Reply to Reviewers

Dear Dr. Ghiggi, thank you for your constructive and valuable comments. We have revised the manuscript in response to your comments and hope that you will find our revised manuscript suitable for publication.

Reviewer 1

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 model 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.

Major Comments

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 select-ed European catchments".

   Author's Response: We agreed and have updated the title, abstract **(L10)** as well as other occurrences throughout the text.
   Title: "A 500-year annual runoff reconstruction for 14 selected European catchments".
   The abstract was modified as,
   L7-10 ➔ "In this study, we have used reconstructed precipitation and temperature data, Palmer Drought Severity Index and available observed runoff across fourteen European catchments in order to develop annual runoff reconstructions for the period 1500–2000 using two data-driven and one conceptual lumped hydrological model."

2. Instead of "long-term", I would suggest using the term "multi-century"

   Author's Response: L7, long-term was changed to 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

   Author Response: We agreed and hence, we have made changes throughout the manuscript (e.g. **L190 and L195).**

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 catchment that extends thousands of km2 (tens of 0.5 x 0.5 grid cells)?

Author's Response: Thank you for this suggestion. We agreed that it is much easier to present the analysis without other natural proxies for the reasons you are summarizing. Therefore, we removed all proxy-related material and analysis from the manuscript.

5. I would suggest creating a separate section to introduce the scPDSI drought indicator (new Section 2.2?).

   Author's Response: Following the comment, we introduced a new section.
   L82 ➔ Section 2.2 'scPDSI Drought indicator'

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

   Author's Response: Indeed, thank you for this comment. We defined the calibration and validation periods in the beginning of the Methods section.
   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."

7. Clearly state that GR1A is a conceptual lumped hydrological model.

   Author's Response: This was done:
   L134-135 ➔ "The GR1A is a conceptual lumped hydrologic model Manabe (1969), considering dynamic storage and antecedent precipitation conditions."

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".

   Author's Response: We updated the section 3.2.
   L138-139 ➔ "The parameter X is optimized individually for each catchment by maximizing the Nash-Sutcliffe efficiency (NSE) between observed and simulated runoff."

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

   Author's Response: The text was altered in response to the reviewer suggestion:
   L174-176 ➔ "The model development process was repeated several times, minimizing the Root Mean Square Error (BRNN) and Mean Square Error (LSTM) for each catchment individually."

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).

Author's Response: We appreciated the comment because it pointed out some inaccuracies in our methodology description. Indeed, we did not use the actual gridded simulation but the mean value of P and T across the catchment (as now explicitly mentioned in the methods). We have decided to use lag 1 year for our analysis because the correlation between (lagged) annual precipitation and runoff drops to 0 after lag 1.
L152-157 ➔ "We considered combinations of reconstructed forcing, OWDA-based scPDSI, and lagged forcing as an input into the network for both model types. Specifically, the network using only reconstructed forcing is referred to as "P+T", the network with reconstructed forcing and OWDA scPDSI is termed as "P+T+PDSI" and finally the network which includes 1-year lagged forcing is referred to as "P+T+Lag". Please note, that dependence between annual precipitation and runoff at longer time lags was explored as well but since the correlation drops significantly at lags longer than 1, longer lags were not considered in the models."

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

Author's Response: In LSTM method, Mean Absolute Error (MAE) was referred as the loss function and Mean Square Error (MSE) was used to check the internal measure of accuracy. In addition, the training data set ranged from 1900 to 2000 and the validation set spanned the years prior to 1800 until GRDC-Runoff became accessible. While, testing set (25% of trained data) was used to avoid overfitting and determine, when training became halted.
The text was modified accordingly,
L172-176 ➔ "To set the optimal hyperparameters of the models (such as the number of neurons and activation functions, etc.) and to reduce the likelihood of overfitting during the calibration/training, the performance was checked considering an independent (or so-called "testing") set. This was pulled from the calibration data (1900–2000) as a (random) fraction (25%). The model development process was repeated several times, minimizing the Root mean square Error (BRNN) and Mean Square Error (LSTM) for each catchment individually. The model with the best performance was then chosen for further evaluation. "

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!

Author's Response: The rSD was added into the heatmap and runoff reconstructions in the Supplementary material.
You are right - our statement in the previous version was incorrect - the variance is underestimated as clear now from the figure. We revised the text accordingly.
L11-13 ➔ "On the other hand, the validation of input precipitation fields revealed an underestimation of the variance across most of Europe, which is propagated into the reconstructed runoff series."

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.

Author's Response: L179 ➔ The term "relative error in standard deviation" has been replaced by "Standard Deviation Ratio" in the Appendix.

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.

    Author's Response: We agreed. The D metric figures were replaced by BIAS or relBIAS figures.

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 R2 . 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

    Author's Response: The definition of R was corrected and added at lines ➔ L405-406.

    $$R = \frac{\sum_{i=1}^{n}(p_i - \overline{p})(o_i - \overline{o})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}\sqrt{\sum_{i=1}^{n}(o_i - \overline{o})^2}}$$

    Where "p" and "o" referred to predicted and observed value. The definition of D was deleted, since it is not reported anymore. The KGE was simplified and scaling factors have been replaced with 1. In addition, the BIAS and relBIAS metrics were included in Appendix.

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 . . .

    Author's Response: Thank you for this suggestion. This analysis originated in the preliminary exploration of the dataset and we agreed that it is not needed for the scope of the paper. The multi-scale analysis was therefore, removed from the paper.

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.

    Author's Response: Figures 2 and 3 were renamed as Figures 3 and 4, respectively. Hereafter, Relative bias is included in Figure 3. While for Figure 4, we introduced Bias instead of D. In both cases, the "relative error of standard deviation"

was replaced with the "Standard deviation ratio". Furthermore, we reduced the marker size and adjusted the KGE and NSE thresholds from maximum to minimum ranges and corrected all color scales.

18. Please color code the cells of Table 3 with the same colormaps of Fig 2 and 3

    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.

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] !!!

    Author's Response: Figure 8 originally represents the runoff deficit based on 33% threshold. As a result, the Figure labels were written as runoff deficit[mm/year], and the caption was also modified.
    "The observed and simulated runoff deficit based on the 33rd percentile threshold for 14 selected catchments during the calibration and validation period."

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

    Author's Response: "We appreciate Reviewer remarks. Figure 6 was modified by adding drought events to each panel and further lines were added.
    L282-286 ➔ "The most severe drought year identified by the models was the same in the periods 1500–1800 and 1900–2000 (Figure 7 left and right panels), while for 1800–1900 the models identified either 1865 (GR1A, LSTM) or 1858 (BRNN, 2nd worse for LSTM). Please note that the 1858 low water mark is available at Laufenburg Pfister et al. (2006) near Basel and was regarded as one of the worst winter droughts in the last 200 years."

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.

    Author's Response: We agreed, and Figure 8 is included in the main text to show the two best reconstructed runoff series.

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

    Author's Response: The reference (Moravec et al., 2019) already exists in the manuscript. The remaining suggested references were added to the database and the following text was updated in the subsection.
    L42-44 ➔ "As another example, Caillouet et al. (2017) provides a 140-year data set of reconstructed streamflow over 662 natural catchments in France since 1871 using the GR6J hydrological model, highlighting several well-known extreme

low flow events."

L341-343 ➔ "Monthly runoff anomalies analyzed from the GRUN data set (Ghiggi et al., 2019) show that August 1976 was the fifth driest month between 1900 and 2014, with some of our study catchment also signaling the 1976 yearly drought (e.g Köln-Rhine, Hann-Munden-Wesser, Bodenwerder-Wesser)."

L44-46 ➔ "A multi ensemble modeling approach using GR4J has been applied by Smith et al. (2019) to develop a UK-based historical river flows and examine the potential of reconstruction for capturing peak and low flow events from 1891 to 2015."

L284-286 ➔ "Please note that the 1858 low water mark is available at Laufenburg Pfister et al. (2006) near Basel and was regarded as one of the worst winter droughts in the last 200 years."

23. Facultative (but potentially interesting and very appreciated), I would be curious to see how a temporally aggregated century-long monthly runoff reconstruction such GRUN (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 time series.

Author's Response: The statistical analysis of annual-based GRUN forced by GSWP3 and runoff reconstruction against GRDC runoff (Figure S9) and two time series (Figure S10) were included in the Supplementary Material and are also presented below.

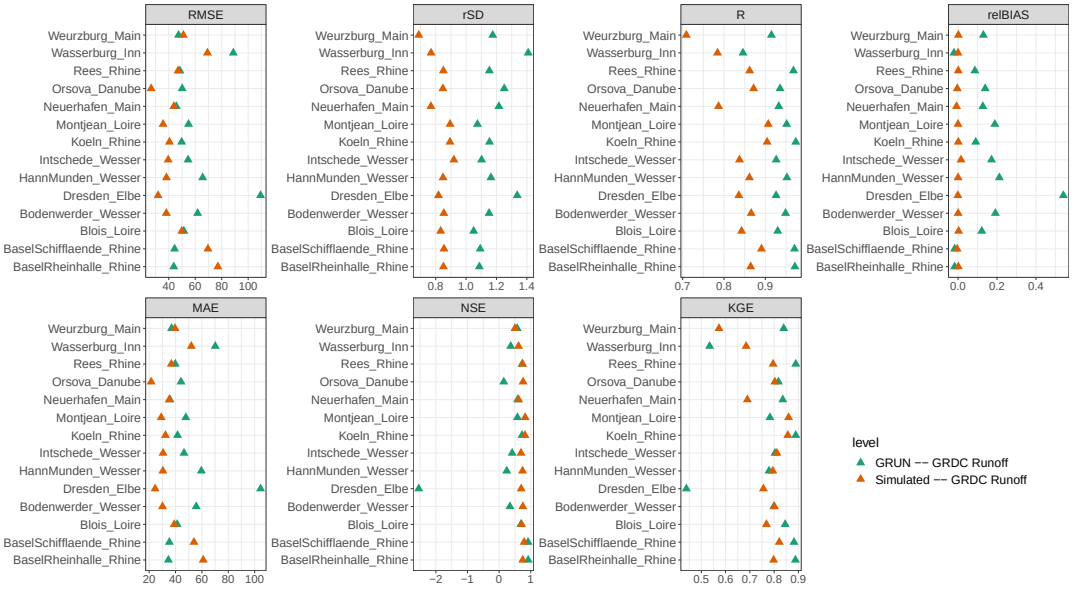

Figure 1. Statistical comparison between reconstructed and GRUN runoff with respect to observed GRDC Runoff for the common period 1902-2000

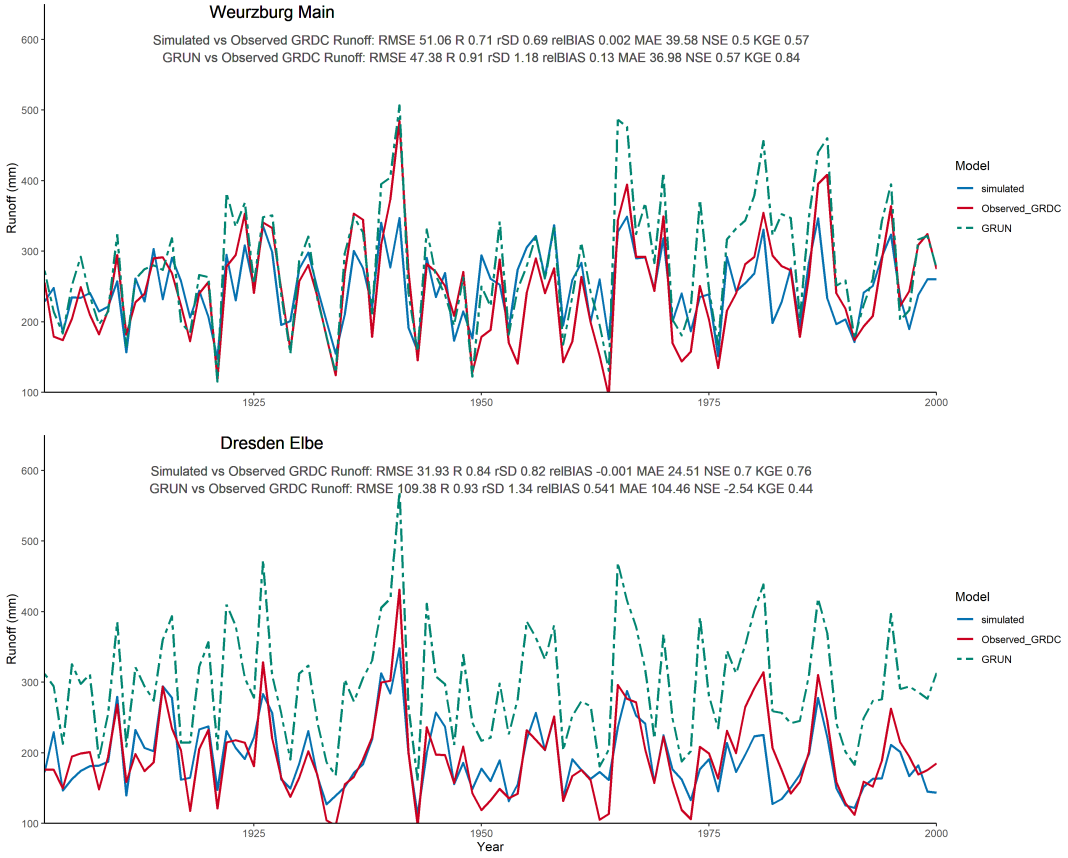

Figure 2. The simulated, observed, and GRUN time series for Dresden and Weurzburg catchments from 1900 to 2000. Model comparisons in terms of statistics can be seen at the top of the Figure.

Text was also added in the manuscript,
L293-300 ➔ "Finally, to check to consistency of our reconstructed dataset, we compared the skill of our simulation with respect to the GRDC runoff observation with that of the GSWP3-forced GRUN monthly runoff averaged over the catchments (Supplementary Fig. S9 and S10). Our reconstruction outperforms GRUN data in RMSE, MAE, relBIAS and NSE in the majority of the catchments, while the correlation to GRDC runoff is slightly higher for GRUN compared to our reconstruction. The variability, which is underestimated by our data-driven models (on average by 16.5%) is overestimated by GRUN (on average by 17.2%). Since the correlation compensates the bias the KGE for our reconstruction and GRUN is comparable. This suggests that GRUN could be used for data-driven model training, provided at least some information on flow characteristics is available in the catchment."

Minor Comments

7: long-term $\longrightarrow$ multi-century annual runoff reconstructions are still lacking (. . .)

Author's Response: L7 ➔ The text was updated

7: Remove: (e.g. monthly, ...)

9: Remove: proxy data (if you follow Important Remark 3)

Author's Response: L8 ➔ The text was removed

25: For the last 40 years ⟶ In the last 40 years

Author's Response: L23-24 ➔ The text was updated accordingly

26: Missing reference for the 45 million loss fact

Author's Response: L24 ➔ Reference was added (Anonymous, 2020).

30-33: To be reformulated, please!

Author's Response:"Continuous records of runoff/discharge series are no longer available, including various multi-year droughts and pluvial periods. On the other hand, proxy-based (typically seasonal or annual) reconstructions are alternatively used, considering various proxy data, such as past tree-rings (Cook et al., 2015; Casas-Gómez et al., 2020; Tejedor et al., 2016; Kress et al., 2010; Nicault et al., 2008), speleothem (Vansteenberge et al., 2016), ice cores, sediments (Luoto and Nevalainen, 2017)"
L25-32 has been changed to,
"While runoff is a key element related to water security, it is difficult to interpret recent hydroclimate fluctuations (multi-year droughts in particular) considering observed runoff records (Markonis and Koutsoyiannis, 2016; Hanel et al., 2018), which are in general seldom available for years prior 1900. In this way, we are missing runoff information on various severe multi-year droughts and pluvial periods, which can be assessed only indirectly using (typically seasonal or annual) reconstructions based on various proxy data, such as past tree-rings (Nicault et al., 2008; Kress et al., 2010; Cook et al., 2015; Tejedor et al., 2016; Casas-Gómez et al., 2020), speleothem (Vansteenberge et al., 2016), ice cores, sediments (Luoto and Nevalainen, 2017) and documentary and instrumental evidence (Pfister et al., 1999; Brázdil and Dobrovolný, 2009; Dobrovolný et al., 2010; Wetter et al., 2011)."

43-44: To be reformulated, please!

Author's Response: "As an example of Hansson et al. (2011), which introduced a runoff series for the Baltic Sea only, between 1550 and 1995 using temperature and atmospheric circulation indices." has been changed to
L39-41 ➔ "As an example, Hansson et al. (2011) introduced a runoff series for the Baltic Sea for the period 1550–1995 using temperature and atmospheric circulation indices."

47-48: To be reformulated, please!

Author's Response:"This can be achieved through a process-based model of varying complexity, with the advantage of following general physical laws – e.g., preserving mass balance, etc. Physical based models:"
L47-52 ➔ "The available reconstructed precipitation and temperature series (or fields) can be used to reconstruct runoff with hydrological (process-based) models (Tshimanga et al., 2011; Armstrong et al., 2020) respecting general physical laws, such as preserving mass balance (e.g. MIKE SHE, Im et al., 2009; or VELMA Laaha et al., 2017) or data-driven methods which are able to capture complex non-linear relationships (for instance support vector machines Zuo et al., 2020; Ji et al., 2021; artificial neural networks ANNs; Senthil Kumar et al., 2005; Hu et al., 2018; Kwak et al., 2020; random forests Ghiggi et al., 2019; Li et al., 2021; Contreras et al., 2021)."

51: Reference Breiman et al., 2001 do not refer to a ML application for runoff/streamflow forecasting. You can find better ones

Author's Response: L52 ➔ Reference was added: Li, Y., Wei, J., Wang, D., Li, B., Huang, H., Xu, B., Xu, Y. (2021). A Medium and Long-Term Runoff Forecast Method Based on Massive Meteorological Data and Machine Learning Algorithms. Water, 13(9), 1308. doi:10.3390/w13091308.

51: Reference Thiesen et al., 2019 and "shannon entropy" are not used for runoff/streamflow forecasting

Author's Response: L52 ➔ Reference has been removed

53: Contrasting (changing) ➔ Changing

Author's Response: L53 ➔ Text was updated

53: Suggestion: limit their application outside boundary conditions observed during model training.

Author's Response: Text has been modified as, **L52-54 ➔** "While, the lack of physical constraints in the data-driven models limits their application under changing boundary conditions (in comparison with those of the model training period), their advantage is that they can often directly use biased reconstructed data as an input series."

55: long-term ⟶ multi-century annual runoff reconstruction for 14 European catchments

Author's Response: L55 ➔ Text has been updated
"The objective of the present study is to provide a multi-century annual runoff reconstruction for 14 European catchments,"

57: Remove: proxy data (if you follow Important Remark 3)

Author's Response: L57 ➔ Text has been updated

57: "other long-term historical data sources" ➔ GRDC and scPDSI

Author's Response: L57 ➔ Following text has been added "Old World Drought Atlas Self-calibrated Palmer Drought Severity Index (scPDSI) reconstruction (Cook et al., 2015). "

57: we use a combination of ➔ we benchmarked the use of

Author's Response: L68-70 ➔ Text has been updated as,
"For validation the reconstructed datasets, we considered the observational data records of precipitation and temperature (Menne et al., 2018), as well as runoff from the Global Runoff Data Center (GRDC; Fekete et al. 1999), which was also used for model calibration."

58: Conceptual HM ⟶ Conceptual lumped HM

Author's Response: L58 ➔ Conceptual HM is revised as "Conceptual lumped hydrological"

59: annual evolution ➔ annual runoff

Author's Response: L60 ➔ Text has been updated

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.

Author's Response: L60 ➔ Suggested statement has been removed

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.

Author's Response:L60 ➔ Suggested statement has been removed

63: Drought identification ⟶ Drought identification methodology

Author's Response: L62-63 ➔ Text has been updated

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

Author's Response: L63-64 ➔ "The accuracy of the employed precipitation and temperature reconstructions, as well as the derived runoff simulations are evaluated in Section 4."

69: data from ➔ scPDSI drought indicator data from

Author's Response: L68 ➔ Suggested term has been defined

69: Remove: natural proxies (if you follow Important Remark 3)

Author's Response: L68 ➔ Text has been removed

75: To be reformulated, please!

Author's Response:"To this end Pauling et al. (2006), reconstructed precipitation (P) was done..."
The above line has been changed to
L74 ➔ "Reconstructed precipitation (P) was derived by Pauling et al. (2006) through principal component regression..."

76. What about subparagraph: 2.1.1 Precipitation, 2.1.2 Temperature?

Author's Response: L72-73 ➔ Subsection Precipitation has been introduced and the text "temperature gridded data" is removed.
L77 ➔ Subsection Temperature has been added.
L79-80 ➔"Reconstructed temperature data was available in the same spatial and temporal resolution as precipitation."
L80 ➔ The phrase "both of these" has been added.

94-96: Consistency: Choose between dataset or data-set

Author's Response: L86, L89... ➔ Suggested term (dataset) has been revised.

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

Author's Response: L97-98 ➔ The sentence was reformulated, as suggested above, so now it should be more clear.

124: Remove "we" ➔ Section 3.4 . . .

Author's Response: L117 has added 'Section 3.4' in the manuscript.

125: Section 3.5 presents the methods to identify annual runoff droughts

Author's Response: L118 ➔ Suggested statement has been updated

132: and the proxy data and

Author's Response: L123 ➔ The text "and the proxy data and" has been removed.

133: validation of individual catchments (Fig.2) ⟶ Fig 2 refers to P evaluation

Author's Response: L124 ➔ It appears that Latex had a typos problem. It is now Table 2 at L124

135: See Important remark 4

Author's Response: L128-129 ➔ Text has been removed and added the following line.
"The catchment average precipitation, temperature and scPDSI were estimated from the corresponding (gridded) data sets by averaging the relevant grid cells over the catchments."

181: Provide the metrics in Capital Case format

Author's Response: L178-181 have been modified with Capital Case format.
"We used a set of seven statistical metrics to assess the performance of simulated runoff, namely: Nash–Sutcliffe efficiency (NSE), Pearson Correlation (R), Standard Deviation Ratio (rSD), Kling-Gupta efficiency (KGE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Bias (BIAS) and Relative Bias (relBIAS). The mathematical formulations of these metrics are provided in Appendix A1."

199-204: This maybe belong more to Section 3.2 and 3.3

Author's Response: L197-201 ➔ We keep the original text since an introduction sentence is essential to describe related analysis.

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."

Author's Response: L212-213 We agreed, the formulation was unfortunate. The statement was modified as suggested.
"Low skill observed at some locations can be explained by the unresolved variability of grid-cell average temperature, especially in regions with complex terrain."

217: Consistency: GOF or gof

Author's Response: L213 ➔ This whole paragraph has been deleted because of aggregated time series analysis was already suggested to remove. But, we kept consistent (GOF) throughout the manuscript.

236-239: Move to Section 3.1

Author's Response: L129-131 ➔ These lines are moved to section 3.1 and was deleted from L224.

240-241: GR1A is not driven by gridded data, but the catchment average value . . . ! Maybe move to Section 3.2.

Author's Response: Section 3.2 has been explicitly stated, as suggested.
L136-137 ➔ "where Q, E and P represent annual runoff, basin average potential evapotranspiration and basin average precipitation, respectively and i denotes the year."

260: Please reformulate (or remove between brackets content)

Author's Response: We removed the brackets' contents and replaced them with the following:
L244-246 ➔ "Across many study locations, the combination of reconstructed forcings and their 1-year lag performed the best in terms of rapid convergence (the number of iterations needed) and high accuracy from all input combinations for both data-driven models (BRNN, LSTM)."

277: I don't get how scPDSI provide better representation of the temporal dependency structure

Author's Response: L260 ➔ This statement regarding scPDSI dependancy was deleted.

287-288: Please reformulate

Author's Response: Following lines has been reformulated Eventually, we decided to utilize that model since the metrics used (NSE, KGE, R, D, RMSE, MAE) to produce better results in one particular model.
L267-269 ➔ "Secondly, we identified the candidate best models for each of the 14 selected catchment, considering the GOFs based on the validation NSE and R greater than 0.5 and 0.7, respectively. The best model for each catchment was finally subjectively selected from those models considering the remaining validation measures (BIAS, rSD, KGE, RMSE and MAE) as well."

294: Please reformulate

Author's Response: The whole statement was revised.
L274-277 ➔ "The latter figure compares the cumulative distribution functions of annual runoff for the periods 1500–1800, 1800–1900 and 1900–2000, as simulated by the BRNN(P+T+Lag) and LSTM(P+T+PDSI) – the two best performing models – and the GR1A (the most deviating simulation from the best model) with the distribution of the observed annual runoff for the Basel-Rheinhalle Rhine catchment."

297: simulations ➔ cumulative distribution of simulated runoff value

Author's Response:
L278-279 ➔ Text has been revised as "The cumulative distribution of BRNN and LSTM simulated runoff values."

319: match, less ➔ agreement, lower

Author's Response: L303, text has been updated as "The agreement between the simulated and observed runoff deficit is lower compared to the annual runoff time series."

354, 358, 362,374,376: runoff ➔ annual runoff

Author's Response: L354, L356, L361, L373, L375 are changed as annual runoff

359: conceptual ➔ conceptual lumped

Author's Response: L357, text has been updated

371-373: Maybe remove?

Author's Response: The text have modified a bit to explain the fact.
L370-372 ➔ "Moreover, proxy records that were used for the derivation of precipitation and temperature input fields are spatially heterogeneous with some regions being better represented than others. This inevitably leads to poor performance over the latter."

374: develop ➔ derive

Author's Response: L373 ➔ Term has been updated

396. Specify g and o before starting describing the metrics.

Author's Response: We chose p as a better symbol for the predicted value instead of g as gridded.
L393 ➔ The terms p_i and o_i refer to the predicted and observed time series at point i respectively.

395: measurement ➔ metrics

Author's Response: L394 ➔ Text has been updated

396: ratio of standard deviations

Author's Response: L395 ➔ Term has been changed as "Standard Deviation Ratio".

435: Remove: and epochs

Author's Response: L433 ➔ Text has been removed

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.

Author's Response: This comment was not entirely clear to us. The loss function (MAE) was in our case minimized during model compilation. However, we improved the description of the model setup and training. See methods and Appendices A2 and A3.

References

Anonymous(2020): , https://www.euronews.com/green/2020/08/03/flowing-low-how-can-europe-use-climate-information-to-manage-dry-river-spells.

Armstrong, M. S., Kiem, A. S., and Vance, T. R.: Comparing instrumental, palaeoclimate, and projected rainfall data: Implications for water resources management and hydrological modelling, Journal of Hydrology: Regional Studies, 31, 100 728, 2020.

Brázdil, R. and Dobrovolný, P.: Historical climate in Central Europe during the last 500 years, The Polish Climate in the European Context: An Historical Overview, p. 41, 2009.

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, Hydrology and Earth System Sciences, 21, 2923–2951, 2017.

Casas-Gómez, P., Sánchez-Salguero, R., Ribera, P., and Linares, J. C.: Contrasting Signals of the Westerly Index and North Atlantic Oscillation over the Drought Sensitivity of Tree-Ring Chronologies from the Mediterranean Basin, Atmosphere, 11, 644, 2020.

Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., and Célleri, R.: Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment, Atmosphere, 12, 238, 2021.

Cook, E. R., Seager, R., Kushnir, Y., Briffa, K. R., Büntgen, U., Frank, D., Krusic, P. J., Tegel, W., van der Schrier, G., Andreu-Hayles, L., et al.: Old World megadroughts and pluvials during the Common Era, Science advances, 1, e1500 561, 2015.

Dobrovolný, P., Moberg, A., Brázdil, R., Pfister, C., Glaser, R., Wilson, R., van Engelen, A., Limanówka, D., Kiss, A., Halíčková, M., et al.: Monthly, seasonal and annual temperature reconstructions for Central Europe derived from documentary evidence and instrumental records since AD 1500, Climatic change, 101, 69–107, 2010.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: Global, composite runoff fields based on observed river discharge and simulated water balances, 1999.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth System Science Data, 11, 1655–1674, 2019.

Hanel, M., Rakovec, O., Markonis, Y., Máca, P., Samaniego, L., Kyselỳ, J., and Kumar, R.: Revisiting the recent European droughts from a long-term perspective, Scientific reports, 8, 1–11, 2018.

Hansson, D., Eriksson, C., Omstedt, A., and Chen, D.: Reconstruction of river runoff to the Baltic Sea, AD 1500–1995, International Journal of Climatology, 31, 696–703, 2011.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., and Lou, Z.: Deep learning with a long short-term memory networks approach for rainfall-runoff simulation, Water, 10, 1543, 2018.

Im, S., Kim, H., Kim, C., and Jang, C.: Assessing the impacts of land use changes on watershed hydrology using MIKE SHE, Environmental geology, 57, 231, 2009.

Ji, Y., Dong, H.-T., Xing, Z.-X., Sun, M.-X., Fu, Q., and Liu, D.: Application of the decomposition-prediction-reconstruction framework to medium-and long-term runoff forecasting, Water Supply, 21, 696–709, 2021.

Kress, A., Saurer, M., Siegwolf, R. T., Frank, D. C., Esper, J., and Bugmann, H.: A 350 year drought reconstruction from Alpine tree ring stable isotopes, Global Biogeochemical Cycles, 24, 2010.

Kwak, J., Lee, J., Jung, J., and Kim, H. S.: Case Study: Reconstruction of Runoff Series of Hydrological Stations in the Nakdong River, Korea, Water, 12, 3461, 2020.

Laaha, G., Gauster, T., Tallaksen, L. M., Vidal, J.-P., Stahl, K., Prudhomme, C., Heudorfer, B., Vlnas, R., Ionita, M., Van Lanen, H. A., et al.: The European 2015 drought from a hydrological perspective, Hydrology and Earth System Sciences, 21, 3001, 2017.

Li, Y., Wei, J., Wang, D., Li, B., Huang, H., Xu, B., and Xu, Y.: A Medium and Long-Term Runoff Forecast Method Based on Massive Meteorological Data and Machine Learning Algorithms, Water, 13, 1308, 2021.

Luoto, T. P. and Nevalainen, L.: Quantifying climate changes of the Common Era for Finland, Climate Dynamics, 49, 2557–2567, 2017.

Manabe, S.: Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the earth's surface, Monthly Weather Review, 97, 739–774, 1969.

Markonis, Y. and Koutsoyiannis, D.: Scale-dependence of persistence in precipitation records, Nature Climate Change, 6, 399–401, 2016.

Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J., and Lawrimore, J. H.: The global historical climatology network monthly temperature dataset, version 4, Journal of Climate, 31, 9835–9854, 2018.

Moravec, V., Markonis, Y., Rakovec, O., Kumar, R., and Hanel, M.: A 250-year European drought inventory derived from ensemble hydrologic modeling, Geophysical Research Letters, 46, 5909–5917, 2019.

Nicault, A., Alleaume, S., Brewer, S., Carrer, M., Nola, P., and Guiot, J.: Mediterranean drought fluctuation during the last 500 years based on tree-ring data, Climate dynamics, 31, 227–245, 2008.

Pauling, A., Luterbacher, J., Casty, C., and Wanner, H.: Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation, Climate dynamics, 26, 387–405, 2006.

Pfister, C., Brázdil, R., Glaser, R., Barriendos, M., Camuffo, D., Deutsch, M., Dobrovolný, P., Enzi, S., Guidoboni, E., Kotyza, O., et al.: Documentary evidence on climate in sixteenth-century Europe, Climatic change, 43, 55–110, 1999.

Pfister, C., Weingartner, R., and Luterbacher, J.: Hydrological winter droughts over the last 450 years in the Upper Rhine basin: a methodological approach, Hydrological Sciences Journal, 51, 966–985, 2006.

Senthil Kumar, A., Sudheer, K., Jain, S., and Agarwal, P.: Rainfall-runoff modelling using artificial neural networks: comparison of network types, Hydrological Processes: An International Journal, 19, 1277–1291, 2005.

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, Hydrology and Earth System Sciences, 23, 3247–3268, 2019.

Tejedor, E., de Luis, M., Cuadrat, J. M., Esper, J., and Saz, M. Á.: Tree-ring-based drought reconstruction in the Iberian Range (east of Spain) since 1694, International journal of biometeorology, 60, 361–372, 2016.

Tshimanga, R., Hughes, D., and Kapangaziwiri, E.: Initial calibration of a semi-distributed rainfall runoff model for the Congo River basin, Physics and Chemistry of the Earth, Parts A/B/C, 36, 761–774, 2011.

Vansteenberge, S., Verheyden, S., Cheng, H., Edwards, R. L., Keppens, E., and Claeys, P.: Paleoclimate in continental northwestern Europe during the Eemian and early Weichselian (125-97 ka): insights from a Belgian speleothem., Climate of the Past, 12, 2016.

Wetter, O., Pfister, C., Weingartner, R., Luterbacher, J., Reist, T., and Trösch, J.: The largest floods in the High Rhine basin since 1268 assessed from documentary and instrumental evidence, Hydrological Sciences Journal, 56, 733–758, 2011.

Zuo, G., Luo, J., Wang, N., Lian, Y., and He, X.: Two-stage variational mode decomposition and support vector regression for streamflow forecasting, Hydrology and Earth System Sciences, 24, 5491–5518, 2020.

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.