"Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling" by F Kratzert, D Klotz, G Shalev, G Klambauer, S Hochreiter and G Nearing

Review by Hoshin Gupta

Brief Summary of the Paper:

[1] This paper presents a novel approach to the problem of catchment-scale modeling.

[2] The classical "hydrological community" approach is to develop conceptual models (CM) of catchment-scale input-state-output behavior that have fixed/rigid structures and parameterizations – that reflect our "scientific/physical" understanding of internal catchment structure (architecture) and functioning (processes and their interactions) – and to then apply these rigid pre-specified structures to different locations by altering the values of the (largely empirical) static parameters that are initially left unspecified (except typically to within "feasible ranges").

[3] Major challenges associated with this CM approach have been discussed in the literature, including the fact that the proposed rigid model structures are difficult to update/correct based on their inability to reproduce observed input-output dynamics sufficiently well (*Bulygina and Gupta 2009, 2010, 2011; Gupta and nearing 2014; nearing and Gupta 2015*), and that the free (optimizable) static parameters of such models have proven challenging to regionalize or relate to observable static data that is expected, based on hydrological understanding, to be (directly or indirectly) informative regarding differences in catchment functioning at different locations.

[4] In contrast, the authors use a machine-learning (ML) approach, based in Long Short-Term Memory Networks, that enables learning, from time-series input-output data, the system structural patterns associated with the observed dynamical system behaviors. So, while classical catchment CM's have "universal structural forms" that have been posed as hypotheses by scientists observing numerous examples of catchments across the world (or across a give region), the ML approach presented herein actually detects and learns the dynamics related attributes of such a universal catchment structure by being given access to time-series data from a great many such catchments.

[5] So, while classical catchment models are highly regularized (structurally constrained) using prior knowledge and the only remaining learning problem is to find values for the model parameters, the machine-learning approach presented herein must ***both*** learn the appropriate system structure and the appropriate location specific parameters necessary for the resulting model to provide good location specific performance. The lack of strong prior regularization means that such models cannot be meaningfully trained on individual catchments and expected to give good performance, because the information necessary to unambiguously learn the "dynamical principles of catchment-scale hydrological behavior" are generally not going to be readily available in any single catchment data set.

[6] In a previous paper [*Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, HESS 2018*], the authors demonstrated that the LSTM type of artificial neural network is suitable for catchment-scale hydrological modeling because of its ability to learn the long-term input-output dependencies that are essential for modelling the storage effects in catchments (e.g., snow accumulation and melt). They also demonstrated that such an approach can be used for regional scale modeling, where a single learned ML model can be used to simulate the discharge at a variety of catchments, and that the single ML models encoding of process behavior at the regional scale actually helps to improve model performance at each individual catchment through transfer learning (i.e., the multiple-catchment data helps to regularize the problem, so that the broader knowledge of catchment-scale behavior serves to improve the stability of local catchment-scale simulations/predictions).

[7] In this paper, the authors extend on that work to:

(1) Demonstrate that such an LSTM can be adapted to be able to capitalize on the availability of observable ancillary data in the form of catchment attributes to produce accurate streamflow estimates over a large number of basins.
(2) Show that the ML model can provide statistically significantly better performance (across a large number of catchments) than several existing CM type hydrology models that embed prior knowledge regarding catchment hydrological structure

(3) Demonstrate the way in which the ML model makes use of information in the ancillary data about catchment characteristics to differentiate between different rainfall-runoff behaviors, thereby enabling the superior performance obtained.

[8] To do so, the authors test two approaches, one in which the LSTM-based model is provided data regarding static catchment attributes as additional inputs at every time step (requiring no modification to the typical LSTM architecture) , and a second that is developed as a modification to LSTM architecture in which the data regarding static catchment attributes is provided separately in a manner that controls (through the input gate) which parts of the LSTM structure are activated for any individual catchment. The call the latter an Entity-Aware-LSTM (EA-LSTM) because it explicitly differentiates between similar types of dynamical behaviors that differ between individual entities (watersheds).

[9] The second (EA-LSTM) approach also differs from the first one in that it allows direct posterior inspection of the ML-based model structure to investigate what the model has actually learned from the static catchment attributes. The authors do this by investigating the nature of the mapping from catchment attribute space into the ML-model learned embedding space in which catchments with similar rainfall-runoff behavior are clustered together, thereby facilitating a (data-driven) catchment similarity analysis.

[10] In brief, the authors show that:

(1) Both the LSTM and EA-LSTM statistically outperformed the *regionally-calibrated* CM-type benchmark models by a large margin, as assessed using the NSE performance metric.
(2) The multi-basin calibrated EA-LSTM even (statistically) outperformed the *individual-basin-calibrated* hydrological models (a more rigorous benchmark), as assessed using the NSE performance metric.
(3) The use of catchment attributes as static input features significantly improves overall ML-based model performance as compared with when the model is not provided with information regarding catchment attributes. While an anticipated finding, the demonstration is both satisfying and convincing.
(4) The newly proposed EA-LSTM approach provides much better interpretability and potential contributions to hydrological understanding and insight regarding catchment similarity compared to the less interpretable traditional LSTM, without significant sacrifice of performance as assessed using the NSE performance metric.
(5) The large boost in ML-model performance obtained by providing information regarding static catchment features is not simply due to *'remembering'* each basin instead of learning a general relation between static input features and catchment specific hydrologic behavior. Adding noise to the catchment attribute data causes only gradual deterioration in predictive performance. Further, striking improvements are seen for basins at the lower end of the performance spectrum which largely represent catchment types that are under-represented in the training data set.
(6) Regional differences in catchment behavior sensitivity to catchment attributes seems consistent with prior hydrological understanding (topography in the Appalachian Mountains, climate indices in the Eastern US, meteorological patterns as we move away towards the Great Plains, etc.). This observed sensitivity ranking is encouraging because most of the top-ranked features are relatively easy to measure or estimate globally from readily-available gridded data products.
(7) Certain groups of catchment attributes did not typically provide much additional information → these included vegetation indices like maximum leaf area index or maximum green vegetation fraction, as well as the annual vegetation differences. Further, most of the soil-related attributes were at the lower end of the feature ranking; this is interesting because soil characteristics are among the hardest features to characterize accurately at a regional scale.
(8) Clustering of "similar" catchments by the values of the EA-LSTM embedding layer provides more distinct results than when clustering by the raw catchment attributes, and seems to be well related to hydrologic behaviors, as assessed in terms of a set of 13 hydrologic signatures, indicating that the EA-LSTM embedding layer largely preserves the information content about hydrological behaviors, while overall increasing distinctions between groups of similar catchments. Further, the EA-LSTM seems able to learn complex interactions between catchment attributes that allows for grouping different basins in ways that account for interactions between different catchment attributes.

[11] In addition, the authors demonstrate that when training ML models to learn system structure regarding dynamical catchment behavior from large data sets (large numbers of catchments), it is important to account for the achievable differences in model performance at each catchment by adjusting the training performance metric. In this regard they propose a modified NSE loss function that seeks to account for the differences in means and variances of the observation data across basins, and that the performance (as assessed using MSE) is typically smaller for basins with low average discharge. By using the ***average*** of the NSE values at each basin that supplies training data as the ML-model training metric (referred to as the NSE*), the authors show that:

(9) Training against the basin-average NSE* loss function improves overall ML-model performance as compared with training against an MSE loss function, especially in the low NSE spectra. In particular there is a significant reduction in the number of basins that are classified as 'catastrophic failures' (i.e., basins with an NSE value of less than zero).

(10) Because the model outputs, and therefore the number of catastrophic failures, differ depending on the randomness in the weight initialization and optimization procedure, running an ensemble of LSTMs can substantively reduce this effect.

My Assessment of the Paper:

[12] I believe that this paper represents a very significant contribution to the Earth System literature related to the development of Dynamical Environmental Systems Models (DESMs). I have alluded to some of the problems associated with the conventional CM approach in paragraphs [2-6] above. In this regard, there has been increasing community interest in the use of both "*large sample*" data sets and the use of "*model-structural-correction-via-data-assimilation*" (learning from data) to extract better understanding about the structure and functioning of hydrological systems, such as catchments.

[13] This paper bridges the challenges of learning from large sample data sets and learning how catchments structures/behaviors can differ at local to regional scales in a very meaningful way. While not addressing the problem of prediction in un-gaged basins directly, the ability of the EA-LSTM to learn from and characterize differences in catchment functioning encoded in catchment attribute data is highly significant, and it would seem that a natural next step would be for the authors to demonstrate that potential by running experiments that seek to demonstrate that predictive ability learned from gaged locations can be transferred to un-gaged locations. I look forward to reading more about this in the future.

[14] As such, I have only a few suggestions to offer the authors. The first is that the current title "*Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling*" presents a rather technical front to what is arguably (in my opinion) a much more significant piece of work. I therefore offer up the possibility for the authors to consider that the introduction and discussion/conclusions sections be somewhat revamped/broadened to reflect the perspectives offered in my above summary of the paper. As indicated, I do think this paper is really more about the interesting challenges of learning and characterizing (via dynamical systems models) the "behavior and functioning" of hydrological systems at the catchment scale in such a manner that ***both*** universal (fundamentally hydrological) principles, and local-to-regional scale uniquenesses of such systems can be learned by accessing the patterns of information encoded in large sample data sets (*Gupta et al 2014*). In this regard the title could also then be generalized to reflect the nature of the conversation about "*Learning Universal, Regional and Local Hydrological Behaviors via Machine-Learning applied to Large Sample data Sets*". Or this more general discussion could be saved for a future publication ☺.

[15] The second is that while the basin-average NSE* loss function does seem to serve the immediate needs of this study, I think that the ML-approach (and more generally hydrological learning from catchment data sets) can benefit from a more thoughtful approach to the problem of model performance metrics. In particular, the use of the observed output data "mean" as a benchmark for constructing the NSE itself, and the use of the output data variance to "normalize" across catchments to obtain somewhat comparable metric values to be averaged (or otherwise summarized in some statistical manner) seems, to me, problematic. In this regard, I think an Information Theoretic approach might ultimately prove to be more meaningful. I point out that the value of the metric, when used as the basis for assessing across different catchment locations, would be much enhanced if it somehow recognized the relative differences in complexity/difficulty associated with modeling the dynamical input-state-output behaviors at

different locations (due to climatic, geological, and other factors). As discussed by *Schaefli and Gupta (2007)*, the problem is at least partly one of appropriate benchmarking in order to make metric values meaningfully comparable. Some types of catchments (such as humid ones perhaps) are relatively easy to model to the level of obtaining high performance (e.g. NSE) values, while others (such as arid ones perhaps) are much more difficult to model … potentially requiring more complex model structures, more data, and perhaps better data quality. Since the challenge here is learning hydrological principles from the data, and some catchment systems are easier to characterize using simpler model structures, it would seem prudent to figure out how to account for this knowledge in the designs of our learning systems, which includes the metrics used as the filter through which information is being extracted.

[16] Finally, I think that the aforementioned issue may also relate to the fact that certain catchment attributes tend to be dominant indicators of differences in catchment behaviors, while others seem to show "lower importance" (sensitivity). It is been well known that "climate" (and one would reasonably expect also "topography") is the dominant indicator of catchment similarity, but this does not really help us to understand what structural differences in catchments drive differences in their behaviors. The finding that soil and vegetation characteristics are low on the "importance" list is interesting, as it suggests that the existing catchment attributes being used may not be sufficiently informative about catchment-scale soil and vegetation contributions to hydrological behaviors. So, is it a problem of poorly encoded soils and vegetation information at the catchment scale, or is really the case that such soils and vegetation do not play as big a role in hydrological behavior as we might expect? It would be interesting to consider how this issue could be better investigated using the ML approach.

[17] In conclusion, I commend the authors for a very interesting and thought-provoking article, and I recommend the paper for publication after only minor revisions, in which the authors can chose to incorporate some of my review comments (or responses to them), or not, as they so choose.

Best Regards

Hoshin Gupta

References:

Gupta HV, C Perrin, R Kumar, G Blöschl, M Clark, A Montanari and V Andressian (2014), *Large-Sample Hydrology: A Need to Balance Depth With Breadth*, special issue of HESS-ESD 'Predictions under change: water, earth, and biota in the anthropocene; Eds: M. Sivapalan, T. J. Troy, V. Srinivasan, A. Kleidon, D. Gerten, and A. Montanari, Hydrology and Earth Systems Science, 18, 1–15, www.hydrol-earth-syst-sci.net/18/1/2014/ doi:10.5194/hess-18-1-2014

Schaefli B and HV Gupta (2007), Do Nash values have value?, Hydrological Processes, 21(15), 2075-2080, simultaneously published online as Invited Commentary in Hydrologic Processes (HP Today), Wiley InterScience, doi: 10.1002/hyp.6825

Bulygina N, and H Gupta (2011), *Correcting the mathematical structure of a hydrological model via Bayesian data assimilation*, Water Resources Research, 47, W05514, doi:10.1029/2010WR009614

Bulygina N and HV Gupta (2010), *How Bayesian Data Assimilation Can be Used to Estimate the Mathematical Structure of a Model*, Stochastic Environmental Research and Risk Assessment, 24:925–937 DOI 10.1007/s00477-010-0387-y

Bulygina N, and HV Gupta (2009), *Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation*, Water Resources Research, 45, special issue on 'Uncertainty Assessment in Surface and Subsurface Hydrology', W00B13, doi:10.1029/2007WR006749.

Gupta HV and GS Nearing (2014), Debates—The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, Invited Commentary, Water Resources Research, 50, doi: 10.1002/2013WR015096

Nearing GS and HV Gupta (2015), The Quantity and Quality of Information in Hydrologic Models, Water Resources Research, 51, 524–538, doi:10.1002/2014WR015895