Reply to Reviewer 2

We greatly appreciate the Reviewer efforts to evaluate our work, and we thank the Reviewer for the remarks that have allowed us to improve further on the presentation and clarify various parts of the previous edition, as described below.

Reviewer 2

"A 500-year runoff reconstruction for European catchments" by Sadaf Nasreen et al.

The manuscript "A 500-year runoff reconstruction for European catchments" by Sadaf Nasreen et al. shows the work and effort that has been done to create a new dataset of long-term runoff reconstruction for various European catchments. While reconstructions of meteorological variables such as temperature and precipitation were already available, this study closes the gap by providing open source runoff reconstructions. This is valuable information as it can provide historical context for upcoming studies, which are interested in assessing present and future extremes such as droughts. To create the runoff reconstruction, a semi-empirical hydrological model (GR1A) as well as two data-driven model (LSTM and BRNN) were used and tested for their suitability. An extensive data collection including precipitation, temperature, drought indices, natural proxy data and runoff observations over the period 1500 to 2000 were used to calibrate and validate the models. The data-driven models showed the most promising results, being able to correct for biases in the input data compared to semi-hydrological model. Furthermore, the separate analysis focussing on droughts showed that the reconstructed timeseries of these models correlated well with the historical documented droughts. The paper was well written and included extensive information on the approach and validation of reconstructed runoff timeseries, especially regarding drought events. The main points for improvement are mainly focussing on additional clarifications regarding certain aspects of the methods. Therefore, I would like to recommend publication after minor revisions. The comments for improvement can be found below.

Major Comments

Section 3.3 Data-driven models: While there is an extensive general explanation on the LSTM model, the BRNN description falls short. More importantly, information necessary to be able to follow as a reader on how the data driven models were trained and the final model parameters are not given or not fully clear. Aspects on the training/testing data, like the type of splitting or how much was withheld for training/testing purposes are not clear. Furthermore, aspects on the input data (for example the gridded+proxy, gridded+PDSI, and gridded+lag) and its preparation (e.g. type of normalization, handling with outliers, etc) would be of interest as well. An additional table (could be in the Appendix) with the input parameters as they are used in the ML models would support the readers understanding. While the Appendix covers the LSTM model structure, similar information on the BRNN is missing. Additional suggestion is to add the final model parameters (e.g. amount of neurons) into the schematization (Fig A.1) as well.
Overall I think adding more specific information on the ML models will only improve the readers understanding and as the data-driven models show the most promising results highlighting and clarifying their use is important.

Author's Response: We have indeed realised that the readers could be confused from the considered models and their description, especially without adding the information about training setup and tuning parameters. As a result, we considerably revised the main text to include the procedures for testing, training, and validation at the beginning of the methods section and improved the text flow. A section on general description of BRNN was added to Appendix A3. The model structure and parameter definitions are also included in the Appendix Table A1. We hope that it is now clear in the revised text.

Section 3 Methods: A schematic overview of the data-preprocessing (3.1), the incorporation of all the different datatypes and sources in the different model types (3.2 and 3.3) as well as the postprocessing steps (3.4 and 3.5) would be a nice addition to this section to not only visualize the general approach of the study but also support the subchapters and the readers under-

standing.

Author's Response:
We appreciated your suggestion. We agreed on this point in order to provide a clear representation of the workflow. We included Figure 2 in the manuscript to illustrate the work flow for data preprocessing, model selection and training technique selection as well as visualization methods.

Minor Comments

line 43: As an example: Hansson et al.... remove "of"

Author's Response: L40 ➔ The text was removed.

Fig 1 (and also Fig 5 and Fig 6): think about changing red or green to a different colour to ensure that colorblind people can follow your figures.

Author's Response: According to the Reviewer suggestion, the colors in Figures 6 and 7 (previously labeled as Fig 5 and Fig 6) were modified. Additionally, the colors in Figure 1 were changed.

Line 75: sentence not flowing, for example move comma in front of reference and remove 'was done'

Author's Response This sentence was properly rephrased.
L74-76 ➔ "Reconstructed precipitation (P) was derived by Pauling et al. (2006) through principal component regression to documented evidence (i.e., memoirs, annals, newspapers), speleothem proxy records (Proctor et al., 2000) and tree-ring chronologies from the International Tree-Ring Data Bank (ITRDB)."

Fig 2 and Fig 3 same range and colorbar per evaluation metric (makes it easier to compare), list min and max values of scale bar for readability

Author's Response:
Figures 3 and 4 (updated figure numbers) ➔ "We have listed the minimum and maximum values for both metrics figures. Some metrics have varying ranges in both figures (For example; relBIAS and BIAS, RMSE for P and T are different), hence, we keep them as an original scale. The color bar in both figure's were made identical but with different ranges."

Section 3.4: possibly some lines on the pros and cons of the GR1A model

Author's Response:
L140-142 ➔ "Compared to other conceptual models from the GR family (GR4J, GR5J), GR1A is simple to use, and allows for analyzing many variants, particularly defining best antecedent rainfall and potentially useful to predict the likelihood of floods and droughts (Mouelhi et al., 2006)."

Section 4.1 (and throughout the rest of the manuscript): be consistent in addressing GOF (now a mix between GOF and gof)

Author's Response: The GOF symbol was made consistent across the manuscript in response to the Reviewer remark.

Section 4.2 the information on calibration and validation should be part of the Methods and the models. Furthermore, it would be nice to move a figure with the time series of one station as seen in Fig S1 from the supplementary to that paragraph

to highlight calibration and validation periods.

Author's Response: The method section has been updated to better explain the calibration and validation phases.
L129-130 ➔ "Data were split into two parts: calibration (1900–2000) and validation (<=1900) to assess the model's accuracy and to select an appropriate model."
In addition, time series of two best runoff reconstruction (Fig. 8) were included in the manuscript.

Line 255: 'greatly increased the performance (NSE from 0.2 to 0.62).' Compared to the values mentioned in prior example of Basel Reinhalle, 0.62 is not listed Table 3 for BRNN(Gridded+PDSI) but 0.57

Author's Response: You are right, there was a mistake that was corrected in the revised manuscript.

Table 3: highlighting the different performances is a nice feature and helps spotting important trends, however the darkest colours make it hard to read the values (same for tables in supplementary). Maybe also add a note in the table description what the colour indication means.

Author's Response: The colormap of Table 3 was updated as similar to Figures 2 and 3. Also, Table 3 was altered to Figure 4 when the legend color bar was added. Likewise, the Supplementary Tables were changed.

Both stations at Basel show higher correlation scores for validation than calibration. Ideas why this is the case?

Author's Response: Table 3 ➔ We do not have any convincing explanation since for the data-driven methods, the calibration exhibits a higher correlation than the validation as expected.

Figure 7: Whitespace around the figure seems to be cut too narrow as the max value for station BaselRheinhalle-Rhine is cut off (130 instead of 1300)

Author's Response: Figure 9 ➔ The figure was corrected in accordance with the suggestion.

Line 347 and Table 5: listing of years does not include 1724, which is also indicated in bold in Table 5.

Author's Response: L324 ➔ Thanks, the year 1724 was added

Section 6: 'using the data set below' move below

Author's Response: The text was corrected.

Appendix: add references to equations and in text (easier to follow in case chapter layout changes)

Author's Response: Thank you, the references were added in equations and text.