General comments

Thank you for your work compiling the CAMELS-DK dataset and the manuscript. The introduction gives an excellent overview of the current state of large-sample hydrological datasets and the challenges of preprocessing and retrieving large amounts of harmonized hydro-meteorological data, which CAMELS datasets help to solve. The data discussion section gives a very good and helpful overview on the hydrologic characteristics and specificities of Denmark, which is especially helpful for someone who is not familiar with the hydrological landscape in Denmark and can provide great insights and ideas for the discussion of one's own study results.

CAMELS-DK is another great addition to the ever-growing family of CAMELS datasets, especially with the incorporation of (simulated) groundwater data and a lot of nested catchments, where relationships between those catchments are clearly stated in the catchment attributes. The addition of many ungauged basins together with simulated discharge timeseries and catchment attributes gives the opportunity to study ungauged basins on a large scale.

Overall, I think the manuscript is a great addition to the growing pool of harmonized hydrological datasets. I hence recommend that the manuscript can be published after “minor” revisions: The only major comment I have is regarding the title:

- title: CAMELS datasets usually state the number of gauged stations in their title, with the number of 3330 Catchments in the title, I would assume that there is data for all of this Catchments available
    - also L103: "CAMELS-DK consistent data for over 3300 catchments" -> the data is not really consistent for all 3300 catchments, as some catchments have observed discharge while the majority does not provide observed discharge, a major difference across the catchments in the dataset

  Although the authors make it clear in their manuscript that the dataset includes ungauged and gauged catchments, I still think the title, particularly in comparison to other CAMELS datasets, is misleading.

Specific (minor) comments

- L19-L20: streamflow observations are typically the core component of CAMELS datasets, so in my opinion, these 304 catchments with streamflow data are really the core of CAMELS-DK, the enhancement maybe are the other 3026 catchments without observed streamflow
- L10-L25 (Abstract): maybe add the temporal range of your timeseries data to the abstract (1989-2019), as this is an essential information about the dataset?
- L126: Is it possible to cite the source of ID15v2.5 and state who / which institution made this division of Denmark into 3351 catchments?
- L158-L161: About the correction based on Stisen et al., 2011 -> I get that you conducted the correction of precipitation with several DMI data products yourself,

correct? Can you elaborate why sensors/gauges under-catch precipitation and why this correction is necessary? Also, can you briefly describe how the correction approach by Stisen et al. works?

- L164-L165: "DMI also provides gridded datasets of wind speed, air temperature and potential evapotranspiration (Scharling, 1999a). All variables are provided at daily timesteps; however, at an original resolution of 20 km2." -> By all variables you refer to wind speed, air temperature and potential evapotranspiration, correct? In Table 1 you only list temperature and pet, not wind speed. Also, did you calculate the pet with the Makkink formula yourself or has this already been done by the DMI? Perhaps you could expand on the section about the climate variables temperature, pet and wind speed and also indicate how you processed the data from the DMI (as you did in L154: "aggregated to catchment scale")
- L165: Give citation of the modified Makkink formula
- L151-165: Can you add a sentence about uncertainty in the climate data? Does the DMI state something about their uncertainty? Does the varying number of 200-500 stations over time affect uncertainty and does Stisen et al. correction affect uncertainty?
- L166 (Table 1):
    - Sources: better add the sources/citations of the datasets to the references and refer to them in the table, this way you can also add when you last visited the websites
    - Source for observed streamflow (Qobs) is the DK-model, should be Danish EPA / Aarhus University?
    - What is "SZ" in the description of some variables?
    - column Qsim is called Qdkm in the dataset
- L170: "Danish EPA" -> what is EPA?
- L171-L172: "Water levels are measured sub-daily (minutes)" -> how many minutes?
- L178: "station may have very limited time series length" -> what was the minimum time series length for CAMELS-DK?
- L186-L187: "Most of the stations have observed streamflow available during the entire years during 1989 to 2005" -> I do not fully comprehend this sentence, the entire years would be 1989-2019, or does this mean that these stations do not have any missing data between 1989 and 2005? You could also give a percentage number for "most of the stations"
- L185-L189: I think you could improve the section about data availability over time by expanding it further and structuring it more clearly. It doesn't really seem finished, and the incomplete sentence gives the impression that you actually wanted to write more here.
- L191 (Figure 2):
    - (a):
        - marker sizes in the legend do not match marker sizes on the map, the marker sizes on the map are good in my opinion, so just increase the marker sizes in the legend, especially for smallest markers.
        - Also, you could change the legend title "Area" to "Catchment area" to clarify what is meant here.
        - You could also think about switching to other colors, as the combination of red and green is not colorblind-friendly.
    - (b):

        - caption: "number of stations have available streamflow" -> "number of stations that have available streamflow"
        - to me, "have data over the entire year" suggests that there is no data missing for these stations over the entire year (all days have data), which should not be the case, looking at panel (c)
    - (c):
        - caption: "different period" -> "different periods"
        - you could also use different colors that are colorbild-friendly
- L210-L211: Why are the streamflow observations used for the DK-model only mostly similar to the timeseries in CAMELS-DK? Can you very briefly state the reasons for that here?
- L255: maybe cite the official source for the ID15 catchments, if possible?
- L257: "Flow direction indicates the downstream catchments, the column was filled with –9999 for most downstream catchments." -> the actual number in the dataset is -99
- L266 (Table 2):
    - unit of longitude and latitude is ° (degree)
    - column names for elevation are different in the dataset (dem_mean, dem_max, dem_median, dem_min instead of ele_mean, ele_max, ele_median, ele_min)
    - cite the 10m DEM via references in the Data source column if possible
- L271: Can you refer to formulas / code you used for the calculation of climate indices? You could also refer to the Github repository of Nans Addor, where you find the code he used to calculate the climate indices (**https://github.com/naddor/camels**)
- L272-L273: Did you also use a threshold of e.g. 5 % missing streamflow data in hydrological years for the inclusion of the year for the calculation of climate indices as in other CAMELS datasets?
- L272-L273: "climatic indices are calculated for water years from 1989 to 2019" -> there are different definitions of "water years", e.g. Oct-Sept, Sept-Aug, please clarify which period you used for a hydrological year in CAMELS-DK
- L275 (Table 3): frac_snow is actually called frac_snow_daily in the dataset
- L282-L283: do you also have a reference for the second nine signatures you mention here?
- L284: You should refer here to Appendix A, where the additional hydrological signatures are listed
- L284-L285: same as for climate indices; did you only use "complete hydrological years" (max. 5% missing streamflow data in a year) for the calculation of indices like other CAMELS datasets?
- L286: Units here are often mm/timestep, I think the unit should better be mm d^{-1} here
- L290-L302: Maybe state something about the limitations and uncertainty in comparing CLC data from different years due to different methodology and satellite data, is it also like that for Basemap04?
- L320: Maybe change the section name to "Hydrogeology and geology"?
- L336: Maybe change the Table name to "Catchment attributes of hydrogeologic and geologic features"?
- L358-L370: I think it would be great if you would include the NSEs and KGEs of your simulation results for each catchment in the dataset, this way others could compare

their model result to yours, which improves the usage of CAMELS-DK as a benchmark dataset. This way, it would also be possible to see how trustworthy the modeled results for individual catchments are (just an idea for improvement)

- L390 (Figure 6):
    - You could change the colors red and green here to make to figure colorblind-friendly
    - caption: "Climatology of precipitation observed streamflow, and..." -> "Climatology of precipitation, observed streamflow, and..."?
    - I would also prefer a more meaningful y-axis label here ("dtp" -> e.g. "simulated shallow groundwater level")
- L393: "Fig. 6 shows the climatology (1990 - 2019) of precipitation, observed streamflow and depth to phreatic layers" -> so are these the mean values of precipitation, observed streamflow and depth to phreatic layers over time? Maybe clarify this in the text.
- L425 (Figure 7):
    - again, I would prefer more meaningful variable names for axis and scale labels (e.g. "groundwater abstraction" instead of "DKM_gwe")
    - the x-axis label for the years is missing
    - Is there a reason for the inverted arrangement of panel (b)? I think the figure would be tidier if the map was also on the left and the time series on the right.
    - The resolution / dpi of Figure 7 is noticeably worse than that of the other figures
- L437-L438: "The Python script of processing the time series and landscape attributes based on original datasets are provided in a folder named ‘Python’." -> Is it possible that this Python script is incomplete? It is just a very small script for getting accumulated basins, and accessing the attribute files and the timeseries files. It is not the code for processing the original datasets for CAMELS-DK, by your wording I would expect e.g. code for processing the DMI meteorological source data to the CAMELS-DK timeseries
- L450: Where do you provide the original raster data? -> maybe rephrase, that you give instructions on how to access the source data in the Data availability section. As mentioned above, I also could not find Python scripts for data processing

Dataset

- In general you have lots of decimal places for all your variables, you could think about rounding to 2-3 decimal places
- the shapefiles "CAMELS_DK_304_gauging_catchment_boundaries.shp" and "CAMELS_DK_3330_catchment_boundaries.shp" yield different catchment boundaries for the same catch_id. I get that the 3330 shapefile contains the ID15 boundaries and the 304 shapefiles contains the ID15 boundary of the catchment + all upstream catchments, but you should clearly state that in the data description, as this can get confusing if someone is not aware of that or uses the wrong shapefile.
- the files CAMELS_DK_signature_obs_based.csv and CAMELS_DK_signature_sim_based.csv are actually the same

- CAMELS_DK_soil.csv uses column name "Id15_model" instead of "catch_id" for the catchment id, you should change that in the soil attributes file to be consistent across the dataset and make it easier to work with it
- the file CAMELS_DK_topography states via column "gauged_type" that there should be observed timeseries data for catch_ids [13210020, 13231401, 16200602, 32211251, 32230800, 35320416, 35321469, 57220534] but there are no csv files in Dynamics/Gauged_catchments/ for these IDs, instead I find observed timeseries data for catch_ids [13210113, 13231400, 16200624, 32230635, 35320540, 57220030, 71270810, 72300455], where the topographic attributes state it should not exist.
  The former list of “gauged” IDs from the topographic attributes are also missing in the stations shapefile and the gauging catchments shapefile, so I guess your topographic attributes are incorrect / out of date?
- In the shapefile CAMELS_DK_georegion.shp you have a georegion with the name “Thy” which are just big rectangles covering Denmark and the surrounding area. I think you can delete that from the shapefile?
- topographic attributes column “gauge_record_pct”: with a value of 100, I would expect no NaN values in the entire Qobs timeseries, but in your case you only refer to the time period for which you have discharge data (1989-2019) while the meteorological variables extend to 2023, so you always have NaNs at the end of the observed discharge timeseries. I think you should make that clear in the data description: “the time percentage with available observed streamflow” -> e.g. “the time percentage with available observed streamflow in the period 1989-01-02 - 2019-12-31”
- Data description:
  - L35-L36: "The hydrometeorological time series and landscape attributes for each catchment are calculated using the python package rioxarray." -> I think this is correct for the meteorological timeseries and landscape attributes, not the observed hydrological timeseries?
  - L102: Source for observed streamflow (Qobs) is the DK-model, from the manuscript the source should be Danish EPA / Aarhus University
  - L102 (Appendix A): column Qsim is called Qdkm in the dataset
  - L107 (Appendix B1): column names for elevation are different in the dataset (dem_mean, dem_max, dem_median, dem_min instead of ele_mean, ele_max, ele_median, ele_min)
  - L109 (Appendix B2): frac_snow is actually called frac_snow_daily in the dataset
  - L113 (Appendix B4): same comments as for technical manuscript comments L303
  - The tables in the appendix should contain all columns in the csv files, Appendix A misses the time column, Appendix B2-B6 miss the catch_id column

Technical corrections

- L12-L13: the acronym CAMELS actually stands for "Catchment Attributes and Meteorology for Large-sample Studies"

- L30: "basis of supporting water resource management..." -> "basis of support for the management of water resources..." or "basis for supporting water resources management..."
- L37: "Bing able" -> "Being able"
- L52: "catchments datasets" -> I think the word "datasets" can be removed here
- L55: "conined" -> "coined"
- L77-L78: "Especially, the lack samples from low-lying, small-size, and groundwater-dominated catchments." -> e.g. "Especially a lack of samples from low-lying, small and groundwater-dominated catchments."
- L83: "CAMLES" -> "CAMELS"
- L89: "CMAELS" -> "CAMELS"
- L90: "have already contribute" -> "have already contributed"
- L102-L103: webpages should better be added to the references, not cited in-text, in any case, add “last visited …” information
- L103: "CAMELS-DK consistent data for over 3300 catchments" -> e.g. "CAMELS-DK provides consistent data for over 3300 catchments"
- L107: "catchments attributes" -> "catchment attributes"
- L114: "CAMELS dataset" -> "CAMELS datasets"
- L119: "enrich existing CAMELS database" -> "enrich the existing CAMELS database"
- L120: "which respect to" -> "with respect to"
- L166 (Table 1):
    - Units: according to the ESSD submission guidelines, units must be written exponentially (e.g. "mm/d" -> "mm d^{-1}")
    - also the case for Table 2, 3, 4, 6, Appendix A, Figure 3, 6, 7
- L171: "through the surface water database of the (https://odaforalle.au.dk/)" -> sentence is not complete, also better to move urls / webpages to references or add “last visited …” to the url
- L175: "are accessible through (https://odaforalle.au.dk/)" -> again, better to move the url to the references and cite, e.g. "are accessible through the ODA platform (Aarhus University, 2024)", add “last visited …” in any case
- L187: "However, many hydrological stations ." -> incomplete sentence
- L215: again, better cite webpages via the references, add “last visited …”
- L219: "attention has been paid the representation" -> "attention has been paid to the representation"
- L250: "are also provide" -> "are also provided"
- L255: "The dataset is organized based on the ID15 catchment." -> "The dataset is organized based on the ID15 catchments."
- L255-L256: "eight-digital name" -> "eight-digit name"
- L262: "monitoring data of the data"? You mean percentage of available observed streamflow data in the time period?
- L273: "which is consistency to the availability of observed streamflow" -> "which is consistent with the availability of observed discharges."
- L303 (Table 5):
    - "Yeas" -> typo in the header
    - Description:
        - You could add "area" at the end of each description; e.g. "Percentage of agriculture" -> "Percentage of agricultural areas", "Percentage of urban" -> "Percentage of urban areas"

        - "Percentage of deciduous forest in" -> the "in" at the end is wrong
    - You have the full citation of Adhikari et al. in the Data sources and References column, move that to the References section at the end and just cite here
    - "Data source and References" column could be renamed to just "Data source" to be consistent with all other tables
- L304: "based on a regression modelling" -> "based on a regression model"
- L314-L316: "Several hydraulic parameters from the dataset of 3D Soil Hydraulic Database of Europe at 1 km and 250 m resolution, such as the water content at field capacity and saturated hydraulic conductivity (Table 6)." -> this sentence is not complete
- L336 (Table 7): First data source are two DOIs; add the full citation with DOIs to the references and cite here
- L353: (Figure 3):
    - order of labels is incorrect (a, b, c, d, f, e)
    - labels (b) and (f) overlap with plotted outliers -> move labels a little bit to the left or to the upper left corner outside the axes
- L401: "clay percentage in Fig. 6d" -> I think you want to refer to Fig. 3d here
- L415: "blow" -> "below"
- L420: again, better cite Jupiter via References
- L446: "The dataset is developed to assistant machine learning studies" -> "The dataset is developed to assist machine learning studies"
- L448: "ang" -> "and"
- L462: use exponential notation for units
- L463-L464: "indicating that a slightly smoother streamflow hydrograph simulated by the DK-model" -> "indicating a slightly smoother streamflow hydrograph simulated by the DK-model"
- L465: "challenged to capture accurately by the model" -> "challenged to be captured accurately by the model"
- L474-L475: "The ID15 catchment shapefile is provided with this dataset, previous version, ..." -> "The previous version X of the ID15 catchment shapefile is provided with this dataset, ..."