as all information is smoothed anyway. However, this always depends on what we are interested in. However, if we study catchments across a range of sizes, then we would argue that resolution does play a role, as we can see that locally the differences can be large. Further, we would argue that you should try and have one consistent dataset for all catchments and not select individual datasets for each catchment. Further, as we show in the estimated snowfall analysis, the resolution does matter, and we can reach better performance when we move to higher resolution.

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.
We will rephrase this.

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).
Thank you for the suggested reference. We will check the reference and will decide whether it fits.

Line 597: Errors of > 3 °C during winter could also explain why the snowfall metrics are not well reproduced by some datasets.
Yes. This is exactly what we argue in the following sentence. We could make this more explicit to snowfall here. For example, we could mention snowfall as an example.

Conclusion
What the word "variability" (Lines 615, 616) refers to is not totally clear to me: is it temporal variability? spatial variability? Both?
This is temporal variability. We will clarify this.

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.
Thank you for the suggestion. We can certainly extend the recommendations aspect more.

The last sentence is vague and could be removed since all datasets have limitations (CERRA included).
Yes, this is true. We will think about either rephrasing or deleting the sentence.

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?
Indeed, the performance grading is subjective and relies on expert judgment. We will try to make this more transparent or maybe rephrase "performs well/poor/satisfactory" to "limitations apply yes/no/partly". The rephrasing might fit better to the limitations part of the figure. Moving the figure to the discussion section might be a good idea.

Technical corrections
Thank you very much for your eye for details. We will directly incorporate the following comments and only reply directly to the ones where we disagree.

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 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 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…"

Line 106: “To shed light on the question of which reanalysis products are most suitable…”
We don´t think that including “of” after “question” is necessary in this context.
Line 192: “To calculate catchment averages metrics, we use two complementary…”.
We might think of a different formulation.