Czech University of Life Sciences Prague
Faculty of Environmental Sciences

Department of Water Resources and Environmental Modelling
Czech University of Life Sciences
Kamycka 129, 165 21 Prague 6, Czech Republic
Phone: +420 224 382 147, Fax +420 234 381 854
e-mail: michalkova@fzp.czu.cz, www.fzp.czu.cz

Dated: 4/15/2022

Christof Lorenz
Editor-In-Chief
Earth System Science Data
Karlsruhe, Germany

Dear Dr. Christof Lorenz,

We'd like to express our gratitude to you and the Reviewers for your thoughtful comments on our article, "A 500-year annual runoff reconstruction for 14 chosen European catchments. "

Those comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches. We carefully reviewed the comments and made changes that we hope will be accepted.

We have carefully addressed Reviewer #1 comments, particularly those regarding the proposed acronyms, which have been revised throughout the manuscript. All heatmaps were newly generated, taking into account Reviewer #1 colour choice selection. Additionally, the text was revised and polished as it is required.

The very useful and remarkable comments from Reviwer #3 is also higly appreciated. We have added the trend anlysis in the discussion part of the manuscript. we analysed the trends in the decadal runoff anomalies calculated from the reconstruction over several time periods. The reconstructed annual runoff for 1500-2000 for each catchment was first aggregated to 10-year time scale and divided by mean annual runoff. Our finding shows that there is no systematic trend throughout the whole 1500-2000 period. For a number of catchments there is a clear period of sustained above (Orsova-Danube and Dresden-Elbe) or below (Blois and Montjean Loire) average runoff during ca 1600-1800, while for the rest the persistence is clearly weaker although the low runoff signal is still visible (BaselRheinhalle, Baselschifflaende and Koln Rhine). Additionally, the linear trends calculated over 1500-2000 period are significantly negative for 13 (out of 14) catchments but there are also periods when most of the catchments show increasing significant trend, for instance the trend is significantly positive for 12 catchments for the 1800-2000 period. Looking at the most recent 1950-2000 period, the trend is negative for seven catchments. Please note, that since the most recent period is not included in the developed reconstruction, any possible climate change impacts are difficult to interpret. Another suggestion from Reviewer #3 to determine the similarities between the developed reconstruction and precipitation/temperature reconstruction at high elevations and domain boundaries.

The answer to such a comment is that the low skill for some catchments cannot be attributed solely to bias in reconstructed precipitation and temperature (as described in Sect. 4.1), but rather to low station coverage in some (especially northern) parts of Europe, resulting in biassed basin-average precipitation and temperature estimates.

The following are the main corrections in the paper and point by point responses to the reviewer comments:

Yours Sincerely,
On behalf of all coauthors,
Sadaf Nasreen

Reply to Reviewers

Dear Dr. Ghiggi, thank you very much for your constructive and valuable comments. We revised the manuscript carefully and hope that it now satisfies the proposed requirements for successful 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.

Minor Comments

1. In Figure 1 ensure that the color of the runoff station markers is different from the land mask color. In the first submission, it was green and was displaying nicely. You might also slightly reduce the marker size.

   Author's Response: The figure was revised and the color of runoff variable was changed into green, and the marker size was reduced.

2. Section 3.1 "Data preprocessing" still require some rearrangement.

   Author's Response: Several changes were made to the 'Data preprocessing' section in order to improve readability and coherency. There are a few phrases that were omitted:
   ➔ "we created two datasets "
   ➔ "models (the GR1A hydrologic model, the BRNN and LSTM data-driven models) that were used for 1500–2000 runoff simulation."
   And datasets were replaced with databases (catchment and observational databases). Additionally, the specified model names (GR1A, LSTM and BRNN) were eliminated from the section.

3. The workflow schema of Figure 2 requires a bit of polishing. Please address the following point:
   - What the "Testing = 25%" of calibration data" refers to? It's never mentioned in the manuscript.
   - You specify "check quality" for "Catchment dataset" and "Forcing validation dataset" but you not describe it the procedure in the paper. Either describe or remove it...
   - You specify QQplot but they are not presented in the manuscript.
   - Please use relBIAS instead of PBIAS for consistency across the manuscript.
   - Start with a Capital letter each bullet

   Author's Response: The 'check quality' was eliminated, and testing 25% percent is described in 'Method' section. QQplot was replaced by a scatter plot. relBIAS was replaced by PBIAS. In addition, the first letter of each bullet was capitalized.
   Line 174-175 ➔ The testing set was for each learning exercise extracted from the calibration data (1900-2000) as a random fraction (25%).

4. Ensure acronymic consistency across the text, figure and tables. For example, use [P,T, Lag] instead of the also appearing "P+T+Lag", (P+T+Lag), (P, T, Lag) and (P, T, Lag P, Lag T) (in Figure 2).

Author's Response: Thank you for the remarkable suggestions. All figures and tables captions were changed to include the proposed acronyms.

5. In Section 3.5 you mention that "annual drought duration and severity were then calculated" but you never present the results. Consider removing the sentence or present such results. Also, please clarify in the section how you computed the annual runoff/streamflow deficit. Please clarify the threshold quantile you selected also in the caption of Table 4.

   Author's Response: The sentence "annual drought duration and severity were then calculated" was deleted. The section was introduced with the definition of runoff/streamflow deficit which is as follows,
   The cumulative difference between runoff and the threshold was determined for each identified drought year, called as runoff/streamflow deficit. The threshold of 5% was chosen to represent extreme events during the last 500 years, and it was revised in Table 4.

6. Figure 3 and 4 can be further improved
   - It looks really strange to me to see the minimum and maximum values of the stations skill to appear at the extrema of the colorbar. This is not common practice!
   - You might want to consider centering the colormap for relBIAS around 0.
   - You should enlarge the extent/bounding box of each plot to not mask some stations (i.e. in Sicily).
   - Instead of repeating "Temperature"/" Precipitation" above each colorbar, put it as a title over the figure

   Author's Response: The colorbar extrema in Figures 3 and 4 have been corrected. The colorbars on the precipitation and temperature maps are evenly spaced. When dealing with the relBIAS comment, the centre is already zero, and the scale smaller than zero has a maximum number of divisions. As a result, we chose not to have the colorbar uniformly spaced, but rather to use prime number spacing (7,5,3,2,1). As a result, the range of relBIAS values is -1.6 to 0.2. Furthermore, each plot enclosing box has been expanded to include a station in Sicily. Furthermore, the temperature and precipitation text above each colorbar has been eliminated, and is now only mentioned in the caption text.

7. Figure 5 would be easier to interpret with a diverging colorbar centered at the (NSE?) value that you consider satisfactory (i.e. blue-yellow-red colormap, with colors tending to red when below the satisfactory threshold).

   Author's Response: The figure was modified and referred to as Table 3. Following the Ghiggi suggestion, the colormap was changed accordingly.

8. In the legend of Figure 6, it appears the term "gridded" that should be replaced by P, T to avoid confusion related to the nature of the model inputs.

   Author's Response: To maintain consistency across the paper, the legend for Figure 5 (formerly Figure 6) was substituted with P, T.

9. In Figure 7, you should consider reducing slightly the size of the point markers. Also consider changing the color of the GRDC observations (i.e. in black) to highlight that the CDF is available only in the 1900-2000 period for Basel Rheinhalle-Rhine catchment.

   Author's Response: According to Ghiggi recommendation, Figure 7 was adjusted. The size of the point markers was decreased, and the colour of the GRDC observations was changed to black.

10. In the text you refer as Figure 9 providing QQ-plots, while it contains scatterplots of modelled vs observed runoff. Please correct it.

    Author's Response: The caption text of Figure 8 (formerly referred to as Figure 9) was updated.

11. In Sect 4.4 you refer to Figure 9 instead of Figure 10

    Author's Response: The figure reference was corrected.

12. In both Figure 9 and 10, please set the aspect ratio of each plot to 1 and ensure to set the same axis limits for the x and y axis. It helps in see under/overestimation patterns in the scatterplots! Please also consider reducing the marker size of the plot for an improved visualization.

    Author's Response: Both figures were updated with same axis limits and reduced the marker size. Since changing the aspect ratio to one shrinks the figure dramatically, we've decided to leave it at its original size.

13. Instead of discussing about minimum/maximum or low/high runoff deficit values, please use the terms small(est)/ large(st) runoff deficit.

    Author's Response: Terms were modified according to the proposed suggestions.

14. The definition of KGE in Appendix 1 is wrong.
    - (rSD 2) should be replaced with (rSD -1)
    - (beta 2) should be replaced with (beta 1).
    - Also note that (beta - 1) correspond to relBIAS

    Author's Response: Thanks for pointing out the mistake, The definition of KGE was updated.

L56 ➔ precipitation (P), temperature (T)

Author's Response: L56 ➔ The text was updated.

L61-62 ➔ The structure of the paper is as follows: the considered hydroclimatic reconstructions, drought indicator and observed data are introduced in Sect. 2.

Author's Response: L60-61 ➔ Sentence was modified as:
"Section 2 introduces P and T hydroclimatic reconstructions, the scPDSI drought indicator as well as precipitation, temperature and runoff observations."

L62 ➔ 'hydrological and data-driven' maybe not worth specifying here

Author's Response: L61 ➔ It was removed.

L63 ➔ "precipitation and temperature reconstructions"

Author's Response: L62 ➔ The text was updated as: "P and T reconstructions."

L67 ➔ Herein, we used precipitation (Pauling et al., 2006) and temperature (Luterbacher et al., 2004) reconstructions

Author's Response: L66-67 ➔ Sentence was modified as per suggestion:
"This section present the data used in this study. To force the models, we investigate the use of precipitation (Pauling et al., 2006) and temperature (Luterbacher et al., 2004)."

L68-70 ➔ "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." Note that Menne et al., 2012 refers to GHCN-D. You used only GHCN-M right? So Menne et al., 2018

Author's Response: L68-70 ➔ The above sentence was modffied as:
"For validating the runoff reconstructions, we used runoff from GRDC Fekete et al. (1999). The accuracy of atmospheric forcing reconstruction used as model input was assessed using the observational data records of P and T from the Global Historical Climatology Network (GHCN;Menne et al., 2018)."

L88-89 ➔ Remove "(....was used to verify the accuracy of the precipitation and temperature reconstructions."

Author's Response: L88 ➔ The text was deleted.

L91 ➔ Reformulate: V4 version were included in the preliminary analysis

Author's Response: L91 ➔ The following text was added in the manuscript.
"(Menne et al., 2012) were used to assess the reconstruction accuracy of the P and T fields as an input into the considered models"

L91 (Menne et al., 2012).

Author's Response: L92 ➔ The reference was deleted

L92 found ➔ selected

Author's Response: L92 ➔ The text was updated.

L116 ➔ Used hydrologic (Sect. 3.2) and data-driven models (Sect. 3.3) for runoff simulation are introduced in the second part.

Author's Response: L116-117 ➔ The text was modified as Ghiggi suggestion.
"The hydrologic and data-driven models used to generate the runoff reconstructions are presented in Sect. 3.2 and 3.3 respectively."

Figure 2 ➔ Calculate P,T catchment average
—Associate GRDC runoff data
—Quality checks? Not described in the paper ... describe or remove ...
—Same for Forcing validation dataset
—Dataset is more raw data
—Database is more something processed one —caption: A schematic overview of the study work-flow.

Author's Response: The 'Quality checks' was eliminated, and the text was modified according to the proposed suggestions. In addition, the figure 2 caption was updated.

L120 ➔ Not scientific writing. "We prepared two datasets".
Describe instead the datasets/database you generated
You could provide them some names to facilitate their call across the text (i.e. CatchDB, ObsDB)

Author's Response: We appreciate Ghiggi comment regarding introducing the call names.
As the datasets are only mentioned once in the 'Data pre-processing' section, we just changed the dataset into the database to

keep clarity and readability. As a result, the text was altered as follows:
L121 ➔ Two databases were considered to analyse and develop the annual runoff reconstruction.
L124 ➔ Second database was further divided into two parts....

L124-125 ➔ Remove "(The GR1A hydrologic model, the BRNN and LSTM data-driven models)"

Author's Response: L125 The text was deleted.

L125 ➔ Scientific writing !
Several input variables were considered for inclusion in ...
You need to clean out a bit the sentences here ...
Lot of sentence are already introduced in the previous sections ...

Author's Response: L123-125 The paragraph was modified as,
"The second database was created as the basis for runoff reconstruction containing the observed runoff data for 21 selected catchments (Table 2) and the corresponding input variables of the models used to generate the multi-century runoff reconstructions. Several input variables were considered for inclusion in models, such as reconstructed precipitation and temperature and Old World Drought Atlas scPDSI."

L144 ➔ Remove "Data-driven methods"

Author's Response: L144 The text was deleted

L151-155 ➔ Reformulate: reconstructed precipitation and temperature fields is referred to as [P,T]

P,T, PDSI

P and T forcing
Ensure to always use such acronym across the manuscript !
Maybe for consistency with tables and figures you could use [P,T,Lag]
Do not repeat this. Say instead: and therefore were not included in presented analysis.

Author's Response: L152-156 The suggested acronyms were used in the whole manuscript and the following text was modified as:
"Specifically, the network using only reconstructed precipitation and temperature fields is referred to as $[\mathrm{P},\mathrm{T}]$, the network with reconstructed forcing and OWDA scPDSI is termed as $[\mathrm{P},\mathrm{T},\mathrm{PDSI}]$; and finally the network which includes 1-year lagged P and T forcing in addition to actual P and T is referred to as $[\mathrm{P},\mathrm{T},\mathrm{Lag}]$. We also considered and explored lag times longer than 1 year. However the correlation between precipitation and runoff drops significantly at lag times longer than 1 year, and therefore were not included in presented analysis."

L158-159 ➔ Reformulate: In this structure, LSTM allows to learn a long-term dataset and controls the overfitting problem (Chen et al., 2020).

Author's Response: L159 The following text was updated in the manuscript. LSTM is known for efficient simulation of time series with long-term memory (Van Houdt et al., 2020).

L160 "a given" ➔ the

Author's Response: L160 The text was updated.

L167-169 implement the initial values of the ➔ initialize using
Reformulate: Initial weights are set up based on a prior distribution function during model training. By applying Bayesian formulation, weight parameters keep updating prior probability distribution to the posterior probability distribution.

Author's Response: L165-171 The text was modified to simplify the description as:
BRNNs are based on the recurrent neural networks, which are often used to model time-series data (Wang et al., 2007), and extend them with Bayesian regularization (Okut, 2016) to account for uncertainty related to network parameters and input data (Zhang et al., 2011).

L184 hydrological droughts ➔ annual hydrological droughts

Author's Response: L186 The text was changed.

L188-189 ➔ Annual drought duration and severity (the cumulative difference of runoff and the threshold) were then calculated for each identified drought year. — This results are not presented right? This should be removed... or eventually added as perspective in the conclusion ... the presented dataset can be used to investigate ...

Author's Response: The text was modified as:
L189-190 ➔ "After that, the difference between runoff and the threshold was determined for each identified drought year, called as runoff deficit."
Furthermore, we added the following sentence in the conclusion section.
L377-378 ➔ "the presented dataset can be used to investigate (...)"

L214 goodness of fit (GOF) ➔ Introduce the acronym

Author's Response: L215 The term "goodness of fit (GOF)" was defined.

L224 Reformulate: gridded reconstruction of P and T

Author's Response: L225 The text was changed as:
"(...)was driven by catchment average P and T and calibrated using observed annual runoff."

L228 Reformulate: These (relatively poorer catchment skills in northern Europe)

Author's Response: L229-234 The text was changed as:
The catchments with relatively poor skills are located in northern Europe, which is in line with the previous findings by Seiller et al. (2012), who noted that the lumped hydrological models often exhibit larger uncertainties and fail to capture the extreme catchment values (both high and low) in those regions. The low skill for some of the catchments cannot be easily attributed only to bias in reconstructed precipitation and temperature (described in Sect. 4.1) but rather to low station and proxy coverage at some (especially northern) parts of Europe, leading to biased basin-average precipitation and temperature estimate.

L239 0.57 and 0.59 for validation ➔ maybe just report validation to simplify the reading?

Author's Response: L244 The text was modified as:
"(...)(NSE 0.76 for calibration and 0.57/0.59 for validation, for BRNN/LSTM respectively)"

Caption Figure 5 satisfactory validation ➔ Maybe directly say: with NSE < 0.5 over the validation period !

Author's Response: The caption of Table 3 (previously named as Figure 5) was changed as:
"with NSE > 0.5 over the validation period".

L263 validation NSE of 0.5

Author's Response: L268-269 The text was modified as: As a threshold, we considered validation NSE greater than 0.5 for at least one model.

L270 combination of reconstructed forcing with lagged values results ➔ models employing also reconstructed forcing with 1-year time lags results (...)

Author's Response: L278 The following text was modified:
"The combination of reconstructed forcing with 1-year time lag results."

Figure 7 ➔ Put in black ...
➔ There are not enough GRDC data in 1800-1900 to build a meaningful CDF right.
➔ Reduce the size of the markers to increase plot clarity

Author's Response: According to Dr. Ghiggi recommendation, the colour of the observed GRDC data in Figure 6 (formerly referred to as Figure 7) was changed to black, and the legend acronyms [P,T], [P,T, Lag], and [P,T, PDSI] were modified. The marker sizes were also lowered.

L283 the models ➔ the models in the period 1500-1800 appears to be 1669 while in the past century (1900-2000) to be 1921.

Author's Response: L293-294 The following text was updated:
"The models in the period 1500-1800 appears to be 1669 while in the past century (1900-2000) to be 1921. "

L287 I think it would be useful to note out that in the period 1500-1800 the CDF has very much lower/higher runoff values for BRNN and GR1A. LSTM seems to extrapolate less...
A comment of forcing uncertainty might be worth ..
GR1A simulate a minima of 250 mm/year of Rhine in Basel, while the the observed minima in the past century is above 500 mm/year „,

Author's Response: L289-291 The following text was modified as:
Our finding shows that GR1A simulates a Rhine minima of 279 mm/year in Basel, whereas the observed minima in the past century is greater than 532.6 mm/year, inferring that CDF has significantly lower/higher runoff values between 1500 and 1800 for BRNN and GR1A, whereas LSTM appears to extrapolate less.

L289-290 ➔ Fig. 9 does not show the QQplot But predicted vs observed runoff !! QQ plots have 0-1 (units on both axis)

Author's Response: L294-295 Thank you for pointing out the error, the text 'scatter plot' was updated.

L295 ➔ for the comparability

Author's Response: L299 The following text was updated:
"to enable comparison (Supplementary Figs. S4 and S5)."

L296-297 ➔ the correlation

Author's Response: L2307 The text was updated as: the correlation (reproduction of interannual dynamics).

L302 I have difficulty in observe this. You should put the axis of the plots in log scale !
And have the same x and y limits !!!!

Author's Response: There's no need for a log scale because it will transfer low flow values. While, the limits of axes were made equal.

L308 In the next step ➔ Use another word ....

Author's Response: L319 The following word was updated: "Furthermore"

L310 maximum/minimum deficit ➔ I am not sure I understand what you refer to.
The minimum runoff value and the q0.05 value?

Author's Response: L321 The text was updated as:
"the large deficit values for catchments (below 5th percentile)"

L312 ...also more severe in terms of hydrologic shifting. Please reformulate?

Author's Response: L323-324 The following text was updated
"more severe in terms of changing hydrologic conditions."

L320 "maximum" ➔ maybe better to use the word "largest" deficit

Author's Response: L331 The text was changed.

L321 severe ➔ extreme

Author's Response: L326 The text was changed.

L322 Alternatively ➔ In the Koln-Rhine catchment , 26 remarkable ...
Why specific to a catchment in the middle of a general analysis

Author's Response: L333 There was discussion of two significant Rhine river catchments in the manuscript, Basel Rheinhalle (1669) and köln (1686), both of which had the highest runoff deficits in the 17th century (see Table 3). The text was updated according to the suggestion.

L323 "In addition" ➔ The 1616 is considered the ...

Author's Response: L334 The text was changed.

Figure 10 ➔ Same axis limits / ratio !!!

Author's Response: The axis was modified in response to Dr. Ghiggi suggestion.

L334 high ➔ large

Author's Response: L345 The text was updated.

L336 including London ➔ why specific London ? Is southern England no? Or is where the analysis of Cook et al., 2015 has focused on?

Author's Response: L334 The text was modified and the citation (Bonacina, 1923) was added. The OWDA map demonstrates the 1921 severe drought that hits southern England and central Europe, also detailed in (Bonacina, 1923).

L337 I would rewrite as:
The low river level of the Rhine, Thames and Loire river has been also documented by .... respectively
photographs from the De Telegraf documented all these rivers?
van der Schrier documented the Loire only?

Author's Response: L350-352 Also reported in newspapers, The Rhine River (Switzerland), Molesey Weir, on the Thames River (United Kingdom), and Loire River (France) all have low river flows in 1921 (van der Schrier et al., 2021).

L339 "The precipitation totals were recorded as the lowest since 1774, and the year was also ranked top (in terms of deficit rainfall) in the Great Alpine region (Haslinger and Blöschl, 2017), where the rainfall deficit began in winter 1920/21 and lasted until autumn 1921."
This should be moved above after Cook et al., 2015 above when discussing of rainfall ...

Author's Response: L348-350 The text was shifted.

L341-342 ...with some of our study catchment ➔ in agreement with some of our catchment reconstructions signaling the 1976 as a yearly drought in the ....

Author's Response: L353-354 The text was modified accordingly.

L346 or other references ➔ and previous works

Author's Response: L357 The text was modified accordingly.

L348 This might be the case as ➔ This sentence should be removed in my opinion. It appears like a surprise / just luck that your reconstruction agree with previous work ... :)

Author's Response: L358-359 The text was updated as, "Because the tree-ring proxies involved in the developed reconstruction were the same, which could reveal the true nature of hydroclimatic shifts."

Author’s Response: L359 The text was deleted.

L348 developed reconstruction ➔ were used to derive the P+T reconstruction used to force our models ?

Author’s Response: L358 We used P, T, PDSI, and Lag-forced data to develop a runoff reconstruction

L349 Still ➔ Remove
It’s important to note that the presented runoff reconstructions might have missed notably documented dry events.
Maybe I would add a word to the fact that analyzing year values, extreme dry summers could be compensated by very wet spring/autumn periods ... which mask the sub-yearly dry period. It’s a problem of resolution, not methodology ...

Author’s Response: L362-366 The text was deleted and replaced with the following lines:
"It is important to note that the presented runoff reconstructions might have missed notably documented dry events." And the proposed reason was included in the concluding section at Lines 395-397. "Finally, since the runoff reconstruction is annual, the dry summers can be compensated by wet winters masking the sub-annual dry periods. However, this should be regarded as a resolution not methodology related problem."

L354 Following ➔ After careful validation

Author’s Response: L376 The text was replaced as: "After comprehensive validation of the simulated series, "

L354 we provided ➔ this work provides

Author’s Response: L377 The text was replaced.

L364 correlated ➔ correlates

Author’s Response: L387 The text was replaced.

Table 4 Minimum ➔ Largest

Author’s Response: Table 5 We decided to remove the "Minimum deficit (year)" column as it provided no meaningful information to the analysis. And, the text of table was changed and improved.

L366 series ➔ runoff

Author’s Response: L389 The text was replaced.

L374 are ➔ were

Author’s Response: L399 The text was replaced.

L374-375 (as was proven in validation).➔ Remove this is not necessary in the conclusion

Author’s Response: L399 The text was removed.

L394 deviation ➔ Deviation (SD)

Author’s Response: L419 The text was removed.

L396 with SD the standard deviation ➔ maybe not necessary ...

Author's Response: L421 The text was removed.

L401 n ➔ I would put n out of the sum maybe

Author's Response: L425 The term was modified.

L418 KGE 1? Why 2? Note that with your current formula (beta-1) ... is equivalent to relBIAS ! KGE is isually defined as mean/sd of o/p or p/o as in your case... ????

Author's Response: L442 The term was modified. relBIAS took the place of $\beta - 1$, and 1 replaced by 2.

L424 Tensorflow capital case. If you want to keep Tensorflow ... add the reference Abadi et al. 2016

Author's Response: L439 The following reference was added as: "Tensor flow (Abadi et al., 2016)."

L433-434 The checkpoint algorithm is also applied to test the model's accuracy level. Finally, the best output of the model is saved.

Author's Response: L456-457 The text was modified as : "Model checkpoints is used to save the model having minimum loss during the training ... "

L444-445 After getting the optimal model, the data is further evaluated the performance on testing data and predicted runoff values for the previous 500 years. ➔ This does not belong to the appendix

Author's Response: L467 The text was removed.

Table A1 Runoff ➔ runoff

Author's Response: Table A1 The text was updated.

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 version. In the following, we provide detailed replies to all comments and discuss changes to the main manuscript. We hope that we have properly addressed all the comments and suggestions.

Reviewer 2

This paper reconstructed annual runoff for 14 European basins for the period 1500 to 2000. Pre-existing temperature and precipitation reconstructions were first assessed against observations. Two data-driven models and a lumped hydrological model were then used to predict annual runoff.

Major comments

In the Introduction, the authors mention that there has been a decease in annual river flow at some river gauges in Europe over the last decade by up to 22%. This is said within the context of climate change. The authors now have time series of annual runoff for several basins in Europe stretching from 1500 to 2000. It would be interesting to see if there are any trends in annual runoff over that period, and if the effects of climate change are evident in the latter part of the timeseries.

Author's Response: Thank you for the remarkable comment. An additional trend analysis was performed and following text was added in the discussion part of the manuscript.
"Finally, we analysed the trends in the decadal runoff anomalies calculated from the reconstruction over several time periods. The reconstructed annual runoff for 1500–2000 for each catchment was first aggregated to 10-year time scale and divided by mean annual runoff. The resulting series are shown in Appendix Figure 1. It is clear that there is no systematic trend throughout the whole 1500–2000 period. For a number of catchments there is a clear period of sustained above (Orsova-Danube and Dresden-Elbe) or below (Blois and Montjean Loire) average runoff during ca 1600–1800, while for the rest the persistence is clearly weaker although the low runoff signal is still visible (BaselRheinhalle, Baselschifflaende and Koñ Rhine). The linear trends calculated over 1500–2000 period (Table A2) are significant negative for 13 (out of 14) catchments but there are also periods when most of the catchments show increasing significant trend, for instance the trend is significant positive for 12 catchments for the 1800–2000 period. Looking at the most recent 1950–2000 period, the trend is negative for seven catchments. Please note, that since the most recent period is not included in the reconstruction any possible climate change impacts are difficult to detect."

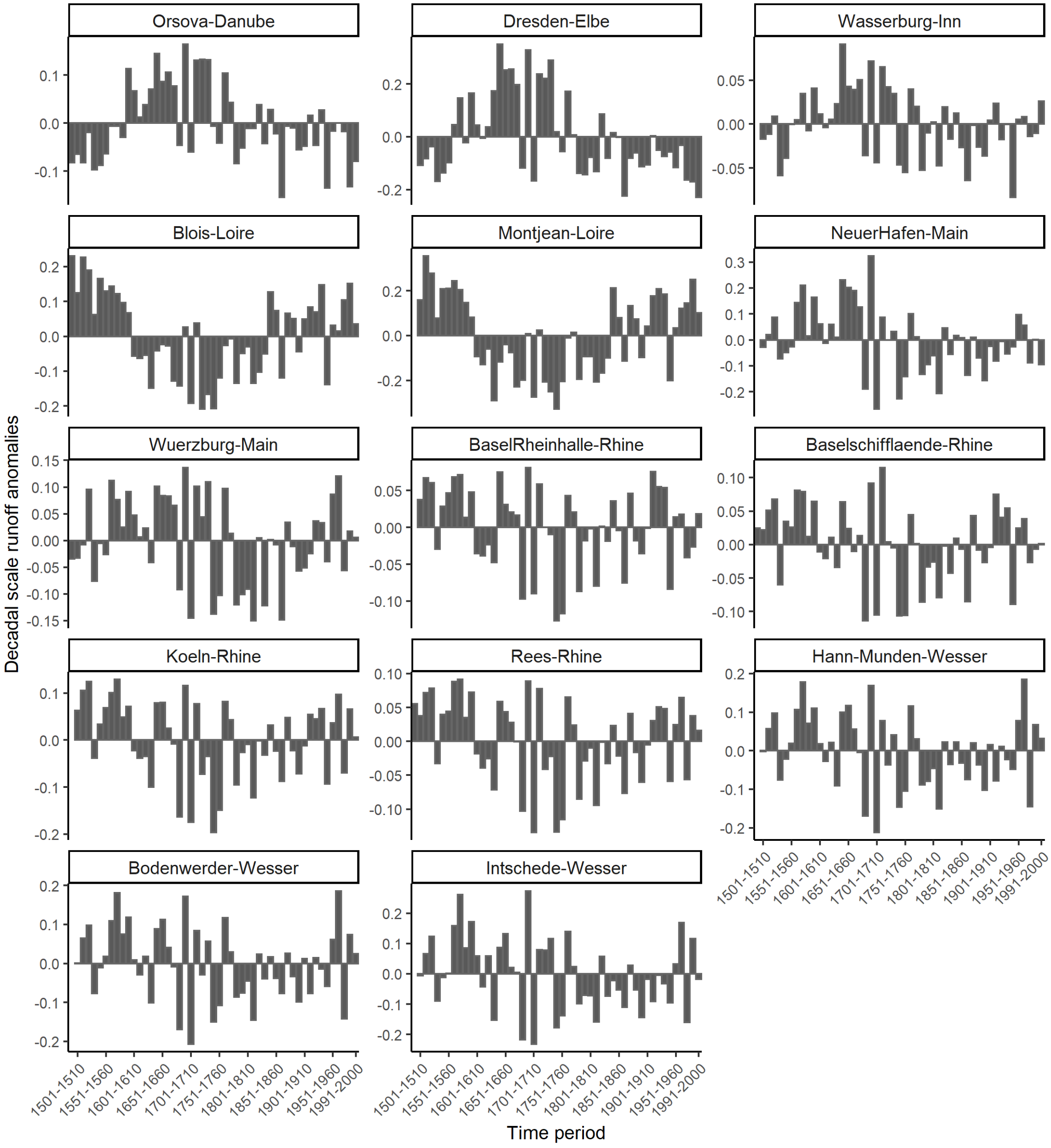

Figure 1. Decadal fluctuation of runoff anomalies in selected catchments over the past 500 years.

Table 1. The average, minimum and maximum slope of the runoff anomalies over different periods.

Time period	Number of catchments with (+/-) trends	Sign of trend	Average (min, max) slope
1950-2000	7	-	-2.573 (-4.374, -0.912)
1950-2000	2	+	2.031 (1.447, 2.615)
1900-2000	4	-	-0.875 (-1.711, -0.480)
1900-2000	4	+	0.664 (0.443, 0.852)
1800-2000	2	-	-0.419 (-0.540, -0.300)
1800-2000	12	+	0.450 (0.066, 1.307)
1700-2000	3	-	-0.425 (-0.813, -0.045)
1700-2000	10	+	0.388 (0.152, 1.288)
1600-2000	6	-	-0.317 (-0.859, -0.096)
1600-2000	5	+	0.305 (0.045, 0.821)
1500-2000	13	-	-0.136 (-0.347, -0.046)
1500-2000	0	+	

For the GR1A model, can you describe how X was optimized? Was the optimization manual or through some sort of algorithm?

Author's Response: For optimization of GR1A X parameter we used a gradient based minimization over a predefined interval as available in the AirGR package. This is now mentioned in the text.

Line 253: 'In the validation period, the differences between the models are more visible,'. Please discuss the possible reasons.

Author's Response: For the calibration period the models are optimized to match the observed runoff as much as possible. Therefore, provided the models are flexible enough and the data appropriate the simulated runoff should not differ dramatically between used models. In the validation period, given different input data the models are demonstrating their generalization skill which naturally differs from model to model. We added a remark on this.

In section 4.1 you mention that the reconstructions of temperature and precipitation differ from observations under certain circumstances: boundary of domain and high elevations. In section 4.3, you exclude the runoff simulations of seven basins due to the relatively poor performance of the models. I wonder if you can analyze if there are any similarities between the seven excluded basins. Since many basins have headwaters in mountainous regions, and since the reconstructions of rainfall and precipitation are relatively worse at high elevations, what does this mean for reconstructing annual runoff based on reconstructions of temperature and precipitation?

Author's Response: We added into Sect. 4.2, following explanation: The low skill for some of the catchments cannot be easily attributed only to bias in reconstructed precipitation and temperature (described in Sect. 4.1) but rather to low station coverage at some (especially northern) parts of Europe, leading to biased basin-average precipitation and temperature estimate.

Minor comments

Line 23: Water use by the power sector in the UK has dropped off considerably over the last 10 years as coal-powered plants are discontinued. I think maybe just revisit this statement for relevance over the last decade.

Author's Response: Thank you for this comment. Since the reference was not appropriate anyway, we dropped the sentence.

Figure 1: The color of the GRDC runoff gauges is too similar to the background map so they are hard to see. Throughout the paper, use a thousands separator and use the minus symbol rather than the hyphen when listing negative numbers. For example in Table 2.

Author's Response: The color of GRDC runoff gauges was changed to green and Table 2 symbols were updated.

Can Figures 3 and 4 be combined by using different shapes for the precipitation and rainfall symbols?

Author's Response: In the earlier stages of the manuscript preparation we have been considering this. However, the presented information was not clear enough (the symbols cannot be too small since their fill color gives the information on the error but then it leads to significant overplotting). Therefore we leave the presentation in two figures.

Line 242: ', the data-driven ? exhibited'

Author's Response: Text was updated as "data driven methods"

Figure 5 is actually a table. I would suggest formatting it as such. Also, for the 14 selected catchments with good GOFs in this table, can you group them together rather than using a bold outline.

Author's Response: In response to a reviewer suggestion, Figure 5 has been renamed Table, and the best ones have been grouped together in a rectangular block.

You mention a 'subjective' process in line 273. Can you somehow try and formalize the process of selecting the best model for each catchment?

Author's Response: We agree that more objective selection procedure would be beneficial. On the other hand, the performance of the best models was comparable and somewhat subjective choices (e.g. the considered metric, their number etc.) are unavoidable. Therefore we left this topic for future applications. We added a note in the manuscript: The best model for each catchment was finally selected from those models considering the remaining validation measures (relBIAS, rSD, KGE, RMSE and MAE) as well. Specifically, we picked the models with consistent good validation measures. This choice is partly subjective and more formal selection should be explored further. On the other hand, the candidate models were all performing comparably in most cases.

Figure 9 and 10: can you add the gradient of the line in each subplot as an indication of model bias?
Author's Response: The figures 8 and 9 were updated to include gradient lines.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265–283, 2016.

Bonacina, L.: The European drought of 1921, Nature, 112, 488–489, 1923.

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

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European seasonal and annual temperature variability, trends, and extremes since 1500, Science, 303, 1499–1503, 2004.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An overview of the global historical climatology network-daily database, Journal of Atmospheric and Oceanic Technology, 29, 897–910, 2012.

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.

Okut, H.: Bayesian regularized neural networks for small n big p data, Artificial neural networks-models and applications, pp. 21–23, 2016.

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.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrology and Earth System Sciences, 16, 1171–1189, 2012.

van der Schrier, G., Allan, R. P., Ossó, A., Sousa, P. M., Van de Vyver, H., Van Schaeybroeck, B., Coscarelli, R., Pasqua, A. A., Petrucci, O., Curley, M., et al.: The 1921 European drought: Impacts, reconstruction and drivers, Climate of the Past Discussions, pp. 1–33, 2021.

Van Houdt, G., Mosquera, C., and Nápoles, G.: A review on the long short-term memory model, Artificial Intelligence Review, 53, 5929–5955, 2020.

Wang, W., Gelder, P. H. V., and Vrijling, J.: Comparing Bayesian regularization and cross-validated early-stopping for streamflow forecasting with ANN models, IAHS Publications-Series of Proceedings and Reports, 311, 216–221, 2007.

Zhang, X., Liang, F., Yu, B., and Zong, Z.: Explicitly integrating parameter, input, and structure uncertainties into Bayesian Neural Networks for probabilistic hydrologic forecasting, Journal of Hydrology, 409, 696–709, 2011.