Deep Learning, Hydrological Processes and the Uniqueness of Place.

Keith Beven
Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ.

Corresponding author: Keith Beven k.beven@lancaster.ac.uk

Keywords: Machine learning, hydrological models, process representations, hysteresis

Funding information: Funded is acknowledged from NERC Grant No. NE/R004722/1

Looking backwards

One of the things that we learn from the history of science is that, with some notable exceptions beloved of philosophers of science, knowledge and understanding progress over time. Looking back we see that understanding of the natural world has (mostly) progressed. Sometimes alternative theories have awaited experimental confirmation; sometimes a new experimental technique has led to significant theoretical advances. We hope, of course, to see some of that progression, and to make a contribution to it, over the time scale of our own careers in science. It is therefore somewhat disconcerting to have something you wrote more than 30 years ago cited (in Nearing et al., 2020) as if the comments were relevant today. Things should have changed, even in hydrology.

The context is that of the availability of the new techniques of machine learning and deep learning and their application to hydrological data. Nearing et al. (2020) suggest that in many respects not much has actually changed since I wrote about the need for a new paradigm in hydrological modelling in 1987 (Beven, 1987). They go on to suggest that machine learning and deep learning can produce models that perform just as well, if not better, than conceptual and process-based hydrological models, including for catchments treated as ungauged (see also Kratzert et al., 2019a,b).

Should this be considered surprising? Not necessarily in the case of individual catchments – if there are consistent anomalies or epistemic uncertainties in catchment data that mean, for example , that water balance constraints are not well met, then a DL model can compensate for those anomalies in ways that a conceptual model, constrained by water balance cannot. If there are consistent anomalies between the conceptual structure of a hydrological model in a particular catchment and the nature of the hydrological processes in that catchment then again a DL model might well be able to capture that behaviour better than a deficient process description (although it is worth noting that DL models are also subject to choices in structure and multiple hidden parameters; that is what gives them flexibility in fitting the training data). Nearing et al. (2020) point out that there are techniques for incorporating conservation constraints into physically-constrained DL models (see also Wang et al., 2020), but given the epistemic uncertainties in water and energy balances, then this might not necessarily be advantageous in obtaining better DL predictions if, for example, the observational data do not themselves provide consistent mass and

energy balance closure. Indeed, recognising this, and how to respond to it, might already represent an advance (see, for example, the discussion of Beven, 2019).

Deep Learning and the Ungauged Catchment

Somewhat more interesting is the result that for catchments treated as ungauged, a DL model can provide better hydrograph predictions than the types of methods traditionally used in such situations, at least given enough training data sets. In fact, Kratzert et al. (2019a) suggest that the DL methods seem to perform better when many quite different catchment data sets are used in training. They can also be made specific to particular types of catchment characteristic and forcing data inputs (e.g. using modelled precipitation inputs or radar products rather than that observed from raingauges). Should this also not be considered surprising? DL trained on a wide range of catchment conditions and responses does have the potential to reduce the impact of anomalies in the data set, which could lead to overfitting in the case of conceptual models calibrated on single catchments and the extrapolation of overfitted parameters to the ungauged case. In doing so they might be able to identify what Young (2013) calls the dominant modes of response in the application of his Data-based Mechanistic (DBM) modelling strategy. The dominant modes of response for catchments are, after all, rather simple. It rains and, if the antecedent conditions are wet enough, there is a consequent hydrograph. The difficulty in all hydrological modelling is in the identification of how the processes that control the volume of that hydrograph change with the rainfall (or snowmelt) inputs and antecedent conditions. Timing of the hydrograph seems to raise less difficulties (though see the discussion of Beven, 2020). In this respect, therefore, the problem has not really changed since the introduction of the Rational Method by Mulvaney (1851) and Kuichling (1896). The coefficient of the rational method could equally be the predictand of a machine learning analysis using catchment characteristics and antecedent states as inputs. As pointed out by Nearing et al. (2020) that could also be set up as a data assimilation exercise to improve the predictions of the next event.

Kratzert et al. (2019a,b) and Nearing et al. (2020) have reported that the only catchment characteristics that seem to have predictive capability are vegetation type and seasonality, while it has been difficult to detect (and therefore predict) any changes in response, since the past data do not necessarily contain the information relevant to expected future changes. Something does seem surprising here. A priori we would expect the geology of an ungauged catchment to have an important effect on its response characteristics, especially the recession characteristics and relationship between storage and runoff coefficients. It is clear also that this is not simply a matter of proximity or regional catchment characteristics, recession characteristics can vary dramatically over short distances and with scale in some regions, depending on changes in geology and soils (see, for example, Oudin et al.; 2010; Jensco et al., 2011; Bergstrom et al., 2016; and the study using the CART data analysis method of Fang and Shen, 2017). This information is not available to their DL algorithms (other than something called “geological permeability”), so how does it do so well on predicting the responses of ungauged basins where the geological characteristics are not well defined?

There would appear to be two answers to this question. One is that, in fact, it does not do that well. The distribution of NSE values for predictions of catchment treated as ungauged case ranges from <0 - >0.95 for the best Long Term Short-Memory Network (LSTM) DL model (Kratzert et al. 2019a,b). So DL has not solved the ungauged catchment problem, but it did show better performance than the two conceptual models that were compared (and that will be subject to similar problems of the input data and, from the choice of conceptual assumptions, the potential to have quite the wrong structure for specific catchments). Clearly there must be other issues at play here.

Secondly, the sensitivity coefficients for the DL show distinct variations in behaviour across the catchments in the training data set. This reflects the range of hydrological responses across catchment characteristics and scales (although there is no need to specify classes using this methodology; Kratzert et al., 2019b also show that the LSTM model performs somewhat better than a k-means clustering algorithm for several hydrological signatures). The essence of predicting an ungauged catchment is then to map that catchment into the DL model space. This is analogous to the mapping suggested by Beven (2000) as a way of assessing the uncertainty in parameterisations of a conceptual model (and then thinking about how that uncertainty might be further constrained). Such a mapping is necessarily based on the catchment characteristics that are available to the DL as indices. But most catchment indices are only surrogate variables and geology, for example, is one characteristic that is not very well reflected in indices that can be used for prediction (a Base Flow Index cannot serve that purpose, since it will not be available a priori for ungauged catchments but must itself be predicted). In this respect the use of a DL approach might indeed be advantageous because the nature of the training data can compensate for the lack of direct hydrological relevance of such indices (and in doing so have more flexibility than, for example, the multiple regressions to estimate the model parameters used in other more traditional regionalisation approaches).

Mapping unique catchments into a model space

This idea of mapping a catchment (gauged or ungauged) into the model space was originally raised by Beven (2000, 2001, 2002a,b) in the context of uniqueness of place in hydrological simulation. It was extended in Beven and Freer (2001) in the application of a conceptual hydrological model. The idea was that, given our limited knowledge of catchment characteristics and understanding of catchment responses, it would be impossible to map that catchment to a single point in the model space (this could apply to both model structures and parameter sets). The uncertainty in model predictions does not then come from the model space itself (even if providing stochastic rather than deterministic outputs), it comes from the mapping of an area of reality into the model space given the available information. The concept then allows consideration of how that mapping might be made more precise, perhaps by applications of likelihoods, hypothesis testing and model rejection, and the collection of additional data.

It is interesting to consider the application of a similar concept to DL models. Kratzert et al. (2018) give an example of these where a pre-trained LSTM model based on regional data (effectively defining a generalised model space) can be refined by more limited data for an individual basin; this gives better results than only training on the limited data for the

catchment itself, even though the regional model might include catchments with a wide range of behaviours. In addition, as Nearing et al. (2020) suggest, DL models, given enough training data, can be trained to represent not only deterministic responses but also the parameters of the uncertainty in outputs (variances or quantiles). The DL model space might also therefore be deterministic or stochastic, but the range of outcomes is well defined given a set of inputs. The nature of that space will, however, be dependent on how the DL has been trained and what objective functions have been used.

In predicting an ungauged catchment, with its unique characteristics, we then have the problem of mapping that catchment, for which we have limited knowledge, into that model space. We should surely not expect that the mapping will be to a precise point in that space, but that there will be some uncertainty associated with our lack of knowledge. This is one way of structuring our understanding of how that catchment might function when that mapping is applied in a thoughtful way. To some extent a stochastic DL should be able to reflect the variation seen for somewhat similar gauged catchments in the training data set, given a big enough sample, even though all those gauged catchments have their own uniqueness of place poorly reflected in characteristic indices. This will be the case for those ungauged catchments that are similar in the way defined by the training process (remembering that for all machine learning methods extrapolating beyond the range of the training data for nonlinear systems may not be well constrained). In this sense therefore DL models provide a model space constrained by the training data into which the characteristics of some new ungauged catchment can be mapped with the aim of predicting the response of that catchment without worrying about any process information. Nearing et al. (2020) posit that DL might be able to do this better than the hydrologist, though again information about the effect of geology on hydrological response might be particularly limited (for both!).

More process information could, of course, be used if it was available in some form for all the training data catchments and the catchment of interest. In applying process-based models we have traditionally expected the right sort of response to be predicted by specifying the physical parameters of a catchment. That this has proven rather difficult should not actually be surprising, it was already underlying my call for a new paradigm in modelling in 1987. Since then, computational power has hugely increased, but knowledge of catchment processes and characteristics in most catchments has not (with some notable research catchment and critical zone observatory exceptions). The difficulty in predicting the future responses of catchment management strategies that might change those characteristics should therefore be even less surprising. It is evident that DL can then only help us in that when such changes are reflected in the indices used to control the pathways through the DL network (and when there are informative relevant cases in the training data).

Looking Forwards

The success of DL over conceptual and process-based hydrological models (and the continued failure to get acceptable simulation results for some catchments, see also Oudin et al., 2008 using more conventional methods) would then appear to pose a number of important questions for the future of hydrological science.

1. Why are some catchments so consistently poorly simulated?
2. How far can DL results be interpreted to derive inferences about processes and scale in particular catchments? Can we learn about a more "correct", possibly scale dependent, physics using ML/DL methods?
3. What process information might prove useful in improving DL (and other) models, particularly in making predictions about a changed future?
4. How best to allow for local information reflecting uniqueness of place?

Question 1 points again to lack of knowledge. That might be lack of knowledge about inconsistencies in the data or lack of knowledge about how to represent processes. That DL models are not able to compensate for that lack of knowledge in some catchments suggests that the data issues might be the first thing to look at (see also Beven, 2019; Beven et al., 2019). Indeed, experience suggests that applying data learning methods can reveal anomalies of interest in a data set (e.g. Iorgulescu and Beven, 2004). We need to do better in assessing actual catchment inputs, observed outputs, and evaluating the importance of data uncertainties and unmeasured fluxes in fitting models of any type. DL models might be informative in this respect, in that they might suggest where data inconsistencies are most significant relative to the constraints of water and energy balances.

The importance of such data inconsistencies also means that we should be careful about inferences made in addressing Question 2. Inconsistencies and anomalies will affect such inference, even if we might be prepared to accept that DL might yield better scale dependent functional descriptions than conceptual model formulations. Of course, if all we need is to have better forecasting models, then this is not important. We can simply take advantage of the better predictive capability. It is more important if we want to learn about processes with a view to predicting how those processes will change in future (and particularly when there may be no representative surrogate of the future in the training set). I think that there are situations where this might be possible. For example, it might be possible to learn about a generalised form of hysteresis in storage discharge relationships directly from analysis of the data. Both observational data (eg. Beven, 2006) and model predictions (e.g. using the MIPs model in Davies et al., 2015) suggest that the hysteretic behaviour might be complex and state dependent, especially in catchments with large storage volumes, but might provide more insight than using a simple functional relationship between storage and discharge (see for example the variability that is neglected in doing so in Figure 6 from Kirchner, 2009).

It might also be that making inferences from DL models might be easier if process information about catchment responses could be incorporated more directly (Question 3). Indeed, Nearing et al. (2020) (in a Section called "Skip the hydrologist?") suggest that it needs to be demonstrated that any form of process information can actually benefit prediction models based only on DL. This has long been discussed in the context of conceptual models, but there is then a significant problem of incommensurability of observed and predicted variables. Incommensurability is less of a problem for DL models, it is only required that there should be useful information that can be extracted from whatever process information could be made available. How that information is used (or

not) would be useful making inferences about process (if, of course, we have already addressed the issues of Question 1).

But what process information might be useful (given the resources to make it available)? Experience suggests that point information on either states or characteristic parameters is not so very useful (e.g. the distributed water table information of Lamb et al. (1998), Blazkova et al., 2002a; and Freer et al., 2004). More integrative measures such as patterns of saturation (Güntner et al., 1999; Blazkova et al., 2002b), bulk subsurface storage (e.g. Güntner et al., 2017), connectivity of hillslopes (Hopp and McDonnell, 2009; Jensco et al., 2011; Tetzlaff et al., 2014; Bergstrom et al., 2016) or residence time distributions (Benettin et al., 2017; Harman, 2019) might be more useful. They require considerable effort, but perhaps this effort would be justified if we really want to improve our predictive models and learn more by inference from DL approaches. It would be interesting, for example, to see a DL study of a catchment where the time scales of hydrographs and water table responses (and also residence times over time) are quite different to see what insights could be gained (see McDonnell and Beven, 2014; Beven, 2020).

But in addressing how to generalise in this way, in Questions 1 to 3, we will not entirely escape the issues of Question 4. How to deal with uniqueness of place in creating models of everywhere, to my mind at least, is one of the most interesting questions in hydrology (Beven, 2000, 2007; Blair et al., 2019). Others will not agree of course. Focusing on uniqueness of place is not a way to understand any universal laws of hydrology (in so far as they might exist, see Dooge, 1986). It might rather be viewed as trying to understand the anomalies of those laws. But there may be significant process (or data) information in those anomalies.

Uniqueness of place and hydrological understanding

I am writing this in Mallerstang, the upper Eden valley in Cumbria. In the last few days I have walked over both sides of the valley. There is an elevation difference of about 500m between valley bottom and the broad peaks on either side. The geology is Carboniferous, with limestones and gritstones predominating, but it also shows plenty of evidence of the remnants of the last glaciation with valley bottom and lateral moraines and occasional erratics sitting on limestone pavement. Streams and springs appear from the limestone under very wet conditions, but the response is generally flashy. The limestone does not here provide large storage volumes that give large volumes of baseflow, but might still mean that the subsurface divide might differ from the surface divide. Between the villages of Ravenstonedale and Newbiggin-on-Lune, at the divide between the valleys of the Eden and the Lune, there is limestone overlain by glacial drumlines: the divide is difficult to locate and it is quite possible that it changes location with the sequence of wetting and drying.

The Eden in Mallerstang has mostly a mobile gravel bed, with some bedrock outcrops and is underfit with respect to its overbank areas, suggesting larger discharges in the post-glacial period. Tributaries are incised into the valley bottom moraines in places. There is a sharp change in direction of the river in the headwaters from south west to northerly flowing, suggesting that these headwaters were captured from the River Ure and Wensleydale at the end of the last glaciation. Vegetation is mostly rough pasture, with some improved pasture

in the valley bottoms, and eroded peat up on the fells. Large areas of the valley sides have been planted with young trees in the last couple of years. These areas are fenced, keeping out the sheep, so that the moorland grasses and reeds have grown much denser (increasing storage and roughness and making the walking harder!). Because of changes in farming subsidies, stocking densities of sheep on the fells have also declined. The hydrology will, therefore, be changing, depending on the variability of the inputs (and how many of the young trees survive on the fells).

So this is a typical upland catchment of the English Pennines, complex in its history and characteristics, and a bit different from adjacent and similar catchments. It suffers from similar problems of not knowing exactly what the patterns of inputs are. The most upstream flow measurement site (apart from the small Fellside Beck) is an Environment Agency compound Crump weir at Kirkby Stephen where the catchment area is 65.8 km^2. The subcatchment at Fellside Beck is currently being monitored in one of the areas planted with young trees. There are other, "similar", tributaries along the valley, that are not being monitored and will differ in detail in their shape, topography, surficial and subsurface geology, soils and patterns of tree planting. Only some of the relevant spatial characteristics can be assessed from terrain maps and remote sensing. But we want to make predictions about the future responses of those catchment areas without (as yet) really having sufficient information to do so.

One approach to the problem is to try and build in what information we do have into a distributed model of the processes. As Nearing et al. (2020) point out, however, there have been millions of dollars invested in such models without real demonstration that this approach is successful (this was already the case in 1987 but much more has been spent since). As I pointed out in Beven (2006), I do not think that is only a matter of a lack of local information in applying such models; there remain issues about how the processes are represented. DL can certainly bypass this problem. We can monitor at different scales (plot to 1st order to higher order streams); we can use the resulting data with DL (once we have enough data at least) to define a predictive model. The smaller scale might contain information useful at predicting the larger scale (at least under current conditions), but as we know only too well, other processes and sources of variability also come into play. Given sufficient (good quality) data we might indeed be able to identify which sites, or periods of time, stand out as anomalous and that need further investigation of their uniqueness. And where there is the possibility of collecting additional data, there may be the potential to evaluate which data would be most valuable in defining uniqueness and testing models and parameter sets as hypotheses about that place (e.g. Beven, 2018).

Learning from Deep Learning

The question then is whether this will make process models redundant (as suggested by Nearing et al. 2020) or whether we can analyse the structure of DL models to make inferences about what improved local process models should look like? The first may well prove to be the case when all we are interested in is prediction. But to predict the impacts of scenarios for future water management, where in some cases spatial patterns of implementation might be important (for example, avoiding synchronicity of sub-catchment peaks in flood management), the first would require that the training data incorporate

existing information about what those future scenarios might look like, though this will necessarily be at catchments with somewhat different characteristics. That suggests that there will still be a role for the second strategy of using DL models to improve process information and understanding. And, clearly, this is essential to being within scope for *Hydrological Processes*!

This could prove to be a fascinating, but rather difficult, area of research. Attempts to learn from DL models have, to date, been rather crude (e.g. the correlations between DL states and conceptual model states in Kratzert et al., 2019b do not really yield much in the way of process understanding). There is also the question of having multiple approaches to DL available (e.g. Shen, 2018), and different ways of training their (many) parameters. Training does not, after all, differ all that much from optimisation and calibration strategies other than in name. There will be similar problems of data uncertainties, parameter uncertainties and equifinality. But there is real potential to provide new insights about how catchments work by the use of DL methods. Given the limitations of the data we have, however, it remains to be seen how far that potential can be realised.

Acknowedgements
This paper was prepared as a contribution to the Q-NFM project led by Dr. Nick Chappell of Lancaster University (NERC grant no. NE/R004722/1).

References.

Benettin, P., Soulsby, C., Birkel, C., Tetzlaff, D., Botter, G. and Rinaldo, A., 2017. Using SAS functions and high-resolution isotope data to unravel travel time distributions in headwater catchments. *Water Resources Research*, *53*(3): 1864-1878.

Bergstrom, A., K. Jencso, and B. McGlynn, 2016, Spatiotemporal processes that contribute to hydrologic exchange between hillslopes, valley bottoms, and streams, *Water Resour. Res.*, 52: 4628–4645, doi:10.1002/2015WR017972.

Beven, K.J. 1987, Towards a new paradigm in hydrology, In: *Water for the Future: Hydrology in Perspective*, IAHS Publ. No. 164: 393-403.

Beven, K. J. 2000, Uniqueness of place and process representations in hydrological modelling, *Hydrology and Earth System Sciences*, 4(2): 203-213.

Beven, K. J. 2001, On landscape space to model space mapping, *Hydrological Processes (HPToday),* 15: 323-324.

Beven, K. J. 2002a, Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system, *Hydrol. Process.,* 16(2): 189-206.

Beven, K. J. 2002b, Towards a coherent philosophy for environmental modelling*, Proc. Roy. Soc. Lond. A,* 458: 2465-2484.

Beven, K J, 2006, The Holy Grail of Scientific Hydrology: $Q_t = H(\underset{\leftarrow}{S}\underset{\leftarrow}{R})A$ as closure, *Hydrology and Earth Systems Science,* 10: 609-618.

Beven, K. J. 2007, Working towards integrated environmental models of everywhere: uncertainty, data, and modelling as a learning process. *Hydrology and Earth System Science,* 11(1): 460-467.

Beven, K. J. 2018, On hypothesis testing in hydrology: why falsification of models is still a really good idea, WIRES

Water, DOI: 10.1002/wat2.1278.

Beven, K. J., 2019, Towards a methodology for testing models as hypotheses in the inexact sciences, Proceedings Royal Society A, 475 : 20180862. http://dx.doi.org/10.1098/rspa.2018.0862

Beven, K. J. 2020, A history of the concept of time of concentration, *Hydrology and Earth System Sciences*, in press

Beven, K. J. and Freer, J. 2001, A Dynamic TOPMODEL, *Hydrol. Process.,*15(10), 1993-2011.

Beven, K. J., Asadullah, A., Bates, P. D., Blyth, E., Chappell, N.A., Child, S., Cloke, H., Dadson, S., Everard, N., Fowler, H. J., Freer, J., Hannah, D.M., Heppell, C., Holden, J., Lamb, R., Lewis, H., Morgan, G., Parry. L., Wagener, T., 2019, Developing observational methods to drive future hydrological science: can we make a start as a community?, *Hydrological Processes, 34*(3): 868-873.

Blair, G.S., Beven, K.J., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., Dean, G., Hunter, N., Edwards, E., Nundloll, V., Samreen, F., Simm, W., Towe, R., 2019, Models of Everywhere Revisited: A Technological Perspective, *Environmental Modelling and Software*, https://doi.org/10.1016/j.envsoft.2019.104521

Blazkova, S, Beven, K, Tacheci, P and Kulasova, A. 2002a, Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): the death of TOPMODEL?, *Water Resources Research*, 38(11), W01257, 10.1029/2001WR000912

Blazkova, S., Beven, K. J., and Kulasova, A. 2002b, On constraining TOPMODEL hydrograph simulations using partial saturated area information, *Hydrol. Process.,* 16(2): 441-458.

Davies, J. and Beven, K. J. 2015, Hysteresis and scale in catchment storage, flow, and transport, *Hydrol. Process.* 29(16): 3604-3615, DOI: 10.1002/hyp.10511

Dooge, J. C. I. (1986). Looking for hydrologic laws. *Water Resources Research*, 22 (9S): 46S-58S.

Fang, K., & Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resources Research*, 53: 8064–8083. https://doi.org/10.1002/2016WR020283

Freer, J.E., McMillan, H., McDonnell, J.J. and Beven, K.J., 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*(3-4): 254-277.

Güntner, A., Uhlenbrook, S., Seibert, J. and Leibundgut, C., 1999. Multi-criterial validation of TOPMODEL in a mountainous catchment. *Hydrological Processes*, *13*(11): 1603-1620.

Güntner, A., Reich, M., Mikolaj, M., Creutzfeldt, B., Schroeder, S. and Wziontek, H., 2017. Landscape-scale water balance monitoring with an iGrav superconducting gravimeter in a field enclosure. *Hydrology and Earth System Sciences*, 21(6), 3167- .

Harman, C.J., 2019. Age-Ranked Storage-Discharge Relations: A Unified Description of Spatially Lumped Flow and Water Age in Hydrologic Systems. *Water Resources Research*, *55*(8): 7143-7165.

Hopp, L. and McDonnell, J.J., 2009. Connectivity at the hillslope scale: Identifying interactions between storm size, bedrock permeability, slope angle and soil depth. *Journal of Hydrology*, *376*(3-4): 378-391.

Iorgulescu, I. and Beven, K. J., 2004, Non-parametric direct mapping of rainfall-runoff relationships: an alternative approach to data analysis and modelling, *Water Resources Research*, *40 (8), W08403,* 10.1029/2004WR003094

Jencso, K. G., and B. L. McGlynn 2011, Hierarchical controls on runoff generation: Topographically driven hydrologic connectivity, geology, and vegetation, *Water Resour. Res*., 47, W11527, doi:10.1029/2011WR010666.

Kirchner, J. W. 2009, Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res*., 45, W02429, doi:10.1029/2008WR006912.

Kratzert, F., D.Klotz, C.Brenner, K. Schulz, and M. Herrnegger. 2018, Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, *Hydrology and Earth System Sciences*., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. 2019a, Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23 (12), 5089-5110.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. 2019b, Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*., *55*, 11,344–11,354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S. and Klambauer, G., 2019c, Neural Hydrology–Interpreting LSTMs in Hydrology. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 347-362). Springer, Cham.

Kuichling, E., 1889, The Relation between Rainfall and the Discharge in Sewers in Populous Districts; *Trans. Amer. Soc. Civ. Engrs.*, 20: 1-56.

Lamb, R., K.J. Beven and S. Myrabø, S., 1998, Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model., *Advances in Water Resources*, 22(4), 305-317.

McDonnell, J. J. and Beven, K. J. 2014, Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities, and residence time distributions of the headwater hydrograph, *Water Resour. Res*., 50, doi:10.1002/2013WR015141.

Mulvany, T.J. 1851, On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. *Proc. Inst. Civil Eng. Irel*. 4: 18–33.

Nearing, G. S., Frederik Kratzert2 Alden Keefe Sampson, Craig S. Pelissier, Daniel Klotz, Jonathan M. Frame, Hoshin V. Gupta, 2020, What Role Does Hydrological Science Play in the Age of Machine Learning?, *available at* https://eartharxiv.org/3sx6g/.

Oudin, L., Andréassian, V., Perrin, C., Michel, C. and Le Moine, N. 2008, Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, *44*(3).

Oudin, L., A. Kay, V. Andréassian, and C. Perrin, 2010, Are seemingly physically similar catchments truly hydrologically similar?, *Water Resour. Res*., 46, W11558, doi:10.1029/2009WR008887.

Shen, C. 2018, A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54, 8558–8593. https://doi.org/10.1029/2018WR022643

Tetzlaff D, Birkel C, Dick J, Geris J, Soulsby C. 2014, Storage dynamics in hydropedological units control hillslope connectivity, runoff generation and the evolution of catchment transit time distributions. *Water Resources Research.* DOI: 10.1002/2013WR014147.

Wang, N., Zhang, D., Chang, H., and Li, H., 2020, Deep learning of subsurface flow via theory-guided neural network, *Journal of Hydrology*, 584: 124700

Young PC. 2013, Hypothetico-inductive data-based mechanistic modeling of hydrological systems. *Water Resour. Res.* 49, doi:10.1002/wrcr.20068.