Dear Zhongbo Yu, editor of HESS,

We would like to thank you and anonymous Referees #1 and #3 for the constructive comments and suggestions on our manuscript.

We addressed all the concerns of you and the reviewers in detail, and revised the manuscript accordingly. A marked-up manuscript version is provided, showing all the modifications that we have made.

Here we respond to the comments of you and anonymous Referees #1 and #3 in detail, which is followed by information regarding the additional modifications we made in the manuscript. Your and the reviewers' comments are marked in ***black italic,*** and our responses are provided in regular font.

Response to your comment:

Reviewers provided positive comments on the paper. Authors are required to address all comments on the revision, particularly to further elaborate the need and new contribution of the study, to include additional in depth analysis.

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

Response to Referee #1's comment:

Thank you for the recommendation to accept our manuscript.

Responses to Referee #3's comments:

The article presents the development of an artificial neural network that aims at connecting water table depth anomalies to precipitations anomalies. The topic is of interest.

Thank you for the positive comment.

However, I have some doubts on the study.
The main one is that this neural network is constructed and assessed based on the simulation of a spatially distributed model. Therefore, there is no assessment on real observations.

The study presents a methodology of deriving a LSTM network model for groundwater table depth anomalies (wtd_a) from precipitation anomalies (pr_a). Indeed, we utilized anomalies derived from simulation results of a terrestrial model (i.e., the TSMP-G2A data set) since: (i) groundwater table depth (wtd) observations are sparse and hard to obtain over Europe; (ii) the TSMP-G2A data set has been proved to have good agreement with hydrometeorological and GRACE observations in different European regions (see Line 205-211 in the Version 2) [Furusho-Percot et al. (2019) and Hartick et al. (2021)]. Therefore, we argue that the TSMP-G2A data set is a good reference data set to establish the methodology.

We made the following modifications in the marked-up manuscript version:

- In Line 220-222, we added "Similar results were obtained by Hartick et al. (2021), who compared anomalies of total column water storage from the TSMP-G2A data set with the novel GRACE-REC data set and obtained R from 0.69 to 0.89 in the different

PRUDENCE regions" to support our argument. The related reference citation was added in the reference section (Line 587-589).

- In Line 464-468, we added "This study presents a methodology of deriving a LSTM network model for wtd_a from pr_a based on simulation results from a terrestrial model (i.e., the TSMP-G2A data set). As demonstrated in Furusho-Percot et al. (2019) and Hartick et al. (2021), the TSMP-G2A data set shows good agreement with hydrometeorological and GRACE observations in different European regions. Therefore, we argue that the TSMP-G2A data set is a good reference data set to establish the methodology" to clarify our statement.

Moreover, it is not clear which is the advantage of using the neural network compares to the spatially distributed model. I guess it is the computational time, but, it is not clearly stated.

Thank you for pointing out the missing information.

Compared with physical-based models (or spatially distributed models), the advantages of using the proposed LSTM networks are requiring: i) less computational time; ii) minor background knowledge. In this study, we used pr_a and wtd_a to construct the proposed networks, without any additional physical background knowledge that are required by traditional physically-based models, thus highly reducing the efforts of data collection and processing. Moreover, when the LSTM networks are available, only pr_a data are needed to estimate wtd_a, which are available from operational forecasts.

In Line 145-148 in the marked-up manuscript version, we added "In addition, compared with traditional physically-based models, the proposed LSTM networks require less computational time and background knowledge to perform the simulations. Moreover, when the proposed LSTM networks are available, we only need the pr_a data to estimate wtd_a, which are available from bias corrected operational forecasts and reanalysis data sets" to state the advantages of using the proposed LSTM networks compared with physically-based models in this study.

Additionally, only the water table depth anomaly is estimated. But what is the meaning of it?

wtd_a reflects anomalies in groundwater storages (stated in Line 43-45 in the Version 2). wtd_a is calculated with respect to the deviation of the *wtd* value from the climatological average for a specified time period normalized by the climatological standard deviation, which is a measure of groundwater drought.

For clarity, we added the aforementioned information in Line 46-48 in the marked-up manuscript version.

Can we expect similar situation in two places with a similar water table depth anomaly?

Yes, we calculated wtd_a values at the individual pixel level, and the anomalies, which are unitless normalized values, allow us to compare groundwater drought conditions in different locations.

For clarity, we added "at the individual pixel level over Europe" in Line 239 in the marked-up manuscript version.

A focus is made on drought, but, what is the general relationship between water table depth anomaly (wtda) and drought? ll these have to be explained and demonstrated

Thanks for pointing out the missing information.

As mentioned in our previous response, wtd_a is a measure of groundwater drought. Here, we define $wtd_a \geq 2$ corresponding to extreme drought, $1.5 \leq wtd_a < 2$ corresponding to severe drought, $1 \leq wtd_a < 1.5$ corresponding to moderate drought, $0 \leq wtd_a < 1$ corresponding to minor drought and $wtd_a < 0$ corresponding to no drought, following McKee et al. (1993).

We added this information in Line 249-251 in the marked-up manuscript version.

Special comments
Introduction: it is not clear how the water table depth anomaly is computed. Please explains and/or refers to equation 7… same for precipitation anomaly.

The value of wtd_a (or pr_a) is the deviation of the value of wtd (or precipitation) from the climatological average for a specified time period normalized by the climatological standard deviation.

We added this information in Line 46-48 and Line 55-56 in the marked-up manuscript version for wtd_a and pr_a, respectively.

Section 2.1: the scheme presented as a complete subsurface water balance is in fact a rough simplification. There is only one aquifer layer, no connection with the rivers, no groundwater abstraction…

Thanks for pointing out the misused terminology. It is true that the scheme has some simplifications, that is, considering only one unconfined aquifer, neglecting the connection between surface water and groundwater and removing the anthropogenic impacts. This study only focused on the connection between pr_a and wtd_a over Europe, and we wanted to show a generic scheme that gives an explicit relationship between the fluctuation of water storage in the unconfined aquifer and precipitation. That's why we made the above simplifications.

In the marked-up manuscript version, we made the following modifications:

- Deleted "complete" in Line 114 and Line 123, respectively;
- Added ", and the impact of anthropogenic activities such as groundwater abstraction is neglected in Line 115-116 to supplement the description of the scheme simplification.

Section 2.2 I am not a neural network expert and would have appreciated a justification of the best suitability of LSTMs. Do other neural networks have memory?

Yes, all types of recurrent neural networks (RNNs) can be considered to have memory. The most commonly used ones are standard RNNs and LSTM networks. LSTM networks are a special type of RNNs that overcome the exploding and vanishing gradient issues of standard RNNs. They have the ability to exploit long-term dependencies between sequences, which is expected in the response of wtd_a to pr_a.

In the marked-up manuscript version, we added ", similar to other RNNs" in Line 137 to state that RNNs except for LSTM networks also have memory. In addition, we added "The response of wtd_a to pr_a is expected to exhibit a long time lag, especially in case of deep aquifers, and thus LSTM networks are an appropriate type of networks to use here." in Line 144-145 to justify the suitability of LSTM networks.

Section 2.3 I'm not convinced of the added value of the wavelet transform.

In this study, cross-wavelet transform (XWT) was used to analyze the variance pattern in the pr_a-wtd_a relationship at the individual pixel level in time and frequency domain. With XWT analyses, we are able to explain decreasing test performance (i.e., performance combination C2, with high training R^2 score and very low test R^2 score) at some selected pixels by a changing temporal pattern in the pr_a-wtd_a relationship.

In the marked-up manuscript version, we added the added value for XWT analysis in this study in Line 192-193, and modified the content in Line 193-196 for clarity.

Section 2.4 Figure 3 shows the mean water table depth simulated by the spatially distributed model on average on 20-year. From this figure, it appears clearly that rivers have a significant impact although they are not taken into account in scheme 1.

We agree that there is a strong hydraulic connection between surface water and groundwater in areas close to surface water, but not in areas far from surface water, such as recharge areas. In this study, we presented a generic scheme of subsurface water balance to illustrate the relationship between precipitation and groundwater for all pixels on the European continent.

Moreover, the range of value is huge, and this figure is very difficult to read.

Thank you for point it out. We changed the color scale of Fig. 3.

Section 3:.1 Why do you focus on wtda>1.5 ? What do you expect it means?

As mentioned in the previous response, $wtd_a > 1.5$ corresponds to strong drought, which leads to significant societal, economic, and ecological impacts. That is why we focused on $wtd_a > 1.5$ in Sect. 3.1.

We added "(i.e., a strong drought) " in Line 317 in the marked-up manuscript version for clarity.

Aquifer level is not completely free to vary: a lower limit can be fixed by a bedrock, and the top surface (topography) can be the upper limit. Thus, I'm not sure that wtda>1.5 represent the same situation everywhere…

As mentioned in the previous response, the wtd_a values are calculated at the individual pixel level, which are the deviations of the *wtd* values from the climatological average for a specified time period normalized by the climatological standard deviation. The influence of the spatial variability of aquifer levels was removed in the calculation of the wtd_a values. The wtd_a values are unitless and normalized, allowing comparison between different locations over Europe.

The figure 5 is very pale, and it's hard to see something…

Thank you for point it out. We changed the color scale of Fig. 5 and other figures related to European wtd_a maps (i.e., Figs. B1 and B2).

Section 3.2 I'm not sure to be able to interpret figure 6. Indeed, we have no idea which percent of the domain is represented by each x-axis value. The pdf of R2 by region will have been more interesting

The focus of this section was to study the impact of different local factors on the network test performances. In Fig. 6, we showed averages and standard deviations of the test R^2 scores and RMSEs for the categories based on different intervals of yearly averaged *wtd*, *ET*, θ and S_w. For statistical significance, we only considered categories with ≥ 50 pixels. By comparing averages of the test R^2 scores and RMSEs in different categories of local factors, we were able to identify their various impacts on the network test performance. For example, we found that the network tended to have poor test performance (i.e., low R^2 score and high RMSE) at pixels with a big value of yearly averaged *wtd*. Such information is hard to interpret through the PDF of R^2 by region. Moreover, Fig.7 and Table 4 provide similar knowledge as the PDF of R^2 by region, showing the percentages of the selected pixels with a test R^2 score ≥ 50% in different PRUDENCE regions. Therefore, we would like to keep Fig.6.

Figure 7 is also hard to read…

Thank you for point it out. We changed the color scale of Fig. 7.

Figure 8: percentage of the pixel that have a negative evapotranspiration. I guess there are not numerous pixels… How can we interpret this figure?

Negative *ET* values were only found in Region Scandinavia (SC). The pronounced freezing and sublimation processes often happen in this region, which explains the negative *ET* values, as stated in Line 382-383 in the Version 2.

For clarity, we changed "this" to "negative *ET*" in Line 404 in the marked-up manuscript version.

Section 3.3 again, not sure there is an added value of the wavelet transform..

Please refer to the response related to Sect. 2.3 to recall the added value of XWT in this study.

In the marked-up manuscript version, we adjusted the content in Line 415-417 to clarify the added value of the XWT.

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.

Hartick, C., Furusho-Percot, C., Goergen, K. and Kollet, S.: An Interannual Probabilistic Assessment of Subsurface Water Storage Over Europe Using a Fully Coupled Terrestrial Model, Water Resour. Res., 57(1), doi:10.1029/2020WR027828, 2021.

McKee, T. B., Doesken, N. J. and Kleist, J.: THE RELATIONSHIP OF DROUGHT FREQUENCY AND DURATION TO TIME SCALES, in Proceedings of the 8th Conference on Applied Climatology, pp. 179–184, American Meteorological Society, Anaheim., 1993.

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

1. There was an error in the data of pr_a and wtd_a values due to the wrong numeric precision setting in the calculation. We reran the LSTM networks with corrected data and updated related figures, tables and text (i.e., Fig. 5-C1, Table 4-C1, Line 374-375, Line 400, Line 402) as well as the shared datasets. The conclusions did not change based on the corrections.

2. We fixed mistakes in the calculation of yearly averaged *wtd* and soil moisture (θ) and recalculated regional averages and standard deviations of yearly averaged *wtd* and θ. We corrected relevant values in Table 1, 5 and C1, Line 229-233, Line 367-368.

3. We fixed the correction of α and updated the relevant values in Table 5 and C1.

4. In Line 7, deleted “mainly” and modified “domestic water use” to “public and industrial water supply”.

5. In Line 13-14, modified “To set up the methodology” to “In the proposed methodology”.

6. In Line 16, changed “They” to “The data”.

7. In Line 18, changed “the spatial distribution of” to “spatially distributed”.

8. In Line 48, modified the content for clarity.

9. In Line 71, changed “literatures” to “literature”.

10. In Line 81, deleted “,”.

11. In Line 130, changed “In case of” to “In the case of”.

12. In Line 163, changed “Numbers” to “The numbers”.

13. In Line 198, added “the”.

14. In Line 214, added “,”.

15. In Line 225, added “from”.

16. In Line 318, changed “and” to “on”.

17. In Line 326, changed “2.5” to “2”.

18. In Line 392, deleted “,”.

19. In Line 399, added “for”.

20. In Line 413, added “the”.

21. In Line 459, changed “gave a hypothesis” to “hypothesize”.

22. In Line 470, changed “re-analyses” to “reanalyses”.

Best regards,
Yueling MA

On behalf of all the authors