The manuscript by Nasreen et al., investigate the reconstruction of annual runoff timeseries for 14 European catchments over the period 1500-2000. In the first part, the authors evaluate the validity of an existing precipitation and temperature reconstruction dataset against GCHN stations. In a second step, they evaluate the use of 2 data-driven models and a lumped hydrological models to predict annual runoff. In a third section, they provide an overview on years with low annual runoff occurred in the selected 14 European catchments during the last 500 years.

Main comments:

I believe the manuscript need improvements in the methodology description, some additional analysis especially related to model and input forcing evaluation, as well as a bit of text reorganization and polishing in some sections. Here below I list my comments.

Important remarks

1. The authors should make it clear through the entire text that the manuscript deal with annual runoff reconstructions. Neither the title and the abstract mention it. I would suggest starting by modifying the title with "A 500-year annual runoff reconstruction for 14 selected European catchments".
2. Instead of "long-term", I would suggest using the term "multi-century"
3. When speaking about droughts in the text, please always specify the scale of the considered drought:
   - Runoff drought → Annual runoff droughts
   - Drought duration → Multi-year drought duration
   - Runoff drought severity → Annual runoff drought severity
4. I wonder if it's wort to keep in the text everything related to the "natural proxy data".
   - The inclusion of such data does not improve the model at all. You could simply add a sentence in the model discussion saying that "the inclusion of additionally proxy data has been investigated but did not provide benefits to model accuracy". I would suggest to just focus on the improvement provided by adding drought indicator (scPDSI).
   - In the text you mention multiple times the use of "natural proxy data" that clutter and complicate the reading in the introduction, Section 2.2 and 3.1.
   - In Section 2.2. you speak about data standardization. For what reason? Then you applied the normalization to 0-1 for model training as described in the Appendix?
   - For GRDC stations where there is no close proxy data which data did you use? There is skill reported for all catchments in Table 3 for "Gridded+Proxy" models
   - Are you adding a column for the precipitation natural proxy, and one column for the temperature proxy? This is not explained in the text.
   - In Section 3.1 you say: "selected the raw proxy data from inside the catchment or within a 100 km buffer around the catchment.". The following question arises:
     - ➔ If more than one proxy in the 100 km outside the catchment do you take the average value of the proxies?
     - ➔ If no proxy in the 100 km radius, which value do you assign to the catchment?
     - ➔ Is it representative a single proxy within a catchments that extends thousands of km^2 (tens of 0.5 x 0.5 grid cells)?

5. I would suggest creating a separate section to introduce the scPDSI drought indicator (new Section 2.2?).
6. In Section 3.1 please introduce how you define the calibration and validation set. Currently they are defined only in the results Section 4.2
7. Clearly state that GR1A is a conceptual lumped hydrological model.
8. In Section 3.2, please specify in more detail how the X parameter of GR1A is optimized and if it is optimized independently for each catchment. Only in the result Section 4.2 I can read "for each catchment separately".
9. In Section 3.3, please specify if a NN model is trained for each catchment, or a single model is trained for all catchments
10. In Section 3.3, I would avoid the use of term "Gridded". All models received as input is the sum/average catchment P and T. Maybe the word "Forcing" is more appropriate. "Gridded" erroneously make thinking to "distributed" or gridded simulations.
    Additionally, at line 159, please specify that the "lagged forcing" refers to 1-year lag data. Currently is just specified at line 259. Also provide explanation why you didn't use additional temporal lags (i.e. 2 and 3 years).
11. In Section 3.3 please clarify what is currently described at line 175-179.
    - "Best performance" at L175 refers to which metric? MAE?
    - "To reduce the likelihood of overfitting during the calibration/training, a fraction of the calibration data was used to check the performance of an independent (or so-called "testing") set"
      ➔ Which fraction?
      ➔ I am confused. Training/Calibration: 1900-2000; Verification/Test set: (prior 1900). The model tuning/validation set is a fraction of 1900-200 data?
      ➔ Please use the term "test set" only for data not used for model training and hyperparameter tuning
12. L12: "On the other hand, the data-driven models have been proven to correct this bias (referred to underestimation of variance)".
    In the main text, but also in the supplementary you don't provide the rSD metric on the annual runoff evaluation against GRDC. It is therefore difficult to verify such statement. On my experience, data-driven models are good in coping with conditional bias in the data, less in conditional variance. I would be surprised if you overestimate the variability of runoff. Please provide the rSD metric also in the runoff evaluation. In Fig.5, I see all models to underestimate the variance!
13. What you define as rSD in the Appendix is the "ratio of standard deviations" and not the "relative error in standard deviation" as referred to at line 181.
14. In both the evaluation of P, T and R, you don't provide information related the bias. Please provide the BIAS (mean difference between pred. and obs.) or the relative BIAS (BIAS/mean(obs)). You could report BIAS instead of D.
15. Please correct the definition of the skill metrics in the appendix.
    - The definition of R at line 404 is wrong!
    - At line 412, the coefficient of determination is equivalent to R^2. And "decided improvement" is maybe a too strong word …

- In the equation of the index of agreement (maybe IoD), which sometimes appears as D and sometimes as d, in the denominator there is a missing "i" subscript within mean(o).
- At line 421, alpha corresponds to rSD, r to R. There is lot of repetitions. I guess you could remove also the scaling factors "s" within KGE since I guess you use 1 for all of them.
- Maybe add the BIAS or relBIAS metric

16. I would suggest removing entirely the analysis of the impact of aggregating ed time-scale analysis. I believe it does not have anything in relation to the objective of the manuscript, and introduce plenty of questionable sentence
    - L220: "The RMSE decrease with increasing temporal aggregation because the RMSE depend on the number of observations. "
      → I would eventually argue that RMSE decrease because aggregating over time smooth (aka) decrease the variance.
    - L222: "Except for correlation which shows relatively stable values over aggregations, it is evident that the reconstruction skill decreases the greater the (aggregation) time-scale".
      → It means that for the reconstruction skill you refer to NSE or KGE.
      → The rSD is expected to decrease when averaging over time
      → If NSE and KGE decreases, if the correlation is relatively stable, and the RMSE decrease, the source of the decrease is increasing bias. But you don't provide information on it …
    - Eventually, the caption of Figure 4 should be completely reformulated. It does not describe the figure content, it does not mention if it refers to P or T evaluation; it refers to GRDC instead of GHCN, ..
      →"Fig. 4. Benchmarking GOF accuracy of P (or T) reconstruction against GHCN stations at various temporal scale …
17. Figure 2 and 3 should be revised.
    - Please add the BIAS or relBIAS metric (eventually replacing IoD)
    - Please correct the colorbar limits to facilitate comparison. I suggest setting to 1 the max value for Index of Agreement, NSE and KGE colormaps.
    - KGE should be bounded to 0 as far as I know. But I see negative values !!!
    - NSE below 0 means that the long-term mean of the station time series would provide better accuracy than using the reconstruction. Maybe set lower limits of NSE also to 0 (unbounded left)
    - Strangely the KGE colormap as a single step value of 0.3 (0.3-0.6). I guess there is a code mistake here !!!
    - Slightly reduce the marker size to reduce a bit the superimpositions of the circles.
18. Please color code the cells of Table 3 with the same colormaps of Fig 2 and 3
19. I am not sure I understand what is represented in Fig. 8. Is a comparison between GRDC vs simulated runoff values in the Q0-Q33 range? If yes, the axis label should be runoff [mm/year] !!!
20. I find really interesting the analysis in 4.4. Maybe you could highlight the value of some your statements, by adding an interesting figure or by for example plotting some drought years labels close to their cdf points in Fig 6.

21. I think that an additional plot with the "best" reconstruction of one or two time series (selected and zoomed) from Fig S1 and S2) could be a nice addition to the main manuscript.
22. I would like to draw your attention to the fact that there are a couple of works related to century and multi-century hydrological reconstructions that are not currently present in your references and would be worth adding:
    - Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A., and Graff, B.: Ensemble reconstruction of spatio-temporal extreme low-flow events in France since 1871, Hydrol. Earth Syst. Sci., 21, 2923–2951, https://doi.org/10.5194/hess-21-2923-2017, 2017.
    - Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth Syst. Sci. Data, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.
    - Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2021). G-RUN ENSEMBLE: A multi-forcing observation-based global runoff reanalysis. Water Resources Research, 57, e2020WR028787. https://doi.org/10.1029/2020WR028787
    - Moravec, V., Markonis, Y.,Rakovec, O., Kumar, R., & Hanel, M.(2019). A 250-year European drought inventory derived from ensemble hydrologic modeling. Geophysical Research Letters, 46. https://doi.org/10. 1029/2019GL082783
    - Smith, K. A., Barker, L. J., Tanguy, M., Parry, S., Harrigan, S., Legg, T. P., Prudhomme, C., and Hannaford, J.: A multi-objective ensemble approach to hydrological modelling in the UK: an application to historic drought reconstruction, Hydrol. Earth Syst. Sci., 23, 3247–3268, https://doi.org/10.5194/hess-23-3247-2019, 2019.

    Related to Rhine Drought, this one is also very interesting:

    Christian Pfister, Rolf Weingartner & Jürg Luterbacher (2006) Hydrological winter droughts over the last 450 years in the Upper Rhine basin: a methodological approach, Hydrological Sciences Journal, 51:5, 966-985, DOI: 10.1623/hysj.51.5.966

23. Facultative (but potentially interesting and very appreciated), I would be curious to see how a temporally aggregated century-long monthly runoff reconstruction such G-RUN (Ghiggi et al., 2019, 2021) (i.e. forced by GSWP3) would compare to your annual time series during the calibration period. I guess it could require a couple of day of work, but I am intrigued to know if an ad-hoc catchment based annual runoff reconstruction provides better results than annual catchment runoff derived from gridded monthly runoff timeseries.

Minor corrections
7: long-term → multi-century annual runoff reconstructions are still lacking (…)
7: Remove: (e.g. monthly, ….)
9: Remove: proxy data (if you follow Important Remark 3)
25: For the last 40 years → In the last 40 years
26: Missing reference for the 45 million loss fact
30-33: To be reformulated, please!

43-44: To be reformulated, please!
47-48: To be reformulated, please!
51: Reference Breiman et al., 2001 do not refer to a ML application for runoff/streamflow forecasting. You can find better ones 😉
51: Reference Thiesen et al., 2019 and "shannon entropy" are not used for runoff/streamflow forecasting
53: Contrasting (changing) → Changing
53: Suggestion: limit their application outside boundary conditions observed during model training.
55: long-term → multi-century annual runoff reconstruction for 14 European catchments
57: Remove: proxy data (if you follow Important Remark 3)
57: "other long-term historical data sources" → GRDC and scPDSI
57: we use a combination of → we benchmarked the use of
58: Conceptual HM → Conceptual lumped HM
59: annual evolution → annual runoff
59: "We pay particular attention to low flows during drought years."
The models are not optimized to pay particular attention to negative annual runoff anomalies so I would avoid such sentence.
60: "Using long-term data on climatic conditions and runoff may provide an efficient technique of visualizing droughts and low flow periods". Please reformulate or remove.
63: Drought identification → Drought identification methodology
63. To be reformulated. Suggestion: The accuracy of the employed precipitation and temperature reconstructions, as well as the derived runoff simulations, is evaluated in Section
69: data from → scPDSI drought indicator data from
69: Remove: natural proxies (if you follow Important Remark 3)
75: To be reformulated, please!
76. What about subparagraph: 2.1.1 Precipitation, 2.1.2 Temperature?
94-96: Consistency: Choose between dataset or data-set
104: "The runoff series from the GRDC were selected based on the condition of data availability, at least 25 years prior to 1900." → Only GRDC runoff time series with at least 25 years of data prior 1900 were selected
124: Remove "we" → Section 3.4 …
125: Section 3.5 presents the methods to identify annual runoff droughts
132: and the proxy data and
133: validation of individual catchments (Fig.2) → Fig 2 refers to P evaluation ☹
135: See Important remark 4
181: Provide the metrics in Capital Case format
199-204: This maybe belong more to Section 3.2 and 3.3
212: "Some stations indicated a worse performance and could not adequately capture the observed temperature variability". → Very likely, is not the station that has bad skill, but the reconstruction 😊 → "Low skill observed at some GHCN stations can be explained by the unresolved variability of grid-cell average temperature, especially in regions with complex-terrain."
217: Consistency: GOF or gof
236-239: Move to Section 3.1
240-241: GR1A is not driven by gridded data, but the catchment average value … ! Maybe move to Section 3.2.
260: Please reformulate (or remove between brackets content)
277: I don't get how scPDSI provide better representation of the temporal dependency structure

287-288: Please reformulate
294: Please reformulate
297: simulations → cumulative distribution of simulated runoff values
319: match, less → agreement, lower
354, 358, 362,374,376: runoff → annual runoff
359: conceptual → conceptual lumped
371-373: Maybe remove?
374: develop → derive
378-379: I suggest removing the aggregation time scale analysis and results
396. Specify g and o before starting describing the metrics.
395: measurement → metrics
396: ratio of standard deviations
435: Remove: and epochs.
435-437: Suggestion: The Huber Loss is employed to minimize the mean absolute error between observations and predictions. Model checkpointing is used to keep track of model weights evolutions during training and select the best model weights when the allocated max number of training iterations is reached.

For further questions / discussions:
Gionata Ghiggi: gionata.ghiggi@epfl.ch