Comments/Text of Anonymous Referee 1 (AR1) posted in blue, our text in black with old passages in red and the new passage in green.

2^nd Review of: Rainfall-runoff modeling using Long-Short-Term-Memory (LSTM) Networks
By Kratzert et al, submitted to HESS, 2018

The revised version of this manuscript is greatly improved, and I appreciate all the efforts the authors took to make improvements and corrections. This time when I read it, the experimental setup and results were overall more convincing, and the explanations of the LSTM method was a lot better for readers not extremely familiar with machine learning methods. Below I list some minor corrections that are mainly related to wording/typos/grammar, but otherwise I feel this paper is an interesting and novel contribution and would be ready for publication.

We are glad that our efforts in the first revision satisfy AR1 and we would like to thank AR1 again for his comments and suggestions (of the first and this review), which helped us to improve our manuscript.

All corrections regarding working/typos/grammar are changed in the revised manuscript. Below, we provide our answers and corrections regarding every remark that affects an entire sentence or paragraph.

1. Page 10, Line 11: recommend to rewrite sentence to remove parenthesis

   The relatively flat South Atlantic-Gulf region contains more homogeneous basins (similar to the New England region), but is in contrast not influenced by snow.

   The relatively flat South Atlantic-Gulf region contains more homogeneous basins, but in contrast to the New England region is not influenced by snow.

2. Page 11, Lines 4-5: re-write awkward sentence that starts “As final model”

   As final model (and as the model we used for comparison), the model with the lowest RMSE in the calibration period is chosen.

   Of these 10 models, the one with the lowest RMSE in the calibration period is used for validation.

3. Section 2.5.2: At the beginning of this section, you mention “2 ideas” but the second idea (ungauged basins) comes very late after the first – should briefly state the 2 motivations early in the first paragraph, then spend next two discussing them in more detail.

We agree and adapted the beginning of Section 2.5.2 as follows:

Our second experiment is motivated by two different ideas: (i), deep learning models really excel, when having many training data available (Hestness et al., 2017; Schmidhuber, 2015), and (ii), regional models as potential solution for prediction in ungauged basins.
Regarding the first motivation, having a huge training data set [...].

4. Page 13, Line 33: “fewer epochs”

   We agree that “for a few number” should be changed. However, “fewer epochs” as suggested by AR1 does not contain the information we want to state with this phrase. Here, it is important for us to state that we do not train the networks for fewer epochs compared to the previous experiments, but instead that fine tuning is done for a comparatively small number of epochs. If fine-tuning is done for more than just a few epochs, the network would again start to overfit to the specific catchment it is fine-tuned for.

   Therefore we changed the sentence as follows:

   Then, the pre-trained network is further trained for a few number of epochs…

   Then, the pre-trained network is further trained for a small number of epochs...

5. Page 14, Line 23: recommend to re-write sentences to omit “;”. Also the phrase “Afterwards, we start by” is contradictory.

   We now present the results of our experiments and discuss the following points; at first, we give an illustrative comparison of the modelling capabilities of traditional RNNs and LSTMs to hightlight the problems of RNNs to learn long-term dependencies and its effect on the task of rainfall-runoff modelling. Afterwards, we start by presenting how well our LSTM network can model runoff processes of single catchments. Therefore, we analyze the results of Experiment 1, for which we trained one network separately for each basin and compare the results to the SAC-SMA + Snow-17 benchmark model.

   We start presenting our results by showing an illustrative comparison of the modelling capabilities of traditional RNNs and the LSTM to highlight the problems of RNNs to learn long-term dependencies and its deficits for the task of rainfall-runoff modelling. This is followed by the analysis of the results of Experiment 1, for which we trained one network separately for each basin and compare the results to the SAC-SMA + Snow-17 benchmark model.

6. Page 15, Line 32: rephrase “it shows very well the problem”

To conclude, although only based on an illustrative example, it shows very well the problem RNNs have with learning long-term dependencies and why they shouldn’t be used if (e.g. daily) discharge is predicted only from meteorological observations.

Although only based on a single illustrative example that shows the problems of RNNs with long-term dependencies, we can conclude that traditional RNNs should not be used if (e.g. daily) discharge is predicted only from meteorological observations.

7. Page 17, Lines 1-3: This was brought up and addressed in the previous round of comments, but here I still feel that the “surprise” could be toned down, and this aspect could be posed more as a potential benefit of this type of model, in that it is able to simulate long-term processes. E.g. instead of noting your surprise compared to what you expected, discuss that feature as a notable benefit of the LSTM approach, where the example shows how it can learn long-term dependencies with ease.

   This is a somewhat surprising result, since we were expecting that the correct reproduction of snow accumulation and snowmelt processes might be challenging for the LSTM approach. However, from our results it seems that the model can easily learn these long-term dependencies, i.e. the time lag between precipitation falling as snow during the winter period and runoff generation in spring with warmer temperatures.

   This clearly shows the benefit of using LSTMs, since the snow accumulation and snowmelt processes are correctly reproduced, despite their inherent complexity. Our results suggest that the model learns these long-term dependencies, i.e. the time lag between precipitation falling as snow during the winter period and runoff generation in spring with warmer temperatures.

8. Page 21, Line 9: re-word phrase “while there exist some basins”

   While there exist some basins in the eastern part…

   While some basins exist in the easter part...

9. Page 23, Line 16 – Page 24, Line 7: This paragraph seems a bit casually written compared to the rest of the paper – contains several typos and grammar errors and should be somewhat re-written.

   Neural networks (as well as other data-driven approaches) are often criticized for their “black-box-ness”, not only in the hydrological community. Yes, this criticism is very justifiable – at least in science the question of how and why a specific model or method works well or not is important. Looking behind the scene is what makes our work and science attractive. In this context, we want to conclude with a visualization of a preliminary analysis of a cell state of the applied LSTM. Figure 15 shows the evolution of the value of a single cell state (c t , see Sect. 2.1) in the LSTM over the period of one input sequence (which equals to one year in this study) for an arbitrary catchment used in this study, exhibiting snow accumulation and melt in spring. Very surprising and interesting temporal dynamics are evident. We can see that increases and decreases, as well as the fluctuations between time step 60 and 120 of the cell state value match pretty good with the dynamics of the temperature curves (use the gray vertical lines in Fig. 15 for guidance). As an example we can see that the cell state increases with temperatures falling below 0°C (approx. time step 60) and a fast depletion as soon as the daily minimum temperature increase above the freezing point (time step 200). These seasonal dynamics are exactly what we expect, when we think about snow accumulation and melt on the catchment scale. Thus, the LSTM unintentionally generated observable snow dynamics within a cell state, suggesting that there is more to find behind the scenes

Finally, we want to show the results of a preliminary analysis in which we inspect the internals of the LSTM. Neural networks (as well as other data-driven approaches) are often criticized for their “black box” like nature. However, here we want to argue that the internals of the LSTM can be inspected as well as interpreted, thus taking away some of the “black-box-ness”.

Figure 15 shows the evolution of a single LSTM cell (c_t, see Sect. 2.1) of a trained LSTM over the period of one input sequence (which equals 365 days in this study) for an arbitrary, snow influenced catchment. We can see that the cell state matches the dynamics of the temperature curves, as well as our understanding of snow accumulation and snow melt. As soon as temperatures fall below 0°C the cell state starts to increase (around time step 60) until the minimum temperature increases above the freezing point (around time step 200) and the cell state depletes quickly. Also the fluctuations between time step 60 and 120 match the fluctuations visible in the temperature around the freezing point. Thus, albeit the LSTM was only trained to predict runoff from meteorological observations, it has learned to model snow dynamics without any forcing to do so.