AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

COMMENTARY
10.1029/2020WR028091

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

What Role Does Hydrological Science Play in the Age of Machine Learning?

Grey S. Nearing[1], Frederik Kratzert[2], Alden Keefe Sampson[3], Craig S. Pelissier[4], Daniel Klotz[2], Jonathan M. Frame[1], Cristina Prieto[5], and Hoshin V. Gupta[6]

[1]Department of Land Air & Water Resources, University of California Davis, Davis, CA, USA, [2]LIT AI Lab and Institute for Machine Learning, Johannes Kepler University, Linz, Austria, [3]Upstream Tech, Natel Energy Inc., Alameda, CA, USA, [4]NASA Center for Climate Simulation, NASA Goddard Space Flight Center, Greenbelt, MD, USA, [5]IHCantabria Instituto de Hidrulica Ambiental, Universidad de Cantabria, Santander, Spain, [6]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA

Abstract This paper is derived from a keynote talk given at the Google's 2020 Flood Forecasting Meets Machine Learning Workshop. Recent experiments applying deep learning to rainfall-runoff simulation indicate that there is significantly more information in large-scale hydrological data sets than hydrologists have been able to translate into theory or models. While there is a growing interest in machine learning in the hydrological sciences community, in many ways, our community still holds deeply subjective and nonevidence-based preferences for models based on a certain type of "process understanding" that has historically not translated into accurate theory, models, or predictions. This commentary is a call to action for the hydrology community to focus on developing a quantitative understanding of where and when hydrological process understanding is valuable in a modeling discipline increasingly dominated by machine learning. We offer some potential perspectives and preliminary examples about how this might be accomplished.

Key Points:
- Hydrology lacks scale-relevant theories, but deep learning experiments suggest that these theories should exist
- The success of machine learning for hydrological forecasting has potential to decouple science from modeling
- It is up to hydrologists to clearly show where and when hydrological theory adds value to simulation and forecasting

Correspondence to:
G. Nearing,
gsnearing@ua.edu

Citation:
Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning?. *Water Resources Research*, *57*, e2020WR028091. https://doi.org/10.1029/2020WR028091

Received 5 JUN 2020
Accepted 25 OCT 2020
Accepted article online 13 NOV 2020

1. Beven's Clouds

On April 27, 1900 William Thomson (Lord Kelvin) gave his "Two Clouds" speech ("Nineteenth-Century Clouds over the Dynamical Theory of Heat and Light") at the Royal Institution, in which he argued that "The beauty and clearness of the dynamical theory, which asserts heat and light to be modes of motion, is at present obscured by two clouds." The two open problems in physics that Kelvin referred to were the failure of the Michelson-Morley experiment to detect the luminous ether ("how could the earth move through an elastic solid, such as essentially is the luminiferous ether?"), and the ultraviolet paradox ("the Maxwell-Boltzmann doctrine regarding the partition of energy"). Within a decade, Einstein had proposed fundamentally novel insights that led to two paradigm shifts that define modern physics to this day—the transformation of these two "clouds" into relativity and quantum mechanics.

In 1987, Keith Beven gave what might be considered hydrology's version of the Two Clouds speech at a symposium of the International Association of Hydrological Sciences (IAHS) (Beven, 1987). He took a perspective inspired by Thomas Kuhn's theory of scientific revolutions (Kuhn, 1962) to argue that "[t]he extension of laboratory scale theory to the catchment scale is unjustified and that a radical change in theoretical structure (a new paradigm) will be required before any major advance can be made in [predicting catchment-scale rainfall-runoff responses]." He proposed that two things would be necessary to push the field of surface hydrology into a new period of "normal science": (i) scale-relevant theories of watersheds ("[h]ydrology in the future will require a macroscale theory that deals explicitly with the problems posed by spatial integration of heterogeneous nonlinear interacting processes") and (ii) uncertainty quantification ("[s]uch a theory will be inherently stochastic and will deal with the value of observations and qualitative knowledge in reducing predictive uncertainty.")

Unfortunately, hydrology has not had its Einstein (with all due respect to Einstein, 1926, 1950). Nine decades from the establishment of the Hydrology section of the American Geophysical Union and after more than a half-century of computer-based hydrological modeling (Crawford & Burges, 2004), Blöschl et al. (2019) listed as one of the 23 "Unsolved Problems in Hydrology": "what are the hydrologic laws at the catchment scale and how do they change with scale?"

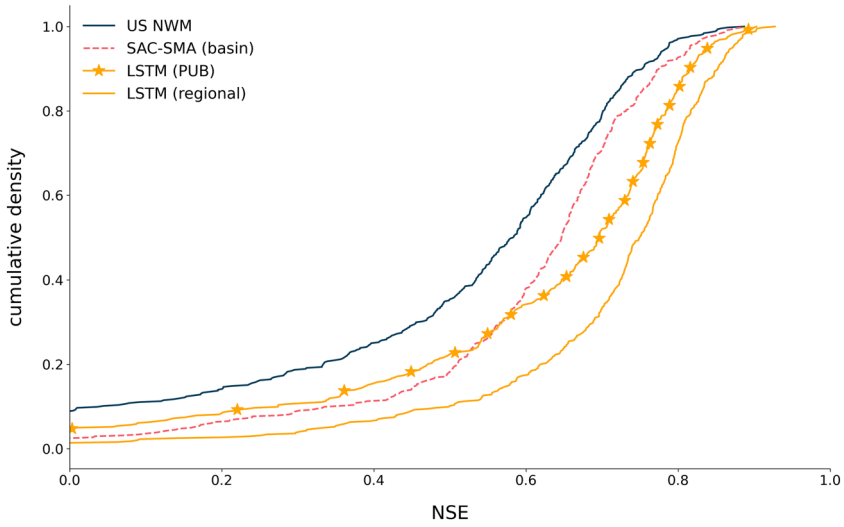

Figure 1. Results from Kratzert et al. (2019) showing the empirical and cumulative distributions of model performance (Nash Sutcliffe Efficiencies) over a 15-year test period in 531 CAMELS catchments. SAC-SMA is the Sacramento Soil Moisture Accounting model, NWM is the National Water Model Reanalysis, and LSTMs are Long Short-Term Memory networks (a type of deep learning architecture). The PUB LSTM was trained on data from out-of-sample catchments, whereas the Regional LSTM was trined on data from all catchments.

2. Tilting at Windmills

There are several potential reasons why the search for scale-relevant theories in hydrology has been unsuccessful, but lack of effort is not one of them (e.g., Beven, 2006b; Blöschl & Sivapalan, 1995; Dooge, 1986; Peters-Lidard et al., 2017; Sivapalan, 2006). One potential reason is simply that there might be no scale-relevant theories to find—it is possible that macroscale watershed behaviors are dominated by heterogeneity, meaning that there is little consistency across different basins. As summarized by Hrachowitz et al. (2013), "Beven (2000) highlighted the varying importance of different hydrological processes, active at different time scales in different catchments, and thereby emphasized uniqueness of place as a consequence of the variability of nature."

Alternatively, it could be the case that there are consistent macroscale patterns in hydrologic behaviors across watersheds, but we lack sufficient observations (type, scale, and scope) to discover these similarities. Again, as summarized by Hrachowitz et al. (2013), "[i]t was realized that increased physical model realism (and complexity) requires both more input data and more model parameters, which are rarely available with sufficient detail to account for catchment heterogeneity at the required resolution."

Uniqueness of place and lack of data are, in our experience, two of the most common hypotheses about why hydrology lacks both scale-relevant theories of watersheds. The alternative to such hypotheses is that these theories could exist and that there is enough information in available observation data that we could have discovered them, but that hydrologists simply have failed to do so. Prior to last year, it is fair to say that as a community we did not know which of these reasons was the cause of our lack of success. However, with the accelerating development of modern machine learning (ML) and deep learning (DL) in particular, we know that the reason is the third one listed: watershed-scale theories (and models) could have been derived from currently available observation data, but the hydrology community simply failed to do so.

The reason that we know this is because general models can be learned with DL. In a large sample study using 30 years of data from several hundred basins in the continental United States, DL gave better daily streamflow predictions on average in *ungauged* basins than traditional hydrology models when calibrated to long data records in *gauged* basins (Kratzert et al., 2019). That study used benchmarks based on (i) a modern process-based model that was the product of several millions of dollars of development funding and (ii) a conceptual model calibrated separately for each individual basin (Figure 1). These DL models have been benchmarked against a number of conceptual and process models calibrated both locally and regionally

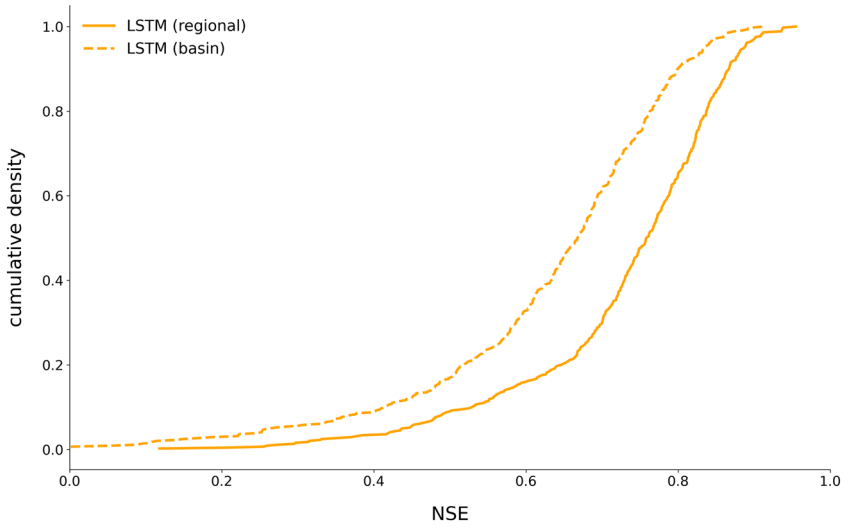

Figure 2. Cumulative distributions of NSE values over the same 531 CAMELS basins used by Kratzert et al. (2019) from a single model trained over all basins (single CONUS LSTM) versus separate models trained at each basin (per-basin optimized LSTM).

using a variety of metrics and hydrological signatures (Kratzert et al., 2019, 2020). The fact that DL learned to predict in unseen basins better than traditional models in gauged basins indicates that there exists interbasin consistency that we should be able to exploit and develop into a watershed-scale theory of rainfall-runoff behavior.

The problem of prediction in ungauged basins (PUB) (Hrachowitz et al., 2013; Sivapalan et al., 2003) is fundamentally a problem of extrapolation. Unlike both conceptual and process-based hydrology models and also unlike shallow ML models that the hydrological science community has used in the past (e.g., Hsu et al., 1995), DL models work *better* when trained on multiple catchments than when trained on individual catchments (Figure 2; also see the more thorough data scaling analysis by Gauch et al., 2019). This means that DL models learn relationships from a large sample of hydrological variability and are able to translate those learned relationships into better predictions in any individual basin. In contrast, traditional hydrology models are best when calibrated to individual basins, and performance always degrades when transferring to other basins or when using regional calibration.

It is often claimed that one of the reasons hydrology models do not extrapolate well is because they are over parameterized. Hrachowitz et al. (2013) reported that "several authors (Kirchner, 2006; McDonnell et al., 2007; Wagener et al., 2007) (These references are apparently incorrect in the quoted manuscript.) expanded on and strongly reiterated (Klemeš, 1986b) arguments that models which perform adequately well during calibration, but fail to predict the hydrological catchment response in validation, frequently do so because they do not sufficiently represent the real-world processes that control the catchment response. Rather, their often high number of parameters together with the limited number of constraints (including both calibration objectives and calibration criteria) resulted in high degrees of freedom, i.e., poorly conditioned parameter estimation problems, so that models "behaved more like mathematical marionettes." The problem is that this is not true. DL models generally have several orders of magnitude more degrees of freedom than calibrated conceptual models, and it is this *lack* of regularization that allows them to learn general and transferable hydrological relationships. Using DL as a benchmark demonstrates that it is the regularization in the traditional models (i.e., the hydrological theory that the model structures are based on) that is actually the cause of their lack of generality and transferability, rather than this being a problem of overparameterization. To summarize, this benchmarking between DL and traditional hydrology models demonstrates three things. First that hydrologists could have developed general, scale-relevant theories of watersheds from available data, but failed to do so. Second, that our understanding of why such theories do not exist were incorrect—neither uniqueness of place nor lack of data was a valid reason for this failure. Third, that our understanding of why our existing models perform poorly in extrapolation is also incorrect—

this is not due to a lack of regularization or to overparameterization, but instead due to bad theory—the regularization (structure) that does exist in these models actively hurts us.

3. Overlapping Magisteria: Faith and Fact in Hydrology

The type of benchmarking result outlined in the previous section is not new—hydrologists have been benchmarking ML models against both calibrated conceptual models and process-based models for at least a quarter century, and it has always been the case that ML generally performs better (e.g., Abramowitz, 2005; Best et al., 2015; Hsu et al., 1995; Nearing, 2013; Nearing et al., 2016, and many others). Todini (2007) framed the issue like this: "physical process-oriented modellers have no confidence in the capabilities of data-driven models outputs with their heavy dependence on training sets, while the more system engineering-oriented modellers claim that data-driven models produce better forecasts than complex physically based models. The key phrases in this sentence are "confidence in" and "better forecasts"—one is a statement of belief and one is a statement of fact.

As an applied science, hydrology is motivated by both *epistm* and *techn* (Parry, 2003). On one hand (*techn*), hydrologists are tasked (and funded) to tackle acute societal needs for managing water resources and mitigating water-related hazards. On the other hand (*epist*), many of us are curiosity driven and operate under the assumption that increased understanding of natural systems leads to increase predictability and control. These two objectives cannot be cleanly separated. Whether motivated by societal relevance versus pure curiosity the fact remains that scientific hypotheses are tested by their ability to make accurate predictions.

Because of the success of DL, not just in making better hydrological predictions, but at fundamentally changing the nature of the watershed simulation problem, we see potential for a growing decoupling between epistm and techn in hydrology. This worries us. One one side, we see growing interest from the ML community to attack hydrological informatics problems with sometimes greater and sometimes lesser collaboration with hydrologists. Given that their models generally work better, we might wonder how much they need us? On the other hand, hydrologists have for decades dismissed the success of ML for hydrological prediction as nonphysical (e.g., Blöschl et al., 2019; Kirchner, 2006; Sellars, 2018). The gap between epistm and techn will grow unless we work actively to close it. We can do this by attacking two objectives: (1) learn how to use DL to advance the science (i.e., extract hydrological insight from DL model), and (2) show value in hydrological theory against a backdrop of successful DL (i.e., inject hydrological insight into DL models). Despite the rapidly accelerating pace of ML and DL research in the hydrological sciences, we see relatively little explicit and systematic work against these two problems (there is some, but not much). While the previous generation of hydrologists (e.g., Eagleson, 1991) made concerted effort toward making hydrology stand as a distinct branch of geoscience, our generation must work to recouple the scientific and practical aspects of the discipline. If we do not, it will be epistm that suffers, since at least some of the first-order problems that society asks hydrologists to address can apparently be done with relatively little hydrological science.

Related to extracting insight from trained models, it is often said that ML is a black box. While there is—arguably—some sense in which this is true, there is a much more important sense in which we should think about DL models as containing complex, multilayered, structured information that is accessible if we choose to query it. Recognizing this, our job is one of translation: the information we want is in the models, and we must learn how to translate it into something that is human interpretable. In hydrology, new insights from modeling studies sometimes come from probing models with various types of diagnostic tools (e.g., Martinez & Gupta, 2010; Nearing et al., 2018; Ruddell et al., 2019; Yilmaz et al., 2008), many of which are equally applicable to DL models. Examples of these tools are things like sensitivity analyses to understand (e.g., spatiotemporal) input contributions (e.g., Sundararajan et al., 2017), counterfactuals to understand cause and effect, (e.g., Pearl, 2013; Ribeiro et al., 2016), or DL-specific tools like embedding layers and feature layer analyses (e.g., Bianchi et al., 2020; Wang et al., 2017). It is also at least feasible to leverage advances in explainable AI (XAI; Samek, 2019) to help develop new scientific theory.

An example of looking for explainability in a trained model is in Figure 3. This figure shows the sensitivity of a time series DL model to past inputs. The model learned to store winter precipitation and release this as runoff when temperature and radiation increased in the spring. Kratzert et al. (2019) showed that a DL time series model trained with inputs of precipitation and daily air temperature and targets of only daily streamflow contained internal states that correlated with snow cover and soil water storage. They showed that these

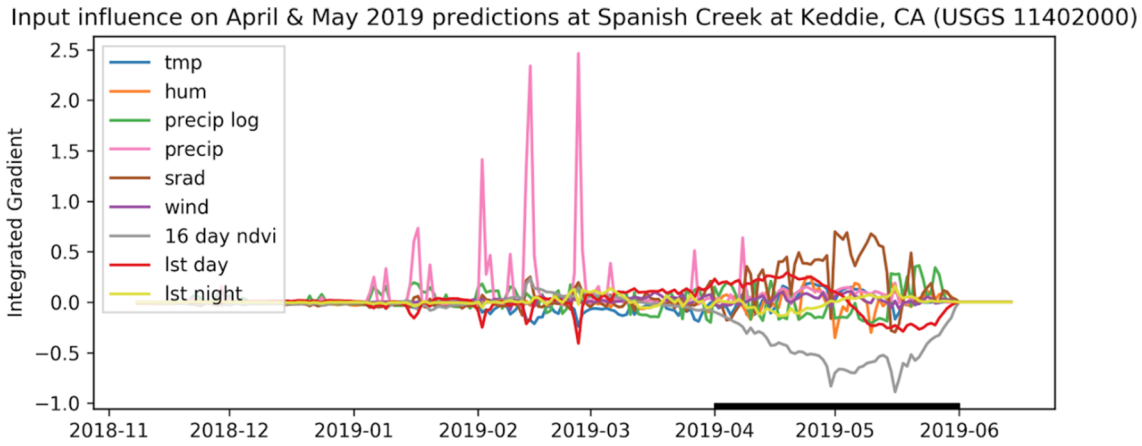

Figure 3. Sensitivity analysis using integrated gradients (Sundararajan et al., 2017) that shows the relative contributions to simulated streamflow during the months of April–May (heavy black shading on the *x*-axis) from the time series of past inputs. The DL model learns to store winter precipitation and responds to increasing temperature and solar radiation in the spring.

"snow" states were sensitive to inputs only when temperatures were below zero. None of this behavior was prescribed a priori - the model learned hydrologically-relevant, interpretable behavior about latent (unobserved) variables.

Looking at the transferability and catchment similarity issues discussed in section 2, Kratzert et al. (2019) constructed a DL network with an embedded feature layer that quantified catchment similarity along a number of learned dimensions (Figure 4). The features extracted from the trained network represent how the DL model transformed observable catchment characteristics into a representation of similarity and diversity in rainfall-runoff relationships. This matrix looks a little like noise, but it is a better representation of catchment similarity than anything human scientists have so far been able to develop. If we want to understand the information encoded in this matrix, then the job ahead of us is to translate this information into a human-interpretable form. Kratzert et al. (2019) used dimensionality reduction to relate first-order features in this similarity matrix with observable catchment characteristics and found that vegetation type and seasonality were the dominant influences.

While ML has been used in hydrology for decades, the ability (at least partially due to computational advances) to arrange shallow learning models into complex structures with feature layers that can learn multiscale patterns opens the door to leveraging diverse (e.g., multicatchment) data in interpretable ways. The idea that ML models are "black boxes" is more of a testament to a lack of inspection, rather than to a limitation of the models themselves. It is worth noting that the DL models used by Kratzert et al. (2019) were invented around the same time (Hochreiter, 1991; Hochreiter & Schmidhuber, 1997) as some of the earliest shallow neural network applications in hydrology (e.g., Hsu et al., 1995). As a discipline, we have not done a great job of keeping pace with developments in ML.

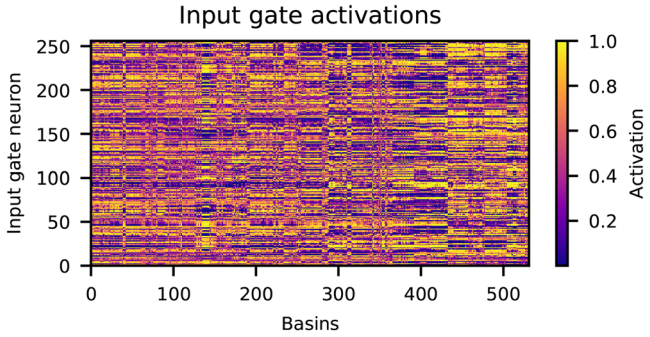

Figure 4. Results from Kratzert et al. (2019) showing a matrix representing catchment similarity as identified by a deep learning model. There are 531 catchments (*x*-axis) and 256 model states (*y*-axis). Each state is activated for any individual catchment to some degree in the range [0, 1], with 0 meaning that the state is not used for that particular catchment. Similar catchments share more of this state space and dissimilar catchments share less.

Perhaps more importantly, we can imagine doing hypothesis testing with DL. One of the major challenges with testing specific processes in complex systems (like watersheds) is that this generally requires simulating the whole system. This is the problem of *holist underdetermination* (Duhem, 1954; Laudan, 1990), whereby auxiliary hypotheses confound the ability to falsify specific hypothesis. Instead of extracting information from trained DL models, we could put hydrological theory into these models and assess improvement (or otherwise). From an ML perspective, this is a regularization problem, and common methods include things like

(i) regularizing the loss function to penalize violations of physical principles like conservation, monotonicity, etc. (e.g., Nabian & Meidani, 2020), (ii) augmenting scientific models with DL structures

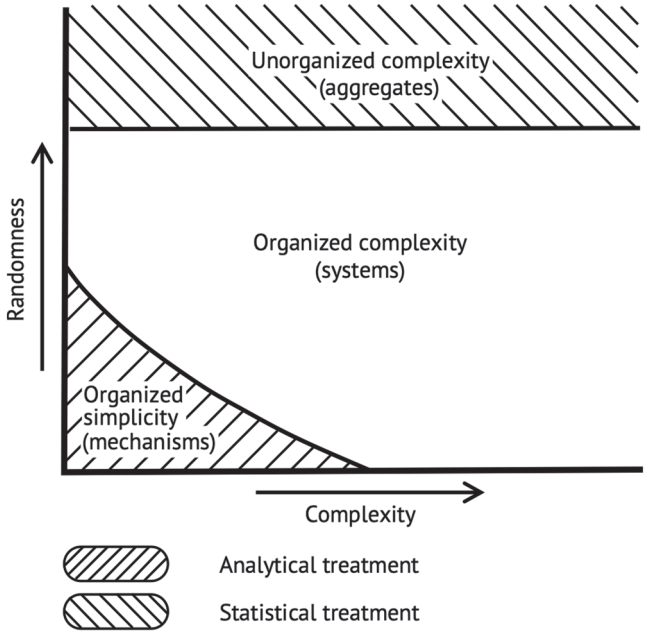

Figure 5. Recreation of an illustration that Dooge (1986) borrowed from Weinberg (1975) to show different types of successful theories in science. Watersheds arguably live in the area of organized complexity, where the complexity (heterogeneity) is at a similar scale to the randomness (lack of information).

(e.g., Pelissier et al., 2020; Rackauckas et al., 2020), and (iii) architecturally constrained neural networks (e.g., Beucler et al., 2019; Daw et al., 2020). This would allow for hypothesis testing against a backdrop, or null hypothesis, that is derived solely from data (Nearing et al., 2020; Nearing & Gupta, 2015).

Hypothesis testing is an example of how we cannot really decouple modeling from science, or techn from epistm. In all versions of the scientific method that we are aware of, hypotheses are tested by comparing observations with predictions, and predictions must come from a model. Cartwright and McMullin (1984) argued that phenomenological laws—not theoretical laws—are the only thing that can actually be tested. The results discussed in section 2 can be interpreted as a hypothesis test that compares the total information content of theory as encoded into hydrological models relative to a null hypothesis derived from data (Nearing et al., 2020), but these experiments do not take the next step and test *specific* (biogeo)physical hypotheses by embedding them as constraints into DL models.

4. Known Unknowns

The second "cloud" in Beven (1987) speech was uncertainty. There has been an enormous amount of attention paid to this topic in the hydrological sciences (e.g., Beven, 2006a, 2009, 2016; Beven & Binley, 2014; Beven et al., 2007, 2008, 2011, 2012; Clark et al., 2011; Kumar, 2011; Mantovan & Todini, 2006; Montanari, 2007; Montanari & Koutsoyiannis, 2012; Nearing, 2014; Nearing & Gupta, 2018; Pappenberger & Beven, 2006; Renard et al., 2010; Stedinger et al., 2008; Todini & Mantovan, 2007; Vrugt et al., 2009); however, we have not had a major breakthrough that led to a paradigm shift. We have suggested previously (Nearing et al., 2016) that the uncertainty literature in hydrology is somewhat detached from the discussion about uncertainty that is taking place in the larger academic (science and philosophy) communities. However, irrespective of that opinion, our community has not developed the stochastic theory of watersheds that Beven (1987) anticipated.

Dooge (1986) offered a discussion about why finding scale-relevant laws is difficult in many branches of science. His argument was that there are two basic categories of scientific theory: mechanistic and aggregate. In the former—mechanistic theories—we track properties (e.g., position and velocity) of individual components of a system, and the resulting model is usually expressed as a system of partial differential equations (PDEs). In the latter—aggregate theories—we rely on ergodic properties like the law of large numbers to derive consistent statistical approximations (e.g., temperature and density) at scales that are much larger than the individual components of a system. The prototypical example of a mechanistic-type theory is Newton's laws, and the prototypical example of an aggregate-type theory is thermodynamics. Dooge borrowed the image in Figure 5 from Weinberg (1975) to illustrate this dichotomy—watersheds live in the middle area of organized complexity, where complexity (heterogeneity) is at a similar scale to randomness (lack of information).

Beven imagined a hydrological theory that is fundamentally stochastic to account for heterogeneity. This is different than how hydrologists typically treat uncertainty. Typical modeling approaches are mechanistic and treat a lack of complete information by adding additional (usually probabilistic) structure to a modeling problem. What we mean by this is that our basic hydrologic theories are largely deterministic, and we represent lack of complete information by adding distributions on top of model inputs, structures, and predictions. This is true even for models based on stochastic PDEs, which necessarily add a distributional component to the structural equation(s). Intuitively, it seems odd that we add *more* structure to a problem to represent a lack of information. Beven's (1987) view of hydrologic theory is compelling in the sense that it would be preferable to have a theory of watersheds that is itself an aggregate-type theory, since at least a significant portion of the variability and complexity in watershed behaviors are due to both landscape and process heterogeneity.

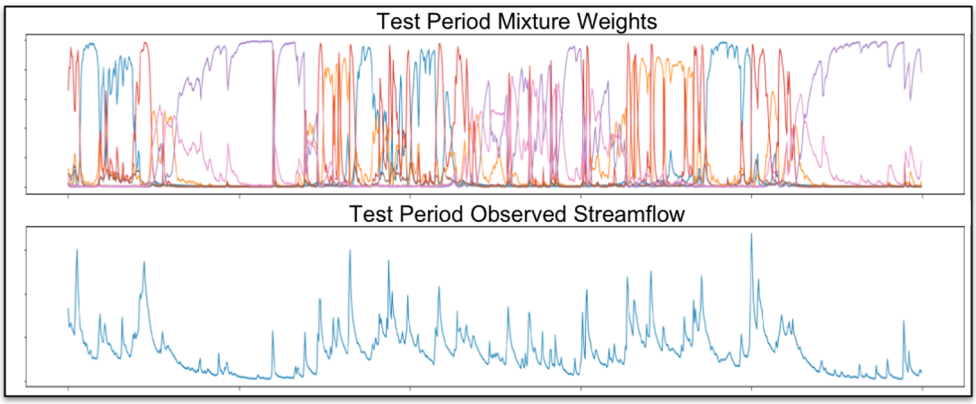

Figure 6. Mixture density weights (6 kernels) predicted by a deep learning model (top) as compared with the corresponding observed hydrograph (bottom). The mixture density weights vary in hydrologically relevant ways; i.e., as a function of peaks (red) and recessions (blue).

ML offers something similar to this in a straightforward way. Instead of predicting the quantities of interest directly, we can predict distributional representations (e.g., probabilistic and fuzzy) directly from input data. This can be as simple as having the output of a DL model be the parameters of a parametric distribution (e.g., a mixture density Bishop, 1994) or the quantiles of a nonparametric distribution (Taylor, 2000). An example of this is shown in Figure 6, which shows the weights of a mixture density over streamflow predicted by a DL model. The training loss function in this case was a likelihood function, and the model did not learn the mixture density parameters directly; instead, it learned how to predict these parameters from dynamic inputs. This figure shows that the individual kernels of the mixture density respond in hydrologically relevant ways; for example, some of the mixture weights have a seasonal cycle, and some are active only in rising or falling limbs of the hydrograph. It is important to understand that the DL model here maps directly from inputs (atmospheric forcings and static basin attributes Addor et al., 2017) to predicted probabilities, rather than sampling a priori probabilities over different model components. There is no need to prescribe any a priori probabilities.

We find an important distinction between *generative* versus *discriminative* models (Nearing et al., 2013). Generative models produce a joint distribution between targets, Y, and inputs, X, and then invert that distribution to obtain conditional predictive probabilities $p(Y|X)$. Discriminative models, on the other hand, map directly onto conditional probabilities. Discriminative models avoid the need to assign any a priori probabilities, and if we believe that we have some information about uncertainties associated with various inputs, these uncertainties can always be used as additional inputs into the model.

Traditional hydrology models, on the other hand, are generative. We must first define all input distributions, and our predicted distributions come from sampling those a priori prescribed distributions. When we use an ensemble to represent uncertainty, for example, the hydrological model or family of models produces a joint distribution between inputs and targets. Although we can sample the predictive conditional by simply looking at one ensemble member, the distribution itself does not exist except as implied by the ensemble where each ensemble member is a joint sample of (X, Y). The bottom line is that in a generative approach, the predicted probabilities are defined in advance by the input or sampling probabilities.

While aggregate theories exist for certain hydrological fluxes (e.g., Singh et al., 2003; Wang & Bras, 2011), most operational models are based on mechanistic theories—hydrologists have not developed an aggregate theory of watersheds. ML does not produce aggregate theories, but it does allow for discriminative modeling.

In addition to predicting probabilities directly, discriminative ML models can take *any type of input*, given sufficient training data. This offers an alternative to inverse methods like data assimilation for integrating ancillary data streams (Nearing et al., 2013). Feng et al. (2020), for example, used the discriminative approach to integrate lagged streamflow values in a (deterministic) DL streamflow model. In principle, it is feasible to add any type of input into one of these models as long as there is sufficient training data. We

no longer need to prescribe the various input distributions directly; instead, these are learned (either implicitly or explicitly) by the DL model from all available data in a way that is dynamic (i.e., changes) in time and place and under different hydrologic conditions.

5. Hydrology Beyond Streamflow

The hydrological sciences are diverse, and the discussion so far has been about catchment hydrology and streamflow. Supposing the reader accepts the arguments we have laid out so far, it is worth asking whether there are implications for other branches of the discipline. The answer is—of course—that we do not know. On one hand, there are major differences between the challenges faced in catchment hydrology versus groundwater or ecohydrology or hydrometeorology, but at the same time, it is difficult to overestimate the impact of DL and AI throughout all types of human endeavors. In hydrometeorology, several studies have shown that even very simple regression models produce better estimates of radiation partitioning than process-based land surface models (Abramowitz, 2005; Best et al., 2015; Nearing et al., 2018). Fang and Shen (2020) showed that DL can produce highly accurate soil moisture forecasts with remote sensing. Hydrometeorology is similar to streamflow hydrology in that observations are (relatively) abundant from satellites and mature sensor networks like FluxNet, etc. These fields are also similar in that the major sources of uncertainty are due to spatial heterogeneity at intermediate scales.

In groundwater, which is often more data limited than surface hydrology, many of the standard methods have close or direct analogs in ML already (e.g., Kriging is just Gaussian process regression Williams & Rasmussen, 2006). It may be the case that there is less potential for a fundamentally new result. One recent study reported that a physically based groundwater model outperformed several shallow ML models (Chen et al., 2020). There have been some relatively small DL studies in groundwater hydrology (e.g., Mo et al., 2019; Sahoo et al., 2017) that did not report transformative results.

It is hard to draw strong conclusions from the existing body of work. In all of these studies (including those by the current authors but with the notable exception of Fang et al., 2018) is a lack of big data. ML does not have the ability to learn multiscale hierarchical patterns in the same way as DL and therefore cannot leverage diversity in big data in the same way. After testing several shallow ML models, Chen et al. (2020) concluded that "the generalization ability of numerical model is superior to the machine learning models because of the inclusion of physical mechanism."

The basic problem is a lack of real investment into this type of effort. There are major programs across hydrologic disciplines to build comprehensive multiscale models, e.g., groundwater (de Graaf et al., 2020), streamflow (Li et al., 2015, 2019), hydrometeorology (Rodell et al., 2004), and many others, but to our knowledge, there is no similar effort to build global AI models. DL does not scale like traditional models—it works differently on large data sets than small data sets—so small pilot studies do not tell us much.

There is no question that we are in a new information age and that modern data science techniques have been transformative across scientific disciplines. The message that we would like to leave the reader with is that hydrologists currently do not know what how transformative this technology will across our discipline. We do not know this because we have not made a serious investment in AI-based hydrology. Our major modeling centers continue to invest primarily in old technologies and old approaches. In the case of streamflow hydrology, this has been a disaster. The point of this opinion piece is that there are clues that maybe the balance of data and theory will not look like what hydrologists anticipate (e.g., references in section 3).

6. Where the Sidewalk Ends

So what could we do about this? The following subsections outline what we see as both immediate needs for expanding DL in hydrology, as well as some ideas about what the longer term future could look like.

6.1. Distributed Modeling

The first immediate need is for spatiotemporal DL models in all areas of hydrology. We simply just need to make serious investments across the discipline to gather the data that each community has—across regions and countries, to the extent possible—and make a serious attempt to develop state-of-the-art AI models.

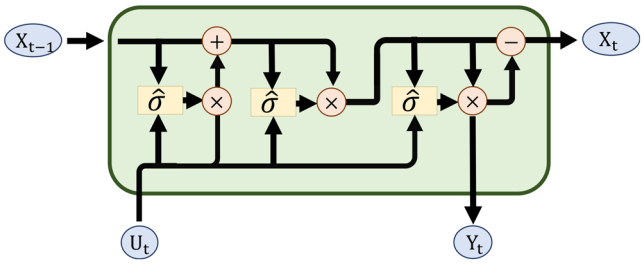

Figure 7. A time-recurrent deep learning network that is architecturally constrained to conserve mass, energy, and/or momentum. U_t is time-dependent inputs, Y_t is time-dependent outputs, and X_t is a vector of N memory states in the network; $\hat{\sigma}$ represents a set of N L1-normalized sigmoid activation functions that produce a set of real values in [0, 1] that sum to unity. These are scaled by the conserved quantities (in the inputs and states) so that the total sum of the time history of inputs plus outputs is always equal to the total sum of the system state. There are three sets of "gates" in this network—an input gate that moves mass (energy, momentum) from inputs to states, a reshuffling gate that moves mass (energy, momentum) between states during each individual time step, and an output gate that moves mass (energy, momentum) from states to outputs at each time step.

We expect that first-order attempts at this type of project will look similar to current models with some explicit spatiotemporal extent/resolution and some number of latent (hidden) variables. Previously, we criticized calls for hyper-resolution modeling, and while the race to higher density, more-of-the-same type models does seems to be a particularly unthoughtful idea, it is nevertheless the case that hydrological processes have both spatial and temporal components. We expect that within the next 1–2 years, the community will develop several distributed DL watershed models (e.g., Moshe et al., 2020). There are various ways that we might incorporate a multitude of different types of spatiotemporal data into trained models. DL allows for complex interactions between different feature layers, and fine-tuning allows modelers to train individual components of a model. We can imagine a model developed by training different feature layers—perhaps themselves multilayer DL models—and piecing these together to represent theory-guided architectures. As an example, we could imagine training a convolutional network to map from remote sensing data like SMAP (Entekhabi et al., 2010) to root-zone soil moisture by training directly on target data from in situ networks like the USDA Soil Climate Analysis Network (Schaefer et al., 2007) and/or FluxNet (Baldocchi et al., 2001). The weights of this trained convolutional layer(s) could then be frozen, and the trained network then used as one (of many) input feature layer(s) into an LSTM (or other time series model) for predicting streamflow (or evapotranspiration or groundwater recharge). In principle, input data streams could be integrated at arbitrary spatiotemporal resolutions so that irregular convolutional networks (e.g., graph convolutions) could be used for routing.

The details of this type of model will need to be worked out, but the potential for, and basic components and principles of, a DL-based integrated hydrology model is relatively clear. There is no fundamental limitation that precludes developing integrated DL hydrology models at multiple temporal and spatial scales. The questions that we anticipate are about what value will come from integrating different types of features and feature layers and about how we might pretrain various feature layers to account for different types and scales of observational data in large, integrated models.

6.2. Theory-Informed ML

As mentioned in section 3, there is a feeling among hydrologists and Earth scientists that models without explicit process representation might be unreliable under changing conditions. Although we do not know if this is really true, one way to approach this is to integrate physical constraints or process-based theory into DL models. The goal is to extract as much information as possible from a combination of theory and data. This is not a new idea—Karpatne et al. (2017) called for *theory-guided data science*, which consists of efforts to integrate scientific consistency into generalizable models. Notably, members of that same group later collaborated on development of a DL model that is architecturally constrained to not violate prescribed monotonicity relationships (Daw et al., 2019).

A simple and general way to enforce conservation constraints (e.g., mass, energy, and momentum) in a DL architecture is to L1-normalize a set of bounded ($\in [0, 1]$) activation functions and scale by the conserved quantity. This concept can be integrated into almost any type of neural network architecture, including into the long short-term memory networks used by Kratzert et al. (2019) and Kratzert et al. (2019). This concept is illustrated in Figure 7, and the result is a model that learns nonlinear input-state-output relationships that obey arbitrary and interacting conservation principles.

Another approach for directly combining process understanding with ML is to incorporate the ML models inside of a dynamical systems model. A basic approach was outlined by Ghahramani and Roweis (1999), where—effectively—an empirical model is trained on the analysis states resulting from data assimilation (e.g., by a Kalman-type filter). We can generalize this idea as follows:

Suppose that we have a dynamical systems model that solves a set of PDEs:

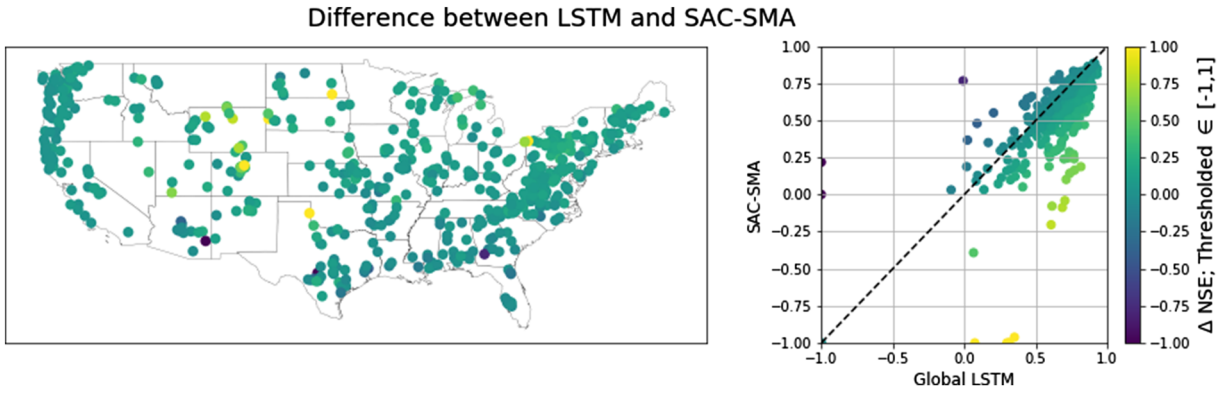

Figure 8. An illustration from Kratzert et al. (2019) that compares a deep learning model (LSTM) against a calibrated conceptual model (SAC-SMA) over 531 CAMELS basins. The deep learning model does better on average, but not in every catchment, indicating that there is at least potential to improve by incorporating some of the conceptual constraints from SAC-SMA.

$$\frac{dX}{dt} = f(X,\ U,\ \theta), \tag{1}$$

where X is modeled system states, U is time-dependent boundary conditions, θ are model parameters, and function $f(\cdot)$ is the total divergence (inputs less outputs). A discrete-time approximate solution might then be

$$X_t = f^*(X_{t-1},\ U_t,\ \theta). \tag{2}$$

We can augment the $f^*(\cdot)$ state-transition function with a learned component, $g^*(\cdot)$, as follows:

$$X_t = f^*(X_{t-1},\ U_t,\ \theta) + g^*(X_{t-1},\ U_t,\ \theta), \tag{3}$$

where $g^*(\cdot)$ is any ML model. As above, $g^*(\cdot)$ can itself be probabilistic so that Equation 3 is a discrete-time solution to a set of stochastic PDEs. The challenge is to learn the $g^*(\cdot)$ function given that we cannot expect to have direct observation pairs (X_t, X_{t-1}) of all system states to use for supervised learning. As an example, Nearing and Gupta (2015) applied the data assimilation approach by Ghahramani and Roweis (1999) to the HyMod conceptual rainfall runoff model, and Pelissier et al. (2020) applied a similar technique to the Noah-MP land surface model for soil moisture accounting.

Another example of potential for theory-guided data science in hydrological workflows is for data assimilation itself. Significant information loss often results from assigning the distributions and parameters of a probability-based assimilation algorithm (Nearing et al., 2018), and many assimilation algorithms require that the model and observation be in the same climatology (Kumar et al., 2012), meaning that these algorithms only treat stochastic error. One potential way to mitigate these problems is to use ML to learn relationships between model states and assimilated observations (e.g., Kolassa et al., 2018). As an example of this, Nearing (2013) derived the fixed form of the Kalman-type gain and its associated adjoint that results from assimilating with a Gaussian process observation operator. We see theory-guided data science, and more specifically, physics-informed ML, as a likely strategy for simultaneously leveraging what we do know from scientific theory about catchment behavior with the now undeniable ability of DL for extracting patterns and information directly from data. There is some indication that this might be useful: Figure 8 shows a comparison between the performance of a DL model applied to CAMELS basins versus a calibrated conceptual model. These data are from Kratzert et al. (2019), and the takeaway message is that while the DL model is better overall, it is not better everywhere. Kratzert et al. (2019) could not find any relationship(s) between observable catchment characteristics and the difference in performance between these two models, but it is nevertheless apparent that there is at least the potential to improve by adding some elements of hydrologic theory to the DL architecture.

6.3. Skip the Hydrologist

Clark et al. (2016) gave an account of the sources of uncertainty (information loss) in a hydrological modeling chain. These are things like uncertainty in meteorological forcings from global circulation models

(GCMs), downscaling forcings to the watershed scale, errors in the hydrological model structure, parameter uncertainty, etc. Each of these represents a step in a chain of information from the GCM dynamical core (i.e., Navier-Stokes approximations and data assimilation) to streamflow or other hydrological variables. Every step in this modeling chain introduces uncertainty. DL has the potential to let us skip at least several steps in this type of modeling chain by developing relationships directly between high-quality data sources.

Take as an example the largest source of hydrological error, which is typically precipitation data. This is true whether we are using the output of weather or climate models, interpolated gauge data, or remote sensing data from radar and/or satellites. The problem is exacerbated by downscaling. The major precipitation-related uncertainty in a GCM is due to parameterization of subgrid cloud formation processes. There have been recent successes using ML to parameterize cloud physics and cloud formation (e.g., Gentine et al., 2018), which could help mitigate these issues to some extent, but we still have to feed these uncertain precipitation fields into a hydrology model that is subject to both parameter and structural uncertainties.

We could think about the problem in a different way. The four-dimensional pressure, wind, and temperature fields that result from Euler solutions in the dynamical cores of GCMs are relatively accurate, at least as compared with the accuracy of parameterized precipitation fields. We could, in principle, use DL to extract information directly from states of the dynamical core about terrestrial hydrological variables. For example, we could in principle develop four-dimensional convolutions to regress directly from GCM fields and digital elevation maps to pixel classifiers over satellite-derived maps of flood inundation and thereby skip sources of information loss from (i) subgrid convection parameterizations, (ii) GCM downscaling, (ii) lack of scale-relevant theories of watersheds, (iii) parameter equifinality, (iv) rating curves, etc. It is possible (perhaps likely) that this type of model would give more accurate inundation forecasts at similar lead times relative to state-of-the-art hydrology models, since this would skip uncertainties related to cloud physics parameterizations, downscaling, watershed parameterizations, etc. All of these things could be learned implicitly by a DL model.

The point is that DL offers at least the potential to make societally relevant hydrological forecasts without any type of hydrological model or hydrological process understanding at all. Because DL allows for almost arbitrarily complex relationships and has demonstrated to extrapolate well out of sample, it might be the case that successful water resources and water hazard predictions might not require anything that looks even like a simple hydrology model. This is all speculative, but the point is that the idea about hydrological understanding being necessary for reliable forecasting discussed in section 3 may not be true even in the most superficial sense. This is an extreme and hypothetical example, but one that is worth (1) trying experimentally and (2) being aware of as we calibrate our expectations about the role of hydrological theory and hydrological science in the context of big data and <L.

6.4. Observations and Benchmarks

Beven (2006b) proposed that the search for closure schemes at the watershed scale is the *second* most important problem in the discipline, with the most important being to improve observation capabilities. We agree completely. As was the case in 1987, the first and foremost job of hydrologists are and will continue to be related to improving observational capacity. The approaches discussed in this article only increase the need for observation data related to as many aspects of the water cycle as possible.

Shen et al. (2018) noted that past progress in the field of ML can be partially attributed to the culture of using public data sets and benchmarking new methods against previous state of the art. There have been calls for consistent practices related to hypothesis testing, model intercomparison, and model rejection (e.g., Beven, 2018). While some of the philosophical counterarguments to this are compelling (e.g., Baker, 2017; Nearing et al., 2016, 2020), without *some* community standard for benchmarking, it is difficult to track progress in the field in an objective way.

This means that we need two things. First are better centralized data repositories. The community is aware of this (Gupta et al., 2014; Shen et al., 2018), and there are several such efforts happening in the field right now (e.g., Addor et al., 2017; Hoffman et al., 2016; Newman et al., 2015). We expect that this issue will sort itself out in the near future. Still, our opinion is that one of the best investments that could be made in the discipline right now is to develop standardized and easily accessible big data repositories. The second thing we

need is the willingness to use those data repositories. Just like in previous decades when the community responded to calls for making uncertainty quantification required for every modeling study (Pappenberger & Beven, 2006), we need a community standard that requires all new modeling papers to include large-scale benchmarking against standard, centralized data sets.

Hydrological modeling is currently a field of ivory towers where legacy and affiliation guide the choice of model (Addor & Melsen, 2019) as opposed to empirical rigor (Beven, 2018). Different modeling groups largely work on their own models, and while there have been ad hoc intercomparisons (e.g., Best et al., 2015; van den Hurk et al., 2011), this is not routine, and the hydrology community does not keep a list of current performance scores on standard test problems, as is standard in other communities (e.g., CMIP and ML-Perf).

7. A White Whale

During the community contribution phase of the IAHS "Unsolved Problems in Hydrology" effort (Blöschl et al., 2019), one of the suggested questions was: "Does Machine Learning have a real role in hydrological modeling?" In contrast, we suggest that the existential question for our discipline right now is: "What role will hydrological science play in the age of machine learning?" van den Hurk et al. (2011) challenged that "it must be demonstrated that the model physics actually adds information to the prediction system." This is exactly the question that needs to be answered in order to understand how and where hydrological theory has a role to play in a world dominated by data. We see at least potential for DL to help address this by allowing us to decouple different parts of hydrological theory while still retaining scale-relevant predictive systems learned (partially) from data.

Very likely, the future of hydrology will be a mix of AI and physics-based approaches, but we have a hard time envisioning a future where transformative data science approaches like DL become simply another tool in the hydrologist's toolbox. We see it as much more likely that hydrological domain knowledge will become an integral part of guiding and developing fundamentally AI-based systems and analyses (e.g., section 6.2).

Hydrology has roots—at least in part—as a branch of civil engineering. Klemeš (1986a) argued that "practices of bad science in hydrology cannot be blamed on engineers and other decision makers who 'need numbers.' For if these numbers are not to be based on sound hydrologic science but only on manipulations of arbitrary assumptions and concepts, hydrologists are not needed." The situation has not changed much in the 34 years since that was written: our ability to extract numbers (predictions) from data is advancing rapidly, but we have not improved very much our ability to make predictions from anything resembling hydrologic theory. While our models become increasingly complex, a well-calibrated Sacramento model is still one of the best in discipline.

The reason that we think this is an *existential* challenge is because we see hydrological science becoming increasingly decoupled from state-of-the-art hydrological information systems. Major development groups at governmental institutions internationally continue to dedicate the large majority of effort to the traditional models that have never benchmarked well against ML (e.g., Abramowitz, 2005; Best et al., 2015; Kratzert et al., 2019; Nearing et al., 2018). As far as we can tell, these models are dead on arrival. Barring some major fundamental theoretical discovery or innovation, there is essentially no chance that any incremental advancements will allow these models to catch up to the state-of-the-art hydrological predictions. Simultaneously, there has not been any serious or systematic investment into AI-based hydrology at a meaningful scale, and from what we can see (e.g., see section 3), there is still strong resistance in the hydrology community toward adopting these approaches in a serious and fundamental way. Coupled with the fact that DL experiments demonstrate that hydrologists lack even a basic understanding of why their models fail (section 2), this causes us to worry.

Our fear is that if the hydrological sciences community refuses to make a serious investment into the technology that works, then someone else will. This will mean a further decoupling between hydrological science (such as it is) and the societal value that this science is supposed to support. To be clear, the current authors do *not* want to see that happen, but we are not impressed with the reaction we are seeing in the community. Our message in this opinion piece is to stop assuming that the world needs our theories and expertise and start demonstrating—quantitatively and systematically—the value of individual components of that expertise against the backdrop of a growing importance of big data.

Data Availability Statement

No data were generated as part of this project.

Acknowledgments
Authors from the Johannes Kepler University were partially supported by a Google faculty research award. Grey Nearing at the University of Alabama was supported by the NASA Advanced Information Systems Technology program (award ID 80NSSC17K0541). Jonathan Frame at the University of Alabama was supported by the NASA Terrestrial Hydrology Program (award ID 80NSSC18K0982).The author from IHCantabria acknowledges the financial support from the Government of Cantabria through the Fnix Programme.

References

Abramowitz, G. (2005). Towards a benchmark for land surface models. *Geophysical Research Letters*, *32*, L22702. https://doi.org/10.1029/2005GL024419

Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, *55*, 378–390. https://doi.org/10.1029/2018WR022958

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

Baker, V. R. (2017). Debates hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research*, *53*, 1770–1778. https://doi.org/10.1002/2016WR020078

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., et al. (2001). FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, *82*(11), 2415–2434.

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

Beucler, T., Pritchard, M., Rasp, S., Gentine, P., Ott, J., & Baldi, P. (2019). Enforcing analytic constraints in neural-networks emulating physical systems. arXiv preprint arXiv:1909.00912.

Beven, K. (1987). Towards a new paradigm in hydrology. IN: Water for the Future: Hydrology in Perspective. IAHS Publication.

Beven, K. (2000). Uniqueness of place and process representations in hydrological modelling.

Beven, K. (2006a). On undermining the science? *Hydrological Processes: An International Journal*, *20*(14), 3141–3146.

Beven, K. (2006b). Searching for the holy grail of scientific hydrology: Q t=(s, r, δt) A as closure. *Hydrology and Earth System Sciences*, *10*(5), 609–618.

Beven, K. (2009). Comment on equifinality of formal (DREAM) and informal (GLUE) bayesian approaches in hydrologic modeling? by jasper a. vrugt, cajo jF ter braak, h gupta and bruce a. robinson. *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1059–1060.

Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, *61*(9), 1652–1665.

Beven, K. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *Wiley Interdisciplinary Reviews: Water*, *5*(3), e1278.

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, *28*(24), 5897–5918.

Beven, K., Smith, P., & Freer, J. (2007). Comment on hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology by pietro mantovan and ezio todini. *Journal of Hydrology*, *338*(3–4), 315–318.

Beven, K., Smith, P. J., & Freer, J. E. (2008). So just why would a modeller choose to be incoherent? *Journal of Hydrology*, *354*(1–4), 15–32.

Beven, K., Smith, P., Westerberg, I., & Freer, J. (2012). Comment on pursuing the method of multiple working hypotheses for hydrological modeling by P. Clark et al. *Water Resources Research*, *48*, W09301. https://doi.org/10.1029/2010WR009827

Beven, K., Smith, P., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*, 3123–3133.

Bianchi, F., Rossiello, G., Costabello, L., Palmonari, M., & Minervini, P. (2020). Knowledge graph embeddings and explainable AI. arXiv preprint arXiv:2004.14843.

Bishop, C. M. (1994). Mixture density networks.

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH)—A community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158.

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290.

Cartwright, N., & McMullin, E. (1984). How the laws of physics lie.

Chen, C., He, W., Zhou, H., Xue, Y., & Zhu, M. (2020). A comparative study among B and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Scientific Reports*, *10*(1), 1–13.

Clark, M., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., et al. (2016). Characterizing uncertainty of the hydrologic impacts of climate change. *Current Climate Change Reports*, *2*(2), 55–64.

Crawford, N. H., & Burges, S. J. (2004). History of the Stanford watershed model. *Water Resources Impact*, *6*(2), 1–3.

Daw, A., Thomas, R. Q., Carey, C. C., Read, J. S., Appling, A. P., & Karpatne, A. (2019). Physics-Guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling. arXiv preprint arXiv:1911.02682.

Daw, A., Thomas, R. Q., Carey, C. C., Read, J. S., Appling, A. P., & Karpatne, A. (2020). Physics-Guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, SIAM, pp. 532–540.

de Graaf, I., Condon, L., & Maxwell, R. (2020). Hyper-resolution continental-scale 3-D aquifer parameterization for groundwater modeling. *Water Resources Research*, *56*, e2019WR026004. https://doi.org/10.1029/2019WR026004

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

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton University Press.

Eagleson, P. S. (1991). Hydrologic science: A distinct geoscience. *Reviews of Geophysics*, *29*(2), 237–248.

Einstein, A. (1926). The cause of the formation of meanders in the courses of rivers and of the so-called Baer's law. Die Naturwissenschaften.

Einstein, H. A. (1950). The Bed-Load function for sediment transportation in open channel flows: United States Department of Agriculture.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704–716.

Fang, K., Pan, M., & Shen, C. (2018). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233.

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *2020*, 399–413.
Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*, e2019WR026793. https://doi.org/10.1029/2019WR026793
Gauch, M., Mai, J., & Lin, J. (2019). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction.
Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*, 5742–5751. https://doi.org/10.1029/2018GL078202
Ghahramani, Z., & Roweis, S. T. (1999). Learning nonlinear dynamical systems using an eM algorithm. In *Advances in neural information processing systems* (pp. 431–437).
Gupta, H., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences Discussions*, *10*, 9147–9189.
Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen (Ph.D. Thesis), Technische Universität München.
Hochreiter, S., & Schmidhuber, J. A. (1997). Long Short-Term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Hoffman, F. M., Riley, W. J., Randerson, J. T., Keppel-Aleks, G., & Lawrence, D. M. (2016). International land model benchmarking (ILAMB).
Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological sciences journal*, *58*(6), 1198–1255.
Hsu, K.-l, Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517–2530.
Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331.
Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362
Klemeš, V. (1986a). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, *22*(9S), 177S–188S.
Klemeš, V. (1986b). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24.
Kolassa, J., Reichle, R. H., Liu, Q., Alemohammad, S. H., Gentine, P., Aida, K., et al. (2018). Estimating surface soil moisture from SMAP observations using a neural network technique. *Remote Sensing of Environment*, *204*, 43–59.
Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). NeuralHydrology–interpreting LSTMs in hydrology, *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 347–362). Springer.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. (2019). 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., Klotz, D., Hochreiter, S., & Nearing, G. (2020). A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences Discussions*, 1–26.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). 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.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
Kumar, P. (2011). Typology of hydrologic predictability. *Water Resources Research*, *47*, W00H05. https://doi.org/10.1029/2010WR009769
Kumar, S. V., Reichle, R. H., Harrison, K. W., Peters-Lidard, C. D., Yatheendradas, S., & Santanello, J. A. (2012). A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, *48*, W03515. https://doi.org/10.1029/2010WR010261
Laudan, L. (1990). Demystifying underdetermination. *Minnesota Studies in the Philosophy of Science*, *14*(1990), 267–297.
Li, H.-Y., Leung, L. R., Getirana, A., Huang, M., Wu, H., Xu, Y., et al. (2015). Evaluating global streamflow simulations by a physically based routing model coupled with the community land model. *Journal of Hydrometeorology*, *16*(2), 948–971.
Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., et al. (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resources Research*, *55*, 6499–6516. https://doi.org/10.1029/2019WR025287
Mantovan, P., & Todini, E. (2006). Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology*, *330*(1-2), 368–381.
Martinez, G. F., & Gupta, H. V. (2010). Toward improved identification of hydrological models: A diagnostic evaluation of the abcd monthly water balance model for the conterminous United States. *Water Resources Research*, *46*, W08507. https://doi.org/10.1029/2009WR008294
McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., et al. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*, W07301. https://doi.org/10.1029/2006WR005467
Mo, S., Zabaras, N., Shi, X., & Wu, J. (2019). Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. *Water Resources Research*, *55*, 3856–3881. https://doi.org/10.1029/2018WR024638
Montanari, A. (2007). What do we mean by uncertainty? the need for a consistent wording about uncertainty assessment in hydrology. *Hydrological Processes: An International Journal*, *21*(6), 841–845.
Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*, W09555. https://doi.org/10.1029/2011WR011412
Moshe, Z., Metzger, A. F., Kratzert, F., Elidan, G., Nevo, S., & El-Yaniv, R. (2020). HydroNets: Leveraging river structure for hydrologic modeling.
Nabian, M. A., & Meidani, H. (2020). Physics-driven regularization of deep neural networks for enhanced engineering design and analysis. *Journal of Computing and Information Science in Engineering*, *20*(1), 011006.
Nearing, G. (2013). Diagnostics and generalizations for parametric state estimation.
Nearing, G. (2014). Comment on a blueprint for process-based modeling of uncertain hydrological systems by alberto montanari and demetris kousoyiannis.
Nearing, G., & Gupta, H. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*, 524–538. https://doi.org/10.1002/2014WR015895
Nearing, G., & Gupta, H. (2018). Ensembles vs. information theory: Supporting science under uncertainty. *Frontiers of Earth Science*, *12*(4), 653–660.
Nearing, G., Gupta, H., & Crow, W. (2013). Information loss in approximately bayesian estimation techniques: A comparison of generative and discriminative approaches to estimating agricultural productivity. *Journal of hydrology*, *507*, 163–173.

Nearing, G., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*, *17*(3), 745–759.
Nearing, G., Ruddell, B., Bennett, A., Prieto, C., & Gupta, H. (2020). Debates: Does information theory provide a new paradigm for Earth science? Hypothesis testing. *Water Resources Research*, *56*, e2019WR024918. https://doi.org/10.1029/2019WR024918
Nearing, G., Ruddell, B. L., Clark, M. P., Nijssen, B., & Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, *19*(11), 1835–1852.
Nearing, G., Tian, Y., Gupta, H., Clark, M., Harrison, K., & Weijs, S. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*(9), 1666–1678.
Nearing, G., Yatheendradas, S., Crow, W., Zhan, X., Liu, J., & Chen, F. (2018). The efficiency of data assimilation. *Water Resources Research*, *54*, 6374–6392. https://doi.org/10.1029/2017WR020991
Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). 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), 209–223.
Pappenberger, F., & Beven, K. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research*, *42*, W05302. https://doi.org/10.1029/2005WR004820
Parry, R. (2003). Episteme and techne.
Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science*, *37*(6), 977–985.
Pelissier, C., Frame, J., & Nearing, G. (2020). Combining parametric land surface models with machine learning. arXiv preprint arXiv:2002.06141.
Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., Van Emmerik, T., Uijlenhoet, R., et al. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and earth system sciences*, *21*(7), 3701.
Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., et al. (2020). Universal differential equations for scientific machine learning. arXiv preprint arXiv:2001.04385.
Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*, W05521. https://doi.org/10.1029/2009WR008328
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
Rodell, M., Houser, P. R., Jambor, U. E. A., Gottschalck, J., Mitchell, K., Meng, C.-J., et al. (2004). The global land data assimilation system. *Bulletin of the American Meteorological Society*, *85*(3), 381–394.
Ruddell, B. L., Drewry, D. T., & Nearing, G. (2019). Information theory for model diagnostics: Structural error is indicated by Trade-Off between functional and predictive performance. *Water Resources Research*, *55*, 6534–6554. https://doi.org/10.1029/2018WR023692
Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resources Research*, *53*, 3878–3895. https://doi.org/10.1002/2016WR019933
Samek, W. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. (Vol. 11700). Springer Nature.
Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USD: A natural resources conservation service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077.
Sellars, S. L. (2018). Grand challenges in big data and the earth sciences. *Bulletin of the American Meteorological Society*, *99*(6), ES95–ES98.
Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *11*, 5639–5656. https://doi.org/10.5194/hess-22-5639-2018
Singh, V. P., Yang, C. T., & Deng, Z. Q. (2003). Downstream hydraulic geometry relations: 1. Theoretical development. *Water Resources Research*, *39*(12), 1337. https://doi.org/10.1029/2003WR002484
Sivapalan, M. (2006). Pattern, process and function: Elements of a unified theory of hydrology at the catchment scale. Encyclopedia of Hydrological Sciences.
Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880.
Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, *44*, W00B06. https://doi.org/10.1029/2008WR006822
Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, pp. 3319–3328.
Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311.
Todini, E. (2007). Hydrological catchment modelling: Past, present and future. *Hydrology and Earth System Sciences*, *11*(1), 468–482.
Todini, E., & Mantovan, P. (2007). Comment on: On undermining the science? by keith beven. *Hydrological Processes: An International Journal*, *21*(12), 1633–1638.
van den Hurk, B., Best, M., Dirmeyer, P., Pitman, A., Polcher, J., & Santanello, J. (2011). Acceleration of land surface model development over a decade of GLASS. *Bulletin of the American Meteorological Society*, *92*(12), 1593–1600.
Vrugt, J. A., Ter Braak, C. J. F., Gupta, H., & Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) ayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1011–1026.
Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, *1*(4), 901–931.
Wang, J., & Bras, R. L. (2011). A model of evapotranspiration based on the theory of maximum entropy production. *Water Resources Research*, *47*, W03521. https://doi.org/10.1029/2010WR009392
Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, *29*(12), 2724–2743.
Weinberg, G. M. (1975). *An introduction to general systems thinking* (Vol. 304). Wiley New York.
Williams, C. K. I., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2). Cambridge, MA: MIT Press.
Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*, W09417. https://doi.org/10.1029/2007WR006716