General comments:

1. How have the equipment and methods used for discharge observations changed over the years? How has this affected the quality of the discharge observations? I would especially like to see a bit more text on how the quality of the measurements is ensured in the field, not just from the postprocessing side. For example, how often/when/if rating curves are updated? I'm assuming the usage of rating curves, but that would be good to state explicitly also in section 3.2.1 (I think it is mentioned briefly in section 4). Has there been an adoption of ADCPs that allow for faster and more accurate bottom geometry and stream velocity measurents? Do the automatic stations measure hourly values that are then used for calculating the daily value? Have the values before automatic stations been measured with limnograph (and are therefore are actual daily averages) or has the value been derived from the water level at a specific time of the day? References would be welcome if available, even if not peer reviewed or in Spanish.
2. The catchment polygons of adjacent basins (that should not overlap) exhibit small overlaps and gaps, likely due to border simplification that has been applied separately to each catchment. I suggest modifying the postprocessing to get rid of these artifacts. This could have an effect when selecting overlapping catchments, for example. Additionally, clean borderlines give an impression of higher quality of data.
3. I understand the need to use simulations for the hydrologic indices due to many gaps and short observation periods. However, it would be interesting to compare the simulated indices to the observed ones for the catchments with longest observation periods and least gaps, as an additional quality check to the simulated data. This might require adjusting the exact indice period to get enough gauges included.
4. Since the land cover data is available annually, it would be highly beneficial to include all the yearly steps instead of just 2025 to enable studies on the effects of land cover changes to discharge. This sort of information has previously been included in CAMELS-CH and CAMELS-FI, for example (Höge et al., 2023; Seppä et al., 2026).
5. Why are only the dominant two geological classes provided? Why not use the geological classification in figure 7 and give areal percentages for each class? This sort of configuration makes it unnecessarily hard to conduct analysis, as the same data column includes multiple different types of rock, making direct (statistical) comparisons to other variables impossible without cumbersome additional preprocessing. Important information is also lost, as two geological classes are not enough to cover most catchments adequately.
6. The current units of water extraction from ADH are not directly comparable to the mean discharge, precipitation or PET of the catchment, which are all in mm d^{-1}. These comparisons could be very useful for determining water stress of the catchment and general level of human influence to the catchment. I suggest adding additional unit converted (using the catchment area) attributes alongside the existing volumetric human intervention indicators (surface_adh_vol, groundwater_adh_vol, reservoir_vol)
7. If the lack of ADH permit cannot be considered as absence of human influence (L503), is there a list of known catchments with minimal/little human influence, that could then be turned into an attribute? This could be very useful, as many researchers want near natural catchments for various tasks.
8. Consider organizing the timeseries.csv into multiple files, one for each variable, one column per catchment and gauge_id as the column name. This would be a more easily usable format, since it allows for easy comparisons of specific variables between gauges over long periods.
9. Has the PISCO-ARNOVIC model been calibrated with data used for evaluation? If yes, then that should be stated explicitly, because it is a less robust form of evaluation than proper separation of calibration and testing periods. Have all the gauges been used for calibration?

If not, it would be interesting to see if there is a performance difference between the gauges used for calibration vs not. I also suggest adding Root Mean Square Error (RMSE) to the pool of metrics, because non of the current metrics show how much the model is typically wrong.

Specific Comments

L74–L75:

- Is there a source that could be cited for the idea of this specific regional division? It seems very natural, but a reasonable division to two, four or five regions could probably also be created.

L238:

- What was done when the catchment delineation of the study differed from that provided by ANA?

L240:

- How is immediate upstream catchment determined when there are multiple non-overlapping options (e. g. Nested group PE 204707)?

L241:

- What was the projection used for the calculation of area and perimeter? Can’t be the prjection the spatial data is provided in, because that’s in degrees.

L266:

- Was the code from https://github.com/naddor/camels used for calculating climatic and streamflow indices? At least the attributes match. If yes, please cite the code.

L355:

- A table in the appendix showing the reclassification of land cover classes (original to reclassified) would be useful.

L378:

- Do I understand correctly that glaciers are lumped together with lakes and wetlands as water bodies? They behave hydrologically quite distinctly, and should therefore be split into separate attributues

L431:

- The machine learning algorithm is not three dimensional itself, rather the inputs and the end product is. The authors of DSOLMap make a similar mistake in their introduction (not in the methods) (López-Ballesteros et al., 2023), which is where I suspect the error to originate.

L563:

- The code used for creating the dataset should be published. Direct quote from the data policy of ESSD: ”The output of research is not only journal articles but also data sets, model code, samples, etc. Only the entire network of interconnected information can guarantee integrity, transparency, reuse, and reproducibility of scientific findings. Moreover, all of these resources provide great additional value in their own right.”. A repository that is messy or provides only partial reproducibility is still better than no repository.

L735:

- Has it been ensured that the spike test doesn't remove real values from flash-flood events in the smallest catchments? In other words, how close do real observations come to the limit?

Technical comments (no need to respond if you implement the change):

Figure 4b & all other relevant figures:

- The text states that aridity exceeds one for many catchments, but the plot caps at one. Please add ’> 1.00’ symbol to the legend of the plots describing the feature in that and all similar cases. Conversely, < should be used when values go below the shown limit of the legend.

References

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth Syst. Sci. Data, 15, 5755–5784, https://doi.org/10.5194/essd-15-5755-2023, 2023.

Seppä, I., Gonzales Inca, C., Uusikivi, J., and Alho, P.: CAMELS-FI: hydrometeorological time series and landscape properties for 320 catchments in Finland, Earth Syst. Sci. Data, 18, 4745–4769, https://doi.org/10.5194/essd-18-4745-2026, 2026.

López-Ballesteros, A., Nielsen, A., Castellanos-Osorio, G., Trolle, D., and Senent-Aparicio, J.: DSOLMap, a novel high-
resolution global digital soil property map for the SWAT + model: Development and hydrological evaluation, Catena, 231,
107339, https://doi.org/10.1016/j.catena.2023.107339, 2023.