First of all, we would like to thank S. Mylevaganam (in the following abbreviated as SM) for the time he took to write this comment. However, before continuing, we would like to express our complete incomprehension for being accused of providing any “favours to related research groups” by citing their papers (see comment 13). The author of the comment does not provide a single piece of evidence to support such a statement. It is strongly offensive, and we would strongly advise the author to step back from such behaviour in the future.

We add the comments from SM in blue and add our replies in black.

1) After decades of entangled and forbidden understanding in the field of hydrology, more systematic research and effort to better understand the underlying processes and the components that form a rainfall-runoff model has made to lease a transition from regression based models to more process oriented rainfall-runoff models (see P-1 LN-1:20). In other words, the quest for more process oriented models prevails to better understand a hydrological system of interest. Otherwise, the efforts to better understand the underlying processes and the components that form a rainfall-runoff model over the past few decades become futile if more regression based models are used instead of process oriented models. What was the reason to spend many decades to understand the components that form a hydrological system? What was the reason to spend many decades to seek a transition to a process based model? Therefore, the reasons for reversing the gear that leads to old school of using regressed equation in lieu of processed oriented models are left unfound.

   Reply: In this comment it is argued that hydrology should be only (and exclusively) performed to better understand a given system; and that pursuing different goals would "invalidate" (sic) this branch of research. We can only partially consent with this statement. Contrary to the statement, we believe the hydrological endeavor is not guided by only a single goal (see e.g. Blöschl (2017), Sivapalan et al. (2003) and Montanari (2013)). As such we believe that progressing into one direction, does not invalidate the research in the others. Following the logic of the statement one would need to conclude that the decades of progression in process based models would invalidate the research of data driven models (say regression based models). This is not the case. Neither theoretically (as lined out above) nor empirically (as we shall demonstrate in the following). In the contribution we show that Artificial Neural Networks have been used in hydrology since the early 90s. They have been experimented with since. To us it seems clear that this branch of research has not made the "quest for more process oriented models" obsolete. And, neither have process oriented models slowed down or invalidated the research regarding data driven models. As part of our research we actually try to bridge the gap between both worlds (see Conclusions P20 L18ff and Fig. 14), so that in the future both approaches can benefit from each other.

2) As per the authors, the computational requirements and high computational costs are some of the striking factors that force to use conceptual models for operational purposes

in lieu of physically based models which have been formulated after many decades of scientific efforts and findings to wade off the lack of understanding of the underlying processes of a hydrological system (see P-1 LN-21; P-2 LN-3). What are those computational requirements and computational costs? It would be more appropriate to list all the computational requirements and computational costs to solve the intended tasks. It would be more appropriate to plot the graph of computational speed with years and technology development to show the current status of technological development to support the hydrological system. As per the authors, the technology development is not at an appreciable level to meet the computational requirement and the computational cost to solve a simple hydrological problem that is explained in the current version of the manuscript.

Reply: We do not see anything that is wrong or contradictory in the lines SM is referring to. Running fully coupled 3D-land surface schemes to describe the water cycle for large river basins/catchments does come at the price of high computational costs. Of course it is in principle possible to apply such a model for each of the CAMELS catchments, but because of the computational requirements (and also data requirements, e.g. 3D soil hydraulic properties, …) this is still impossible for many operational purposes where fast response times are required, when multi-ensemble runs within sensitivity studies or uncertainty estimation frameworks are performed.
Our argument is not that it is not possible to solve the hydrological problem (that we are addressing) with any other method. The point that we want to make is, that often fast model runs are needed (e.g. to describe a hydrological system under investigation or to conduct a comprehensive uncertainty analysis) and that recent developments in AI are worth to be explored.
We will be happy to make that point clearer in a revised version of the manuscript.

3) As per the authors, more conceptual based models are used in operational purposes (see P-2 LN-4). From the reader's point of view, the statement of this nature needs more understanding on the purposes that these conceptual models are used. A conceptual model may perfectly suffice an operational need if the need is well governed by the conceptual model. In other words, the selection of models should be based on the need and the problem to be solved. For example, if the need is about the peal flow, a conceptual model may (depending on the consequences of incorrect estimation of flow magnitude) suit the operational need. On the other hand, if the operational need is about the timing of the peak flow, a conceptual model may not meet the operational purpose. Therefore, is it wise to conclude that conceptual models are applied to meet the operational needs? The manuscripts that are cited need to be thoroughly scrutinized to understand the purposes for which those conceptual models are used in lieu of process based mode

Reply: See the answer to comment (1). While we agree with the general argument of the statement, we do not see how it would improve the understanding of the reader (regarding the subject at hand). We believe that readers will, at this point of the manuscript, not appreciate an extensive discussion about the relative weight of the specific use cases for hydrological models. We therefore disagree with the conclusions that are drawn from the statement (for example, the claim that conceptual models do not suffice for modelling). That said, conceptual models do usually have system states (e.g. soil moisture index, snow water equivalent) or fluxes (e.g. infiltration, groundwater recharge), which can be of interest to the hydrologist. Physical models also simulate these fluxes and states, but in the context of operational and engineering purposes the balance between the effort of setting up the physically based model and the computational expenses vs. potential improved model results often leads to a preference for simplifications. In our manuscript we state the following: “Thus, simplified physically based or conceptual models are still routinely applied for operational purposes” (P2 L4). This is not a conclusion, it is rather an observation from long term experience in water resource management. The references we cite are just examples of different applications in the operational context to underline our statement.

4) As per the authors, the CAMELS dataset (i.e., freely available dataset of 671 catchments with minimal human disturbances across the contiguous United States) contains time series of simulated discharge from the calibrated Snow-17 models coupled with the Sacramento Soil Moisture Accounting Model (see P-9 LN-20). Were the developers of the dataset constrained by high computational requirements and computational costs? How did they develop the time series of simulated discharge for this dataset that represents the contiguous “United States”? What is accomplished by the authors in the manuscript using the proposed LSTM is pending further clarification.

Reply: The method for deriving the time series data of the meteorological forcings is well described in the original publication of the data set (Newman et al. 2015) in Section 2.2. Further, we have no information if the authors of the data set were limited by computational costs and we are unsure about the relevance of this question regarding our manuscript. The same is valid for the second question (the development of the simulated discharge time series) that is well described in Section 3 in Newman et al. (2015). We use their model outputs without modification and therefore see no need for describing their modelling approach in more detail. We explain in P9 L21 - P10 L3 which period of the time series are used by Newman et al. (2015) to calibrate their models and state that we use the exact same period to be able to compare the resulting model outputs. Regarding the last sentence (“What is accomplished by the authors in the manuscript using the proposed LSTM is pending further clarification”): The statement is unclear to us in the context of the remaining comment 4. We tested a new data-driven modelling approach (using the LSTM network) for rainfall-runoff modelling and tested our approach with data from the CAMELS data set. We used the CAMELS data set especially because a) it is publicly available b) contains a large number of different basins c) contains calibrated model outputs (from SAC-SMA + Snow 17) that can be used for comparison. As stated in Newman et al. (2015) the “benchmark application is intended for the community to use as a test bed to facilitate the evaluation of hydrologic modeling and prediction questions” and further “focus on providing a benchmark performance assessment for a widely used calibrated, conceptual hydrologic modeling system. This type of data set can be used for many applications including evaluation of new modeling systems against a well known benchmark system over wide ranging conditions.” To us, this clearly justifies the use of the SAC-SMA+Snow-17 as benchmark for our study.

5) As per the authors, process based models are more data intensive (see P-1 LN-18; P-1 LN-21). Therefore, the temptation to adopt data driven models rooted based on ANN and its branches is fast becoming common. However, as per the authors, the data driven models also heavily rely on extensive data for proper training and validation (see P-10 L-17; P-10 L-20). Without proper training with adequate good quality data, data driven models (e.g., ANN) that reveal no physical meaning of the underlying processes of a hydrological system are also not feasible. Therefore, the authors’ statements need more clarification.

   Reply: We agree that the two passages (P1 L24ff and P10 L17f) in our manuscript might be misleading. Therefore a clarification: The “data-need” for physically-based and/or processed-based models and data-driven modelling approaches is of a different nature:

   a) Physically based models are data intensive, because they need “a-priori” data for the setup; e.g. 3d information about the soil and sub-surface characteristics (P1 L24 ff.) at a high spatial resolution - an information often not available.
   b) Neural networks (in our case LSTM) need many training samples (they get better and better the more training samples are available). By no definition, do these samples need to contain any information about e.g. the sub-surface characteristics at a high resolution. As we show in our study, 15 years of daily data is enough to achieve comparable results with our (rather small) LSTM model. Regarding the data quality: We would state that data quality is less important for Neural Networks as for physically based models (see e.g. Raleigh et al. (2015) for a sensitivity analysis of model forcings of a physically based model). For example Banko and Brill (2001) and Krause et al. (2016) have shown that the size of the training corpus is more important than the quality of the training data. Here one has to differentiate between data quality and quantity.

   We will elaborate this difference more clearly in the revised manuscript.

6) As per the authors, the output (discharge) for a specific time step is predicted from the input x = [x-n, ..., x0] consisting of the last n consecutive time steps of independent variables (daily precipitation, min/max temperature, solar radiation and vapor pressure) and is processed sequentially(see P-4 LN-6). In other words, as per the authors, the rainfall-runoff modeling is represented by the selected independent variables (i.e., daily precipitation, min/max temperature, solar radiation and vapor pressure). Among the selected independent variables, what is the variable that best explains the “infiltration”?

Reply: The major part of the statement succinctly summarizes some of the operational functionality of the implemented LSTM approach. We have nothing to add to this description. The last sentence, however, asks, which of the input variables explains the infiltration process. This question might be interesting, but could as well be interesting for any other type of model used (and in any hydrological context). Similarly, it was not part of our research agenda.

7) In the current version of the manuscript, the distinction between "basins" and "catchments" is not well understood. For example, as per the authors, the CAMELS dataset is freely available and includes meteorological forcing data and observed discharge for "671 catchments" across the contiguous United States (see P-3 LN-15). However, in the subsequent statement(see P-3 LN-16), the authors state that for each basin, the CAMELS data set also includes time series of simulated discharge from the Sacramento Soil Moisture Accounting Model coupled with the Snow-17 snow model. What are basins? What are catchments?

   Reply: We use those terms interchangeably, following the convention laid out by the UNESCO International Glossary of Hydrology (see n. 133 p. 31 in WMO and UNESCO (2012)). We can see that this might be confusing for some readers and are willing to adapt the manuscript and reduce to only one of the two words, if the reviewers or editor wish so.

8) As per the authors, the first 14 years of the 15-year calibration period is the training data and the last, fifteenth, year is the independent validation period (see P-10 LN-10). The selection of training period and the validation period needs more explanation. What is the impact of selecting the last year of the 15-year calibration period for validating the trained model (i.e., LSTM)? Any scientific evidence to show that this type of data selection for training and validating a LSTM model to solve a hydrology related problem works well.

   Reply: Here the preliminary study (to determine the number of training epochs) and the actual experiment (where one model is trained and later evaluated to derive the scores presented in the results section) should not be mixed:

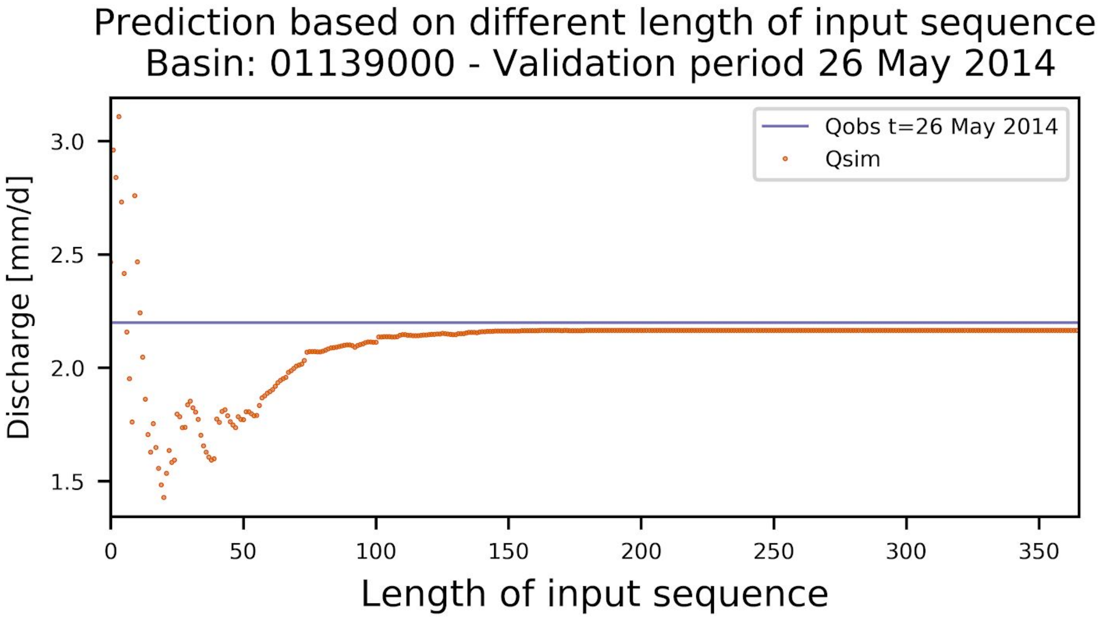
   a) Regarding the preliminary study: Because we wanted to be comparable with the model outputs of Newman et al. (2015) we were limited to the 15-year period available for e.g. the hyperparameter search (the number of training epochs). Thus we split this 15 year period into a new calibration period (the first 14 years) and a new validation period (the last year) and used these two subsets for deriving the number of epochs.
   b) Regarding the actual experiment: With the number determined in (a) we then trained a LSTM on the entire 15 year period (the same period used in Newman et al. (2015) for calibration of their models). All numbers presented in the results

section are derived using the so trained model on the original validation period (all data following the first 15 years of the time series, as in Newman et al. (2015)).

Because, to our knowledge, we are the first to publish rainfall-runoff modelling using the LSTM, we can’t provide references to other publications with the same approach. Further, since we want to test the forecast/simulation ability of the network, it seemed natural to us to not randomly split the calibration data for this preliminary experiment, but to use the first continuous part (14 / 15) for training and the following part (1 / 5) for validation.

9) In the current version of the manuscript, the length of the input sequence, which corresponds to the number of days of meteorological input data provided to the network for the prediction of the next discharge value is set to 365 days in order to capture at least the dynamics of a full annual cycle(see P-6 LN-22). Would this lead to a highly memorized network? What would be the status of the trained network if the length of the input sequence is set to 90 days instead of 365 days? How did you validate your trained network with one year of data (see comment-8) when the length of the input sequence is set to 365 days?

Reply: If we understand SM correctly, the question is: Can a LSTM model, trained exclusively to predict the discharge based on 365 days of meteorological input be used for inference with shorter time series of input and if yes, how does this affect the model output. Then our answer is: Yes, it is indeed possible to evaluate the trained network with shorter time series as input, as the network was originally trained on. Here one has to remember that the parameters of the network are adapted during training based solely on the prediction error (and the subsequent loss function) of the prediction made at the last (365th) day. Figure 1 exemplarily shows the effect of the length of the input sequence on the simulated discharge value of one time step for an arbitrary basin. The blue line shows the observed discharge value at a given time step, the orange points show the network prediction as a function of the number of days provided as input data . The data comes from the validation period, thus the network has not seen the data during training. As we can see for this case and time step, the prediction stabilizes after 150-200 days. This suggests, that for this basin and time step shorter input sequences may suffice for a good prediction. In further studies, systematic investigations should analyse the effects of the length of the input sequence on the simulated discharge, as is stated in the manuscript. If the question is rather, why we have chosen 365 days and not e.g. 90 then our answer regarding comment 9 is: As we state in the conclusions (P20 L13 ff.) and in the Section 2.1 (P6 L22 ff.) the length of the input sequence is one of the hyperparameters and should be investigated more closely in future studies. It was out of scope of this study to have a closer look at each of the hyperparameters, thus we have chosen 365 days to cover a full annual cycle. Regarding the last question of this comment 9: The split is made based on discharge values. Thus the “artificial” validation set we used in the preliminary studies to determine the number of epochs for training are the discharge values of the last, 15-th year (of the original calibration period) with their corresponding 365 previous days of meteorological input.



10) What is the definition of “HUC”? As per the USGS website (https://water.usgs.gov/GIS/huc.html), the HUCs contain either the drainage area of a major river, such as the Missouri region, or the combined drainage areas of a series of rivers, such as the Texas-Gulf region, which includes a number of rivers draining into the Gulf of Mexico. With this definition of HUC, is the development of a more generalized model (see P-10 LN-30) for each of the selected HUCs misleading? Moreover, are the catchments/basins in each HUC ungaged (see P-10 LN-27)? I think, the current version of the manuscript is distant from providing all these details.

Reply: Regarding the first question, if our statement is misleading: We do not agree. The author of the present comment has cited the definition of one HUC correctly. But the basins within one HUC can be quite different, as for example discussed in Section 3.2 (and partly shown in Fig. 11) and also in the conclusions.
Regarding the second question: No, each of the basins is gauged. But as we state in the conclusions (P20 L2ff.) the CAMELS data set could be used to investigate the potential of LSTMs for predicting the discharge in ungauged basins by leaving out some basins in the calibration and then evaluating the model performance on these basins, i.e. cross-validation.

11) The section 2.1 needs an example to illustrate the use of the authors’ mathematical formulations of LSTM. For example, on Wednesday, May 16, 2018, if the precipitation, max temperature, min temperature, and vapor pressure are p unit, l unit, h unit, and v unit, respectively, how the reader of this manuscript uses the developed LSTM model to

determine the intended output (i.e., discharges/runoff?) is needed. Otherwise, the equations that are formed and welded would lead to rely on mathematicians to decode and understand.

Reply: We do not agree that Section 2.1 needs an explicit example on how to calculate the discharge for e.g. “Wednesday, May 16, 2018” given “precipitation, max temperature, min temperature, and vapor pressure are p unit, l unit, h unit, and v unit, respectively”. We think that the provided equations with Fig. 1 in the manuscript and the entire description in Section 2.1. is sufficient for understanding how the discharge of one day is calculated. This is equally done in publications of hydrological models (e.g. Perrin et al. (2003), Samaniego (2010), Aghakouchak and Habib (2010)). Furthermore, the 6 equations of the LSTM involve only rather simple operations of linear algebra (summation, element-wise multiplication and matrix-vector-multiplication). We do not believe that only mathematicians (but not a hydrologist) could “decode” (sic!) these equations.

12) What is meant by SAC-SMA+Snow-17(see P-1 LN-11)? Is it meant to convey that the outputs of SAC-SMA and Snow-17 are added to determine the final output? What is the output of SAC-SMA? What is the output of Snow-17?

Reply: As stated in P1 L6, the results of this study are compared to the CAMELS benchmark model, which consists of a calibrated, coupled Snow-17 snow model and the Sacramento Soil Moisture Accounting Model. The term “SAC-SMA+Snow-17” underlines this coupling. The results of the single model components are not reported but only the overall runoff prediction generated by the coupled modelling framework.
In Section 3.1 in [4] the interested reader can find a more detailed description of the Snow-17 snow model and the SAC-SMA hydrological model, as well as the references to the original publications of these two models.

13) In the current version of the manuscript, some of the cited manuscripts are questionable. For example, citing Shen et al., 2018 to state that the “potential use and benefits” of DL approaches in the field of “hydrology and water sciences” has “only recently come into the focus of discussion”(see P-2 LN-32). Are the authors citing Shen et al., 2018 based on their relationship with Shen et al., 2018? When that manuscript (i.e., Shen et al., 2018) is under severe criticism from the esteemed referees, does it make sense to give credits for that manuscript? Moreover, the cited manuscripts to support the following statements also lead to confusion. Would it be possible for the authors to state the reason for citing these manuscripts to support the statements?

Reply: We reject the reproach that we cite Shen et al. (in review) due to a relationship with the authors of this publication.This is definitely not the case! Shen et al. (in review) - which is still under discussion at the moment of this reply - is cited next to Marçais and de Dreuzy (2017) as these manuscripts focus on the use of DL approaches in the field of hydrology and discuss possible future applications. Both publications are cited in the same sentence: “In general, the potential use and benefits of DL approaches in the field of hydrology and water sciences has only recently come into the focus of discussion.” The two references are a technical commentary and a HESS Opinion manuscript and thus in our opinion are well suited to underline the recent advent of the discussion on DL approaches in hydrology. In addition, the two comments by the referees of Shen et al. (in review) were not online at the time of the submission of the present manuscript. However, we do not see the “severe criticism” of the manuscript in the comments of the two referees. The major concerns are related to the format of the manuscript (HESS Opinion paper) and the similarity to Marçais and de Dreuzy (2017). Content-wise “severe criticism” seems to be raised only by SM himself.

a) The transferability of model parameters (regionalization?) from catchments where meteorological and runoff data are available to ungauged or data scares basins is one of the ongoing challenges in hydrology (Buytaert and Beven, 2009; He et al., 2011; Samaniego et al., 2010).

   Reply: We cite the following three publications in this context because they all deal with the problem of regionalization in hydrology. Buytaert and Beven (2009) analyse the uncertainty involved in regionalizing hydrological model structures. He et al. (2011) review regionalization methods for continuous streamflow estimation for ungauged catchments. Samaniego et al. (2010) propose a regionalization method which accounts for sub-grid variability, which accounts for many of the difficulties of modern methods.

b) The second motivation is the prediction of runoff in ungauged basins, one of the main challenges in the field of hydrology (Blöschl, 2013; Sivapalan, 2003). A regional model that performs reasonably well across all catchments within a region could potentially be a step towards the prediction of runoff for such basins.
   Reply: In our opinion the references concerning prediction in ungauged basins are well justified; both references explicitly deal with the problem of runoff prediction in ungauged basins. Blöschl (2013) is the result of the IAHS initiative “Predictions in Ungauged Basins (PUB)”.

Regarding the unnumbered minor comments:

A. In Figure 5, would not it be appropriate to show the HUC boundaries instead of the state boundaries.

Reply: We tried to match our visualizations as closely as possible to the figures in Newman et al. (2015), as it facilitates a comparison for the reader.

B. In Figure 5, the precipitation values are given in mm/yr. However, on P-9(see LN-19), the precipitation values are given in mm/year.

Reply: It should be mm/yr everywhere and we are thankful for this correction. We will update this mistake in the next version.

C. Should it be “dataset” or “data set”?

Reply: We will revise the manuscript and change all instances of this word to “data set”.

References:

1. Aghakouchak, Amir, and Emad Habib. "Application of a conceptual hydrologic model in teaching hydrologic processes." International Journal of Engineering Education 26.4 (S1) (2010).

2. Banko, Michele, and Eric Brill. "Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing." *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001.

3. Blöschl, Günter, ed. *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press, 2013.

4. Blöschl, Günter (2017) “IAHS2017 Unsolved Problems in Hydrology” https://www.youtube.com/watch?v=jyObwmNr7Ko&feature=youtu.be

5. Buytaert, Wouter, and Keith Beven. "Regionalization as a learning process." *Water Resources Research* 45.11 (2009).

6. He, Y., A. Bárdossy, and E. Zehe. "A review of regionalisation for continuous streamflow simulation." *Hydrology and Earth System Sciences* 15.11 (2011): 3539.

7. Marçais, Jean, and Jean-Raynald de Dreuzy. "Prospective interest of deep learning for hydrological inference." *Groundwater* 55.5 (2017): 688-692.

8. Krause, Jonathan, et al. "The unreasonable effectiveness of noisy data for fine-grained recognition." European Conference on Computer Vision. Springer, Cham, 2016.

9. Montanari, Alberto, et al. "“Panta Rhei—everything flows”: change in hydrology and society—the IAHS scientific decade 2013–2022." *Hydrological Sciences Journal* 58.6 (2013): 1256-1275.

10. Newman, A. J., et al. "Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance." *Hydrology and Earth System Sciences* 19.1 (2015): 209.

11. Perrin, Charles, Claude Michel, and Vazken Andréassian. "Improvement of a parsimonious model for streamflow simulation." *Journal of hydrology* 279.1-4 (2003): 275-289.

12. Raleigh, M. S., J. D. Lundquist, and M. P. Clark. "Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework." *Hydrology and Earth System Sciences* 19.7 (2015): 3153.

13. Samaniego, Luis, Rohini Kumar, and Sabine Attinger. "Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale." *Water Resources Research* 46.5 (2010).

14. Shen, Chaopeng, et al. "HESS Opinions: Deep learning as a promising avenue toward knowledge discovery in water sciences." (in review)

15. Sivapalan, Murugesu, et al. "IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences." *Hydrological sciences journal* 48.6 (2003): 857-880.

16. WMO and UNESCO. "International Glossary of Hydrology." ISBN 978-92-63-03385-8 http://www.wmo.int/pages/prog/hwrp/publications/international_glossary/385_IGH_2012.pdf