Dear Zhongbo Yu, editor of HESS,

Hereby the authors of the revised paper hess-2020-382 take the opportunity to thank you and anonymous Referees #1 and #2 for the useful comments and suggestions for improving our manuscript.

We addressed all the concerns of you and the reviewers in detail, and revised the manuscript accordingly. In addition, we discovered an issue of pixel-shifting between the atmospheric and land components in the post-processed output of the Terrestrial Systems Modeling Platform (TSMP), which we corrected in the revisions. We ran the proposed LSTM networks again with the corrected TSMP data and updated the results in the revised manuscript. In the new results, which were improved, there is only a single pixel where the network performance followed C3 (i.e., training R^2 score ≤ 0%, test R^2 score ≤ 0%). Thus, we removed the analyses related to C3 in the revised manuscript. The conclusions did not change based on the corrections. Some grammar, spelling, and punctuation mistakes were also corrected. Moreover, we improved the quality of most figures to guarantee colorblind safe.

We provide a marked-up manuscript version to indicate all the modifications that we have made.

Here we respond to the comments of you and anonymous Referees #1 and #2 in detail, which is followed by information regarding the additional modifications as aforementioned. Your and the reviewers' comments are marked in ***black italic,*** and our responses are given and marked in regular font.

Response to your comment:

Authors are required to address all reviewers' comments, particularly to further elaborate the new contribution of the study, to include additional in depth analysis on the results, and to have a thorough editing for conciseness and smooth transition among sections.

Thank you for the comment. We have addressed all the concerns of the reviewers in detail and made corresponding changes to the revised manuscript. For details, please consider our responses to the reviewers' comments below.

Responses to anonymous Referee #1's comments:

Authors used the LSTM network to forecast water table depth in Europe and analyzed the effects of local elements, which is very interesting. Comments are shown as follows:

Thank you for the positive feedback.

1. Line 63-64: The most obvious advantage of ANNs is not using learnable parameters. Some basic machine learning models, such as MLP, can also adapt weights and bias.

Yes, this is true. The MLP (Multilayer perceptron) is a type of feedforward network, belonging to ANNs, so you probably meant models such as linear regression.

The aim of Line 63-64 is not to express that using learnable parameters is the major advantage of ANNs, but to give the reader a brief overview of how ANNs work (i.e., "adapting learnable parameters on the links between neurons") and what they can achieve (i.e., "give an appropriate input-output mapping based on observed data even for complex nonlinear relationships").

In Line 62-64 of the marked-up manuscript version, we rephrased the sentence to "ANNs adapt learnable parameters (i.e., weights and biases) in the links between neurons to achieve an appropriate input-output mapping based on observed data, also for complex nonlinear relationships." for clarity.

2. It is suggested that the authors should describe the relationship between ANN and RNN before introducing the details of RNN. The same problem occurs in the introduction of LSTM. The limitation of RNN is not introduced first.

Thanks for pointing out the structural inconsistency.

Feedforward networks (FFNNs) are commonly used ANNs in previous studies for groundwater level modeling. RNNs are a type of ANNs designed for sequential data analysis, outperforming FFNNs in handling the relationship between sequences. LSTM networks belong to RNNs, capable to better exploit the long-term relationship between sequences than standard RNNs.

We agree and changed the structure of the related content as follows in Line 68-90 in the marked-up manuscript version: firstly, to introduce the popularity of the application of FFNNs in groundwater level modeling and their limitation in handling sequential data; then, to introduce RNNs and point out the drawback of standard RNNs; finally, to briefly introduce LSTM networks.

3. Line 71. Lots of research should be cited related to RNN rather than ANN here.

We cited many studies related to FFNNs and their variants in Line 71 to demonstrate the popularity of their application in groundwater level modeling, and then highlighted the advantage of using RNNs compared to feedforward networks. Since standard RNNs tend to fail to exploit long-term dependencies between groundwater and other hydrometeorological variables, the RNNs applied in previous studies for groundwater level modeling are mostly LSTM networks, the majority of which have been included as the cited papers in the reference section of our manuscript. Further, there is a limited number of studies on groundwater level modeling using RNNs until now.

4. Line 129. Why did the authors say they have the same architecture of hidden neurons as Gers et al. (2000) but without the detailed introduction of Gers et al. (2000) or the architecture?

Gers et al. (2000) is one of the pioneer papers on LSTM networks. We gave a short introduction about the structure of hidden neurons in Line 131-132 in the original manuscript, and illustrated all the components of a hidden neuron in Fig. 2. Moreover, we provided the reader with the pseudocode of the LSTM network displayed in Fig. 2 to help the reader understand how data is transferred in a LSTM hidden neuron. Therefore, we think the current information regarding the structure of a LSTM hidden neuron should be sufficient. Detailed information such as the functions of each component in the hidden neuron might be too technical for the reader, and the one who is interested in the details is referred to Gers et al. (2000).

In the marked-up manuscript version, we made the following modifications to help find the relevant information easier:

- In Line 141-142: we added "which is shown in Fig. 2" after the sentence "In this study, we employed LSTM networks having the same architecture of hidden neurons as Gers et al. (2000)" (i.e., Line 129 in the original version) to point out the figure that shows the architecture of the LSTM hidden neuron;

- We moved the position of Fig. 2 after the first paragraph of Sect. 2.2;
- In Line 162: we added "to detail how data is transferred in the given LSTM network" to highlight the function of the pseudocode.

5. In equation (3), some representation should be shown as subscript

We modified the representation in Eq. (3) and the related representation in Sect. 2.3.

6. Line 194. TSMP should be written in full name when it shows for the first time.

We changed TSMP to the full name when it was introduced for the first time (i.e., Line 14-15 and Line 96-97 in the marked-up manuscript version).

7. The website of data access should be shown in this paper.

Could you please specify which data you are referring to? If you meant the TSMP-G2A data set, we have provided the corresponding DOI in "Code and data availability", which points the reader to the data set. In addition, we also provided an accessible link of the TSMP-G2A pr_a, wtd_a and the LSTM wtd_a data in the revised manuscript (see "Code and data availability", Line 495-496 in the marked-up manuscript version).

8. Figure 5 shows the result of the training dataset. The good performance of the training dataset cannot prove that the model is good. It is suggested that the test dataset should be used to show the result of the model.

We agree that the model cannot be proved to be good based on its training performance. That is why we reported the model performance for both the training dataset and the test dataset. In Fig. 5, we provide the comparisons between the water table depth anomaly maps not only in 2003 (i.e., in the training period) but also in 2015 (i.e., in the test period). We extended the captions of Figs. 5, B1 and B2 accordingly in the revised manuscript.

9. Line 311. It is confusing that the authors say Figure 6 is based on the categories in Table 3, but the categories seem to base on Table 1 in Figure 6.

Table 1 mainly describes the climatologic information of our study regions (i.e., the PRUDENCE regions), and Table 3 presents the intervals of local factors that used for categorification. For example, we categorized the selected pixels in each PRUDENCE region into two categories of yearly averaged snow water equivalent based on the intervals in Table 3, and calculated the average and standard deviation of R^2 scores and RMSEs for each category in each region to obtain Fig.6d.

In the marked-up manuscript version, we modified the following sections for clarity:

- In Line 19: we added "intervals of";
- In Line 257-258: we added "different intervals of" to complement the criterion of categorification;
- In Line 305: we changed "…categorized the pixels into various groups based on yearly averaged…" to "…categorized the pixels into groups based on various intervals of yearly averaged …";

- We changed the caption of Table 3 to "Intervals of yearly averaged *wtd*, *ET*, θ, S_w, and S_t and dominant *PFT* for categorization";
- In Line 344-345: we modified "… scores and RMSEs for the different categories (Table 3) of yearly averaged…" to "… scores and RMSEs for the categories based on different intervals (Table 3) of yearly averaged …".

10. When C3 is shown, it only means that the training process of the model has some problems and needs to be modified further.

Yes, you may be right. After correcting the issue of pixel-shifting in the TSMP-G2A pr_a and wtd_a data, there was only a single pixel where the network performance followed C3 in the new results. Therefore, we removed the analyses of C3 in the revised paper and made associated modifications to the related tables and figures (i.e., Table 5 and C1, Figs. 8, 9 and C1).

11. Line 365. What is the standard of the selection of Pixels 1-3?

We randomly selected pixels that satisfied the representative climatologic characteristics of C1, C2 summarized in Fig.8 (see Line 415-425 in the marked-up manuscript version), and conducted cross wavelet transform at these pixels to verify our hypothesis that there was a time-varying pattern between pr_a and wtd_a at several pixels over the European continent. Here, we only showed Pixels 1-2 and Pixels 3-4 (in Appendix C) as representative examples. For clarity, we added this point "which were randomly selected based on the hydrometeorological characteristics of C1 and C2 summarized in Fig. 8" in Line 434.

Responses to anonymous Referee #2's comments:

In their paper, Ma et al proposed to investigate the link between water table depth anomalies (wtda) and precipitation anomalies (pra) using Long Short-Term Memory networks (LSTM). To test the proposed approach, they use a dataset generated with the Terrestrial System Modelling Plateform (TSMP) over Europe and compare the results provides by both approaches (TSMP and LSTM). The effect of several factors on the performance of the approach are also investigated. Cross-wavelet transform are also used to analyze the response of the network regarding time frequency.

Thank you for the cogent summary of our study and the thorough review.

Overall, the paper is well written and organized. The approach proposed is interesting and its novelty is clearly explained in the introduction as this type of networks used is not commonly used to examine the response of groundwater. The study has a specific focus on response to drought which is of importance for groundwater management.

Thank you for the positive feedback.

The results presented are promising but the presentation/discussion should be improved. In my opinion, the results are not discussed thoroughly especially when the performance of LSTM is not so good. My issues with the paper, and some additional minor comments and corrections, are detailed in the following.

In the discussion of the results, we explained the low LSTM performance through the weak physical link between pr_a and wtd_a and the time-varying pattern between them. We further improved the presentation and discussion by incorporating your comments and suggestions in the revised version to address your concerns.

-The dataset generated with TSMP is the foundation of the proposed methodology as the evolution in time and space of all the variables used in the study are simulated ones. Although the reader is sent to relevant references to have further information, I think some key features need to be presented to make the paper self-consistent.

We agree and, therefore, included some key features of the TSMP in the revised manuscript (see Line 215-230 in the marked-up manuscript version).

I especially would have liked to know how the TSMP was calibrated (or not) against observed values to have an idea of how reasonable or relevant the simulated evolutions are.

The TSMP simulation results, especially anomalies, show good agreement with common reference observational datasets (i.e., E-OBS v19 and GRACE datasets), as presented in Furusho-Percot et al., (2019). We provided a sentence in the original manuscript (Line 200-204) to address your concern about the performance of TSMP, and the reader is referred to related references [e.g., Furusho-Percot et al., (2019)] for detailed information. For clarity, we rephrased the related section and added the Pearson correlation coefficient values between the TSMP-G2A and the observed data as additional information in the revised manuscript (see Line 223-230 in the marked-up manuscript version).

- I think that the figures with maps are very hard to interpret owing to the extension of the study area and the spatial resolution of the approach proposed. The authors state that the agreement is good visually (Line302) which is in my opinion not so evident and not enough.

The maps (e.g., Fig. 5) show the spatial extent and severity of groundwater drought over Europe in a specific month, which is indicated by different colors. This type of presentation is typical for the spatial analysis of a drought event [see e.g., Shukla and Wood (2008), Gumus and Algin (2017), and Van Loon et al. (2017)]. In the interpretation, we focused on the agreement of the spatial patterns with respect to severity and the spatial distribution of dry and wet events. The visual comparison between the two maps in Fig. 5b shows reduced agreement on severity but still good agreement on the spatial distribution of dry and wet events. For clarity, we further extended the analysis and the discussion of Fig. 5 and added more details about the figure (see Line 318-338 in the marked-up manuscript version).

I would have liked (if possible) some indicators to be presented – maybe for each PROVIDENCE regions – to have a quantitative diagnostic rather than a visual one.

The plots of R^2 values as a function of intervals of local factors such as yearly averaged *wtd* (Fig. 6) have provided a quantitative diagnosis of the network performance for each PRUDENCE region. In addition, the results presented in this study are not suitable for a quantitative diagnostic of the spatial agreement, because of the following reasons: 1) the networks were trained at the individual pixel level, and at different pixels, the structures of the optimal networks for the final result calculation are mostly different; 2) we constructed the LSTM networks only at selected pixels but not all the pixels over Europe to save the computational time, leading to discontinuous results at the continental scale.

- Overall, the performance of the LSTM approach is not discussed with enough details. Especially when the performances are poor. Line334–335 is an example where some more

details are needed. Table 4 demonstrates that the agreement is not good in some specific PRUDENCE regions (for instance MD or IB) and no specific explanations are provided. The same goes for the discussion of Figure 8.

Thank you for pointing this out. In the revised manuscript, we improved the discussion of our results by adding more details, especially for poor network performance. Please consider Sect. 3 (Results and discussion) for the changes we made.

- The conclusion is a bit misleading, as it may convey the message that the LSTM approach is relevant all over Europe when the results are very good only in specific conditions (as specified line 406-407). Some rephrasing may be needed here.

We agree and rephrased the relevant sentence to "Using the output of the LSTM networks, we successfully reproduced TSMP-G2A wtd_a maps over Europe for drought months in both the training and testing period (e.g., August 2003 and August 2015) in terms of the spatial distribution of dry and wet events" for clarity (see Line 474-476 in the marked-up manuscript version). In addition, we also made the corresponding change in the abstract (Line 18).

Specific comments:
- Line 84: Should be RNN and not ANN here

LSTM networks are a special type of RNNs, and RNNs belong to ANNs, so it is also correct to state that LSTM networks are ANNs. Further, "The consistency of the temporal pattern between input and target variables" is required by all ANNs for good performance. Therefore, we stated ANNs rather than RNNs in Line 84 in the original manuscript. We clarified the relationship among ANNs, RNNs and LSTM networks in the revised manuscript (see Line 59-90 in the marked-up manuscript version).

- Line 110: It is mentioned here that "Areas with surface water are not taken into account". I wonder if or to what extend this assumption could impact the results of the study.

This assumption does not affect the results in this study, because we constructed the proposed LSTM networks only at pixels without surface water (i.e., river, lakes and reservoirs).

- Figure 6 can be improved: the color legend that specifies the PRUDENCE regions should be bigger and placed elsewhere.

We improved Fig. 6 following your suggestions.

Reference:

Furusho-Percot, C., Goergen, K., Hartick, C., Kulkarni, K., Keune, J. and Kollet, S.: Pan-European groundwater to atmosphere terrestrial systems climatology from a physically consistent simulation, Sci. data, 6(1), 320, 515doi:10.1038/s41597-019-0328-7, 2019.

Gumus, V. and Algin, H. M.: Meteorological and hydrological drought analysis of the Seyhan−Ceyhan River Basins, Turkey, Meteorol. Appl., 24(1), 62–73, doi:10.1002/met.1605, 2017.

Shukla, S. and Wood, A. W.: Use of a standardized runoff index for characterizing hydrologic drought, Geophys. Res. Lett., 35(2), 1–7, doi:10.1029/2007GL032487, 2008.

Van Loon, A. F., Kumar, R. and Mishra, V.: Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time, Hydrol. Earth Syst. Sci., 21(4), 1947–1971, doi:10.5194/hess21-1947-2017, 2017.

Additional modifications (in the marked-up manuscript version):

1. We reran the proposed LSTM networks with corrected input and output data, and updated the results shown in Table 4-5 and C1 and Figs. 5-9, B1-B2 and C1 and some values in Table 1.

2. We fixed the calculation of soil moisture data from the TSMP-G2A data set, and corrected some values in Table 1 and C1.

3. We improved the quality of Figs. 1, 3, 5-9, B1-B2 and C1.

4. Line 4: we changed "Research center Jülich" to "Forschungszentrum Jülich" (i.e., the original German name of the institute).

5. Line 25: changed "*pr*" to "pr_a".

6. We changed the abbreviation of precipitation from "*pr*" to "*P*" to be consistent with the label in Fig. 1.

7. Line 111: removed ",".

8. Line 114: rephrased part of the sentence to "…set, before presenting a…".

9. Line 118: rephrased the sentence to "Note, areas with surface water are not taken into account in this study".

10. Line 119-120: changed "fluxes" to "flows" to ensure the correct unit (i.e., LT^{-1}).

11. We extended the caption of Fig. 1 by adding the description of variables shown in the figure. Fig.1 is totally different from the figure shown in [Maxwell, 2010], and thus we changed "modified from Maxwell, 2010" to "after Maxwell, 2010" to avoid the copyright issue.

12. We changed "$div(\boldsymbol{Q_g})$" to "$\boldsymbol{Q_g}$" in Eqs. (1)-(2) and the relevant representation in Sect. 2.1 to ensure the correct unit (i.e., LT^{-1}).

13. Line 132-133: we deleted the definition of "$div(\boldsymbol{Q_g})$" and added the definition of "$\boldsymbol{Q_g}$" (i.e., "$\boldsymbol{Q_g}$ is the lateral groundwater flow [LT^{-1}]").

14. Line 145-146: corrected the term to "exploding and vanishing gradient issues".

15. We extended the caption of Fig. 2 by adding more details of $c(*)$ and σ.

16. Line 197: added "the".

17. Line 237: changed “ranges” to “range”.

18. Line266: changed “*wtd*” and “*pr*” to “wtd_a” and “pr_a”, respectively.

19. Line 306: changed “result” to “results”.

20. We changed the description of the dashed lines in the caption of Fig. 4 to “blue dashed lines”.

21. Line 344: removed “,”.

22. Line 346-348: we added “For statistical significance, we only considered categories with ≥ 50 pixels. In addition, negative R^2 values at the pixel level were set to zero in the calculation of averages and standard deviations” as additional information of the calculation of averages and standard deviations of the test R^2 scores and RMSEs.

23. Line 354: added “(not shown here)” after “There was no significant influence of S_t and dominant *PFT* on the scores”.

24. We changed “studied” to “selected” in the caption of Table 4.

25. We corrected the caption of Fig. 8 to “Bar plots showing percentages of pixels where the network performance followed the combinations (a) C1; (b) C2 in different regions and intervals of yearly averaged *wtd*, *ET*, θ, S_w, from left to right, respectively. Black dashed lines indicate percentages equal to 30%”, since the figure shows bar plots instead of histograms.

26. Line 434-436: we corrected the sentence to “XWT extracted the common power of the frequency components in the pr_a and wtd_a time series derived from the TSMP-G2A data set (i.e., TSMP-G2A pr_a and wtd_a) at these pixels”, since the previous description is for wavelet coherence but not XWT.

27. Line 437: deleted “very”.

28. Line 450: we corrected the words to “pr_a and wtd_a”.

29. Line 473: corrected “2018” to “2016”.

30. Line 488-489: changed “pr_a” to “other hydrometeorological variables”.

31. Line 494: corrected “contract” to “contact”.

32. Line 593-595: added the reference “Gasper, F., Goergen, K., Shrestha, P., Sulis, M., Rihani, J., Geimer, M. and Kollet, S.: Implementation and scaling of the fully coupled Terrestrial Systems Modeling Platform (TerrSysMP v1.0) in a massively parallel supercomputing environment - A case study on JUQUEEN (IBM Blue Gene/Q), Geosci. Model Dev., 7(5), 2531–2543, doi:10.5194/gmd-7-2531-2014, 2014”.

Sincerely,
Yueling MA

On behalf of all the authors of the revised paper hess-2020-382