Comments/Text of Anonymous Referee 3 (AR3) posted in blue, our text in black.

Benchmarking LSTM

Overall, this paper stands at the forefront of hydrology. There are three aspects of the paper that I like. First, this work show state-of-the-art performance in terms of large-scale streamflow prediction accuracy. This would serve to push hydrologic science forward. Second, the authors implemented a novel LSTM structure to enable a static layer through which they could examine the impacts of different static catchment attributes. Third, they investigated network internal embeddings which is the first time in hydrology which I have seen, and provided some insights (not so perfect, as I would expand on later). These are all novel and I believe the paper should eventually be accepted.

Upon deeper examination I indeed found some issues related to potentially un-robust analysis, points of confusion and lack of clarity, need for more hydrologic insights, and somewhat superficial discussion in the exploration of embeddings. Some relevant citations are also missing. Thus I rate the manuscript a moderate revision. The comments below are not to cast the paper in a negative way, but they are in the hope of helping the authors improve the paper to a strong state before publication.

Major comments:

1. Hydrologic understanding: the discussion of the clustering and embeddings was, shall I say, not entirely satisfying. I liked the novelty of the visualization and the construct of the LSTM to enable this. It helped us understand a bit more about how LSTM works. However, I craved for a bit more hydrologic understanding. The discussion in section 3.4 was a bit sporadic and not so memorable. The take-home message appears to be “the EA-LSTM is able to learn complex interactions between catchment attributes that allows for grouping different basins”. Stopping here does not help with the long-standing criticism of machine learning as a blackbox. I had hoped to gain some deeper hydrologic insights, e.g., why different basins were grouped together? What is the characteristic of each cluster and how are these clusters different from previous catchment clustering schemes, e.g., (Berghuijs et al., 2014; Carrillo et al., 2011; Fang & Shen, 2017; Sawicz et al., 2011; Toth, 2013; Troch et al., 2013)? To go deeper it may not need additional work, but more thoughts about the results.

To avoid misunderstandings, we would like to clarify that we see the main contribution of this paper as i) demonstrating of the LSTM-based modeling approach for large-scale hydrological modeling in general (building upon the results of Kratzert et al., 2018) ii) introducing the EA-LSTM and iii) benchmarking vs. a large set of well-established hydrological models.

With this premise we would like to address the comment regarding Section 3.4: Within our EA-LSTM, we include an embedding layer (the static LSTM input gate), that can ‘learn’ about catchment diversity purely from discharge data. Analyzing the (physically) meaningfulness of

the learned embedding can be seen as a (fairly simple) form of explainable AI. Although, as said above, our paper is mainly intended as a modeling paper, we think that Section 3.4 has the following benefits (copied from our answers to Reviewer #1):

- The blackbox i somewhat opened. The section provides at least some intuition about what happened in the embedding layer of the input gate in the EA-LSTM (namely grouping of basins in a way that matches expectations).
- It provides an example of the kind of analysis that are possible with the EA-LSTM in general and potentially opens the door for many follow-up studies in the future. Such studies could either concentrate more on the interpretability of LSTM-based models or try to extract new hydrological understanding from the learned groupings.

Given the above mentioned scope and the benefits of the section, we would like to avoid extending the hydrological interpretation of Section 3.4. Especially, because here we are not analyzing the model performance, but rather just examine the intrinsic properties of the model. Additionally, as the reviewer herself/himself cites, doing a full-fledged cluster analysis is the work of many individual publications themselves and would clearly be out of scope here.

2. More robustness: I'm afraid many of the attributes in Table 4 are correlated in space and it may be not very robust to draw conclusions from them especially for attributes that are not the highest ranking. For example, does geological permeability really stand position #9? Can we take it that permeability is the second important factor amongst non-climatic factors? This is somewhat surprising and is worth more discussion, but I'm afraid it might just be due to coincidence. To see so the authors could remove some basins (randomly or removing a spatial cluster) or attributes (as the factors tend to have interaction in these kinds of factor analysis) and train again and see how this table react to the perturbation.

First off , we would like to state that any spatial correlation in physical catchment features is real information that can and should be leveraged by regional models. We even explicitly did not include latitude/longitude inputs to our model in this study so that only real, physically-based information is leveraged directly by the EA-LSTM.
This is discussed, for example, by Addor et al (2018), and our findings are in line with the results of said publication. The results may thus be less surprising than indicated in this comment. In this context we would also like to mention that we did an independent robustness analysis by perturbing the features with gaussian noise (see L395ff), which shows the reliance (and robustness) of the model with respect to changes of the features.

We do however agree that the results do not form a particularly strong ranking. Regarding this point, it is important for us to emphasize that in the original contribution did not claim anywhere that the absolute rank of any particular feature has a meaning. This was a model sensitivity analysis, which is common for modeling studies. The only conclusions that we drew from this sensitivity analysis are:

- "*..the most sensitive catchment attributes are topological features [...] and climate indices [...].*" (L 422f)
- "*Certain groups of catchment attributes did not typically provide much additional information. These include vegetation indices [...], as well as the annual vegetation differences. Most soil features were at the lower end of the feature ranking*" (L 423ff)

These seem to be valid conclusions of a sensitivity analysis like this.

That said, our experiments suggest that the obtained qualitative ranking of the feature groups (like climatic, topological, soil and vegetation) is rather robust. To strengthen upon this statement, we added below the results of the same analysis for all 8 repetitions of the same model settings (the EA-LSTM optimized with the basin average NSE) as used in Table 4. As we can see from these tables, the qualitative ranking of these feature groups remained similar. We hope that the reader does focus on exact rankings or exact sensitivity values of any particular feature but rather on the overall image of Table 4 - which is why we grouped these into categories in the first place.

We also agree with the reviewer that this might not be clear from the way the manuscript is currently written, and thus the results could be questioned as being a *coincidence.* We will therefore update the manuscript to clarify regarding this point.

p_mean	0.682248	p_mean	0.737930
aridity	0.564276	elev_mean	0.620351
area_gages2	0.504591	area_gages2	0.520789
elev_mean	0.459893	frac_snow	0.481682
high_prec_dur	0.406671	clay_frac	0.471752
frac_snow	0.405564	aridity	0.407523
high_prec_freq	0.382006	gvf_max	0.335463
slope_mean	0.370855	geol_permeability	0.296405
geol_permeability	0.352949	soil_depth_pelletier	0.291952
carbonate_rocks_frac	0.339022	pet_mean	0.290072
clay_frac	0.330383	slope_mean	0.282145
pet_mean	0.310769	low_prec_freq	0.280580
low_prec_freq	0.299585	soil_depth_statsgo	0.274308
soil_depth_pelletier	0.273934	soil_conductivity	0.274178
p_seasonality	0.272786	silt_frac	0.259220
frac_forest	0.267421	high_prec_dur	0.259150
sand_frac	0.255156	p_seasonality	0.250047
soil_conductivity	0.243641	low_prec_dur	0.248930
low_prec_dur	0.219104	sand_frac	0.243023
gvf_max	0.213809	carbonate_rocks_frac	0.235574
gvf_diff	0.212412	frac_forest	0.232315
lai_diff	0.208096	gvf_diff	0.222452
soil_porosity	0.194036	high_prec_freq	0.202322
soil_depth_statsgo	0.191936	lai_diff	0.192319
lai_max	0.190274	lai_max	0.168555
silt_frac	0.183365	soil_porosity	0.167466
max_water_content	0.158722	max_water_content	0.142825

elev_mean	0.592299
p_mean	0.576201
aridity	0.506481
area_gages2	0.474934
frac_snow	0.447427
clay_frac	0.443380
carbonate_rocks_frac	0.432280
slope_mean	0.400347
geol_permeability	0.368472
pet_mean	0.368429
soil_depth_pelletier	0.342860
gvf_max	0.329133
sand_frac	0.314407
high_prec_freq	0.301476
soil_conductivity	0.295581
high_prec_dur	0.279304
gvf_diff	0.279032
p_seasonality	0.276549
silt_frac	0.267486
low_prec_freq	0.233236
soil_porosity	0.212450
soil_depth_statsgo	0.212345
frac_forest	0.193351
lai_max	0.192644
lai_diff	0.185409
max_water_content	0.171502
low_prec_dur	0.167153

elev_mean	0.687521
p_mean	0.675504
aridity	0.524162
low_prec_dur	0.427454
soil_depth_pelletier	0.403478
clay_frac	0.401946
frac_snow	0.396246
area_gages2	0.384397
high_prec_dur	0.383352
slope_mean	0.378002
gvf_max	0.345176
low_prec_freq	0.342597
pet_mean	0.327042
geol_permeability	0.323855
p_seasonality	0.322263
frac_forest	0.320437
silt_frac	0.292820
high_prec_freq	0.273888
max_water_content	0.245695
soil_depth_statsgo	0.245633
sand_frac	0.203782
soil_conductivity	0.201373
gvf_diff	0.195265
carbonate_rocks_frac	0.188165
lai_diff	0.173871
lai_max	0.141322
soil_porosity	0.113825
dtype: float64	

p_mean	0.683777
elev_mean	0.535805
aridity	0.474985
area_gages2	0.473146
frac_snow	0.430077
high_prec_freq	0.429197
slope_mean	0.406997
soil_depth_pelletier	0.387183
carbonate_rocks_frac	0.375872
clay_frac	0.357481
geol_permeability	0.344788
gvf_max	0.327458
gvf_diff	0.324827
pet_mean	0.320391
low_prec_freq	0.310324
p_seasonality	0.291569
high_prec_dur	0.273754
silt_frac	0.272043
low_prec_dur	0.241519
soil_depth_statsgo	0.226876
max_water_content	0.219675
sand_frac	0.215917
soil_conductivity	0.214915
frac_forest	0.196826
soil_porosity	0.189546
lai_max	0.173699
lai_diff	0.155421

elev_mean	0.614620
p_mean	0.600607
frac_snow	0.515789
aridity	0.506338
area_gages2	0.439643
soil_depth_pelletier	0.366733
slope_mean	0.346615
clay_frac	0.343064
carbonate_rocks_frac	0.329496
gvf_max	0.318784
high_prec_freq	0.314248
p_seasonality	0.309494
low_prec_freq	0.305372
geol_permeability	0.285597
sand_frac	0.285117
high_prec_dur	0.272177
low_prec_dur	0.235999
pet_mean	0.231612
silt_frac	0.230806
gvf_diff	0.225315
soil_conductivity	0.222055
frac_forest	0.194465
soil_depth_statsgo	0.189208
soil_porosity	0.177579
lai_diff	0.166245
lai_max	0.157240
max_water_content	0.143066

elev_mean	0.624811	p_mean	0.690424
p_mean	0.607064	elev_mean	0.563552
aridity	0.483825	frac_snow	0.557419
area_gages2	0.437059	aridity	0.513616
p_seasonality	0.421215	area_gages2	0.464579
slope_mean	0.390884	high_prec_freq	0.374508
frac_snow	0.390787	high_prec_dur	0.359950
high_prec_freq	0.382907	gvf_diff	0.333755
high_prec_dur	0.349156	soil_depth_pelletier	0.325056
gvf_max	0.349014	geol_permeability	0.324826
geol_permeability	0.335745	pet_mean	0.324743
soil_depth_pelletier	0.323830	clay_frac	0.322207
carbonate_rocks_frac	0.309022	slope_mean	0.317398
gvf_diff	0.288544	gvf_max	0.311988
pet_mean	0.287186	sand_frac	0.305872
clay_frac	0.284030	low_prec_dur	0.281842
low_prec_freq	0.245233	p_seasonality	0.278623
frac_forest	0.224514	silt_frac	0.271964
soil_conductivity	0.219839	carbonate_rocks_frac	0.267376
silt_frac	0.212018	soil_conductivity	0.259286
low_prec_dur	0.208801	low_prec_freq	0.237494
sand_frac	0.183467	lai_diff	0.178141
soil_depth_statsgo	0.180952	frac_forest	0.177963
lai_diff	0.178751	lai_max	0.177861
max_water_content	0.166533	soil_porosity	0.176054
soil_porosity	0.146432	soil_depth_statsgo	0.158091
lai_max	0.145553	max_water_content	0.127484

3. Details for reproducibility: one of the selling points of the paper was the high performance. Hence it imperative that the results are reproducible. Are the transformations applied for input and output? How many layers of LSTM were used (in comparison with authors' HESS 2018 paper, this choice seemed ad hoc?)? How was the ranking for Table 4 done indeed? This was a local method, so what is the origin for perturbation?

All information demanded by the reviewer are already reported in the manuscript:

- "Are the transformations applied for input and output?" L 247 "*All input features (both static and dynamic) were standardized (zero mean, unit variance) before training*". However, we agree that such an information should probably be placed in the data section (Section 2.4) and will update the manuscript accordingly.
- "How many layers of LSTM were used [...]?" The number of LSTM layers (one LSTM layer) is specified alongside the other network details in the Appendix B (L 562).
- "How was the ranking for Table 4 done[...]? " The details on how to derive the feature ranking is explained exhaustively in Section 2.6.2 "Robustness and Feature Ranking". Concretely, regarding the ranking of Table 4: "*Further, since we predict one time step of discharge at the time, we obtain this sensitivity measure for each static input for each day in the validation period. A global sensitivity measure for each basin and each feature is then derived from taking the average absolute gradient (Saltelli et al., 2004).*" and then L. 420f "*Table 4 provides an overall ranking of dominant sensitivities. These were*

derived by normalizing the sensitivity measures per basin to the range (0,1) and then calculating the overall mean across all features"

- "... what is the origin for perturbation?" We used the optimized parameters as starting value. There might be some confusion here. We do not solve the gradient computation by numerical approximation, but rather calculate the gradients analytically through backpropagation. So if at all, the true values for the static catchment attributes can be seen as the origin of perturbation. In the original manuscript this is explained in 2.6.2 "Robustness and Feature Ranking" L.251f.

4. Share more experience please: there are many choices which were unexplained, and the community would benefit from the authors providing more discussion of what worked and what did not during their experiments. How did other objective functions do? What if you don't do ensemble averaging? How large are the impacts of hyperparameters, e.g., hidden layers and learning rates? These do not necessarily need figures and could be answered by a couple of sentences. Some minor points below are related to this.

Sadly, we do not know how we can do this. We tried to provide as much information as possible. And, to our knowledge, no choices in our network architecture or training procedure remained unexplained. Appendix B explains the hyperparameter search settings. We did not experiment with different learning rates and can't share any experiences on this question. Furthermore, we did not test any other objective functions than the two reported in this paper. Hyperparameter search was performed using MSE (the machine learning community standard for regression tasks).
The only thing that comes to mind is that we did not report the results of all considered configurations and if wished we can update the Appendix B accordingly with a short description. As a short summary: The median model performance (across the basins) remains more or less stable between most configurations, while the most variance can be observed in the mean NSE. Two layers did not provide any meaningful improvement, that would justify the additional computational cost. However, our hyperparameter search was not exhaustive and at no point in the manuscript we claim to have found the best possible architecture for this task.

5. The authors should also expand on why climatic factors showed up on top of table 4. It appears other static basin physical attributes were not important at all. Does this suggest catchment co-evolution? A potential indication of overfitting (to climatic factors that obviously vary), and more discussion is begging to be done here.

The climatic factors show up on the top of the table, since through the method of Morris they have the highest gradient. We don't know of any experiment that would tell us *why* climate factors appear there (i.e. why they have the highest gradient), except hydrological intuition. (This is not different than any sensitivity analysis for any type of hydrologic model - sensitivity analyses do not answer questions about 'why' certain features are more sensitive) As such, these results in isolution do not suggest catchment co-evolution. They tell us that the model

uses certain features more heavily than others. However, these findings are also in line with the results reported by Addor et al. (2018), as we state in L 428 "*It is worth noting that our rankings qualitatively agree with much of the analysis by Addor et al. (2018).*"

Also, this table doesn't suggest that physical attributes are unimportant, just that they are not as important as climate features. Again, this agrees with previous literature, as cited. This intuition that climate-related factors are the dominant drivers of hydrological systems, for example, models are often tested in terms of their ability to predict departures from the Budyko curve. We therefore do not see any indication of overfitting from this analysis.

Minor points:

1. I'm at a loss to understand the opening statement about streamflow being an out-standing problem. At what point is this problem solved vs not solved? Is there a hard threshold? Did the present work solve this problem?

To clarify: The sentence in question reads: "*Regional rainfall-runoff modeling is an old but still mostly out-standing problem in Hydrological Sciences"*. Here, *out-standing* is referring to *regional* modeling, not to streamflow modeling in general. There is no hard threshold to determine when a problem like this is solved (and we believe that the sentence does not imply that either; as a matter of fact we added the word "mostly" to avoid such a conclusion). However, the benchmarking in our paper with state-of-the-art regionalization methods and the fact that the proposed LSTM-based modeling approach significantly (and by far margins) outperforms these models, suggest that there is (or at least was) still significant room to improve how the community addresses this problem.
We believe that most readers will not be puzzled by the provided formulation and will therefore leave it unchanged.

2. L73, "which part of the network are used for a given basin"—this sentence is difficult to interpret at this point. What does "used for" mean here.

We added some clarity to this sentence: "*Concretely, we propose an adaption of the LSTM where catchment attributes explicitly control which parts of the LSTM state space are used for a given basin*"

3. L76, "similarly behaving". Is this referring streamflow responses or attributes? (only the former would be called a behavior, but this work didn't seem to include streamflow response in the clustering part)

"Behavior" here refers to the similarity in the rainfall-runoff dynamics, as suggested by the reviewer. This is also stated implicitly in the two sentences directly preceding the one in question (L74f) "...*it can learn how to combine different parts of the network to simulate different*

types of rainfall-runoff behaviors. In principle, the approach explicitly allows for sharing parts of the networks for similarly behaving basins..."

4. L78, "embedding". This is a natural language processing jargon. Quite difficult for hydrologists to comprehend. I think it would be reader friendly if the authors spend two sentences explaining this word. My understanding is that embeddings are not just hidden layer activations, but a mapping of inputs to an ordered hidden space that has meanings. For example, the hidden layers of machine translation layers form an embedding. Each ranked item in the embedding in NLP can be related to a linguistic concept.

Historically, "embedding" is not a term from the field of natural language processing, but rather a general mathematical concept. Maybe the reviewer is confusing this term with "word embeddings", which is a term-of-art from natural language processing, but is not what we are referring to. More importantly, L. 77f defines the term embedding exactly: "*..our adaptation provides a mapping from catchment attribute space into a learned, high-dimensional space, i.e. a so-called embedding*".

5. L117 "some amount of information" is fuzzy. Is it about catchment attributes or about streamflow responses? This is critically important as the two have very different meaning regarding what would be done. From reading the later parts, here you seem to refer to static Attributes.

We changed the previous sentence to: "*..our objective is to build a network that learns to extract information that is relevant to rainfall-runoff behaviors from observable catchment attributes.*" so that the context is hopefully clearer.

6. L122, regarding using static attributes as a constant array. It would be relevant to cite (Fang et al., 2017) which used this setup and was already distinguishing different landscapes using static attributes as inputs to LSTM. It occurs this paper should at least be mentioned in the present one.

Using static attributes as constant input is not something we are claiming is novel. More specifically, this method has been applied many times before in the field of machine learning (e.g. Karpathy and Fei-Fei 2014, Wen et al. 2015, Wen et al. 2016). The technique was not originally proposed by Fang et al. (2017) and their manuscript is not working on the same topic as our manuscript (rainfall-runoff modeling), we therefore do not see this as an especially appropriate reference to cite in this case.

7. L134-135. This is an interesting setup. It's worth mentioning that, from Eq 9 & 11, what was selected by the input gate were not only x_d but also h from the last step.

It is not entirely clear what the reviewer wants to suggest. If this refers to the fact that that h[t-1] is used in the forget and output gate, as well as the cell update (g[t]), then they are right. The

input gate however, does not get any information of x_d[t] in our proposed EA-LSTM and neither from h[t-1]. We hope by changing the following sentence "*..while the dynamic and recurrent inputs control what information is written..*" we can resolve the confusion.

8. L158 – what happened when you used other loss functions?

We are unsure about the exact intent of this question. We used two loss functions in this manuscript and compared the results. From a hydrological modelling perspective it seems obvious that different loss functions might provide different optimization results. Designing and choosing (good/correct) objective functions is an old and important problem in hydrology. It is highly non-trivial, yet unsolved and surrounded by many discussions. However, it is also not the focus of this contribution and we therefore view the testing of more loss functions as out of scope.

9. L171 "25,000 km2" – is it really appropriate to model those with an area of 25,000 km2 the same as other smaller basins?

Although results of experiments not shown in this manuscript suggest there is no problem with doing so, in this manuscript only basins with an area smaller than 2000km² were used. As stated in L. 174 we use the same 531 basins as Newman et al. (2017): to cite their manuscript: "*We subset the complete Newman et al. (2014) basin list to remove...basins larger than 2000 km²*". That said, we agree that we missed to state this clearly in our manuscript and therefore adapt L 176 to add the following sentence "*Furthermore, out of the 671 basins, only those with an area smaller than 2000km² were kept.*"

10. L194 – "favor of"

Corrected, thank you.

11. L222, regarding the ensemble averaging, readers deserve to know, how big is the spread? What if you don't take the average? Sometimes the ensemble mean gets a better R2 but it misses peaks.

Yes, it is true that taking the ensemble mean will reduce variance. We've not explored more complex ensemble techniques, of which many exist. However, we see testing different ensembling strategies as out-of-scope for this paper.

12. It is unclear what "six different settings and eight different models" are.

This is explained in the preceding sentences.

- Regarding the “six different settings” L 219f “*All three model configurations were trained using the squared-error performance metrics discussed in Sect. 2.3 (MSE and NSE*). This resulted in six different model/training configurations.*”
- Regarding the “eight different models” L 221f “*To account for stochasticity in the network initialization and in the optimization procedure (we used stochastic gradient descent), all networks were trained with n = 8 different random seeds*”

The phrase quoted by the reviewer from L. 224 (immediately following the two sentences quoted) pulls these sentences together “*In total, we trained and tested six different settings and eight different models per setting for a total of 48 different trained LSTM-type models*”

13. L261. might be useful to say you extracted gradients from the learned network after training (correct?), as some readers are unfamiliar with how this is done. However, these gradients are time-step dependent.

Indeed, we calculated the gradients w.r.t. the static inputs from the trained model, since we are interested in analyzing the robustness and feature ranking of a trained network, not of a randomly initialized one. In L. 244 we stated this fact for the model robustness “*To estimate the robustness of the trained model to uncertainty in the catchment attributes…*”. We will add a similar sentence to the feature ranking to avoid possible confusion around analyzing untrained models.

14. Also, why is it called global sensitivity test? It is also local, around a origin for perturbation.

Citing Campolongo et al. (2015) from their introduction “*The Morris method is simple to understand and implement, and its results are easily interpreted. Furthermore it is economic in the sense that it requires a number of model evaluations is in the number of model factors. The method can be regarded as global as the final measure is obtained by averaging a number of local measures (the elementary effects), computed at different points of the input space.*” To clarify the result of Eq. 14 (or Eq. 15 in our case), this is not a global measure in the sense that the entire space of possible values is considered, but in the sense that more points are considered to derive the sensitivity (see Saltelli et al., 2004). This is reflected in our statement in L. 263f: “*A global sensitivity measure for each basin and each feature is then derived from taking the average absolute gradient (Saltelli et al., 2004)*”

15. L264 better say "the average of absolute gradients across all basins and all time steps", and----why absolute?

Here, we are still referring to a global sensitivity measure for each individual basin. Therefore, “*for each basin*” is correct in this sentence. The averaging across multiple basins is then applied to derive the values in Table 4 (see answer to major comment #3), after normalizing the sensitivity measures to the range (0,1) per basin. Absolute, because otherwise oscillating (positive, negative) gradients, have the potential to cancel and (erroneously) suggest that the

respective feature(s) are unimportant. Furthermore, taking absolute values is the proposed method for deriving the global sensitivity measure from these local points and is referred to as μ* in the literature (e.g. Saltelli, 2004; Campolongo et al. 2011).

16. L267-268 “represent xxx into xxx"? the sentence does not make grammar sense. please fix. This is obviously an expansion of from 27 to 256. Why would this be really necessary?

We do not see a grammatical error in this sentence. Embedding can be used as a noun, which makes the phrase “*[this] vector […] represents an embedding of xxx into yyy*” grammatically correct.

This transformation is necessary, since the resulting input gate must be a vector of 256-dimensions - the same size as the LSTM has cell states. This is basically the same as in every other gate where e.g., the 5-dimensional dynamic inputs (the 5 meteorological variables) have to be transformed into a vector of 256-dimensions for the forget and output gate and the cell update respectively.

17. Table 2. this value is indeed the highest I have seen. Good work!

Thank you.

18. L380. Why 447 basins now? What are missing?

The first sentence in Section 3.2 (L.363) explains this: “*The results in this section are calculated from 447 basins that were modeled by all benchmark models*”.

It’s important to reiterate that we used benchmark models that were run by the respective model development groups. We did not run our own benchmark models. This is critical because we want to give the benchmark models the highest possible chance of success - the presumption being that the respective model development groups are the most well-qualified to run their own models. Notice that this is a common strategy in model intercomparison and model benchmarking studies (e.g., Best et al., 2015)

19. L410 Unsure how this answers the question if the network just remembers. The logic is Confusing.

We think that the general results of the robustness analysis (as shown in the boxplot in Fig. 6) indeed address whether the network is simply remembering basins. If we understand the reviewer correctly, s/he is referring to pure overfitting against the static attributes and that the LSTM simply remembers all 531 catchments individually. If the LSTM simply remembers all 531 catchments individually, there would not be slow degradation in performance (as seen as increase in the variance of the boxplot over increasing level of additive noise) but rather a more

drastic performance drop, when not using the exact catchment attributes for each basin. We will add a sentence that better describes this result.

20. L414, mean precipitation, etc --- aren' these supposed to be climatic inputs rather than static? (can we not let the network generalize it from the forcing data)?

Mean precipitation, high precipitation duration etc. are indeed climatic inputs, but also static inputs since these are aggregated values over the time series (see Addor et al. 2017). The network would only be able to derive statistics like mean precipitations internally from the time length we derive as the input for predicting a single day (here we use an input sequence length of 365 days).

21. Table 4. Echoing a major point raised above. What further conclusions can be drawn from the fact that climatic attributes take the most important positions? catchment Co-evolution theory?

Indeed, what could be inferred here? It's a good question. Certainly this type of speculation is far outside the scope of this paper. We are not prepared to speculate on climate-driven catchment co-evolution, but we suspect that the 30-year data record in CAMELS is not long enough to address this question

22. L454 "before vs. after the transformation into the embedding layer". This is a good comparison, although later there didn't seem to be much comment on this comparison

There are a few comparisons made throughout the analysis:

- L 453ff "*In all cases with cluster sizes less than 15, we see that clustering by the values of the embedding layer provides more distinct catchment clusters than when clustering by the raw catchment attributes*"
- L 462ff "*In both the k = 5 and k = 6 cluster examples, clustering by the EA-LSTM embedding layer reduced variance in the hydrological signatures by more or approximately the same amount as by clustering on the raw catchment attributes. The exception to this was the hfd-mean date, which represents an annual timing process (i.e., the day of year when the catchment releases half of its annual flow). This indicates that the EA-LSTM embedding layer is largely preserving the information content about hydrological behaviors, while overall increasing distinctions between groups of similar catchments*"
- L 471ff "*Although latitude and longitude were not part of the catchment attributes vector that was used as input into the embedding layer, both the raw catchment attributes and the embedding layer clearly delineated catchments that correspond to different geographical regions within the CONUS*"

For each of the steps of the cluster analysis (silhouette plots, variance reduction and cluster results shown on the map of the USA), we actually gave a direct comparisons between the results using the embedding of the EA-LSTM or using the raw catchment attributes. We are unsure what kind of additional comments are expected from the reviewer.

23. UMAP—might be good to briefly explain what it does. Is it just PCA?

We agree that the explanation of the UMAP method could be extended in Section 2.6.3 and will update the manuscript accordingly.

24. L479 Honestly, it's not that easy to see which cluster you are talking about. could use some annotation on the plot.

This is a good idea and we will update the plot accordingly.

25. L489 I found this discussion, as a take-home message, to be somewhat superficial, and unsurprising. I'd appreciate somewhat more in-depth discussion about the hydrology.

We understand and appreciate the reviewer's perspective (more is usually better), however it's hard to see from this comment what the reviewer finds missing in our analysis. We did give quite a lot of hydrological discussion in the context of analyzing the embedding layer - does the reviewer see something in our analysis that is missing? Is there something they might hope to learn that we didn't explore? We would love suggestions about how to improve this analysis, but just asking for more is not really an actionable suggestion.

Regarding the reviewers' suggestion that our conclusions were not surprising, I guess surprising is somewhat subjective. We were generally happy that (a) the model performed as well as it did against benchmarks, and (b) that the similarity analysis generally agreed with previous literature. This means that the model at least appears to be giving the right answers for the right reasons.

26. Figure 11 these colors do not mean anything. It is a bit confusing. Why not use a at least partially consistent color scheme?

These colors actually do mean something (and it was actually somewhat difficult to get the colors to match on the various plots). These colors present the results of several clustering analyses, and are categorical labels. Therefore, we chose to color the basins in a categorical color palette, where each color reflects one cluster class. This makes a categorical color-scheme necessary, since there is no intrinsic ordering (excluding continuous, sequential and diverging color schemes). Furthermore, we made sure that the clusters between the different subplots are more or less colored similarly, so it is easier to compare between the subplots. What else does the reviewer meant by "partially consistent color scheme"? Consistent with what? Certainly the color scheme is consistent between subplots in the figure.

27. Figure 12 better annotate axes even if they don't mean much

We decided to exclude the 2D-coordinates of the UMAP embedding because, as the reviewer suggested herself/himself, they do not mean anything. We would therefore argue that they are probably more confusing and distracting and the reader could ask what why this basin has an embedding coordinate of (4,2) and the other basin only of (0,-0.5) (both are arbitrary sets).

28. L524. I am confused why this is called regional, as the LSTM was trained with all basins over CONUS. What would constitute a model that is not regional?

Regional modeling in Hydrology has a very specific meaning. The alternative is a local model (i.e., one that is calibrated to a specific basin). The second reviewer suggested that this is potentially a universal rainfall-runoff model that could be applied to basin groups of any scale (small regions, US scale, continental scale or even globally), but our intent was to draw a connection with what is a named (and well-defined) problem in Hydrology.

29. L529-530. It either goes against a belief or it does not. Can't go "somewhat against". And, the logic here is not quite clear. This paper is not about parameter identification. The fact that the network works does not imply that parameters can be identified. First the LSTM parameters cannot be interpreted. Second, even very different parameters could give you similar predictions.

First, it's actually possible for two opinions to partially disagree or somewhat disagree.

Secondly, we are not sure if we understand the comment about parameter identification:

- The technical correctness of the statement "*the LSTM parameters cannot be interpreted*" depends on one's understanding of interpretation (for a lengthy discussion on this topic we refer to Lipton, 2016). The LSTM parameters are (maybe) not one-to-one translatable into physical properties as some of the hydrological model parameters, however this criticism is no less valid for conceptual models: it might be questionable what a e.g., catchment-wide infiltration value represents.

- However, the function of each individual parameter in the trained LSTM could indeed be interpreted, to see if a certain weight e.g., thresholds to specific temperatures in the input. The huge number of parameters however, makes such work difficult. This is not really related to the point of the sentence in question, however, which is about deficiencies in hydrology models, not about the interpretability of LSTM parameters.

- We are also not sure if we understand the second point of the review in this context. Indeed, different parameters can give similar predictions and overall performances, as

we have shown in the paper. However, what is the point here regarding our statement that we think traditional large-scale hydrological models can be structurally improved?

References:

Addor, N., Newman, A.J., Mizukami, N. and Clark, M.P., (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences (HESS)*, *21*(10), pp.5293-5313.

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N. and Clark, M.P., (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), pp.8792-8812.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., ... & Ek, M. (2015). The plumbing of land surface models: benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425-1442.

Campolongo, F., Cariboni, J., Saltelli, A., & Schoutens, W. (2005). Enhancing the Morris method. In *Sensitivity Analysis of Model Output. Proceedings of the 4th International Conference on Sensitivity Analysis of Model Output (SAMO 2004)* (pp. 369-379).

Campolongo, F., Saltelli, A., & Cariboni, J. (2011). From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, *182*(4), 978-988.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137).

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B. and Nearing, G., (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), pp.2215-2225.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004): Sensitivity analysis in practice: a guide to assessing scientific models, *pp. 94–100, Wiley Online Library*.

Wen, T. H., Gasic, M., Mrkšić, N., Su, P. H., Vandyke, D., & Young, S. (2015, September). Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1711-1721).

Wen, T. H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P. H., Ultes, S., ... & Young, S. (2016, November). Conditional Generation and Snapshot Learning in Neural Dialogue Systems.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2153-2162).