AR1

General comments This article discusses the potential benefit of deep learning models to let emerge knowledge about water science systems from hydrological data. The paper is well written, the opinion is clearly stated and the authors present their arguments based on their expertise and their understanding of deep learning techniques. I'm wondering to what extent this is new and original compared to the opinion paper of Marçais & Dreuzy (2017 - see reference below). For example, the figure presented in this former article expressly conveys the idea that DL methods could enhance the unraveling of hydrological properties from data which is the core of this current article.

Marçais, J., & de Dreuzy, J. R. (2017). Prospective interest of deep learning for hydrological inference. Groundwater, 55(5), 688-692.

The roles taken by these papers are very different. We are cheerful to see others echoing the same enthusiasm for DL in hydrology. We would welcome others to have discussion and work together on this topic. In terms of the paper, though, we see significant differences between these articles, which is summarized in the table below. In our opinion, the Marcasis and de Dreuzy 2017 (MD17) paper was a timely and welcomed first "call into the wild" (despite that one of the co-authors of this Opinion paper has had a DL paper, Tao et al. in 2016). While it has fulfilled the mission, that paper was very brief and had a different focus. It did not explain why DL could unravel hydrologic properties. It did not mention interrogative studies which is a crucial part of our argument. It also did not discuss what we need to do as a community to incubate such research. In this article we gather from our past working to voice some enthusiasm as well as challenges.

However, prompted by the reviewer (and other comments), we will revise this Opinion paper significantly to emphasize our main points, which are: (1) DL+interrogative study is a valuable research avenue; (2) what challenges face the community and what we can do together to incubate DL research; (3) water resources present unique challenges and opportunities for DL.

Table. Difference between papers

Paper	Unique ideas
This HESS Opinion	(As its title indicates, this is truly an opinion paper. We need to assume readers have access to Shen's review paper) 1. Opinion: DL is not a hype. Supported by a review of its solid progress, winnings of competitions and adoption in daily uses. 2. Proposition of the complementary, data-driven scientific avenue: the integration of interrogative studies into the avenue. 3. Following unique Opinions are about what we can do as a community: a. scientific methods: hypotheses come from machine learning. We do not pose an opinion before doing data mining. Difference from earlier ML: now we have DL to automatically extract features. b. call for open competition of DL in hydrology with criteria focusing on both performance and explainability c. collecting big data through data sharing and citizen scientists 6 (**to be enhanced**). Water science provide unique challenges and opportunities for DL. 7 (**to be added**): Roadmap toward DL-supported science discovery. & Practice challenges and research thrust as a community
Marcasis and Dreuzy 2017	Main points: DL can be used for prediction issues; it may contribute to initial choice and alternatives of physical model structures; model reduction; emergent system properties; calibration. (however, each was only mentioned in one sentence). Test on hydrologic numerical data; benchmarks
Shen. 2018 Review	1. Technical details on ML and DL 2. Trans-disciplinary review of DL applications and experiences in sciences 3. Technical details of progress: interpreting DL and GANs 4. (revision) prospects for DL to help tackling grand challenges facing water sciences: inter-disciplinarity, human dynamics, data deluge (from novel sources), scaling and equifinality issues, non-unique inversions and high-dimensional, multi-modal data.

I also feel that DL techniques and especially why it does work so well is still not understood by computer scientists and mathematicians. However, this article can give the impression that the "DL reasons of success" are now understood (see specific comments) paving the way for knowledge discovery in water sciences through its use. I would consider being more cautious about that as the understanding of the specific properties of DL models compared to more traditional statistical learning models is still an active area of research. This does not mean that DL has not to be widely tested for hydrologic purposes.

There have been some studies that looked at why DL is powerful, but the point is taken. We will add clarification to this regard on the lines of 'why DL works so well is not fully understood. there are some suggestions… However in water research it needs to widely before trusted.'

Specific comments Page 1 L.20 Could you specify articles where DL shows capacities for scientific discovery?

It was included in the review paper Shen2018. There were quite a few examples. Here, we will include some summary of these in the revised Opinion article.

Page 2 L.9-16. The paragraph gives the impression that DL is a "plug and play" model whereas to my knowledge building a DL model still requires intensive computer scientists' knowledges and requires use of GPUs.

Good point. We will add the following sentence to this sentence:

"*While showing many advantages, DL models will require substantial amount of computing expertise. The tuning of hyper-parameters, e.g. network size, learning rate, batch size, etc., often require a priori experiences and trial and error. The computational paradigm is also substantially different from ordinary hydrologists' educational background.*"

Page 2 L.24. I don't think that generalization capacities of DL come from its interpolation capability. Indeed, classical neural networks have been proven (see citation below) to be universal interpolators but they do not generalize well. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5), 359-366. Page 3 L.25. I agree that increase in environmental data opens new opportunities for data-driven techniques in general and particularly for DL techniques. Along with the development of spatialized, remote sensing data, I would also insist on the development of environmental observatories that collect a lot of time series, monitoring data even though they are site specific. These two types of data are complementary to advance through knowledge discovery in hydrology.

The original sentence was "*Moreover, the differentiable nature allows for greater success for interpolation and mild extrapolation, contributing to the strong generalization capability of DL.*". The differentiability "contributes" to the generalization ability, but that is not the sole reason. Other factors include improved architecture, regularization, big data, weights sharing etc., which were mentioned earlier. To avoid any confusion, this sentence will be revised as

"*Moreover, the differentiable nature allows for greater success for interpolation and mild extrapolation, partially contributing to the strong generalization capability of DL.*"

Page 6. L.6-17. This paragraph is intended to bridge the link between interrogative techniques brought in DL by the "AI neuroscience" subdiscipline and the potential of DL for knowledge discovery in water sciences. If the arguments tend to prove that such interrogative techniques enlighten the way the architecture of DL works, it does not explain the success of DL in itself. For example, the sentence

L.13: "activations of recurrent neural networks can be visualized to show the control domain of certain cells, which explains its functioning" is not correct. This only explains the functioning of the architecture of the DL, not the reason of success of such a method. There is some literature exploring the need for explanation of DL techniques. For Convolutional Neural Networks (CNNs), their understanding can be linked with wavelet theory (see reference below). Especially their capacity to extract invariants through a lot of different scale in high dimensional datasets but this is still a subject of active research. This capacity could explain their generalization capabilities especially for image datasets.

The interpretive study does not solely focus on "why DL was successful", and this is not really the point. The interpretive studies answer "what has DL learned". We would argue that for scientists the second question is more important than the first. To clarify, we will revise this part to separate out the two question separately.

Mallat, S. (2016). Understanding deep convolutional networks. Phil. Trans. R. Soc. A, 374(2065), 20150203. Page 7 L.8-18. It could be interesting to explore how DL techniques can improve hypothesis testing through an exploration of competing process-based models? The Structure for Unifying Multiple Modeling Alternatives (SUMMA) (see reference below) could be a start to generate process-based models with alternative hypotheses. For example, process-based models could be used to feed DL models with numerical generated data.
Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... & Arnold, J. R. (2015). A unified approach for process-based hydrologic modeling: 1. ˇ Modeling concept. Water Resources Research, 51(4), 2498-2514

Page 10 L.20-Page 11 L.24. I would add to this list the fact that water sciences provide to DL a unique challenge because hydrologic data are intrinsically heterogeneous. Building a model able to integrate these heterogeneous data might be the key toward knowledge discovery in water sciences and toward big progresses in AI.

Good point! This part will be expanded to include this point.

Our plan of re-organization:

Section 1. Overview

current overview, with more discussion about promising attributes of DL

Section 2. The emergence of a complementary research avenue

More examples of DL success, some potential uses of DL in hydrology. Some possible interrogative study methods to show the promise.

Section 3. Challenges and opportunities

Expand on original section 4. There are many old and new challenges, many of which cannot be resolved by individual research groups: regionally-imbalanced dataset; strong heterogeneity and contextual variables; partial observations; computational challenges; data access; myriad configurations and "tricks"; lacking training data, especially unlabeled data; problem complexity; missing dynamics; large variation in performance based on DL configurations; Non-stationary world and increasing extremes are beyond previous observations.

Section 4. A community roadmap to DL-powered scientific advances in hydrology. How to solve challenges raised in Section 3 with a community-based approach

(i) synergy between PBM and DL
(ii) readily **accessible large dataset** with uniform formats: earth observations and monitoring networks. assimilate large amount of data to learn true patterns.
(iii) community-shared baseline DL models and data-processing pipelines
(iv) Open and transparent modeling **competitions** in water to facilitate algorithm comparisons, with evaluation on both **performance** and **interpretation** → we need to recognize the significant roles played by competitions in the development of DL research.
(v) Develop a baseline suite of DL interpretation and visualization software that support mainstream DL models, especially those that interpret the hidden layers.

Sections 3 & 4 will be greatly enhanced.