**Reply to review by Franziska Clerc-Schwarzenbach**

Thank you for taking the time to review the CAMELS-LUX dataset and the accompanying manuscript. Please find our comments below.

**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.

*-> We have further investigated the dataset quality. The issues regarding the time shifts could be traced down to the raw data. For those stations, we asked for new data of higher quality and resolution, which we incorporated. We were also cross-checking the data for further systematic errors. In an updated version of CAMELS-LUX, we (1) kept NaN CIN values as they are, instead of filling the gaps, (2) updated discharge values from 2021, that were corrected in the raw data, (3) recalculated P station values from higher resolved raw data, and (4) added air temperature station data. An updated version (v2.0) of CAMELS-LUX will be uploaded to zenodo.*

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.

*-> The "basin_id.csv"-file was used in the LSTM code. We agree that it has no added value to the dataset in the way it is. We therefore consider deleting it from the folder.*

• 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.

*-> Yes, it is the minimum and maximum of the hourly time series during all complete years [2005-01-01 - 2020-12-31]. The labelling is a relic from the data extraction method via annual values. I renamed the parameters without "annual".*

• 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.
*-> 2003 is wrong - it should instead say 2005. And then it is indeed from January 2005 - December 2020. The annual mean values were calculated from complete years only, so the data in 2004 and 2021 were not considered for the annual aggregations. The calculations were done in the same way for all variables in the CAMELS-LUX dataset. The incomplete years at the beginning and end of the time series were neglected. The minimum and maximum air temperatures refer to the absolute hourly minimum and maximum within the time period in which the annual means were calculated." The text is adjusted as "The values are calculated based on the hourly data available between 2005-01-01 and 2021-12-31."*

• 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/Epot. However, in the data description, the term for the aridity index is stated as Epot/P. Please check and clarify this. Related to this, if the sums for Epot 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 Epot/P > 1, please check if all values and calculations related to this are correct.
-> I had calculated Epot/P and the values are mostly above 1, which is what the data suggests (). The values used for calculations were the complete years, i.e. mostly 2005-2020 and then rounded. In the updated file, I will recalculate P/PET and the runoff ratio from the climate attribute file to be easier reproducible and understandable. I will also retain from the wording "aridity index" and recalculate and rename it as P/Epot.

**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.
*-> We can confirm Ather Abbas and your findings that this error is limited to the one file only. A note about this was added to the data description on Zenodo. This has been corrected in the updated version of the dataset.*

• 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.
*-> The timestamps throughout all variables and the differing temporal aggregations are set at the end of cumulation intervals. I have added this information to subsection 2.1 Time series data - General information.*
*-> In the new dataset version, the number of decimals for Qspec was increased to 4.*

• 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).
*-> Due to technical data base constraints at the time of the data collection, air temperature data were not available. As the air temperature is rather homogeneously distributed and snow accumulation and melt effects are minor in hydrological simulations throughout the year in Luxembourg, and especially with respect to extreme events during the summer, we moved on with the easier available ERA5 data. Yet, we agree, that the highly resolved air temperature data would be a nice feature, along with the higher resolved precipitation station data, and added both to the new version of the data set.*

• 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.
*-> While there are some more localized data from the radiosonde at Idar-Oberstein, we would like to limit the dataset to the ERA5 model results. The ERA5 dataset presents a spatially distributed dataset of rather homogeneous quality as well as consistency throughout parameters and catchments. I have added to the atmospheric data description the following sentence: "This coarse grid as well as the fact that the data is extracted from a global model limits the data quality despite the data being assimilated to measurements and regridded." And for the soil moisture data description: "Note, that also the ERA5-Land data is extracted from a global model and its quality limited to tendencies. Yet, we found the data to be sufficiently accurate on the catchment scale."*

• 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.
*-> Thank you. I decapitalized the abbreviations in Table 4 according to 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".
*-> Thank you for the hint. As new data sets will keep being published, I will refer to only a few examples within Europe.*

• Please note that "Caravan" is not an abbreviation, therefore, the name of the community data set is not written in capital letters.
*-> Thank you. It's adjusted.*

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".
*-> The space is removed and the citation split.*
In Caravan, the ERA5-Land data are used, this needs to be corrected in the end of the first paragraph.
*-> Okay.*

• 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.
*-> I have added the following sentence for clarification "We included proxy parameters for atmospheric instability (convective available potential energy (CAPE), convective inhibition (CIN), K-index), moisture (specific and relative humidity, total column water vapour (TCWV)) and storm motion and organisation (wind speed (WS), low-level wind shear (LLS) and deep-layer wind shear (DLS))."*

**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 second paragraph in subsection 2.1 I have added: "The catchments for each gauge were extracted from the constructed DEM and its respective stream network described in section 4."*

• 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?
*-> Yes, this is what it means. There might be high values in the station data that might have been missed in the radar. In that extreme case, the event would only be represented in the station dataset. Yet, this figure nonetheless gives an idea about the general over- or underestimation of high rainfall events in radar data. I tried to clarify the caption.*

• 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?
*-> We agree that the radar and station data are of higher quality. For the study to investigate thunderstorms in Meyer et al. 2022, the radar data was used. The precipitation data from ERA5 were added to this dataset as potential additional information for the comparison of rainfall products or the LSTM that might find useful information in it.*

**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/Epot) 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/Epot" 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 Epot/P. This needs some clarification.
-> I will adjust the axis labels accordingly. For further explanations see the comment above.
I used Epot/P and am changing and relabelling it to P/Epot.

• 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.
*-> Okay, I have switched the subsections for climate and hydrology, which leads to first introducing the major basins as well as the map.*

• 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.
*-> Okay, I have taken out the first lines from subsection 3.4 and extended subsection 3.1 by a few words from the former subsection 3.4 paragraph.*

**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.
*-> Okay, I will limit the range to 0.01*
▪ 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.
*-> I have deleted the confusing equations and added a definition for r and n in the text "The resulting vector (r) is standardized by the number of cells (n) surrounding the calculated grid cell."*

**Catchment behaviour**
• If I interpret Fig. 6 (left column) correctly, the data points 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?
*-> Yes, I will replace the 0% values with NA values to remove the points. The other idea of only including the dominant geology per catchment would cut a lot of data points. "Limestone & dolomites" as well as "Surface deposits" would completely vanish from the Figure.*

**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.
*-> I have removed the sentence "Yet, relative humidity decreased slightly." in line 319.*

• 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.
*-> Okay. This paragraph was split into 2 (which got lost in the original typesetting). The second part about the regional model is now rewritten and extended a bit.*
*"This CAMELS-LUX dataset is used to set up a regional Long-Short-Term-Memory (LSTM) model for Luxembourg. Including the atmospheric data as additional input parameters in the modelling process allows the LSTM to identify atmospheric conditions that favour thunderstorms. This might support the model when describing related flood generation, even if the triggering precipitation is not accurately represented in the data. Moreover, the regional approach of comparing and relating 56 catchments, allows for the inclusion of catchments with limited data. A lack of data, as it is the case for some catchments of particular interest, i.e. where flash floods occurred in 2016 and 2018, might be compensated by knowledge gained from the surrounding catchments."*

**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 "stream flow", or "discharge", or "runoff", but to not mix these terms
*-> I have eliminated "stream flow", but still mix discharge and runoff, as I find these two are slightly different.*

• I would claim that "meteorologic" should be replaced with "meteorological" (to be honest, I noticed because Microsoft Word complained when I copied the subtitle).
*-> Yes, agree.*

• 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.
*-> I agree it should be made consistent, while to my knowledge it can be both.*

• 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.
*-> Yes, I have tried to spot and correct all inconsistencies.*

• There is a quotation mark in the end of line 108 that should not be there.
*-> I only see ", but will make sure to not have further quotation marks in the text.*

• Make sure that the way to write a date is consistent.
*-> Okay.*

• 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.
*-> Thank you. This can now be found in subsection 2.5.*

• 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.
*-> Yes.*

• 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).
*-> Thank you!*

• 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.
*-> I have replaced the words with variables in eq. 6.. Yet, to calculate the parameters such as WS, LLS, DLS, PET, IDPR, TWI, VRM, I prefer to leave the abbreviations as they are. Else, I think that I have used only single lettered variables.*

• 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.
*-> I have removed the numbering and changed the classes to consistent & without capitalization to match Fig. 5.*

• In Figs. 4, 5, and 8, please add the units to the axis labels.
*-> Okay.*

• Over the whole manuscript, please be consistent in the capitalization of directions (e.g., “East ern” vs. “eastern”).
*-> I have capitalized them all.*