Review of the data set and manuscript "CAMELS-LUX: Highly Resolved Hydro-Meteorological and Atmospheric Data for Physiographically Characterized Catchments around Luxembourg", submitted to ESSD by Judith Nijzink et al.

General comments

First and foremost, I want to sincerely thank the authors for the effort they have put into creating this new CAMELS data set. It is initiatives like this that advance large-sample hydrology and the goal of better data availability in different regions. Having said that, I consider the CAMELS-LUX data set and accompanying manuscript a valuable contribution to ESSD. The authors provide a data set as well as a benchmark model for Luxembourg and the surrounding regions. This is a great extension of the LSH landscape.

However, I think that there are some points that should be improved before final publication, and I list them below. I hope that these are helpful. I am happy if I can contribute to further improving this data set and article with these thoughts and hope that I am doing the authors a useful service.

Specific comments – data set

General

The data set is easy to access and deposited in a well-structured Zenodo archive. The data description is helpful and gives a good overview of what is included in the data set. I also appreciate the availability of one folder containing the shapefiles for the boundaries of the catchments as well as the locations of the stream gauges.

I carefully checked the data set and noted some errors and things that were not clear to me. I hope that this helps to ensure the high quality of the data set. However, having found quite a few issues in the data set, it is my concern that there are further issues that I have not noticed. Therefore, I ask the authors to carefully check the data set again from their side. This will help to achieve the goal that users of the large-sample data set do not have to worry about the reliability of the data set.

Static attributes

- In the data description, you state that the file "basin_id.csv" was added to match the structure of other CAMELS data sets. I am not sure if I ever saw such a file, purely listing the IDs, and I am sure that I never used it. This does not mean that it cannot be helpful, but I was wondering if the value of this file would not be higher if some other data were given in it, such as for example the identifier used by the state agency (that I could not find anywhere else and may be valuable for some applications). This file is the only one that does not contain a header, I think this is a source of errors and should be changed.
- Please specify (in the data description file, for example), what you mean with "min. annual hourly air temp." and "max. annual hourly air temp.". Is this just the minimum and maximum value recorded per catchment for the whole time series? The "annual" confuses me here.
- In the data description file, it is stated for the total annual specific discharge that it was calculated for 2003-2020. Are January 2003 to December 2020 meant? Why is this the case, if the data range from November 2004 to October 2021? From the text, I would understand that the averages are taken from all available data per catchment. If this is not the case, please clarify. For *Qspec* as well as for *prad_sum*, my own calculations for the mean annual sums (over the whole time series) do not match with the values given in the climate attribute file for the examples I tested (ID1 and ID5). These are just examples, please check the calculations for all variables and all catchments and make sure that it is clear what exactly you included in the calculations.

- In the manuscript (see also the corresponding comment below), you state that a humid catchment has an aridity index > 1, which makes me assume that you calculate the aridity index as P/E_{pot}. However, in the data description, the term for the aridity index is stated as E_{pot}/P. Please check and clarify this. Related to this, if the sums for E_{pot} and the radar-based precipitation indicated in the climate attributes file are used to calculate the aridity indices, these do not equal the aridity indices stated in the attributes file, the same applies for the runoff ratio. I assume that this comes from rounding errors when the aridity indices and runoff ratios are calculated for each timestep and then averaged. However, I recommend you to aim for consistency within the attributes file, as this may lead to confusion otherwise. Also, I am surprised to find so many catchments with $E_{pot}/P > 1$, please check if all values and calculations related to this are correct.

Time series

- As already mentioned in the community comment by Ather Abbas, there are duplicate rows in the time series file with a resolution of 15 minutes for catchment 16. I could not find any other duplicate rows in all catchments and all temporal resolutions.
- I assume that the *Qspec* is always given for the time interval of interest (15 minutes, 1 hour, and 1 day, respectively), starting at the time given in the "Date" column. Please state this somewhere, for example in Table 1 of the data description file. This would be helpful to understand which hourly data belong to which daily data, and which data with a 15 min resolution belong to which hourly data. Due to rounding differences, this cannot clearly be identified from the data themselves. Related to this, as the *Qspec* values for the data with a 15 min resolution are often very small, I think it would be beneficial to include more than just three decimals. This would increase the added value of the data with a high temporal resolution, currently there are often many rows with exactly the same value following each other.
- I wonder if there is no other source of temperature data for Luxembourg than the global ERA5 data. While temperature may not be the most influential variable in hydrological modelling, it could be very relevant for other applications. And as the data are global, they are probably not the most accurate ones for these small areas. Please motivate why you use these data, or – this is the solution that I would actually prefer – use a more regional source for temperature data and the dependent variables (also in the static attributes).
- Similarly, the soil moisture and atmospheric data stem from global sources, which adds quite some uncertainty to them. However, for these parameters it may be less simple than for temperature to find an alternative data source. Therefore, I think it would be valuable, if possible, to include some information on the reliability of these data, and on what they can be used for and what not.
- The abbreviations given in Table 4 of the data description file do not always match with the column names actually used in the data set.

Specific comments – manuscript

Introduction

- It is great that you include different CAMELS data sets in the introduction. However, I would either mention a few examples and state that these are just some examples, or make sure that all existing data sets are included. Currently, data sets like CAMELS-SE (Teutschbein, 2024), CAMELS-FR (Delaigue et al., 2025), CABra (Almagro et al., 2021), LamaH-Ice (Helgason and Nijssen, 2024), or BULL (Senent-Aparicio et al., 2024) are not on the list but already published. Potentially, there are more that I don't have on top of my mind right now. Related to that, note that the Indian data set is called "CAMELS-IND", not "CAMELS-INDIA".
- Please note that "Caravan" is not an abbreviation, therefore, the name of the community data set is not written in capital letters. Please also make sure that you distinguish between ERA5

(Hersbach et al., 2020) and ERA5-Land (Muñoz-Sabater et al., 2021) and note that there is no space between "ERA" and "5". In Caravan, the ERA5-Land data are used, this needs to be corrected in the end of the first paragraph.

- In the very end of the introduction, you mention the atmospheric data that you included in the data set. I think it would make sense to already provide the reader with the information on the type of data that you refer to here, otherwise, this is not immediately clear.

Hydro-meteorologic(al) time series

- Where from or how did you obtain the catchment areas? I could not find this information in the manuscript. I think it would fit well in subsection 2.1.
- In the caption of Fig. 2, the statement "the rain rates were then sorted and correlated" is not clear to me: Does that mean that one point does not necessarily describe the same rain event in the x- and the y-direction?
- For the atmospheric data used to investigate thunderstorms, why do you use the precipitation data from ERA5? Wouldn't it be more favourable to use the precipitation data from the radar or rain gauges described earlier as it can be expected that these are of a higher quality?

Physiographical setting of Luxembourg

- When the aridity index is introduced, please state which way around you use it (the statement that an aridity index larger than 1 stands for humid catchment make me think that you use P/E_{pot}) as this as well as the reciprocal value of it are often seen in literature. In Fig. 3, analogous to the runoff ratio, you could then also give "P/E_{pot}" instead of "AI" as an axis label to avoid confusion. See also the comment regarding this in the comments on the data set, where you actually state that you calculated it as E_{pot}/P. This needs some clarification.
- The major river basins that are given on the x-axis in Fig. 3 are not introduced yet at this point of the manuscript. Please consider switching Fig. 3 and Fig. 4 to have a map first.
- The first part of subsection 3.4 is basically the same as subsection 3.1, please check if you could take out some redundancy there.

Topography and derived morphometric parameters

- I think that it is great that you include these morphometric parameters in the data set. I have a few remarks about the VRM, though:
    - Please give a reference for the definition of rugged landscapes ("greater than 0.01-0.02"). In addition, I would claim that for a definition, it would be better to either state greater than 0.01 or greater than 0.02, but not greater than a range. Otherwise, values within the range remain undefined.
    - I find the definitions currently not sufficient to understand the measure: The denominator n remains undefined. In the paper by Sappington et al. (2007), n is defined as the number of cells in the neighbourhood (used to calculate r). Please state the definition of n in your article and indicate what number you used. Furthermore, it is also not clear what sums you are calculating in the definition of r. Equations 11-15 (which stem from Figure 2 in the above-mentioned paper) are not clear. The way they are stated right now, x and y are defined by themselves in Equations 14 and 15. In addition, I think that including a multiplication with 1 in a definition is misleading. Please revise this part about the VRM and consider including a sketch to visualize what is being calculated.

Catchment behaviour

- If I interpret Fig. 6 (left column) correctly, the datapoints that align on an x-coordinate of 0% do not contain much information, or in other words, the information content of all the other points (i.e., the geology types that are actually present in a catchment) are much more important. Therefore, I recommend you to not include the geology types that are not present in a catchment,

allowing for a better visibility of the other datapoints. As an additional idea, would it make sense to only include the dominant geology type per catchment?

Data set application

- In line 319, relative humidity is stated to decrease slightly, but in line 321, I understand that relative humidity remains stable due to an increase in air temperature and an increase in atmospheric moisture content. I think that this needs some clarification.
- There is a lot of information in the second part of the first paragraph in subsection 6.1. I would appreciate if you could elaborate a bit more on the regional model, how the atmospheric data helps to describe flood generation, and how catchments with limited data can be included, if possible, to make this part easier to follow for the reader.

Technical corrections and typos

This list may not be exhaustive, but here is what I think needs some improved regarding technical issues and spelling mistakes:

- To increase consistency over the whole manuscript, I suggest you to either use the term "streamflow", or "discharge", or "runoff", but to not mix these terms.
- I would claim that "meteorologic" should be replaced with "meteorological" (to be honest, I noticed because Microsoft Word complained when I copied the subtitle).
- To my knowledge, "data" are always plural, so data "is" not available, data "are" available, for example. You could improve the occurrences where you use it in singular to enhance consistency.
- Units like "mm/h" should be written as "mm h^{-1}", this needs to be adapted for example in Fig. 2b and the accompanying text. Later you also use expressions like "mm/month" and "mm per year", this should also be adapted for consistency.
- There is a quotation mark in the end of line 108 that should not be there.
- Make sure that the way to write a date is consistent.
- In subsection 2.3, make sure that the different equations are given in the same format, and make sure that subscripts are not only subscripts in the formulae, but also in the text. In the description of the components of the formulae, please be consistent if you give the unit in brackets or separated by a comma. Please provide the unit for all components, where applicable.
- In line 114, for example, note that an average *annual* value should be given in mm, while an average value should be given in mm a^{-1}, in my opinion.
- For the Penman-Monteith equation, the reference to "Allen et al., 2015" is wrong (linking to some ResearchGate entry). If I investigated this correctly, you want to refer to the FAO56 report (Allen et al., 1998).
- In general, I see multilettered variables in equations critical, as this is mathematically not correct. Please consider if you can work with subscripts instead. Similarly, e.g., in Eq. 6, please do not use words but parameters in equations.
- Please make sure that the figures are ordered according to their mentioning in the text (currently, Fig. 5 is mentioned before Fig. 3).
- For the list of geology classes given at the end of subsection 3.4, please aim for more consistency (& or and, capitalization). Furthermore, are the numbers of the classes required? For the land use groups, you don't give numbers.
- In Figs. 4, 5, and 8, please add the units to the axis labels.
- Over the whole manuscript, please be consistent in the capitalization of directions (e.g., "Eastern" vs. "eastern").

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration - Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56, FAO, Rome, Italy, 1998.

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., and Troch, P.: CABra: A novel large-sample dataset for Brazilian catchments, Hydrol. Earth Syst. Sci., 25, 3105–3135, https://doi.org/10.5194/hess-25-3105-2021, 2021.

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., and Andréassian, V.: CAMELS-FR dataset: a large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking, Earth Syst. Sci. Data, 17, 1461–1479, https://doi.org/10.5194/essd-17-1461-2025, 2025.

Helgason, H. B. and Nijssen, B.: LamaH-Ice: LArge-SaMple DAta for Hydrology and Environmental Sciences for Iceland, Earth Syst. Sci. Data, 16, 2741–2771, https://doi.org/10.5194/essd-16-2741-2024, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Sappington, J. M., Longshore, K. M., and Thompson, D. B.: Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert, J. Wildl. Manag., 71, 1419–1426, https://doi.org/10.2193/2005-723, 2007.

Senent-Aparicio, J., Castellanos-Osorio, G., Segura-Méndez, F., López-Ballesteros, A., Jimeno-Sáez, P., and Pérez-Sánchez, J.: BULL Database – Spanish Basin attributes for Unravelling Learning in Large-sample hydrology, Sci. Data, 11, 737, https://doi.org/10.1038/s41597-024-03594-5, 2024.

Teutschbein, C.: CAMELS-SE: Long-term hydroclimatic observations (1961–2020) across 50 catchments in Sweden as a resource for modelling, education, and collaboration, Geosci. Data J., gdj3.239, https://doi.org/10.1002/gdj3.239, 2024.